Xia et al. BMC Systems Biology 2010, 4(Suppl 2):S6
http://www.biomedcentral.com/1752-0509/4?issue=S2/S6

BMC Systems Biology

PROCEEDINGS Open Access

Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces

Zheng Xia[1], Ling-Yun Wu[2], Xiaobo Zhou[1], Stephen TC Wong[1*]

From Optimization and Systems Biology
Zhangjiajie, China. 20 - 22 September 2009

Abstract

Background: Predicting drug-protein interactions from heterogeneous biological data sources is a key step for *in silico* drug discovery. The difficulty of this prediction task lies in the rarity of known drug-protein interactions and myriad unknown interactions to be predicted. To meet this challenge, a manifold regularization semi-supervised learning method is presented to tackle this issue by using labeled and unlabeled information which often generates better results than using the labeled data alone. Furthermore, our semi-supervised learning method integrates known drug-protein interaction network information as well as chemical structure and genomic sequence data.

Results: Using the proposed method, we predicted certain drug-protein interactions on the enzyme, ion channel, GPCRs, and nuclear receptor data sets. Some of them are confirmed by the latest publicly available drug targets databases such as KEGG.

Conclusions: We report encouraging results of using our method for drug-protein interaction network reconstruction which may shed light on the molecular interaction inference and new uses of marketed drugs.

Background

Developing a new drug is an expensive and time-consuming process that is subject to a variety of regulations such as drug toxicity monitoring and therapeutic efficacy. Meanwhile, there are thousands of FDA-approved drugs in the market and drugs in later phases of clinical trials. Finding the potential application in other therapeutic categories of those FDA-approved drugs by predicting their targets, known as drug repositioning, is an efficient and time-saving method in drug discovery [1]. Additionally, predicting interactions between drugs and target proteins can help decipher the underlying biological mechanisms. Therefore, there is a strong incentive to develop powerful statistical methods that are capable of detecting these potential drug-protein interactions effectively.

Various methods have been proposed to address the drug-target prediction problems *in silico*. One common method is to predict the drugs interacting with a single given protein based on the chemical structure similarity in a classic classification framework. Keiser *et al.* [2,3] proposed a method to predict targets of proteins based on the chemical similarity of their ligands. This kind of approach, however, does not take advantage of the information in the protein domain. Another widely-used method is molecular docking [4] which requires the non-trivial modeling of 3D structure of the target protein. Unfortunately the 3D structures of many proteins are not available [5], e.g., very few GPCRs have been crystallized.

Recently, some new approaches are proposed to perform drug-target prediction using both the chemical (drug chemical structure) and genomic (protein structure) spaces information [3,6,7]. In [6] the two spaces are encoded together by defining a pair wise kernel which is then fed to the support vector machine (SVM) for classification. The drawback of this kernel

* Correspondence: stwong@tmhs.org
[1]Bioinformatics and Bioengineering Program, The Methodist Hospital Research Institute, Weill Medical College, Cornell University, Houston, TX 77030, USA
Full list of author information is available at the end of the article

framework is that there will be a huge number of samples to be classified (i.e., number of drugs multiplies number of proteins) which poses significant computational complexity. Another problem is that the negative drug-protein pairs are selected randomly without experimental confirmation. Yamanishi *et al.* [7] developed a bipartite graph model where the chemical and genomic spaces as well as the drug-protein interaction network are integrated into a pharmacological space. In the bipartite model, the known interactions in the training data are labeled as +1 while all other unknown drug-protein pairs in the training data are assumed as non-interactions with label 0. Then three different classifiers are available: new drug candidate versus known target protein, known drugs versus new target protein and new drug candidate versus new target protein candidate. More recently, Bleakley and Yamanishi [8] proposed a state-of-the-art bipartite local model (BLM) by transforming edge-prediction problems into well-known binary classification problems. Nevertheless, the first flaw of the bipartite model, like the kernel SVM method [6], is that the unknown interactions of the drugs and proteins in the training data are all assumed non-interaction and cannot be inferred. We also prefer only one classifier to predict whether one drug-protein pair interacts or not. Lastly, all the methods did not utilize a wealth of unlabeled information to assist prediction.

