Assessing Drug Target Association using Semantic Linked Data

Bin Chen[1], Ying Ding[2], David J. Wild[1,*]
1 School of Informatics and Computing, Indiana University, Bloomington, IN, USA
2 School of Information Science and Library, Indiana University, Bloomington, IN, USA
∗ E-mail: djwild@indiana.edu

Abstract

The rapidly increasing amount of public data in chemistry and biology provides new opportunities for large-scale data mining for drug discovery. Systematic integration of these heterogeneous sets and provision of algorithms to data mine the integrated sets would permit investigation of complex mechanisms of action of drugs. In this work we integrated and annotated data from public datasets relating to drugs, chemical compounds, protein targets, diseases, side effects and pathways, building a semantic linked network consisting of over 290,000 nodes and 720,000 edges. We developed a statistical model to assess the association of drug target pairs based on their relation with other linked objects. Validation experiments demonstrate the model can correctly identify known direct drug target pairs with high precision. Indirect drug target pairs (for example drugs which change gene expression level) are also identified but not as strongly as direct pairs. We further calculated the association scores for 157 drugs from 10 disease areas against 1683 human targets, and measured their similarity using a 157×1683 score matrix. The similarity network indicates that drugs from the same disease area tend to cluster together in ways that are not captured by structural similarity, with several potential new drug pairings being identified. This work thus provides a novel, validated alternative to existing drug target prediction algorithms. The web service is freely available at: http://chem2bio2rdf.org/slap.

Author Summary

Modern drug discovery requires the understanding of chemogenomics, the complex interaction of chemical compounds and drugs with a wide variety of protein target and genes in the body. A large amount of data pertaining to such relationships exists in publicly-accessible datasets but it is siloed and thus impossible to use in an integrated fashion. In this work we have integrated and semantically annotated a large amount of public data from a wide range of databases, including compound-gene, drug-drug, protein-protein, drug-side effects and so on, to create a complex network of interactions relating to compounds and protein targets. We developed a statistical algorithm called Semantic Link Association Prediction (SLAP) for predicting "missing links" in this data network: i.e. compound-target interactions for which there is no experimental data but which are statistically probable given the other relationships that exist in this set. We present validation experiments which show this method works with a high degree of accuracy, and also demonstrate how it can be used to create a drug similarity network to make predictions of new indications for existing drugs.

Introduction

Understanding the interaction of drugs with multiple targets can identify potential side effects and toxicities [1–3], as well as identifying possible new applications of existing drugs [4–8]. Many efforts have been made to integrate drug-target interactions in a large scale [9–12]. A variety of computational approaches have been previously explored for predicting drug-target interactions, including molecular docking [3,13,14], ligand-based predictive models [15,16], phenotype similarity (side effect similarity [17] or gene expression profile similarity [18]) and chemical ontology similarity [19]. Some similarity measurements have been combined to elucidate drug targets [20]. Network analysis based on the topology of

known drug target network has also been utilized for drug target prediction, but is currently limited to small data sets [21, 22].

Recent advances in the Semantic Web [23] have enabled the creation of large heterogeneous networks of experimental and other data in life sciences (for example: Chem2Bio2RDF [24], LODD [25], Bio2RDF [26], OpenPHACTS (openphacts.org), Linked life data (linkedlifedata.com), Linked Open Data (linkeddata.org)), where the nodes can include physical and abstract entities (compounds, protein targets, substructures, side effects, diseases, pathways, tissues, gene ontology terms and so on), and the edges (or links) represent various relations between objects such as drug-drug interactions, and drug target interactions, protein-protein interactions and so on. The ability to easily integrate heterogeneous datasets in a meaningful fashion makes semantic technologies attractive, although it is only recently that supporting technologies have adequately matured to make them useful in the biological sciences: in particular the advent of fast triple stores for data storage, the SPARQL query language (http://www.w3.org/TR/rdf-sparql-query/) for searching, and the OWL ontology language (http://www.w3.org/TR/owl-features/) for the description of ontologies. Despite remaining deficiencies which are being addressed in the Semantic Web community (including difficulty weighting edges and maintaining provenance information) there are now many examples of successful use of semantics in the life sciences [27]. In contrast to hyperlinked data, semantic linked data encodes explicit meanings of nodes and links, allowing traversing from one node to another via particular kinds of relationship. Prediction of links not in the dataset, based on the existing links, is widely used in social networking, in which it is assumed that two nodes are similar if they share similar topology (e.g., a certain number of neighbors, and similar shortest paths) [28–30]. For example, in a coauthorship network, two authors are similar in terms of research interests if they coauthor lots of papers, hence their potential collaboration could be predicted (it should be noted that social networks generally only deal with positive relationships; drug discovery data is different in that negative relationships such as inactivity are important).

In this work, we sought to use such semantic methods to integrate and annotate the data in relation to drug target interaction, constructing a heterogeneous network composed by over 290k nodes and 720k edges. We further developed a statistical model called Semantic Link Association Prediction (SLAP) to assess the association of drug target pairs and to predict missing links. An association score is calculated based on the topology and semantics of the neighborhood. We demonstrate that SLAP can correctly identify known drug target pairs from random pairs with high accuracy and can also identify indirect drug target relations (e.g., the change of gene expression level). The association scores of a drug against a set of targets constitute a biological signature that allows assessing the similarity of drugs in the context of the whole system. The resulting drug similarity network clusters drugs from the same therapeutic indication in ways not observed using chemical structure similarity, and can also be used to identify potential new indications for existing drugs.

Results

Semantic Linked Data

The SLAP pipeline is shown in Fig 1. A heterogeneous network consisting of 295,897 nodes and 727,997 edges was constructed from 17 public data sources pertaining to drug target interaction. Every node and edge was semantically annotated using a systems chemical biology/chemogenomics ontology previously developed in our labs (unpublished). The nodes were grouped into 10 classes which are linked by 12 types (Fig. 1b). A single node is an instance of a corresponding class, for example: a node for the drug Troglitazone (labeled as 5591 in Fig. 2) is an instance of class *Chemical Compound*. We term paths of nodes and edges that share the same semantics (but different data) path patterns - each path is an instance of a path pattern. Table 1 shows 6 path pattern examples between *Drugs* and *Targets*. In Fig. 2, the path from node 5591 (Troglitazone) to node PPARG (Glitazone receptor) via ACSL4 (Long-chain-

fatty-acid CoA ligase 4) and 446284 (Eicosapentaenoic acid) is an instance of the path pattern 1 in table 1. We can interpret this path as indicating Troglitazone could bind to ACSL4 which shares compound Eicosapentaenoic acid with target PPARG. With the assumption that two nodes are associated if they link to at least one other node, or their linked nodes are linked, their relations can be assessed by the analysis of the links (or paths) between the two nodes [31]. The strength of their relation in the network can be measured by the distance, the number of shortest paths and other topological properties between the two nodes. In our example of the relationship between Troglitazone and target PPARG, several paths provide "evidence" of a relationship: Troglitazone and Rosiglitazone both are hypoglycemic drugs and the latter is the ligand of PPARG; Troglitazone binds to ACSL4 which shares pathway(PPAR signaling pathway), ligand (Eicosapentaenoic acid) and GO term (response to nutrient) with PPARG. A total of 1684 paths (length$\leq$ 3) belonging to 10 path patterns contribute to their relation.

