Published in final edited form as:
J Proteome Sci Comput Biol. ; 1: . doi:10.7243/2050-2273-1-6.

Characterizing protein domain associations by Small-molecule ligand binding

Qingliang Li[1], **Tiejun Cheng**[1], **Yanli Wang**[1,*], and **Stephen H. Bryant**[1,*]
[1]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA

Abstract

Background—Protein domains are evolutionarily conserved building blocks for protein structure and function, which are conventionally identified based on protein sequence or structure similarity. Small molecule binding domains are of great importance for the recognition of small molecules in biological systems and drug development. Many small molecules, including drugs, have been increasingly identified to bind to multiple targets, leading to promiscuous interactions with protein domains. Thus, a large scale characterization of the protein domains and their associations with respect to small-molecule binding is of particular interest to system biology research, drug target identification, as well as drug repurposing.

Methods—We compiled a collection of 13,822 physical interactions of small molecules and protein domains derived from the Protein Data Bank (PDB) structures. Based on the chemical similarity of these small molecules, we characterized pairwise associations of the protein domains and further investigated their global associations from a network point of view.

Results—We found that protein domains, despite lack of similarity in sequence and structure, were comprehensively associated through binding the same or similar small-molecule ligands. Moreover, we identified modules in the domain network that consisted of closely related protein domains by sharing similar biochemical mechanisms, being involved in relevant biological pathways, or being regulated by the same cognate cofactors.

Conclusions—A novel protein domain relationship was identified in the context of small-molecule binding, which is complementary to those identified by traditional sequence-based or structure-based approaches. The protein domain network constructed in the present study provides a novel perspective for chemogenomic study and network pharmacology, as well as target identification for drug repurposing.

Keywords

Protein domain; drug repurposing; domain network; promiscuous drug; drug target identification

Correspondence: ywang@ncbi.nlm.nih.gov and bryant@ncbi.nlm.nih.gov.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Conceived and designed the experiments: QL YW and SB. Analyzed the data: QL and TC. Wrote the paper: QL, TC and YW.

Background

Protein domains are evolutionarily conserved units in protein sequence, structure and function, which can be recombined in different arrangements to create new proteins in biological organisms [1–6]. The interactions between protein domains and other molecules play a fundamental role in molecular recognition in living organisms. Small molecule binding domains are of particular interest, as many of them represent targets for biologically important ligands including drugs [7,8]. Studies on small molecule-protein domain interactions have received increasing attention for their potential to advance chemogenomics research and drug development [9–11].

Many studies have investigated the interactions between small molecules and protein domains. For example, Yamanishi *et al.,* [12] used the canonical correspondence analysis method to investigate the rules governing the recognition of chemical substructures and protein domains. Bender *et al.,* [13] built a statistical model on chemical structures and protein domains to triage the affinity chromatography data. Wang *et al.,* [14] used protein domains and therapeutic information to predict drug targets. Besides, Kruger and Overington [15] incorporated protein domain information to analyze small-molecule bindings of homologous proteins in human and rat. Collectively, the underlying assumption of these studies is that small molecule-protein recognitions are accomplished through small molecule-protein domain interactions. However, due to the lack of accurate binding site information, these interactions are usually assumed according to the presence of protein domain(s) within a protein, yet a specific connection between a domain and its ligand is not guaranteed. This strategy may work well for single-domain proteins, while it may fail for multi-domain proteins that are usually observed in human genome. To address such issues, Kruger and Overington [15] proposed to derive small molecule-domain interactions based on the observed frequency in single-domain proteins. However, the results based on such empirical assignment are nonetheless compromised. Meanwhile, proteins are conventionally grouped into individual families based on sequence or structure similarity [1–6,16]. The inter-relationship across such families, especially for small-molecule binding, is seldom studied, though they are important for understanding the regulatory roles of small molecules in biological systems.

In the present study, we attempted to address such issues by first collecting the physical interactions between small molecules and protein domains derived from the experimentally determined structures in Protein Data Bank (PDB) [17] and then characterizing the protein domain inter-relationship with respect to small-molecule binding on a large scale. As PDB contains protein 3D structures and accurate structural information of protein-ligand interactions, several secondary databases have been developed to include small molecule-protein domain links as recently reviewed by Bashton and Thornton [18]. For example, the PDBLIG [19] database associates small molecules contained in PDB to the CATH domains [1]. Likewise, PROCOGNATE [20] links small molecules in PDB to three distinct domain databases including CATH, SCOP [2] and Pfam [3,4], with a special highlight on cognate molecules that are endogenous in living organisms for enzymes [21]. In addition, the Inferred Biomolecular Interactions Server (IBIS) [22] contains detailed description and classification of binding sites between small molecules and proteins. The interactions compiled in IBIS are integrated with the Conserved Domain Database (CDD) [5,6] and PubChem database [23,24], a protein domain annotation database and a chemical structure database, respectively. The above three databases, *i.e.* IBIS, CDD and PubChem, were used in this work to derive pairwise associations between small molecules and protein domains.

By analyzing these small molecule-protein domain interaction data, we identified promiscuous small-molecule ligands that bound to two or more protein domains, which

subsequently led to the generation of an inter-connected protein domain network. By analyzing this network, we found that many protein domains, despite belonging to various families, can bind common or similar ligands. Moreover, tightly connected domains were observed to form modules in the network, which often share similar biochemical mechanisms, or are involved in related biological pathways. This study provides a global view of the complex role of small molecules in biological systems and reveals a novel relationship among protein domains, complementary to the traditional classifications derived solely from protein sequences or structures. Meanwhile, the success of identifying potential targets for marketed drugs based on this network may shed light on network pharmacology study and systematic identification of novel targets for drug repurposing.

Methods

Physical interaction data for small molecules and protein domains

Three databases were used to derive the physical interactions between small molecules and protein domains, including IBIS [22] (updated Oct 25, 2011), CDD [5,6] (version 3.01) and PubChem [23,24]. IBIS contains binding site information of small molecules and proteins in PDB; the CDD database consists of both manually curated protein domain models and those imported from other resources, such as Pfam [3,4], SMART [25] and COG [26,27]; PubChem comprises standardized and validated chemical structures of small-molecule ligands, in which a Compound Identifier (CID) represents a unique chemical structure.

Identification of small molecule-protein domain interactions

For each small molecule-protein interaction, we mapped the binding sites obtained from IBIS to the domain footprint annotations provided by CDD. A flowchart of our approach is shown in Figure 1.

Firstly, we retrieved a total of 88,774 small molecule-protein interactions derived from IBIS, corresponding to 13,851 unique small molecule structures and 67,619 distinct protein sequences. Here, we used the IBIS criteria to define a small-protein interaction that five or more amino acid residues of a protein are within 4Å from its small-molecule ligand (heavy atom). We excluded the ‘non-biological’ interactions marked by IBIS, as most of them were resulted by auxiliary molecules, such as buffers, salts, detergents, solvents and ions, used for crystallization or purification. Moreover, we confined our study to the small molecules with two properties: (1) molecular weight between 100 and 800; (2) containing only organic elements (H, C, N, O, F, P, S, Cl, Br and I) in one covalent unit (*i.e.* non-mixtures). As a result, we obtained a dataset containing 11,582 unique small molecules and 51,594 protein sequences with accurate binding site information.