In this paper, a semi-supervised learning method - Laplacian regularized least square (LapRLS) [9] is employed to utilize both the small amount of available labeled data and the abundant unlabeled data together in order to give the maximum generalization ability from the chemical and genomic spaces.

Further, the standard LapRLS is improved by incorporating a new kernel established from the known drug-protein interaction network (NetLapRLS). In our framework, the known interactions are labeled as +1 and all other unknown pairs are labeled as 0 to indicate they are going to be predicted. Two classifiers are trained on the drug and protein domains respectively and then are combined together to give the final prediction. Compared with a naive weighted profiled method, the proposed drug-protein interaction methods based on LapRLS and NetLapRLS obtain better results than using the labeled data alone. And the proposed NetLapRLS which incorporates drug-protein network information provides superior performance than standard LapRLS.

Results and discussion

Cross validation results analysis

The weighted profile method, standard LapRLS and NetLapRLS were evaluated on the four classes of target proteins including enzymes, ion channels, GPCRs and nuclear receptors. We carried out a ten-fold cross-validation by splitting the golden standard interaction dataset into 10 subsets. Each fold was then taken in turn as a test set and the remaining nine folds are used as training set. For example, there are 54 drugs and 26 proteins in the nuclear receptor data set with 90 known interactions. In each cross-validation, the 80 drug-protein pairs are used as the training data while the remaining 1,324 drug-protein pairs including the 10 positive interactions are designated as the testing data set. Thus the training sample is very small compared with the testing data set. This motivates us to employ the semi-supervised method that can utilize the information from the unlabeled samples to predict drug-protein interaction. The performance is evaluated using receiver operating curve (ROC) analysis [10]. For simplicity, we set $\beta_d = \beta_p = 0.3$, $\gamma_{d1} = \gamma_{p1} = 1$, and $\gamma_{d2} = \gamma_{p2} = 0.01$ for NetLapRLS. These parameters can be better selected by a further cross validation. If γ_{d2} and γ_{p2} are set to be 0, NetLapRLS becomes the standard LapRLS method. Table 1 shows the AUC (area under the ROC curve), sensitivity and specificity. The sensitivity and specificity are defined as TP/(TP+FN) and TN/(TN+FP), respectively. The cutoff for calculation of sensitivity and specificity is set to select the top pairs with the same number of the test set.

From Table 1 and Figure 1, we can see that LapRLS and NetLapRLS methods, which use unlabeled information, provided better performance with respect to AUC score and sensitivity. Among the four data sets, the two semi-supervised learning methods provided the highest sensitivity scores in enzyme data set because there are most known interactions. The known interaction number is a key factor of our semi-supervised methods since the testing data set is much larger than training data set in our cross-validation setup. The proposed NetLapRLS which incorporates the drug-protein interaction network information obtained better result than the standard LapRLS, especially with respective to the sensitivity which is dramatically improved. On the four data sets, the sensitivity from NetLapRLS performed better than LapRLS by 42%, 100%, 108% and 31% respectively and, demonstrated the importance of network information. The improvement in sensitivity of NetLapRLS over LapRLS is most significant in ion channel data set because the inner-connection in the ion channel drug-protein interaction network is most complete according to the proportion of unreachable paths between drugs and proteins [7]. Yildirim *et al.* [11] concluded that there are an overabundance of 'follow-on' drugs from the topological analyses of current drug-protein network, that is, drugs that target already known proteins, i.e., me-too drugs. With the drug-protein network being completed fastly by high-throughput experimental and computational approaches, this network information is becoming critical in drug discovery.