Pattern score distribution

Each path between two nodes may contribute to the relation between them, but the degree of contribution varies depending on path distance and the weight of the edges involved in the path. For example, a gene ontology molecular function term (GO:0005515) shared by proteins is not as informative as a binding term (GO:0005488) in assessing the similarity of two proteins. Thus the weight of the edge linking one protein node to the molecular function node is lower than that linking to the binding node. According to this observation, we developed a statistical model to measure the weight of edges as well as the significance of paths (see methods). The model takes into account the distance and the weight of each edge, and renders a raw score indicating the strength of each path. We found that the raw scores within the same path pattern are normally distributed, while the mean and standard deviation of patterns are different (Fig. S1). Z scores converted from raw scores based on pattern score distribution are used to measure the contribution to the association: the higher the z score, the more contribution the path has. The sum of z scores of all paths is defined as association score indicating the association strength of the drug target pair. The logarithm of association scores of random drug target pairs fit to a normal distribution (Fig. S2), that enables calculation of the significance of a given association score. For our Troglitazone & PPARG example, the p value is 3.35E-5, indicating a strong association.

Pattern importance

A low p value between a drug-target pair indicates a strong probability of association between the drug and target, but it does not necessarily mean the drug and target would interact biologically. Some patterns may be uninformative. We therefore considered each pattern as a feature and assessed each feature alone for its ability to identify drug-target pairs from random pairs across the set. Table 1 lists three informative patterns and three uninformative patterns along with ROC scores. The first two patterns illustrate the drug likely interacts with a protein that shares commonalities in terms of GO or ligand binding profile with an existing target that the drug already is known to interact with. The third pattern indicates that the drug likely interacts with a protein with which another structural similar drug could interact. As a result of this analysis, 12 "uninformative" patterns were removed. The sum of z score of a given pair is the sum of z scores of the paths belonging to the informative patterns.

Association scores of drug target pairs

We randomly selected 1000 known drug target pairs from DrugBank and compared their association scores with 1000 random pairs of drugs and targets sampled from DrugBank. For each drug target pair, their direct link was removed in the score calculation so that their association is only determined by their neighborhood properties. We thus aimed to test the ability of SLAP to correctly identify "missing links" in the data, with the assumption that this might be used, for instance, to profile a group of compounds

against an identified set of targets. As Fig. 3 shows, random pairs have a broad range of scores, but most of them are close to zero. Overall, real drug-target pairs have much higher scores than random pairs (p-value<2.2e-16 using paired t test). We also took all drug target pairs from DrugBank (in total 5607 pairs in which 4508 pairs have at least one path with length ≤ 3). We sampled the same number of random drug target pairs as decoys to check the capability of identifying real drug target pairs by SLAP. We compared SLAP with other link prediction methods adopted in social network analysis [31]. The ROC of SLAP is 0.92, outperforming other methods (i.e., the number of shortest paths, and the number of valid paths)(Fig. 4). As the ratio between true drug target pairs versus random pairs decreases (e.g., ratio=1/12), the ROC scores do not vary very much (ROC score$\approx$0.92) and SLAP still performs much better than others, although the precision goes down considerably (Fig. S5). Even when random pairs are 12 times more than positive pairs, the precision still can reach 0.6 while recall is 0.7. In addition, we noticed using the sum (or max or mean) of raw score of the shortest path (without converting into z scores) performs as a random choice, indicating the importance of introducing random samples. Since several drug target prediction approaches reported that the performances may vary among different target classes [32], we grouped the drug target pairs into 5 classes (Enzyme, Membrane Receptor, Ion Channel, Transporter and Transcription Factor), and found that the score does not have any preference to a particular target class, indicating SLAP is capable of treating different classes of protein targets(Fig. S4).

As far as we are aware, SLAP is the only large predictive network model that has been applied to drug discovery data. However other drug-target prediction methods have been the subject of recent publications [7, 17, 33], and we thus sought to consider how the effectiveness of SLAP compares with these methods. We ran SLAP against 23 drug target pairs (including 15 aminergic G-protein-coupled receptors and 8 cross-boundary targets) predicted and confirmed in using the SEA method [7], a novel drug prediction method based on similarity analysis. 9 pairs of aminergic GPCRs were identified by SLAP (p value<0.05); 1 pair was not decided (p value>0.05); the rest of GPCRs have no mappings in the network (the drug was not found in the network), while only one of eight cross-boundary targets was identified by SLAP (see supplemental table), indicating that, SLAP is not capable of finding surprising pairs (cross-boundary targets). For example, Vadilex, an ion channel drug was predicted in SEA as a ligand of a transporter, a totally different target, but was not identified by SLAP. Nevertheless, SLAP performs considerately well among GPCRs in this case.

In addition, we examined drug target pairs from MATADOR [34] which serves as an external dataset for validation. 1065 direct pairs were collected, of which 444 pairings are not represented in our network. 560 out of 621 known pairs and 170 out of 444 unknown drug target pairs were identified by SLAP (p value<0.05).

Comparison with Connectivity Maps

By calculating association scores across multiple targets, SLAP can be used to build a polypharmacology profile of a drug even when a full data matrix is not available from drug-target experiments. We took all the 164 small molecules from the Connectivity Map (CMap), an online dataset mapping relationships of disease profiles to known drugs [18], and 113 molecules that were mapped to our network were used to build a library. The association scores of these compounds against 1683 targets were calculated, yielding a 113×1683 score matrix. The targets of which max score is <113 (p value<0.01) were eliminated so that each remaining protein is a target of at least one drug. After this filtering, a matrix composed by 113 compounds and 679 targets was built. We used the signature of a given drug to compare it with all the compounds in the library to find the most similar drugs according to Pearson correlation coefficient. Following the CMap approach, 8 queries including 2 HDAC inhibitors, 1 estrogen and 5 Phenothiazines were created and the similar pairs are listed in supplemental table. We set 0.75 as threshold. 21 pairs were identified by SLAP, 19 out of 21 pairs were actually the pairs identified by CMap. SLAP recovered all HDAC inhibitors, but missed two hits (genistein and tamoxifen) for estrogen, however, both hits rank

very high. Two Phenothiazines were not recovered using this similarity threshold, but they are quite similar with the other three Phenothiazines compared to the remaining compounds in the library. The results show that most of hits identified by SLAP are true positive, indicating that the profiles derived from SLAP resemble gene expression profiles being used for target identification .

Assessing drug similarity from biological function

We took 157 drugs from 10 disease areas to determine whether SLAP is able to distinguish drugs from different therapeutic areas. For each drug, we ran SLAP against 1863 human targets and got an association score for each drug target pair. At the end a 157×1683 score matrix was created. We only kept the drugs and targets in which the max score is ≥ 113(p value<0.01) to make sure each drug has at least one valid target and each target has at least one valid drug. The matrix was then reduced to 147×339, followed by the correlation calculation of every drug pairs. Only pairs with coefficient>0.9 were taken to build a network(see methods).