Secondly, we annotated each protein sequence obtained in the previous step with domain footprints, *i.e.* domain positions, by searching against CDD with default parameters. In the retrieved results, we selected the manually curated domain models (CDD accession starting with ‘cd’) ranked on the top of the hit list (if available); otherwise, we used the Pfam models (CDD accession starting with ‘pfam’). At last, we obtained 3,012 distinct protein domains in total. Additionally, we retrieved the superfamily information (CDD accession starting with ‘cl’) for these protein domains from CDD as well.

Thirdly, we mapped each small-molecule binding site obtained in the first step onto a specific protein domain (if possible), according to domain footprint annotations. A small molecule-protein domain interaction was determined if more than 75% of the contact residues were within the domain region. This process produced 13,822 small molecule-protein domain pairs, corresponding to 9,529 unique small-molecule structures and 2,125 distinct protein domains.

Drug and cognate molecules

Some of the small molecules obtained in the previous step are marketed drugs according to the DrugBank [28,29] annotations. These can be easily accessed through the PubChem CIDs, as DrugBank has deposited drug data into PubChem.

A cognate ligand is an endogenous small molecule in biological organisms. To identify such cognate molecules, we used a similar strategy reported by Bashton *et al.,* [20]: a small molecule was 'cognate' if it has a similar compound (with the Tanimoto coefficient above 0.9 by using the PubChem fingerprint [23,31]) in the KEGG Reaction database [30], which consists of detailed annotations of the biological reactions in organisms.

Protein domain network

To study the relationship of the small-molecule binding domains, we constructed a domain network (Figure 1), in which a node represents a protein domain, and an edge links two protein domains if they bind common or similar ligand(s). Here, we considered two ligands similar if their Tanimoto similarity is above 0.9, as calculated by using the PubChem fingerprint [23,31].

To characterize network properties, the following metrics were used:

Node degree (k_i) measures the number of edges connecting to node *i*.

Shortest path ($L_{i,j}$) is defined as the shortest distance or minimum number of steps between any two given nodes (*i* and *j*) over the domain network. The average shortest path (<*L*>) is a mean of the shortest paths of all possible node pairs.

Clustering Coefficient (C_i) is defined as $C_i=2n/k_i(k_i-1)$, where n denotes the number of edges connecting the nearest neighbors (k) of node i [32]. The value of C_i is equal to 1 for a node at the center of a fully interconnected cluster, while the value of 0 indicates a node in a loosely connected group. The average clustering coefficient (<*C*>) over all nodes of a network is a measure of the network's potential modularity.

The network was drawn by using Cytoscape [33,34] (version 2.81) and the network properties were calculated with the igraph library (version 0.5.4, http://igraph.sourceforge.net/).

Results

In this work, we compiled 13,822 small molecule-protein domain interactions (See Method), corresponding to 9,529 unique small molecules and 2,125 distinct protein domains. Originally, we identified 3,012 protein domains in total from these small-molecule binding proteins. Some proteins contained multiple domains and the domains (30%) that had no bound small-molecule ligand (see Method) were excluded in the following study.

Small molecule–protein domain interactions: a many to many relationship

We observed that the number of small-molecule ligands varied by each domain, with a ligand count of five on average. The overall distribution is shown in Figure 2A. The majority of the protein domains bound few small-molecule ligands; however, some domains interacted with hundreds of distinct small molecules, such as the trypsin-like serine protease domain (CDD accession: cd00190), carbonic anhydrase alpha I-II-III-XIII domain (CDD accession: cd03119) and HIV retropepsin domain (CDD accession: cd05482) (Table 1). In addition, we found that, although the small-molecule ligands of many protein domains spread over a wide range in chemical space, they have preferential zones in terms of physicochemical properties as indicated by the molecular weight and octanol-water partition

coefficient (Supplement figure S1-A). For example, the HIV retropepsin like domain (CDD accession: cd05482) tended to bind larger molecules (Supplement figure S1-B); while the trypsin-like serine protease domain was prone to bind relatively diverse ligands (Supplement figure S1-C).

On the other hand, we found that 1,168 out of the 9,529 small molecules, including drugs, were promiscuous because they bound to two or more protein domains. For an example, dexibuprofen (PubChem CID: 39912), a non-steroidal anti-inflammatory drug (NSAID), bound to both of the phospholipase A2 domain (PLA2c, CDD accession: cd00125) and albumin domain (CDD accession: cd00015). The overall distribution of the number of protein domains targeted by small molecules is shown in Figure 2B. It is worth noting that 73% (852) of the promiscuous small molecules were observed to bind multiple domains from different domain superfamilies. For instance, nicotinamide adenine dinucleotide phosphate (NADP, PubChem CID: 5886) bound to 103 distinct protein domains from over 20 domain superfamilies; and adenosine diphosphate (ADP, PubChem CID: 6022) interacted with as many as 191 protein domains, belonging to 57 superfamilies that are widely distributed in a biological system. Especially, 72% (842) of the total 1,168 promiscuous molecules were cognate (endogenous) molecules. These results demonstrate the versatility of small molecules, including cognate molecules and drugs, in regulating biological processes. Therefore, our analysis unveiled a many-to-many relationship between small molecules and protein domains, which led us to further investigate the relationship among protein domains as resulted from interacting with small-molecule ligands.

Pairwise protein domain associations

Based on the observation in the previous section, we noted that about 89% (1,883) of the 2,125 domains were associated with at least one other domain through binding common ligands, producing 79,160 domain pair associations. The rest 11% (242 domains) bound with “selective” ligands that interacted with only one single domain target observed in the current dataset, hence these domains did not demonstrate domain associations regarding to share common ligands. Surprisingly, among the domain pair associations, we found that 86% (67,976) of them were from different superfamilies. This clearly indicates that distinct protein domains may associated with each other in terms of small-molecule binding, despite of the differences in protein sequences or structures.

Furthermore, we investigated the strength of these domain associations. Intuitively, the more ligands sharing between two domains, the stronger the association is. In this study, we not only considered the number of common ligands, but also took similar ligands into account, as we noticed that certain ligands shared significant similarity in structure, such as ADP and adenosine triphosphate (ATP, PubChem CID: 5957). We set a similarity (Tanimoto coefficient) threshold of 0.90 to ensure high-quality domain associations identified. By incorporating ligand similarity, we observed a 6% increase in the number of domain associations identified.

For any two domains, the ligand structures of them were compared in pairwise. The number of similar ligand pairs, named NSLP score, was calculated to represent the strength of a domain association. By systematically evaluating the NSLP score for each domain pair, we found a great variation among the domain association strength (Figure 3). Some domain pairs from the same superfamily tended to have high NSLP scores. For example, the bacterial photosynthetic reaction center complex M domain (CDD accession: cd09291) and bacterial photosynthetic reaction center complex L domain (CDD accession: cd09290) had an NSLP score of 926, both of which belong to the photosynthetic reaction center superfamily (CDD accession: cl08220). Particularly, we observed that certain domain pairs from different superfamilies also had high NSLP scores, indicating considerable similarities

among their ligands. For instance, the nucleoside diphosphate kinase group I domain (CDD accession: cd04413) and canonical ribonuclease A domain (CDD accession: cd06265), despite that they belong to the nucleoside diphosphate kinase superfamily (CDD accession: cl00335) and ribonuclease A superfamily (CDD accession: cl00128), respectively, had an NSLP score of 151, with many being nucleotide derivative ligands. More examples of protein domain associations with high NSLP scores are listed in Table 2.