Table 1 Statistics of the prediction performance

Data	Methods	AUC	Sensitivity(%)	Specificity(%)
* Enzyme	Combining weighted profile	92.2	6	99.9
	LapRLS	95.0	53	99.9
	NetLapRLS	98.3	75	99.9
* Ion channel	Combining weighted profile	90.7	17	99.7
	LapRLS	96.1	36	99.8
	NetLapRLS	98.6	72	99.9
* GPCR	Combining weighted profile	86.9	13	99.7
	LapRLS	93.4	24	99.8
	NetLapRLS	97.1	50	99.8
Nuclear receptor	Combining weighted profile	81.0	11	99.4
	LapRLS	85.0	16	99.4
	NetLapRLS	88.8	21	99.5

The AUC is the area under the ROC curve, normalized to 100. The cutoff for sensitivity and specificity is set to select the number of the interactions in the test data.

Comparison with bipartite local model [8]

Recently, Bleakley and Yamanishi [8] extended Yamanishi's bipartite method [12] to bipartite local model which is considered as state-of-the-art. The predictions from the drug domain and protein domain using SVM are combined together to form a final prediction by a maximum operation. We also employed this kind of integration by a mean operation. However, we used a semi-supervised learning method to handle the classification with small samples labeled which is difficult for traditional supervised classifiers. For instance, in the above cross-validation experiment of the nuclear receptor data set, the semi-supervised classifier is trained on 80 positive samples in order to make predictions on 1,324 unlabeled samples. In the BLM, the ten-fold cross-validation is performed on the drug and protein domains separately. The known interactions between the selected drugs and proteins are labeled as interaction while interactions between the drugs and proteins for training are regarded as non-interaction. Though we consider the undetermined relationship between drug-protein pair should not be labeled as non-interaction, we adopt the cross validation method in the BLM for the sake of comparison in the same condition. The comparison is performed in terms of AUC, area under precision-recall (AUPR), sensitivity, specificity and PPV as shown in Table 2. Sensitivity, specificity and PPV are calculated when the top one percentile in the prediction score is chosen as a cutoff because high-confidence prediction results are more useful in practical applications. We observed that BLM method outperformed our NetLapRLS in AUC and AUPR scores, but the performances of our NetLapRLS are comparable with BLM in sensitivity, specificity and PPV.

Semi-supervised learning method is superior to the traditional supervised learning method when labeled samples are small along with large unlabeled samples available. In this cross validation setup, the unknown interactions are labeled as non-interaction in the training data set. So our NetLapRLS did not get good results in AUC and AUPR scores compared with BLM because most of samples are labeled. However, NetLapRLS still gave good prediction results in sensitivity, specificity and PPV. This indicated that NetLapRLS can provide a list of drug-protein interaction candidates with high confidence.

Enzyme

Table 3 shows the list of the top 5 predicted drug-protein pairs, with annotation given in the KEGG database [13]. Searching the latest version of KEGG drug database and Drugbank [14], we found that the fifth highest scored drug-protein pair (D00097 and hsa5743) in Table 3 is annotated as an interaction. Figure 2 shows the predicted top 50 scoring drug-protein interaction network on the enzyme data using the all known interactions as the training data set.

Ion channel

Table 4 shows the list of the top five predicted drug-protein pairs on the ion channel data set, with annotation given in the KEGG database [13]. In the latest version of KEGG drug database, the targets of drug D00477 (rank 2 in table 4) include SCN1A, SCN2A, SCN3A, SCN4A, SCN5A, SCN8A and SCN9A. The targets of drug D00552 are SCN10A, SCN1A, SCN2A, SCN3A, SCN4A, SCN5A, SCN8A and SCN9A. Thus, our predicted target of D00552 is confirmed (rank 3 in Table 4). The targets of drugs D00477 and D00552 are very similar which can be explained by their common chemical structures in Figure 3. Based on the chemical structure similarity, we predict that SCN10A is also a target of drug D00477 (rank 2 in table 4), as the interaction between SCN10A and D00552 is known. Rank 5 in