Identifying mechanisms of action: Drugs with the same therapeutic indication tend to cluster together (Fig. 5), and we also found that these subcluster by mechanism of action. For example, hypertension drugs, subcluster into ACE inhibitors, thiazide-based diuretics, angiotensin II antagonists, antiadreneric agents and beta(1)-adrenergic receptors (clusters 1-5 in Fig. 5 respectively).

Calculating similarity of drugs by biological function: Mostly, chemically similar drugs have similar biological function. However, small changes of structure may also result in big change of function, or even totally different indications. For example, adding a methyl group to Levodopa, a dopaminergic agent for Parkinson's disease, makes it Methyldopa, an antiadrenergic (Tanimoto coefficient=0.89; Fig. S6b) for antihypertension. They are distinguished by SLAP (similarity<0.3). The antihypertensive effect of Methyldopa is likely due to its metabolism to alpha-methylnorepinephrine (CID:3917). SLAP is still able to distinguish its metabolite from Levodopa (similarity=0.23). Conversely, biologically similar drugs identified by SLAP are not necessarily structural similar. For example, the drugs treating insomnia are quite different in term of structure(Fig. S6a), but they are clustered together by SLAP.

Drug repurposing: Some drugs with very different indications are clustered together. This may suggest some new indications of drugs or possible side effect considerations. For example, Butalbital, a Barbiturate used to treat Migraines, is clustered with nine Insomnia drugs, two of which (Butibarbital and SecoBarbital) are Barbiturates. Barbiturates act as central nervous system depressants, capable of producing all levels of CNS mood alteration including Insomnia. Triprolidine, an HIV drug, is first generation histamine H1 antagonist used in allergic rhinitis (and is clustered with other rhinitis drugs). Cycrimine is a central anticholinergic drug designed to reduce the levels of acetylcholine in the treatment of Parkinson's disease, while its neighbor Carbinoxamine, used for allergic rhinitis, is likely capable of treating mild cases of Parkinson's disease as well (http://www.ebi.ac.uk/chebi/searchId.do?chebiId=3398). It should be noted that since SLAP does not differentiate positive and negative interactions (activation or inhibition), the pairs may present opposite indication. Phenylpropanolamine (an Alpha-1A adrenergic receptor agonist), clustered with Doxazosin (an Alpha-1A adrenergic receptor antagonist for treating hypertension) is known to cause severe hypertension [35] .

Discussion

In this paper we demonstrate the utility of predicting associations based on semantic networks and the SLAP method of association prediction. The method performs extremely well in correctly identifying known drug-target pairs in the data, has been shown to out-perform similar link prediction methods used in social networking, and compares favorably with the established SEA method for predicting new drug-target interactions, as well as with the CMap method for associating drugs with changes in gene expression levels. We introduce the use of a drug-similarity network based on association profiles of drugs

across targets, and use these to propose potential new drug indications, although these indications have not yet been validated experimentally.

The use of large semantically annotated datasets to identify potential relationships from the linked data is a very new area, and we consider this an initial work in this field. There are several limitations to our current version. First, adding more data pertaining to drugs and targets would help identify more pairs. The side effect, disease and chemical ontology data are only linked to a limited number of drugs at present, and protein-protein interaction and protein pathway mapping data should greatly enhance its utility. In particular, the ability to embed compounds into the network for which there is no public information using chemical structure similarity, or new targets into the network using sequence similarity, would enable predictions to be made (albeit more indirectly) for newly synthesized or resolved compounds and targets. Second, as the complexity of path finding increases dramatically with increasing path length, only shortest paths with length ≤ 3 was considered, thus potentially missing important path patterns that have a greater path length. Third, edge weights are defined with the assumption that the probability from one node to its neighbors with same semantic type (e.g., from one drug to its targets) is equal. An important limitation of our current algorithm is that it does not enable differentiation of relationships other than categorical ones defined in the ontology. For instance, binding affinity could be used to weight the edge between drug and target, the edge with lower affinity is expected to have higher probability than that with higher affinity (or inactive interaction). Using such data brings up the issue of comparability between datasets: some chemogenomics datasets such as DrugBank currently do not provide sufficient binding affinities, but the weighting schema can be modified straightforwardly in SLAP once the data is provided. In addition, binding types (agonist/antagonist, activator/inhibitor) can be incorporated to classify and weight edges. Fourth, it should be pointed out that using large public integrated datasets means there is often a fuzziness between "no data" and "inactive data": i.e. we cannot assume that because two items do not have a relationship in the dataset, that they are not related - for instance that a drug cannot inhibit a target.

A key question in employing any drug-target prediction method is the extent to which it requires data completeness - in the extreme a full experimental matrix - to work properly (i.e. if it needs to be trained with consistent known active/inactive information for all compounds against all targets). Our methods does not require such training, indeed its purpose is to suggest potential "missing links" in incomplete data. However, it should be pointed out that the level of data completeness in a set will affect the path lengths, z-scores and associations scores produced. We believe that overall SLAP should be considered a useful tool for predicting that a relationship exists between drugs and targets, and thus as a tool primarily for ideas generation and for suggesting relationships to be probed experimentally. Since its purpose is to predict a relationship, not necessarily indicating a strong physical interaction. We believe it is also useful, as demonstrated in our drug network, for profiling compounds by their target associations (and vice versa) and we plan to explore other types of network that can be derived from SLAP.

All drug target prediction methods only employ single kinds of information or relationship (e.g., substructure, side effect, etc.), these methods are limited due to incompleteness of the data, for instance drug target relation are far from complete [36]. The employment of various data information can compensate for the lack of completeness of individual information. SLAP shows a direction to leverage such information for drug target prediction. Several sample pairs along with their key information are listed in table S3. For instance, the association between pyridoxal phosphate (CID: 1051) and cysteine conjugate-beta lyase 2 (CCBL2) is very strong (p value=1.9E-3), but if we removed gene ontology information, their association would become very weak (p value=0.02); the association between Dexamethasone (CID:5743) and annexin A1(ANXA1) would hardly be captured if substructure information were not considered.

The most compelling advantage of SLAP is its consideration of relations from a system level rather than just by known binding affinity data. Other than direct drug target interactions, SLAP is also capable of recognizing indirect interactions (e.g., the change of gene expression level) from random pairs, although the association scores are often smaller than direct interactions (Fig. S3). It thus allows us to

evaluate drug similarity based on the biological function. The network demonstrates that such similarity measurements not only is able to identify the drug action modes but also could suggest the new use of drugs.