In fact, we found that the majority of the domain associations identified in the present study were across different superfamilies. Hence, we further investigated domain superfamily associations and their strength in the same way as that for the domain association study. As a result, a number of closely related superfamilies were identified, such as the P-loop NTPase superfamily (CDD accession: cl09099) and Rossmann-fold NAD(P)(+)-binding protein superfamily (CDD accession: cl09931) were associated with a NSLP score of 625. Additional examples of superfamilies with significantly strong associations regarding to small molecule binding are listed in Supplement table S1. This analysis demonstrates, to some extent, the deficiency of the conventional classifications based protein sequences or structures, because they cannot well represent such relationship resulted by small-molecule binding. Therefore, it indicates that our work on identifying protein domain associations based on small-molecule binding may complement the conventional approaches in protein family studies.

Protein domain network

In the previous analysis of pairwise domain associations, we not only identified closely related domains with regard to small-molecule binding, but also found some popular domains that were associating with many other domains through binding common or similar ligands. To characterize the global relationship among these protein domains, we built a domain network (see Method), consisting of 2,125 nodes (domains) and 181,145 edges (domain associations) in total. Among these nodes, about 95% (2,009) nodes connected to at least one neighboring node, named ‘connected’, while the rest 5% (116) nodes were singletons that had no edge linking to others, named ‘isolated’. Particularly, we observed that most ‘connected’ nodes (1,992) were in the giant component, the largest connected component of the network. These results suggest that the small-molecule binding domains are comprehensively associated with each other through binding small-molecule ligands.

Among the entire domain network, we observed a power-law like distribution of the node degrees (Figure 4), which indicates that the nodes with higher degree (“hub” nodes) had a lower frequency in general. For example, the canonical ribonuclease A domain (CDD accession: cd06265) and nucleoside diphosphate kinase group I domain (CDD accession: cd04413), connected to as many as 690 and 676 other domains (Supplement table S2 and S3), respectively. Moreover, the shortest path between any two nodes (domains) in the network was 2.9 on average, *i.e.* any two randomly selected domains were separated by less than three steps, which suggests a small-world property of the network [32,35].

Furthermore, we calculated the clustering coefficient [32] of each node and obtained an average value of 0.5 over the network, which implies potential modularity existing in the domain network. A domain module represents a group of domain nodes that are densely inter-connected within a group, but loosely connected to nodes outside the group. When looking into these domain modules, it is not surprising to observe that domains in such modules often shared a similar biochemical mechanism *in vivo* or belonged to the same superfamily. For example, the alpha carbonic anhydrase (CA) domains, including types I-II-III-X-III (CDD accession: cd03119), V (CDD accession: cd03118), IX (CDD accession: cd03150), XII–XIV (CDD accession: cd03126) and VII (CDD accession: cd03149) that catalyze CO_2 hydration to bicarbonate and protons in living organisms, formed a fully inter-

connected module in the network (the red module in Figure 5, referred as the CA module in this work) through binding acetazolamide, the first non-mercurial diuretic drug [36].

In addition, we also found that some domains within a module were involved in relevant biological processes. One such example was the blue module in Figure 5, which consisted of six protein domains including the PLA2c domain (CDD accession: cd00125), prostaglandin endoperoxide synthase domain (PES, CDD accession: cd09816), lipocalin domain (CDD accession: pfam00061), albumin domain (CDD accession: cd00015), the ligand binding domain of peroxisome proliferator-activated receptors (NR-LBD-PPAR, CDD accession: cd06932) and the ligand binding domain of hepatocyte nuclear factor 4 (NR-LBD-HNF4-like, CDD accession: cd06931). These domains were closely inter-connected in the network as they bound various fatty acids or derivatives. Especially, the PLA2c domain, PES domain, lipocalin domain and albumin domain had relatively stronger associations (higher NSLP scores) to each other, in which the first two domains were closely related to prostaglandin biosynthesis in arachidonic acid metabolism pathway and considered as main targets for NSAIDs; while, the latter two were responsible for transporting lipids, fatty acids and their metabolites *in vivo* [37,38]. More interestingly, the NR-LBD-HNF4-like domain was also identified in this module, which was recently 'deorphanized' because it could be regulated by fatty acids [39]. This result suggests that domains involved in relevant biological processes/pathways can be identified through the domain network analysis.

On the other hand, some domains involved in different pathways and superfamilies were also observed to form modules through binding common cognate molecules. For instance, the ligand binding domain of thyroid hormone receptors (NR-LBD-TR, CDD accession: cd06935), TLP-Transthyretin domain (CDD accession: cd05821) and the ligand binding domain of androgen receptors (NR-LBD-AR, CDD accession: cd07073) formed a three-node domain module (the green module in Figure 5), because they bound thyroid hormones, thyroxine (PubChem CID: 5819), triiodothyronine (PubChem CID: 5920) and a derivative, triac (PubChem CID: 5803). Despite of belonging to different superfamilies, the first two domains are known to participate in the thyroid hormone transportation and signaling process; while the NR-LBD-AR domain was recently reported to bind thyroid hormones [40]. In fact, some modules consisting of hundreds of domains, such as the NADP or ATP binding domains, were also observed. Thus, proteins containing these highly associated domains can be effectively regulated by few common molecules *in vivo*.

Notably, domain modules were often inter-connected to some extent, the three modules shown in Figure 5. Even within the fatty acids related module (colored in blue), we can clearly identify a sub-module consisting of the PLA2c domain, PES domain, albumin domain and lipocalin domain, which inter-connected to each other with strong associations. Indeed, these four domains were also observed in larger modules including the ATP related module and NADP related module. To characterize how the domains or domain modules were organized over the entire network, we investigated the distribution of clustering coefficient and node degree. For a node, the higher the clustering coefficient is, the more likely its neighbors are inter-connected. We found that the clustering coefficients were inversely proportional to the node degrees in general (Supplement figure S2), suggesting that the nodes within a module tend to have higher clustering coefficients, and the nodes with relatively lower clustering coefficients but higher degrees are responsible for integrating domain modules. Similar phenomenon was also observed in other networks that were in hierarchical organization [41–43].

In summary, these results indicate that small-molecule binding domains, sharing the same biochemical mechanism (or within one superfamily), being involved in relevant biological pathways, or binding common cofactors, can be identified in the network as domain

modules. The results reveal new relationships of protein domains, which may be hardly detected through conventional protein sequence or structure based approaches.

Protein domain associations for drug target identification

It is widely accepted that many marketed drugs are derived from natural products or known drugs [44–46]. Thus, it is of great interest to study whether the domain associations identified in this work can be used to infer potential drug targets for drug repurposing. Among the small molecule-domain interaction dataset, we found a total of 252 drug-domain pairs, corresponding to 147 marketed drugs and 135 protein domains (Supplement table S4). A domain network showing interactions between drugs and their protein domain targets was built, and a sub-network including the three domain modules discovered in the previous section is shown in Figure 6.