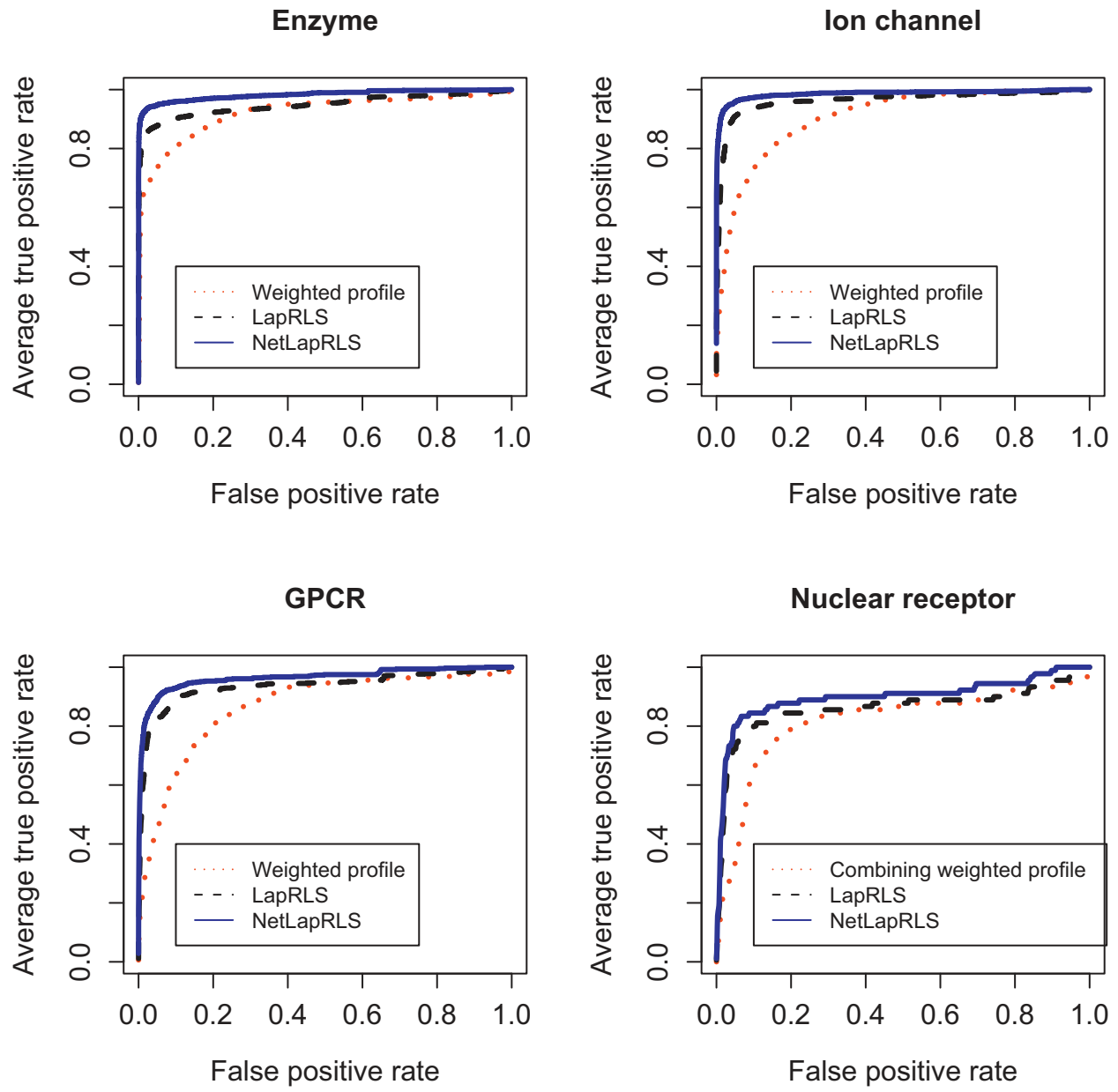

Figure 1 ROC curves of cross validation. ROC curves of the semi-supervised learning and the combining weighted profile methods for four classes target proteins: enzymes, ion channels, GPCRs and nuclear receptors.