Materials and Methods

Network building

We extracted semantic drug-target information from our Chem2Bio2RDF set [24] along with annotations created with our Chem2Bio2RDF ontology [unpublished], to create a semantic drug-target network. We also extracted data contributing to either the similarity of compounds, the similarity of targets or chemical target interaction. For example, two compounds are similar if they share same side effects, same substructures or same chemical ontology terms; two targets are similar if they share same gene ontology terms or ligands, or they function in the same pathway. 10 classes of entities and 12 link types were defined in table S1 and table S2 respectively. For example, a link between a drug and a target via bind type is established if their binding affinity is smaller than 30um if exists. Each node in the network is an instance of one of the classes. The detailed information on the collection of individual nodes and edges are in the supporting materials.

Drug target pairs preparation

Drug target pairs from DrugBank were used to build the network. We took only the pairs in which drugs were small molecules by mapping to PubChem and targets are homo sapiens by mapping to HGNC. A total of 5607 pairs were extracted from the network and serve as a golden standard set. The drug target pairs were grouped into 6 classes according to ChEMBL [37] target classification (i.e., enzyme (2393 pairs), membrane receptor(862 pairs), ion channel(392 pairs), transporter(209 pairs), transcription factor (208 pairs) and others (1543 pairs)). Another benchmark dataset was taken from MATADOR [34] which was not used for network building. We took drug target pairs with direct interaction types and confidence score>800 from MATADOR. 1176 direct MATADOR pairs were used, in which 1065 pairs have at least one path with length$\leq$ 3. 3665 MATADOR indirect pairs were also extracted for evaluating indirect drug target interaction. Indirect interactions are caused by many different mechanisms, such as binding a metabolite of a drug as well as changes in gene expression [34].

Path finding and representation

We extracted paths of length$\leq$ 3. A heap-based Dijkstra algorithm was used to quickly find the paths between two nodes that can achieve a complexity of O(nlogn) [38]. Each path is represented as: node 1–edge 1–node 2–edge 2. The length of a path is the number of edges between two nodes. The paths are visualized in cytoscape [39] and only significant paths(assessed from the statistical models) are visualized.

Path association

Let graph as $G(V, E)$, $P_l(s \rightarrow t)$ as the lth shortest path from node s to t. $e_{i \rightarrow j}$ as the edge from node i to node j. $R_{i,j}$ as the link (relation) type of $e_{i,j}$.

It is assumed that it has an equal probability traversing node i to its neighbor node j within the same type, thus :

$$p(e(i \rightarrow j)) = \frac{1}{\sum_{k}^{n=1} R_{i,n} == R_{i,j}}$$

where k is the degree of node i.

As the probability of each edge is independent, the probability from s to t:

$$p(P_l(s \rightarrow t)) = p(P_l(e_{1\rightarrow 2}, e_{2\rightarrow 3}, ..., e_{m-1\rightarrow m})) = \prod_{i=1}^{m-1} e_{i\rightarrow i+1}$$

where m is the number of nodes in the path. Since p is very small, the logarithm is applied,

$$log(p(P_l(s \rightarrow t))) = \sum_{i=1}^{m-1} log(e_{i\rightarrow i+1})$$

Accordingly, the probability from t to s:

$$p(P_l(t \rightarrow s)) = p(P_l(e_{m\rightarrow m-1}, ..., e_{3\rightarrow 2}, e_{2\rightarrow 1})) = \prod_{i=1}^{m-1} e_{i+1\rightarrow i}$$

$$log(p(P_l(t \rightarrow s))) = \sum_{i=1}^{m-1} log(e_{i+1\rightarrow i})$$

We consider the graph as undirected, then we take the average as the raw score of path l between s and t:

$$log(p(P_l(s,t))) = (log(p(P_l(s \rightarrow t))) + log(p(P_l(t \rightarrow s))))/2$$

Statistical Model

We randomly sampled 100,000 drug target pairs from DrugBank covering 1355 approved small molecular drugs and 1683 human targets, 54,414 pairs have found at least one shortest path with length$\leq$ 3. The sampling yielded 2,344,026 paths, which are categorized into 34 path patterns. The scores of each pattern were fitted to a normal distribution (Fig. S1)and the expected mean and standard deviation were estimated, followed by the calculation of the Z score of every path. Only the paths with Z score greater than 0 were thought as a valid path contributing to the association. The z scores of all the valid paths from s to t were summed up to get its association score, which is later used to measure the strength of the association.

$$raw\ score(s,t) = \sum_{l}^{n} \frac{log(p(P_l)) - \theta(log(P_l))}{\sigma(log(P_l))}$$

where $log(p(log(P_l))) > \theta(log(P_l))$; n is the number of shortest paths between the nodes s and t; $\theta(log(P_l))$ and $\sigma(log(P_l))$ are expected mean, expected standard deviation of the pattern to which P_l belongs.

Some patterns may not be helpful or even noisy for drug target association. We built a test set consisting of drug target pairs from DrugBank and the same number of random drug target pairs sampled from the drugs and targets composing the real drug target pairs. For one pair, raw scores of all the paths within a path pattern were calculated and summed up as a score for that path pattern. The scores were then used to rank the pairs in the test set. The evaluation of each pattern was performed using the area under ROC. We also applied the same procedure to the direct pairs from MATADOR. The patterns with low ROC (< 0.51) were considered as uninformative. The uninformative patterns agreed by the test sets from DrugBank and MATADOR were removed.

The logarithmic association score conforms to a normal distribution (Fig. S2); p value is estimated to show the probability of observing the score by random chance alone. The lower p value shows the stronger relation between two objects.

Model evaluation

A test set was composed of a set of drug target pairs extracted from DrugBank and the same number of random pairs as decoys. Three another test sets were created by increasing the number of random pairs such that the sizes of random pairs are 4, 8, 12 times more than true drug target pairs. For each pair, the paths including the direct link if exists were removed, and the z scores of all valid paths were summed up as the association score. The scores were ranked to generate ROC curves [40], which are widely adopted to measure drug target prediction methods [20, 22, 32, 41]. We also considered Precision and Recall (PR) curve, which shows the ratio of true positives among all the predicted positives under a given recall rate [42]. PR curve is more informative and biologically meaningful while the dataset is imbalanced. The same procedure was also applied to another golden standard dataset collected from MATADOR. Other than using SLAP scores, we considered the number of shortest paths (maximum length 3), the number of valid paths (significant path defined in the model), the sum of raw score of all paths, the max raw score among all paths, and the average raw score of all paths. In addition, we took the pairs validated in experiments in a recent published paper [7] as novel pairs, after manually mapping their drugs and targets to PubChem CIDs and Gene Symbols, we ran SLAP to get the p values of all the valid pairs.

Assess drug similarity

We identified drug-disease pairs from paper [43], then mapped the drugs to PubChem CIDs (which is the default compound identifier in the network). Many drugs have multiple indications, in order to visualize drugs by therapeutic indications, only drugs with one indication were kept. We also only kept the top 10 diseases ordered by the number of related drugs. The association scores of all mapped drugs against a set of human targets construct biological signatures which are later used for measuring drug similarity using Pearson correlation coefficient. The pairs with coefficient$>$0.9 constitute the network. Drug structural similarity was measured by Tanimoto coefficient using MACCS fingerprint.