Based on this network, we successfully identified potential targets for some known drugs, which were retrospectively verified by literature search (shown in Figure 6). For example, in the fatty acids related module (colored in blue), we observed that three NSAIDs, *i.e.* dexibuprofen, indomethacin (PubChem CID: 3715) and diclofenac (PubChem CID: 3033), respectively interacted with several domains (solid lines in grey in Figure 6), including the PLA2c domain and PES domain. Considering the strong associations among domains in this module, one may be interested in repositioning these drugs to other domain members. Some of the predicted drug-domain associations were confirmed by literature mining (dashed line in green in Figure 6). For instance, diclofenac was reported to bind to NR-LBD-PPAR [47], albumin [48] and lipocalin [49]; and indomethacin was found binding to albumin as well [50]. Especially, it has been reported that the NR-LBD-PPAR domain contained proteins, such as peroxisome proliferator-activated receptor gamma, can be activated by many NSAIDs, including ibuprofen (PubChem CID: 3672) and flufenamic acid (PubChem CID: 3371) that produce adipogenesis and peroxisome activity *in vivo* [51]. Thus, we may anticipate more hidden interactions with NSAIDs to be discovered by conducting a systemic assay against all protein domains in this module. Likewise, ethoxzolamide (PubChem CID: 3295) could be successfully repositioned as a ligand for other member domains in the CA module (dashed green line in Figure 6), though it only bound to two domains according to the current dataset (solid grey line in Figure 6). In fact, this CA inhibitor can inhibit almost all CA isoforms in many tissues and organs, producing various inhibitory profiles and clinical applications [36].

Moreover, we could infer potential domain targets from neighboring modules. For instance, the TLP-Transthyretin domain (colored in green in Figure 6), which is responsible for transporting thyroid hormones and retinol in vertebrates, connected to several domains in the fatty acid module (colored in blue in Figure 6), though the associations were relatively weak compared to the ones within the modules. Several drugs, including levothyroxine and diflunisal, were found binding to both the fatty acid module (colored in blue in Figure 6) and the thyroid hormone related module (colored in green in Figure 6) based on the current network, hence it would be interesting to explore whether other drugs can bind to the domains across these two modules as well. From literatures, we found that flufenamic acid, a ligand of the PES domain from the fatty acid module [52], was able to bind to the NR-LBD-PPAR domain of the same module [51], as well as the other two domains, the NR-LBD-AR domain and TLP-Transthyretin domain, in the thyroid hormone related module (dashed line in green in Figure 6). In addition, a plant-derived naphthoquinone, shikonin (PubChem CID: 479503), which did not show interaction with either the fatty acid module or the thyroid hormone related module based on the current dataset, was reported to bind to both NR-LBD-TR domain contained receptors (PubChem AID: 1479) and PES domain contained receptors, including cyclooxygenase-1 and -2 (COX1 and COX2) [53] (dashed line in green in Figure 6). Furthermore, it has also been reported that NSAIDs indeed

compete with thyroid hormone binding *in vivo* [54,55]. Similarly, based on the observed connection between the two neighboring modules (CA module and fatty acid module) due to celecoxib (PubChem CID: 2662), a selective COX2 inhibitor with nanomolar activity against the carbonic anhydrase [56], we successfully verified a hidden interaction of alpha-CA-I-II-III-XIII domain with indomethacin, a ligand of the PES domain [57,58].

Our analysis indicates that additional drug targets may be suggested based on the modules from the domain interaction network. Thus, it demonstrates again that the constructed domain network can be used in drug target identification for drug repurposing.

Discussion

In the present study, we systematically investigated the protein domain associations from the small-molecule binding point of view on a large scale, based on the physical interactions extracted from the PDB structures. To the best of our knowledge, this is the first large-scale study on protein domain associations in with respect to small-molecule binding.

Conventionally, proteins and protein domains are classified into families or superfamilies according to the similarity in protein sequence, structure or biochemical reaction. Thus, proteins from the same family or superfamily are believed to have similar or relevant functions *in vivo.* But, the inter-relationship among families or superfamilies, especially regarding to their interactions with small molecules, has rarely been investigated. In this work, we identified a novel relationship that most small-molecule binding protein domains, despite distributing over different superfamilies, biological pathways, tissues or organs, were comprehensively associated through binding the same or similar small molecules. On the other hand, the development of systems biology provides great opportunities to interfere biological organisms on the system level, for example, modulating multiple targets for disease treatment. The approach in the present study can be used, not only to identify protein targets that are potentially modulatable by small molecules within a pathway, but also to detect the associations among these targets with respect to small-molecule interactions for the multiple-point control of a biological pathway. Notably, this strategy can also identify the inter-connections across biological pathways by using protein domain associations obtained in this study. Through interacting with such protein targets, the involved biological pathways may be affected or regulated by few small molecules including drugs, to generate various biological effects or pharmacological efficacies *in vivo*. Therefore, this study may provide a complementary insight into the complex biological systems from the small-molecule binding point of view, compared to the traditional approaches.

Moreover, the identification of comprehensive associations among small-molecule binding domains coincides with the fact that an increasing number of drugs are found to bind to multiple protein targets [59–61]. The concept of “one drug, one target, on disease” has dominated the field of drug discovery for years, and there has been a long-standing controversy over the count of drug targets in human genome [28,62–68]. Until recently, substantial evidences [69,70] have shown that many successfully marketed drugs, especially those for polygenic diseases (*eg.* cancer, cardiovascular diseases [60] and depression [71]) turn out to interact with multiple targets, though they were originally developed against a single or specific target [60,72]. The mechanisms of action of these drugs for curing polygenic diseases suggest that the count of drug targets may no longer be a substantial question, and the challenge is how to identify potential targets including anti-targets for known drugs, and how to combine multiple drug targets to produce a desirable therapeutic effect. The domain network constructed in this study, though based on an arguably limited dataset of the PDB structures, has demonstrated its capability of inferring potential targets

for marketed drugs. The present study may shed a light on systematic identification of drug targets for drug repurposing and network pharmacology.

In addition, the other side of the coin is that a considerable number of adverse drug reactions are due to drug interaction with unintended targets [73]. Similar to drug repurposing, the strategy reported in this study may be used to predict potential off-target interactions for drugs based on the domain binding profile, and to suggest a group of off-target candidates for drug safety evaluation in preclinical research. In the future studies, we will aim to build an interactive web service and a tool for researchers to explorer protein domain network with additional links to biological pathways, disease information and bioactive molecules including drugs available in public domains.

Conclusions

In this work, we studied the protein domain associations with respect to small-molecule binding on a large scale. Based on the physical interactions of small molecules and protein domains derived from the PDB structures, we characterized the pairwise domain associations, as well as the global relationship from a network point of view. The results indicate that protein domains are widely inter-connected through binding the same or similar small molecules, which can hardly be found via traditional protein sequence or structure based approaches. Most closely related domains further constituted domain modules in the network, through sharing similar mechanism, being involved in relevant biological processes/pathways, or binding common cofactors,. Moreover, using the domain associations identified in this study as guidance, we successfully inferred potential targets for marketed drugs and verified them by literature mining. Collectively, the results reported in the present study, not only provide an insight into the complex role of small molecules involved in biological systems, but also demonstrate a global view of protein domain inter-relationship for small-molecule bindings. The strategy used in this work may shed a light on network pharmacology study and target identification for drug repurposing, as well as chemogenomic research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

1. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure. 1997; 5:1093–1108. [PubMed: 9309224]
2. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995; 247:536–540. [PubMed: 7723011]
3. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A. The Pfam protein families database. Nucleic Acids Res. 2010; 38:D211–222. [PubMed: 19920124]
4. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 1997; 28:405–420. [PubMed: 9223186]
5. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S,

Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 2007; 35:D237–240. [PubMed: 17135202]

6. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita RA, Zhang D, Zhang N, Bryant SH. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 2009; 37:D205–210. [PubMed: 18984618]
7. Stockwell BR. Chemical genetics: ligand-based discovery of gene function. Nat Rev Genet. 2000; 1:116–125. [PubMed: 11253651]
8. Spring DR. Chemical genetics to chemical genomics: small molecules offer big insights. Chem Soc Rev. 2005; 34:472–482. [PubMed: 16137160]
9. Doddareddy MR, van Westen GJP, van der Horst E, Peironcely JE, Corthals F, Ijzerman AP, Emmerich M, Jenkins JL, Bender A. Chemogenomics: Looking at biology through the lens of chemistry. Statistical Analysis and Data Mining. 2009; 2:149–160.
10. Bleicher KH. Chemogenomics: bridging a drug discovery gap. Curr Med Chem. 2002; 9:2077–2084. [PubMed: 12470247]
11. Bredel M, Jacoby E. Chemogenomics: an emerging strategy for rapid target and drug discovery. Nat Rev Genet. 2004; 5:262–275. [PubMed: 15131650]
12. Yamanishi Y, Pauwels E, Saigo H, Stoven V. Extracting Sets of Chemical Substructures and Protein Domains Governing Drug-Target Interactions. J Chem Inf Model. 2011; 51:1183–1194.
13. Bender A, Mikhailov D, Glick M, Scheiber J, Davies JW, Cleaver S, Marshall S, Tallarico JA, Harrington E, Cornella-Taracido I, Jenkins JL. Use of ligand based models for protein domains to predict novel molecular targets and applications to triage affinity chromatography data. J Proteome Res. 2009; 8:2575–2585. [PubMed: 19271732]
14. Wang YY, Nacher JC, Zhao XM. Predicting drug targets based on protein domains. Mol Biosyst. 2012; 8:1528–1534. [PubMed: 22402667]
15. Kruger FA, Overington JP. Global analysis of small molecule binding to related protein targets. PLoS Comput Biol. 2012; 8:e1002333. [PubMed: 22253582]
16. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 2008; 36:D419–425. [PubMed: 18000004]
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
18. Bashton M, Thornton JM. Domain-ligand mapping for enzymes. J Mol Recognit. 2010; 23:194–208. [PubMed: 19810051]
19. Chalk AJ, Worth CL, Overington JP, Chan AW. PDBLIG: classification of small molecular protein binding in the Protein Data Bank. J Med Chem. 2004; 47:3807–3816. [PubMed: 15239659]
20. Bashton M, Nobeli I, Thornton JM. PROCOGNATE: a cognate ligand domain mapping for enzymes. Nucleic Acids Res. 2008; 36:D618–622. [PubMed: 17720712]
21. Bashton M, Nobeli I, Thornton JM. Cognate ligand domain mapping for enzymes. J Mol Biol. 2006; 364:836–852. [PubMed: 17034815]
22. Shoemaker BA, Zhang D, Thangudu RR, Tyagi M, Fong JH, Marchler-Bauer A, Bryant SH, Madej T, Panchenko AR. Inferred Biomolecular Interaction Server--a web server to analyze and predict protein interacting partners and binding sites. Nucleic Acids Res. 2010; 38:D518–524. [PubMed: 19843613]
23. Bolton E, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated Platform of Small Molecules and Biological Activities. Annu Rep Comput Chem. 2008; 4:217–241.
24. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 2009; 37:W623–633. [PubMed: 19498078]
25. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res. 2006; 34:D257–260. [PubMed: 16381859]

26. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997; 278:631–637. [PubMed: 9381173]
27. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003; 4:41. [PubMed: 12969510]
28. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006; 34:D668–672. [PubMed: 16381955]
29. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008; 36:D901–906. [PubMed: 18048412]
30. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 1999; 27:29–34. [PubMed: 9847135]
31. Wang Y, Bolton E, Dracheva S, Karapetyan K, Shoemaker BA, Suzek TO, Wang J, Xiao J, Zhang J, Bryant SH. An overview of the PubChem BioAssay resource. Nucleic Acids Res. 2010; 38:D255–266. [PubMed: 19933261]
32. Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks. Nature. 1998; 393:440–442. [PubMed: 9623998]
33. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc. 2007; 2:2366–2382. [PubMed: 17947979]
34. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–2504. [PubMed: 14597658]
35. Nacher JC, Schwartz JM. A global view of drug-therapy interactions. BMC Pharmacol. 2008; 8:5. [PubMed: 18318892]
36. Supuran CT. Carbonic anhydrases: novel therapeutic applications for inhibitors and activators. Nat Rev Drug Discov. 2008; 7:168–181. [PubMed: 18167490]
37. Varshney A, Sen P, Ahmad E, Rehan M, Subbarao N, Khan RH. Ligand binding strategies of human serum albumin: how can the cargo be utilized? Chirality. 2010; 22:77–87. [PubMed: 19319989]
38. Flower DR, North AC, Attwood TK. Structure and sequence relationships in the lipocalins and related proteins. Protein Sci. 1993; 2:753–761. [PubMed: 7684291]
39. Wisely GB, Miller AB, Davis RG, Thornquest AD Jr, Johnson R, Spitzer T, Sefler A, Shearer B, Moore JT, Willson TM, Williams SP. Hepatocyte nuclear factor 4 is a transcription factor that constitutively binds fatty acids. Structure. 2002; 10:1225–1234. [PubMed: 12220494]
40. Estebanez-Perpina E, Arnold LA, Nguyen P, Rodrigues ED, Mar E, Bateman R, Pallai P, Shokat KM, Baxter JD, Guy RK, Webb P, Fletterick RJ. A surface on the androgen receptor that allosterically regulates co-activator binding. Proc Natl Acad Sci U S A. 2007; 104:16074–16079. [PubMed: 17911242]
41. Clauset A, Moore C, Newman ME. Hierarchical structure and the prediction of missing links in networks. Nature. 2008; 453:98–101. [PubMed: 18451861]
42. Ravasz E, Barabasi AL. Hierarchical organization in complex networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2003; 67:026112. [PubMed: 12636753]
43. Ravasz E. Detecting hierarchical modularity in biological networks. Methods Mol Biol. 2009; 541:145–160. [PubMed: 19381526]
44. Harvey AL. Natural products in drug discovery. Drug Discov Today. 2008; 13:894–901. [PubMed: 18691670]
45. de Sa Alves FR, Barreiro EJ, Fraga CA. From nature to drug discovery: the indole scaffold as a ‘privileged structure’. Mini Rev Med Chem. 2009; 9:782–793. [PubMed: 19519503]