table 4 predicts GABAR2 is one of the targets of drug D00546. This prediction is reasonable because in Drugbank D00546 is annotated to interact with GABAR1 which is very similar with GABAR2 in sequence and function. Figure 4 shows the predicted top 50 scoring drug-protein interaction network on the ion channel data set using the all known interactions as training data set.

GPCRs

Table 5 shows the list of the top five predicted drug-protein pairs on GPCRs data set, with annotation given in KEGG database. Based on the most recent KEGG database, the predictions of rank 2 and 3 in Table 5 are confirmed. Additionally, six predicted new targets (hsa146, hsa147, hsa150, hsa151, hsa152 and hsa155) of drug adrenaline (D00095) from the newly predicted interactions with 50 highest scores are also annotated as an interaction in the latest KEGG drug database. Ranks 4 and 5 in Table 5 predict both D02345 and D00283 target protein DRD3. In Drugbank, D02345 and D00283 are annotated to interact with protein DRD1, DRD2, and DRD4. Because DRD3 is very similar with those

Table 2 Results of BLM and NetLapRLS based on cross validation experiments 5 times

Data	Methods	AUC	AUPR	Sensitivity(%)	Specificity(%)	PPV(%)
* Enzyme	BLM	96.8(0.1)	85.2(0.2)	83.2(0.2)	99.82(0.002)	82.3(0.2)
	NetLapRLS	95.6(0.3)	82.6(0.6)	81.0(0.5)	99.80(0.005)	80.2(0.5)
* Ion channel	BLM	97.2(0.1)	83.2(0.4)	28.0(0.03)	99.96(0.001)	96.4(0.1)
	NetLapRLS	94.7(0.3)	82.5(0.5)	28.4(0.14)	99.98(0.005)	98.1(0.5)
* GPCR	BLM	94.4(0.3)	65.0(1.6)	28.0(0.8)	99.83(0.02)	83.9(2.4)
	NetLapRLS	93.1(0.3)	66.0(1.5)	29.2(0.8)	99.87(0.03)	87.5(2.4)
Nuclear Receptor	BLM	84.1(0.9)	58.4(2.2)	14.0(0.6)	99.89(0.04)	90.0(3.9)
	NetLapRLS	85.6(1.8)	51.6(2.3)	15.1(1.0)	99.97(0.07)	97.1(6.1)

The AUC and AUPR scores are normalized to 100. The cutoff for sensitivity, specificity and PPV is set to choose the top one percentile in the predictoin score as positive.

Table 3 Top 5 scoring predicted drug-protein interactions for the enzyme data set

Rank	Pair	Annotation
*1	D00528	Anhydrous caffeine
	hsa1549	cytochrome P450, family 2, subfamily A, polypeptide 7
*2	D00542	Halothane
	hsa1571	cytochrome P450, family 2, subfamily E, polypeptide 1
*3	D00437	Nifedipine
	hsa1559	cytochrome P450, family 2, subfamily C, polypeptide 9
*4	D00410	Metyrapone
	hsa1585	cytochrome P450, family 11, subfamily B, polypeptide 2
*5	D00097	Salicylic acid
	hsa5743	prostaglandin-endoperoxide synthase 2

proteins in function, our method predicts DRD3 is also the target of drugs D02345 and D00283. This result demonstrated our method employed the information from protein domain. Figure 5 shows the predicted top 50 scoring drug-protein interaction network on the GPCRs data set using the all known interactions as the training data set.