Acknowledgments

We thank Huijun Wang for the assistance in path finding algorithm and Qian Zhu for the web service development. We thank the comments from Alessandro Flammini, Rajarshi Guha, Mohammad Hasan, Xiangnan Kong, Josef Scheiber, Jaehong Shin, Haixu Tang and anonymous reviewers.

References

1. Xie L, Wang J, Bourne PE (2007) In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. PLoS Comput Biol 3: e217.

2. Scheiber J, Chen B, Milik M, Sukuru SCK, Bender A, et al. (2009) Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. J Chem Inf Model 49: 308–317.

3. Xie L, Li J, Xie L, Bourne PE (2009) Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of cetp inhibitors. PLoS Comput Biol 5: e1000387.

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

5. O'Connor KA, Roth BL (2005) Finding new tricks for old drugs: an efficient route for public-sector drug discovery. Nat Rev Drug Discov 4: 1005–1014.

6. Kinnings SL, Liu N, Buchmeier N, Tonge PJ, Xie L, et al. (2009) Drug discovery using chemical systems biology: repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. PLoS Comput Biol 5: e1000423.

7. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, et al. (2009) Predicting new molecular targets for known drugs. Nature 462: 175–181.

8. Dudley JT, Deshpande T, Butte AJ (2011) Exploiting drug-disease relationships for computational drug repositioning. Brief Bioinform 12: 303–311.

9. Garcia-Serna R, Ursu O, Oprea TI, Mestres J (2010) iphace: integrative navigation in pharmacological space. Bioinformatics 26: 985–986.

10. Taboureau O, Nielsen SK, Audouze K, Weinhold N, Edsgrd D, et al. (2011) Chemprot: a disease chemical biology database. Nucleic Acids Res 39: D367–D372.

11. Kuhn M, Szklarczyk D, Franceschini A, Campillos M, von Mering C, et al. (2010) Stitch 2: an interaction network database for small molecules and proteins. Nucleic Acids Res 38: D552–D556.

12. Oprea TI, Nielsen SK, Ursu O, Yang JJ, Taboureau O, et al. (2011) Associating drugs, targets and clinical outcomes into an integrated network affords a new platform for computer-aided drug repurposing. Mol Inform 30: 100–111.

13. Li YY, An J, Jones SJM (2011) A computational approach to finding novel targets for existing drugs. PLoS Comput Biol 7: e1002139.

14. Yang L, Wang K, Chen J, Jegga AG, Luo H, et al. (2011) Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome–clozapine-induced agranulocytosis as a case study. PLoS Comput Biol 7: e1002016.

15. Nidhi, Glick M, Davies JW, Jenkins JL (2006) Prediction of biological targets for compounds using multiple-category bayesian models trained on chemogenomics databases. J Chem Inf Model 46: 1124–1133.

16. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, et al. (2007) Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25: 197–206.

17. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P (2008) Drug target identification using side-effect similarity. Science 321: 263–266.

18. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, et al. (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313: 1929–1935.

19. Ferreira JD, Couto FM (2010) Semantic similarity for automatic classification of chemical compounds. PLoS Comput Biol 6.

20. Perlman L, Gottlieb A, Atias N, Ruppin E, Sharan R (2011) Combining drug and gene similarity measures for drug-target elucidation. J Comput Biol 18: 133–145.

21. Bleakley K, Yamanishi Y (2009) Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics 25: 2397–2403.

22. Zhao S, Li S (2010) Network-based relating pharmacological and genomic spaces for drug target identification. PLoS One 5: e11764.

23. Shadbolt N, Hall W, Berners-Lee T (2006) The semantic web revisited. Intelligent Systems, IEEE 21: 96 -101.

24. Chen B, Dong X, Jiao D, Wang H, Zhu Q, et al. (2010) Chem2bio2rdf: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics 11: 255.

25. Jentzsch A, Zhao J, Hassanzadeh O, Cheung KH, Samwald M, et al. (2009) Linking open drug data. In: In Triplification Challenge of the International Conference on Semantic Systems.

26. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J (2008) Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. JOURNAL OF BIOMEDICAL INFORMATICS 41: 706-716.

27. Wild DJ, Ding Y, Sheth AP, Harland L, Gifford EM, et al. (2011) Systems chemical biology and the semantic web: what they mean for the future of drug discovery research. Drug Discov Today .

28. Jeh G, Widom J (2001) Simrank: a measure of structural-context similarity. In: ACM SIGKDD : 538–543.

29. Aleman-Meza B, Halaschek-Wiener C, Arpinar IB, Ramakrishnan C, Sheth AP (2005) Ranking complex relationships on the semantic web. IEEE Internet Computing : 37–44.

30. Anyanwu K, Maduko A, Sheth A (2005) Semrank: ranking complex relationship search results on the semantic web. In: In WWW 05: Proceedings of the 14th international conference on World Wide Web. ACM, pp. 117–127.

31. Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am Soc Inf Sci Technol 58: 1019–1031.

32. Yamanishi Y, Kotera M, Kanehisa M, Goto S (2010) Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics 26: i246–i254.

33. Vidal D, Mestres J (2010) In silico receptorome screening of antipsychotic drugs. Molecular Informatics 29: 543–551.

34. Gnther S, Kuhn M, Dunkel M, Campillos M, Senger C, et al. (2008) Supertarget and matador: resources for exploring drug-target relationships. Nucleic Acids Res 36: D919–D922.

35. Pentel PR, Asinger RW, Benowitz NL (1985) Propranolol antagonism of phenylpropanolamine-induced hypertension. Clin Pharmacol Ther 37: 488–494.

36. Mestres J, Gregori-Puigjan E, Valverde S, Sol RV (2008) Data completeness–the achilles heel of drug-target networks. Nat Biotechnol 26: 983–984.

37. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, et al. (2011) Chembl: a large-scale bioactivity database for drug discovery. Nucleic Acids Res .

38. Wang H, Ding Y, Tang J, Dong X, He B, et al. (2011) Finding complex biological relationships in recent pubmed articles using bio-lda. PLoS One 6: e17243.

39. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504.

40. Fawcett T (2006) An introduction to roc analysis. Pattern Recogn Lett 27: 861–874.

41. Jacob L, Vert JP (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics 24: 2149–2156.

42. Davis J, Goadrich M (2006) The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. New York, NY, USA: ACM, ICML '06, pp. 233–240. doi:http://doi.acm.org/10.1145/1143844.1143874. URL http://doi.acm.org/10.1145/1143844.1143874.

43. Yildirim MA, Goh KI, Cusick ME, Barabsi AL, Vidal M (2007) Drug-target network. Nat Biotechnol 25: 1119–1126.

Tables

Path patterns	ROC
Chemical/Drug–*bind*–Target–*bind*–Chemical/Drug–*bind*–Target	0.850
Chemical/Drug–*bind*–Target–*hasGo*–GO–*hasGO*–Target	0.824
Chemical/Drug–*hasSubstructure*–SubStructure–*hasSubstructure*–Chemical/Drug–*bind*–Target	0.620
Chemical/Drug–*express*–Target–*hasPathway*–Pathway–*hasPathway*–Target	0.495
Chemical/Drug–*express*–Target–*hasTissue*–Tissue–*hasTissue*–Target	0.501
Chemical/Drug–*express*–Target–*PPI*–Target	0.501

Table 1. Path pattern examples. Edge types are presented as italic. ROC shows the performance of predicting drug target interaction with the pattern alone. The first three patterns are more informative than the last three in their capability to contribute to the associations.