46. Newman DJ, Cragg GM. Natural products as sources of new drugs over the 30 years from 1981 to 2010. J Nat Prod. 2012; 75:311–335. [PubMed: 22316239]
47. Kaur J, Sanyal SN. Modulation of inflammatory changes in early stages of colon cancer through activation of PPARgamma by diclofenac. Eur J Cancer Prev. 2010; 19:319–327. [PubMed: 20485182]
48. Chamouard JM, Barre J, Urien S, Houin G, Tillement JP. Diclofenac binding to albumin and lipoproteins in human serum. Biochem Pharmacol. 1985; 34:1695–1700. [PubMed: 4004886]
49. Chuang S, Velkov T, Horne J, Porter CJ, Scanlon MJ. Characterization of the drug binding specificity of rat liver fatty acid binding protein. J Med Chem. 2008; 51:3755–3764. [PubMed: 18533710]
50. Bogdan M, Pirnau A, Floare C, Bugeac C. Binding interaction of indomethacin with human serum albumin. J Pharm Biomed Anal. 2008; 47:981–984. [PubMed: 18495406]
51. Lehmann JM, Lenhard JM, Oliver BB, Ringold GM, Kliewer SA. Peroxisome proliferator-activated receptors alpha and gamma are activated by indomethacin and other non-steroidal anti-inflammatory drugs. J Biol Chem. 1997; 272:3406–3410. [PubMed: 9013583]
52. Sogawa S, Nihro Y, Ueda H, Izumi A, Miki T, Matsumoto H, Satoh T. 3,4-Dihydroxychalcones as potent 5-lipoxygenase and cyclooxygenase inhibitors. J Med Chem. 1993; 36:3904–3909. [PubMed: 8254620]
53. Landa P, Kutil Z, Temml V, Vuorinen A, Malik J, Dvorakova M, Marsik P, Kokoska L, Pribylova M, Schuster D, Vanek T. Redox and non-redox mechanism of *in vitro* cyclooxygenase inhibition by natural quinones. Planta Med. 2012; 78:326–333. [PubMed: 22174077]
54. Bishnoi A, Carlson HE, Gruber BL, Kaufman LD, Bock JL, Lidonnici K. Effects of commonly prescribed nonsteroidal anti-inflammatory drugs on thyroid hormone measurements. Am J Med. 1994; 96:235–238. [PubMed: 8154511]
55. Barlow JW, Curtis AJ, Raggatt LE, Loidl NM, Topliss DJ, Stockigt JR. Drug competition for intracellular triiodothyronine-binding sites. Eur J Endocrinol. 1994; 130:417–421. [PubMed: 8162174]
56. Weber A, Casini A, Heine A, Kuhn D, Supuran CT, Scozzafava A, Klebe G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. J Med Chem. 2004; 47:550–557. [PubMed: 14736236]
57. Puscas I, Ifrim M, Maghiar T, Coltau M, Domuta G, Baican M, Hecht A. Indomethacin activates carbonic anhydrase and antagonizes the effect of the specific carbonic anhydrase inhibitor acetazolamide, by a direct mechanism of action. Int J Clin Pharmacol Ther. 2001; 39:265–270. [PubMed: 11430635]
58. Puscas I, Coltau M, Pasca R. Nonsteroidal anti-inflammatory drugs activate carbonic anhydrase by a direct mechanism of action. J Pharmacol Exp Ther. 1996; 277:1464–1466. [PubMed: 8667211]
59. Mestres J, Gregori-Puigjane E, Valverde S, Sole RV. The topology of drug-target interaction networks: implicit dependence on drug properties and target families. Mol Biosyst. 2009; 5:1051–1057. [PubMed: 19668871]
60. Frantz S. Drug discovery: playing dirty. Nature. 2005; 437:942–943. [PubMed: 16222266]
61. Hopkins AL, Mason JS, Overington JP. Can we rationally design promiscuous drugs? Curr Opin Struct Biol. 2006; 16:127–136. [PubMed: 16442279]
62. Imming P, Sinning C, Meyer A. Drugs, their targets and the nature and number of drug targets. Nat Rev Drug Discov. 2006; 5:821–834. [PubMed: 17016423]
63. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov. 2006; 5:993–996. [PubMed: 17139284]
64. Golden JB. Prioritizing the human genome: knowledge management for drug discovery. Curr Opin Drug Discov Devel. 2003; 6:310–316.
65. Zheng C, Han L, Yap CW, Xie B, Chen Y. Progress and problems in the exploration of therapeutic targets. Drug Discov Today. 2006; 11:412–420. [PubMed: 16635803]
66. Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov. 2002; 1:727–730. [PubMed: 12209152]

67. Drews J. Drug discovery: a historical perspective. Science. 2000; 287:1960–1964. [PubMed: 10720314]

68. Drews J, Ryser S. Classic drug targets. Nat Biotechnol. 1997; 15:1350–1350.

69. Peterson RT. Chemical biology and the limits of reductionism. Nat Chem Biol. 2008; 4:635–638. [PubMed: 18936741]

70. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. Nat Biotechnol. 2009; 27:157–167. [PubMed: 19204698]

71. Roth BL, Sheffler DJ, Kroeze WK. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. Nat Rev Drug Discov. 2004; 3:353–359. [PubMed: 15060530]

72. Metz JT, Hajduk PJ. Rational approaches to targeted polypharmacology: creating and navigating protein-ligand interaction networks. Curr Opin Chem Biol. 2010; 14:498–504. [PubMed: 20609615]

73. Pouliot Y, Chiang AP, Butte AJ. Predicting adverse drug reactions using publicly available PubChem BioAssay data. Clin Pharmacol Ther. 2011; 90:90–99. [PubMed: 21613989]

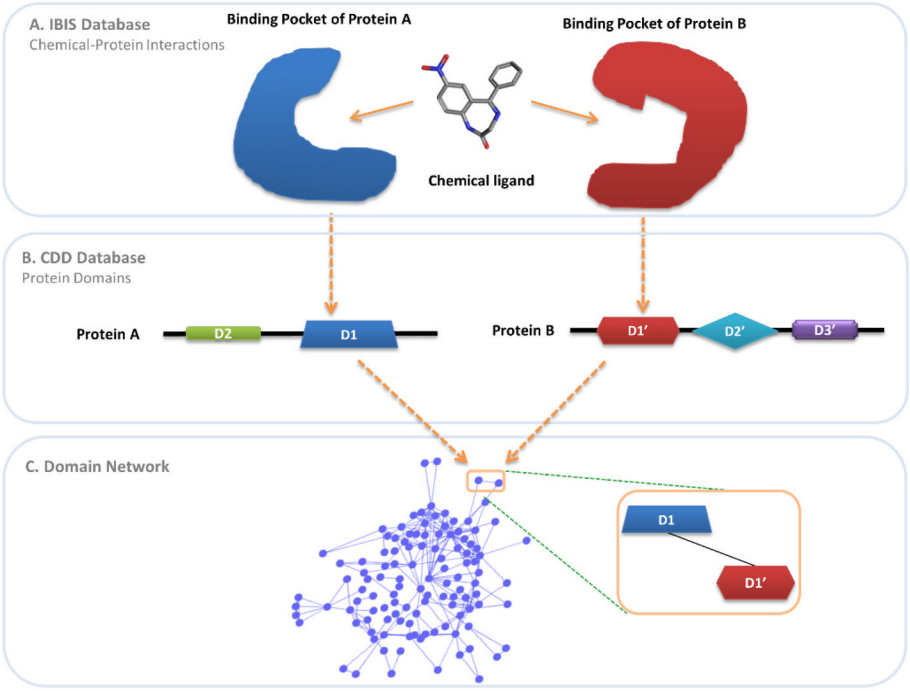

Figure 1. A flowchart illustration of the study on small molecule-protein domain interactions in this study

(A). First, we extracted binding site information of small molecule-protein interactions from IBIS. **(B)** Next, we mapped the binding sites onto the protein domain footprints based on the annotations in the CDD database, to obtain the physical small molecule-protein domain interactions. **(C).** Finally, protein domain inter-associations were constructed and studied based on the small molecules binding to them.