Nuclear receptor

Table 6 shows the list of the top 5 predicted drug-protein pairs on nuclear receptor data set, among which four predictions are about drug D00348. In Drugbank, drug D00348 is annotated to interact with protein (retinoic acid receptor, alpha). The two predicted targets with the

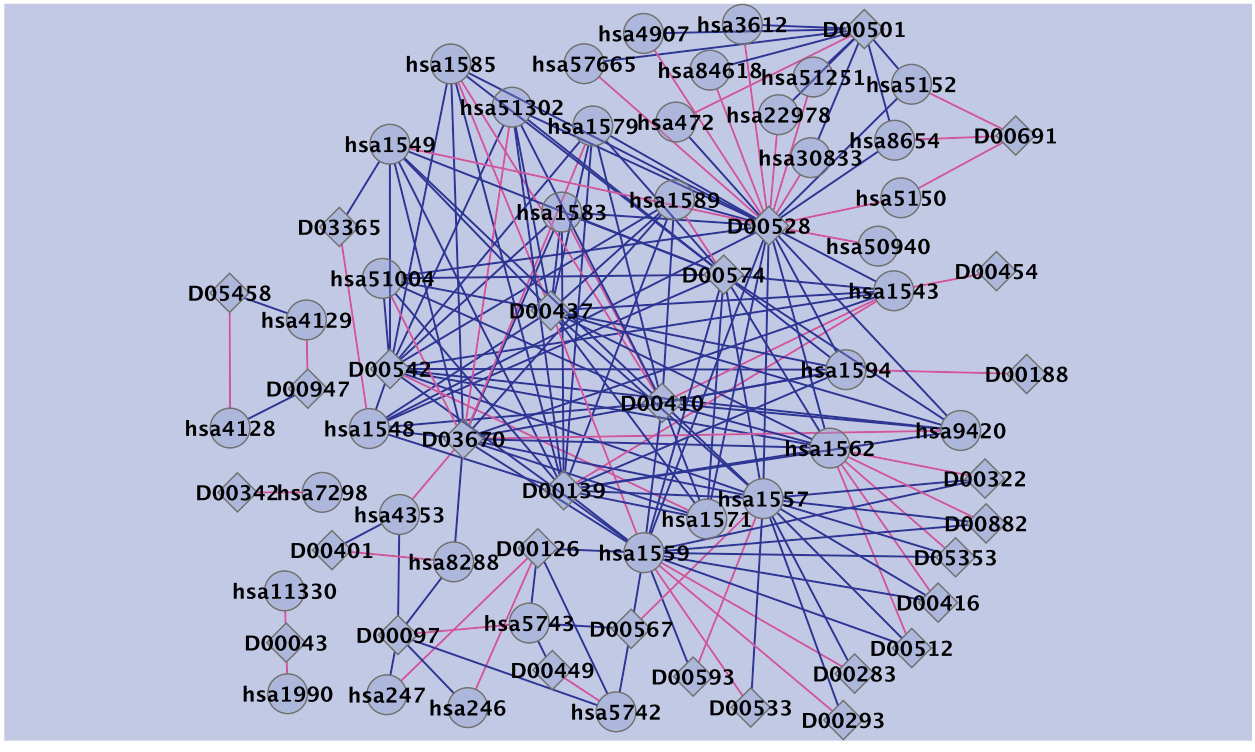

Figure 2 Predicted enzyme interaction network. Diamonds and circles represent drugs and target proteins, respectively. Blue and red lines indicate known interactions and newly predicted interactions with 50 highest scores, respectively.