Figure Legends

Figure 1. SLAP pipeline. An ontology is used to annotate public data sets and integrate them into a semantic linked network. Two nodes are linked by one or more number of paths, but only a small number of significant paths are kept for association estimation. The path significance and drug target associations are assessed by statistical models derived from random samples.

Figure 2. Paths between Troglitazone (label as PubChem ID: 5591) and PPARG with length$\leq$ 3. The nodes and edges are colored by their classes and edge types respectively. Some nodes are annotated additionally to help understand.

Figure 3. Logarithmic association score distribution of drug target pairs.

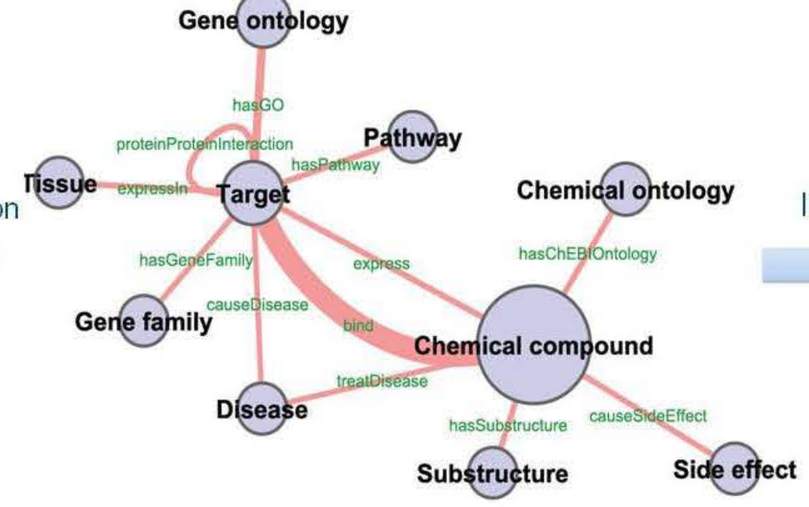
Figure 4. ROC curves among different prediction methods. Valid paths mean their z score>0.

Figure 5. Drug similarity network. Each node presents a drug, and two nodes are linked if their similarity (in terms of biological function) >0.9. The drugs are colored by their therapeutic indication. Five hypertension related clusters are shadowed.

PubChem	ChEBI	DrugBank
UniProt	UniProtKB-GOA	HGNC
SIDER	OMIM	KEGG
HPRD	ChEMBL	TTD
BindingDB	CTD	PDSP