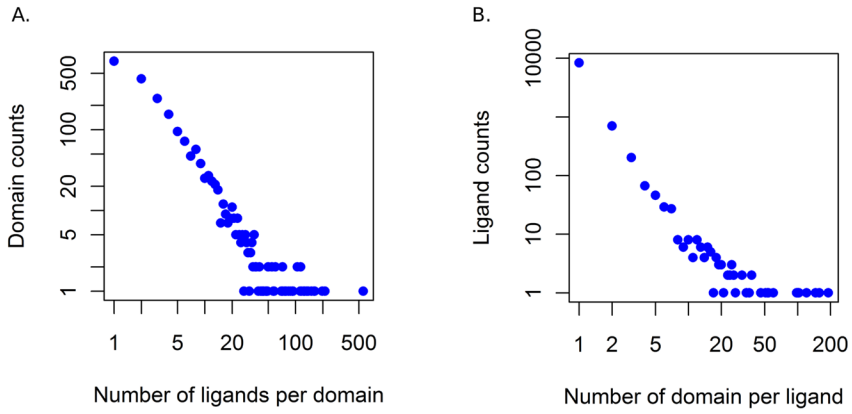

Figure 2. Small-molecule ligand and protein domain associations

(A). Distribution of the number of chemical ligands for protein domains. The majority of domains were associated with one or a few ligands, while a small fraction of domains were targeted by a larger number of ligands. **(B).** Distribution of the number of protein domain targets for small-molecule ligands. Despite that most ligand interacted with one domain targets, a considerable number of ligands bound to two or more domains, *i.e.* promiscuous ligands.

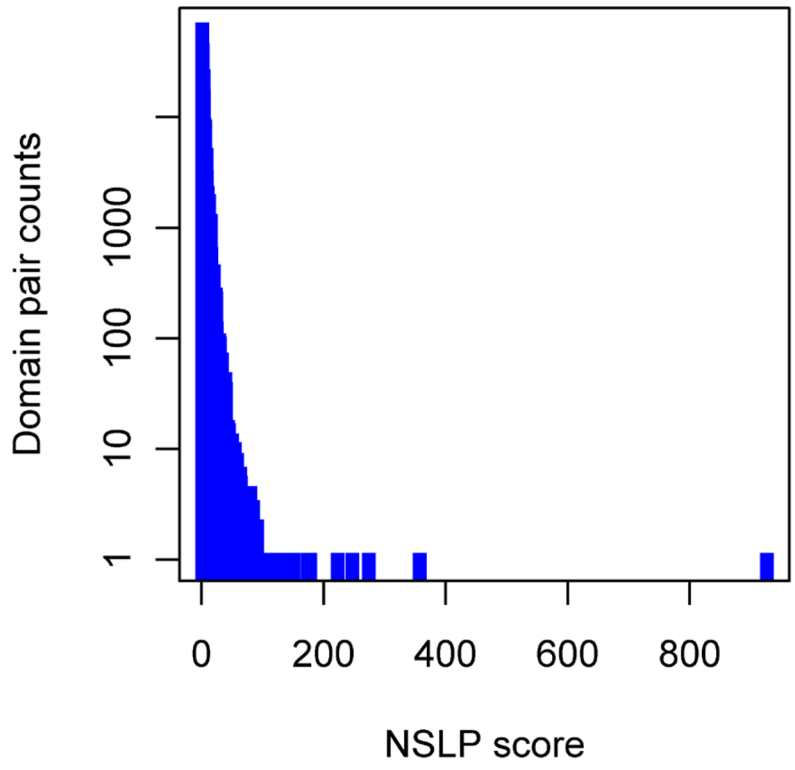

Figure 3. Distribution of the NSLP (number of similar ligand pair) scores between protein domain pairs

It shows that a considerable number of protein domain pairs have a NSLP score above 100, indicating stronger associations between these domains with respect to small molecule binding.

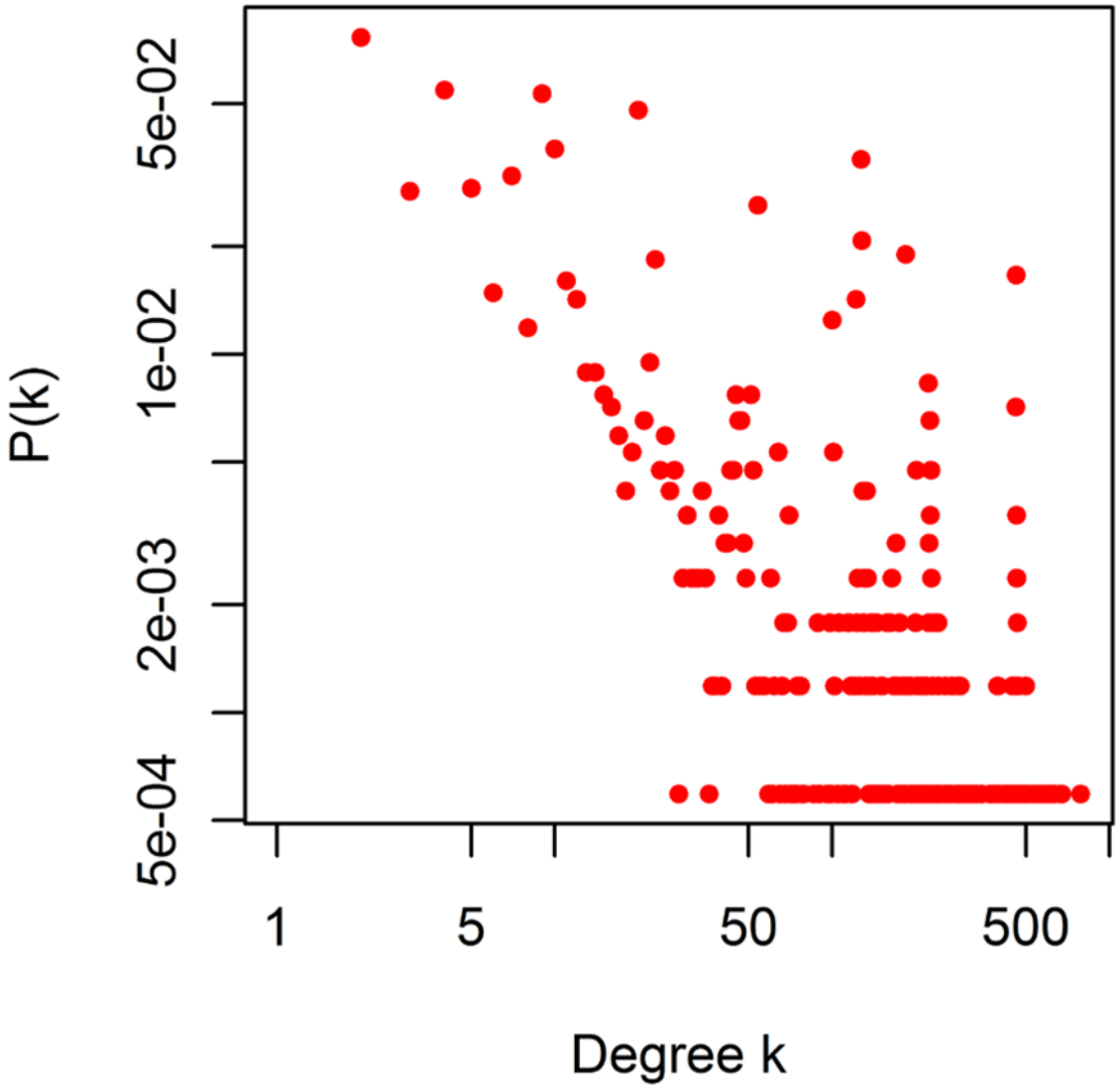

Figure 4. Degree distribution of the protein domain network
P(*k*) denotes the fraction of the nodes with degree (*k*).

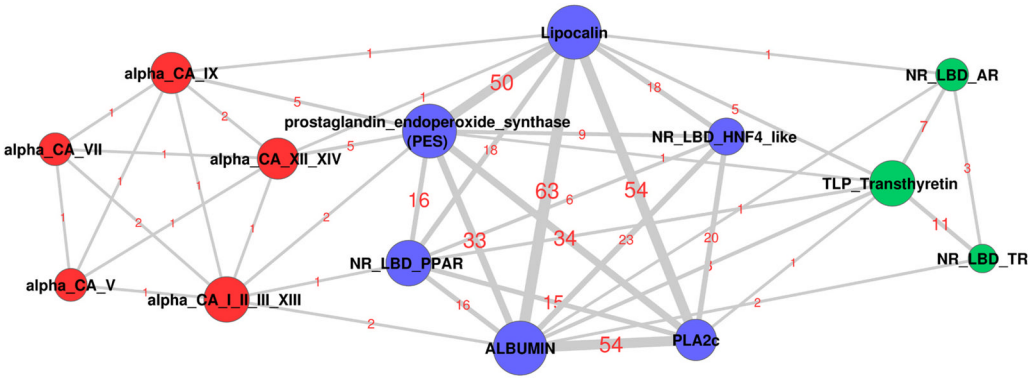

Figure 5. Three selected protein domain modules

The alpha carbonic anhydrase domains, fatty acids related domains and thyroid hormone related domains are colored in red, blue, and green, respectively. An edge connecting two domains represents an association identified based on their binding ligands. The NSLP scores are shown on the edges, and each edge is rendered in proportional to the respective NSLP score.

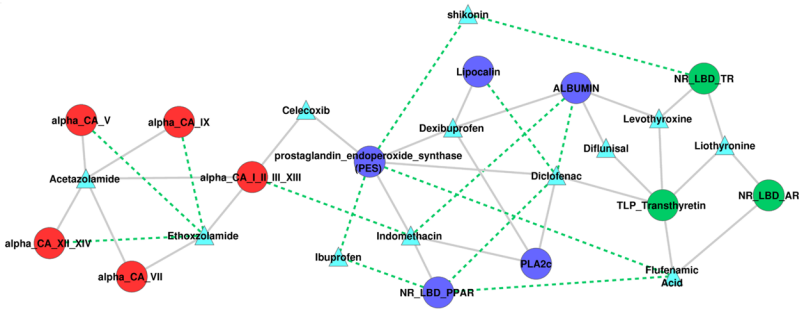

Figure 6. Drug target identification based on protein domain and drug interaction network

The circle nodes and triangle nodes represent protein domains and drugs, respectively. The circle nodes colored in red, blue, green denote protein domains comprising the alpha carbonic anhydrase (CA) domain module, the fatty acids related domain module (five nodes with drug binding shown), and the thyroid hormone related domain module, respectively. An edge connecting a drug and protein domain represents a direct interaction observed between them. The solid grey edges indicate observed interactions of small molecules and protein domains originally derived from PDB structures; the dashed green edges are predicted interactions that are retrospectively confirmed in literatures.

Table 1

Top 20 protein domains binding multiple smallmolecule ligands.

No.	CDD accession	Domain Name	Ligand Count
1	cd00190	Tryp_SPc	559
2	cd03119	alpha_CA_I_II_III_XIII	211
3	cd05482	HIV_retropepsin_like	198
4	cd07860	STKc_CDK2_3	163
5	cd00180	PKc	160
6	cd05473	beta_secretase_like	150
7	cd04300	GT1_Glycogen_Phosphorylase	138
8	pfam02518	HATPase_c	137
9	cd07851	STKc_p38	133
10	cd00209	DHFR	127
11	pfam00061	Lipcolin	125
12	cd04981	IgV_H	123
13	cd00795	NOS_oxygenase_euk	117
14	cd06932	NR_LBD_PPAR	114
15	cd04278	ZnMc_MMP	114
16	cd00312	Esterase_lipase	113
17	cd02248	Peptidase_C1A	105
18	cd00134	PBPb	105
19	cd05123	STKc_AGC	94
20	cd00047	PTPc	92

J Proteome Sci Comput Biol. Author manuscript; available in PMC 2013 June 04.

Table 2

A selected list of protein domain pairs binding to same/similar small molecules.

No.	Domain Name	Domain* (Superfamily)	Domain Name	Domain* (Superfamily)	Ligand pair count
1	Photo-RC_L	cd09290 (cl08220)	Photo-RC_M	cd09291 (cl08220)	926
2	Cytochrome_b_N	cd00284 (cl00859)	Photo-RC_L	cd09290 (cl08220)	357
3	Cytochrome_b_N	cd00284 (cl00859)	Photo-RC_M	cd09291 (cl08220)	274
4	AAT_like	cd00609 (cl00321)	OAT_like	cd00610 (cl00321)	247
5	The Sema domain	cd00925 (cl00017)	Cyt_c_Oxidase_III	cd01665 (cl00211)	223
6	The Sema domain	cd00925 (cl00017)	Cyt_c_Oxidase_I	cd01663 (cl00275)	178
7	Cyt_c_Oxidase_I	cd01663 (cl00275)	Cyt_c_Oxidase_III	cd01665 (cl00211)	174
8	NDPk_I	cd04413 (cl00335)	RNase_A_canonical	cd06265 (cl00128)	151
9	Cyt_c_Oxidase_III	cd01665 (cl00211)	UCR_TM	pfam02921 (cl03782)	139
10	The Sema domain	cd00925 (cl00017)	UCR_TM	pfam02921 (cl03782)	135
11	AAT_like	cd00609 (cl00321)	CGS_like	cd00614 (cl00321)	130
12	Photosystem-II_D2	cd09288 (cl08220)	Photosystem-II_D1	cd09289 (cl08220)	117
13	AAT_like	cd00609 (cl00321)	KBL_like	cd06454 (cl00321)	111
14	NT_POLXc	cd00141 (cl11966)	NDPk_I	cd04413 (cl00335)	109
15	PKc	cd00180 (cl09925)	NDPk_I	cd04413 (cl00335)	108
16	PLPDE_III_AR	cd00430 (cl00261)	AAT_like	cd00609 (cl00321)	107
17	Photo-RC_L	cd09290 (cl08220)	YceI-like domain	pfam04264 (cl01001)	106
18	SQR_TypeC_SdhC	cd03499 (cl00881)	Photo-RC_L	cd09290 (cl08220)	105
19	IgV_L_kappa	cd04980 (cl11960)	IgV_H	cd04981 (cl11960)	101
20	NT_POLXc	cd00141 (cl11966)	RNase_A_canonical	cd06265 (cl00128)	99

Domain*: CDD accession.