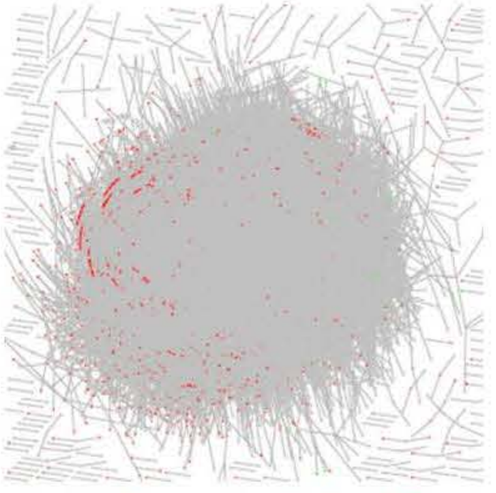
(a) Raw Data Sets

Semantic annotation



(b) Ontological level schema

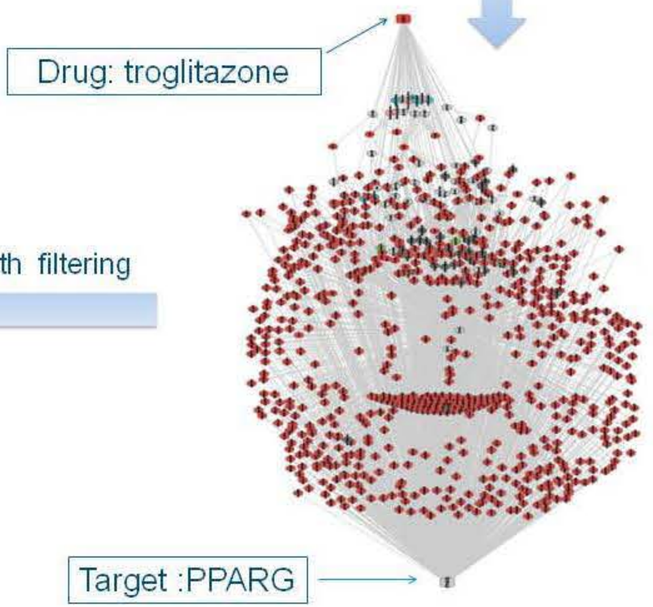
Integration



(c) Semantic Linked Data

Path finding

Drug: troglitazone

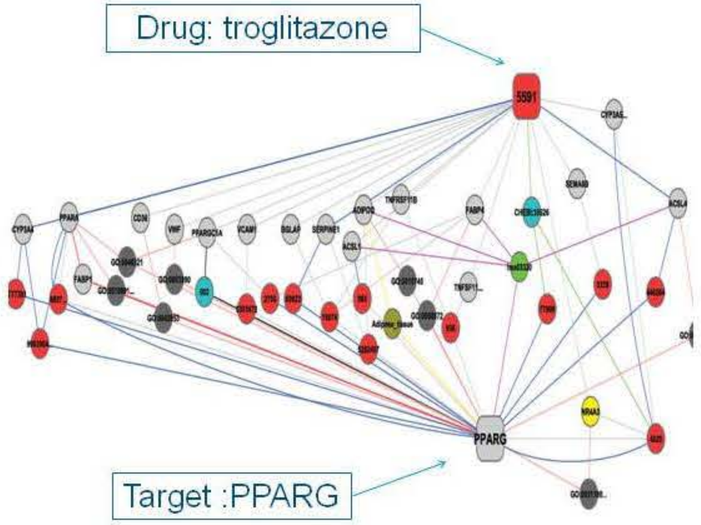


Target :PPARG

(d) Paths between two nodes

Path filtering

Drug: troglitazone



Target :PPARG

(e) Significant Paths between two nodes

ciation assessing

1. Edge weight:

$$p(e(i \rightarrow j)) = \frac{1}{\sum_{k}^{n=1} R_{i,n} == R_{i,j}}$$

2. Path score:

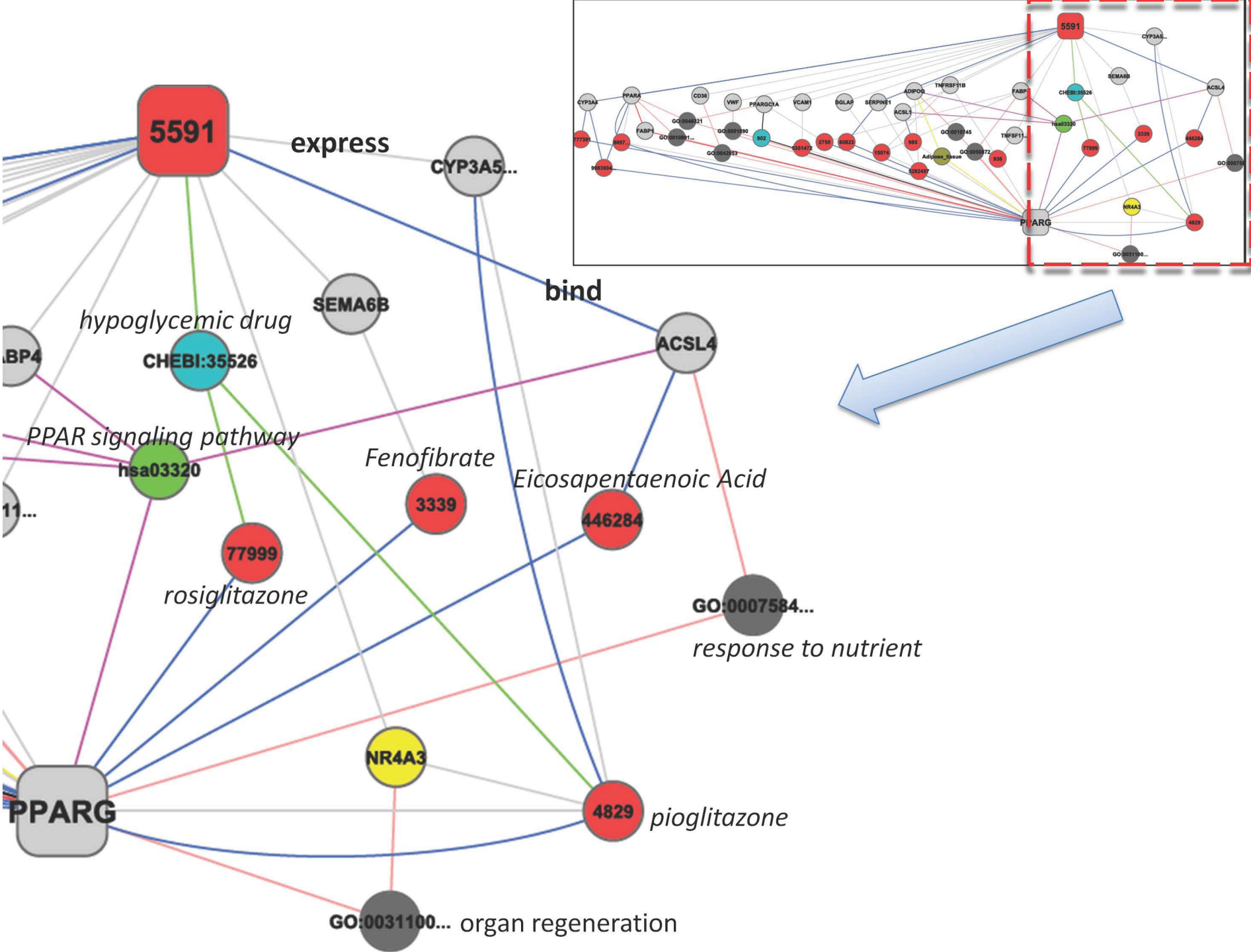
$$p(P_l(t \rightarrow s)) = p(P_l(e_{m \rightarrow m-1}, ..., e_{3 \rightarrow 2}, e_{2 \rightarrow 1})) = \prod_{i=1}^{m-1} e_{i+1 \rightarrow i}$$

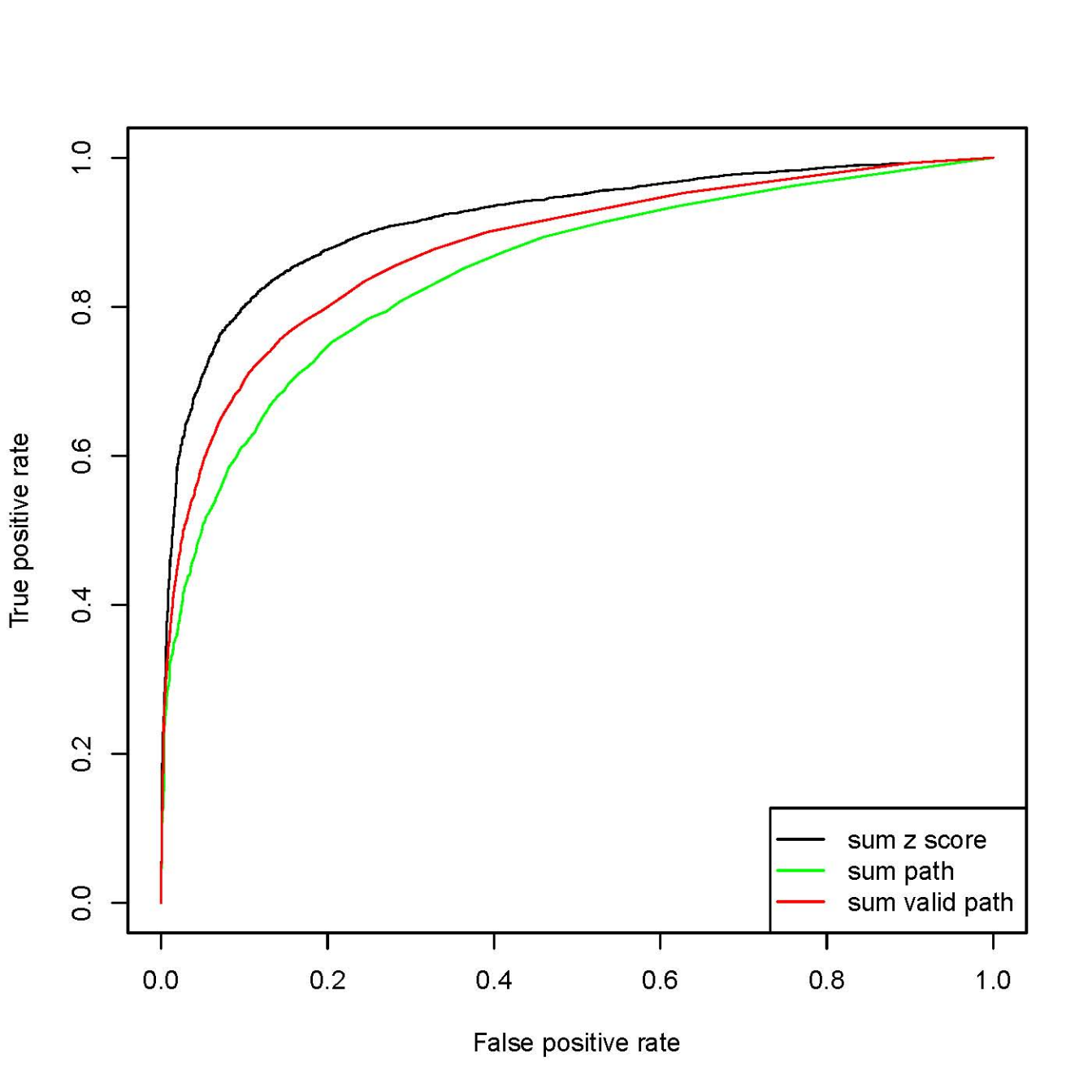
$$log(p(P_l(t \rightarrow s))) = \sum_{i=1}^{m-1} log(e_{i+1 \rightarrow i})$$

3. Association score

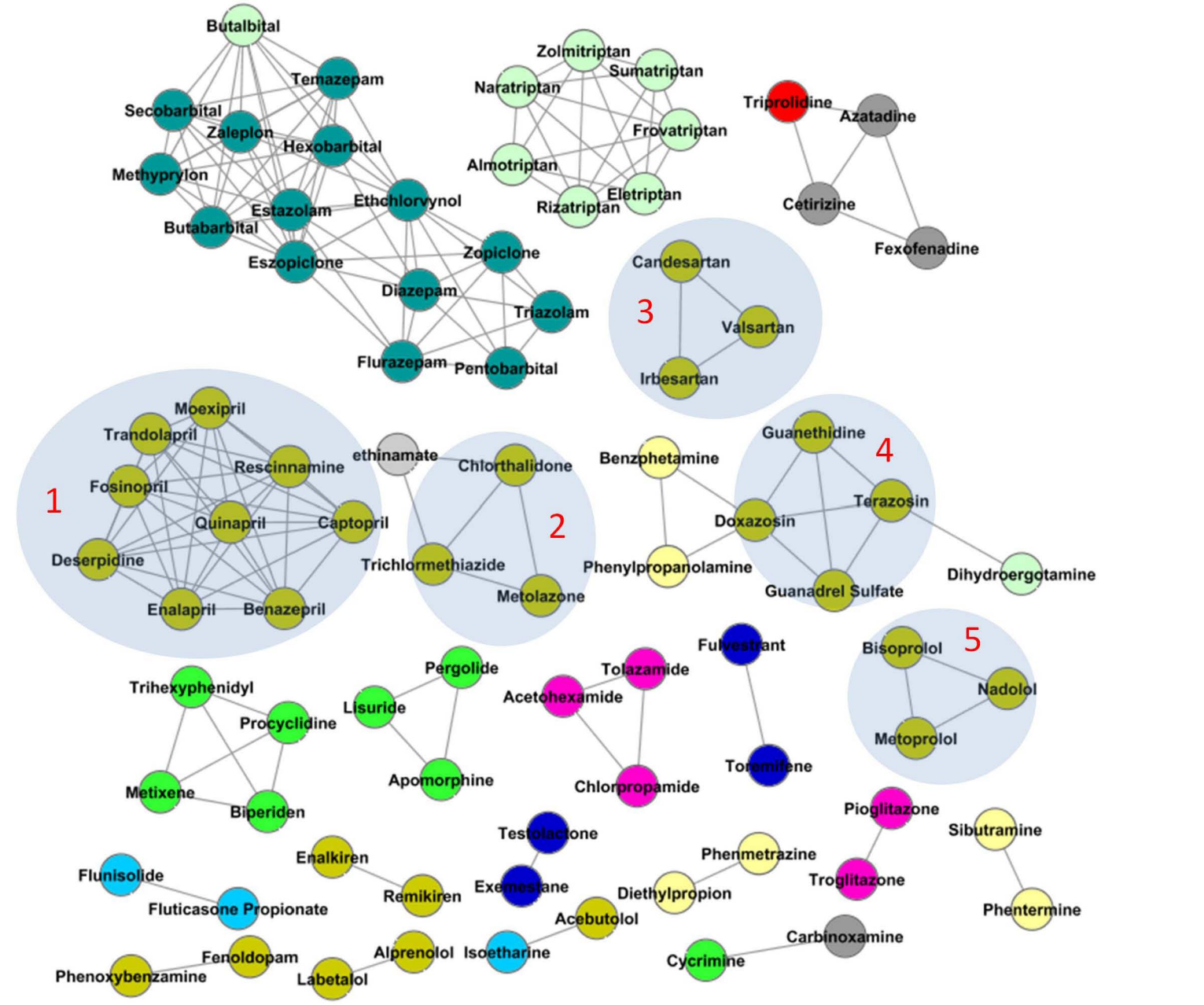
$$raw\ score(s,t) = \sum_{l}^{n} \frac{log(p(P_l)) - \theta(log(P_l))}{\sigma(log(P_l))}$$

(f) Statistical Models


5591
express
CYP3A5...
bind
SEMA6B
hypoglycemic drug
CHEBI:35526
ACSL4
PPAR signaling pathway
hsa03320
Fenofibrate
3339
Eicosapentaenoic Acid
446284
77999
rosiglitazone
GO:0007584...
response to nutrient
NR4A3
PPARG
4829
pioglitazone
GO:0031100...
organ regeneration


True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
sum z score
sum path
sum valid path

log association score
5
0
−5
−10
Drug Target Pairs
Random Pairs


Butalbital
Temazepam
Secobarbital
Zaleplon
Hexobarbital
Methyprylon
Ethchlorvynol
Estazolam
Butabarbital
Eszopiclone
Zopiclone
Diazepam
Triazolam
Flurazepam
Pentobarbital
Zolmitriptan
Sumatriptan
Naratriptan
Frovatriptan
Almotriptan
Eletriptan
Rizatriptan
Triprolidine
Azatadine
Cetirizine
Fexofenadine
Candesartan
3
Valsartan
Irbesartan
Moexipril
Trandolapril
Rescinnamine
Fosinopril
1
Quinapril
Captopril
Deserpidine
Enalapril
Benazepril
ethinamate
Chlorthalidone
2
Trichlormethiazide
Metolazone
Guanethidine
Benzphetamine
4
Terazosin
Doxazosin
Phenylpropanolamine
Guanadrel Sulfate
Dihydroergotamine
Trihexyphenidyl
Procyclidine
Metixene
Biperiden
Pergolide
Lisuride
Apomorphine
Tolazamide
Acetohexamide
Chlorpropamide
Fulvestrant
Toremifene
Bisoprolol
5
Nadolol
Metoprolol
Testolactone
Exemestane
Enalkiren
Remikiren
Flunisolide
Fluticasone Propionate
Acebutolol
Isoetharine
Alprenolol
Labetalol
Fenoldopam
Phenoxybenzamine
Pioglitazone
Troglitazone
Phenmetrazine
Diethylpropion
Sibutramine
Phentermine
Carbinoxamine
Cycrimine
hypertension
Parkinson disease
HIV
Diabetes mellitus
Breast cancer
Obesity
Allergic rhinitis
Asthma
Insomnia
Migraine