Table 4 Top 5 scoring predicted drug-protein interactions for the ion channel data set

Rank	Pairs	Annotation
*1	D00438	Nimodipine
	hsa779	calcium channel, voltage-dependent, L type, alpha 1 S subunit, beta 2
*2	D00477	Procainamide hydrochloride
	hsa6336	sodium channel, voltage-gated, type X, alpha subunit(SCN10A)
*3	D00552	Ethyl aminobenzoate
	hsa6331	sodium channel, voltage-gated, type V, alpha subunit(SCN5A)
*4	D02272	Quinidine sulfate
	hsa3738	potassium voltage-gated channel, shaker-related subfamily, member 3
*5	D00546	Desflurane
	hsa2555	gamma-aminobutyric acid (GABA) A receptor, alpha 2(GABAR2)

highest scores (hsa5915 and hsa5916) of drug D00348 are both from retinoic acid receptor class. Those proteins are probably as the targets of the same protein due to their similarity in sequence and function. Figure 6 shows the predicted top 50 scoring drug-protein interaction network on the nuclear receptor data set with the all known interactions as the training data set.

Conclusions

In this work, we presented a semi-supervised learning method NetLapRLS for drug-protein interaction prediction by integrating information from chemical space, genomic space and drug-protein interaction network space. Our method has no use of the negative samples and predicts the interaction of each drug-protein pair. The results we obtained when predicting human drug-target interaction networks involving enzymes, ion channels, GPCRs, and nuclear receptors demonstrated the superior performance of NetLapRLS. Furthermore, recently added drug-target interactions to the KEGG immediately allowed us to confirm some strongly-predicted drug-target interactions on the four data sets obtained using our method. This enhances the strength of our proposed method for realistic drug-target prediction application.

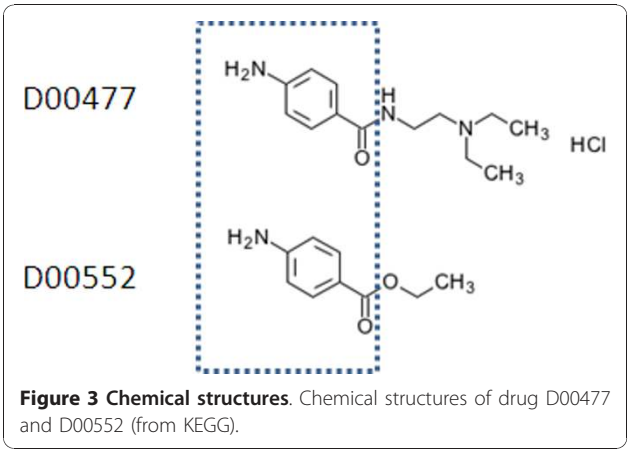

Figure 3 Chemical structures. Chemical structures of drug D00477 and D00552 (from KEGG).

The ideal way to use semi-supervised learning for predicting compound-protein interactions is to incorporate information from different biological spaces by a multi-task kernel and is fed to classical semi-supervised learning. However, the implementation of such a large scale semi-supervised learning method will be computationaly costly. Our future work, will incorporate more sophisticated and biologically relevant information into the kernel similarity, such as side effect [15], to improve the prediction accuracy.

Methods

Semi-supervised learning (SSL) has been attracting much research attention in the machine learning community [16]. SSL provides better prediction accuracy by using unlabeled information. Here we employ a data-dependent manifold regularization framework which uses the geometry of the probability distribution [9]. One of the implementations of this framework is the Laplacian regularized least squares (LapRLS) which is simple and has comparable performance with Laplacian regularized support vector machine.

Consider the drug dataset $=\{d_1,\ldots,d_{n_d}\}$ and the target protein dataset $=\{p_1,\ldots,p_{n_p}\}$ where n_d and n_p are the numbers of the drugs and proteins in the study respectively. An interaction pattern of drug d_i and target protein p_j is represented by a binary label matrix $\mathbf{Y} \in B^{n_d \times n_p}$. If drug d_i is known to interact with target protein p_j, $\mathbf{Y}_{ij} = 1$ otherwise $\mathbf{Y}_{ij} = 0$. Given the 'gold standard' drug-target interactions, the goal is to infer their unknown interactions. Two classifiers will be trained using LapRLS on the chemical and genomic spaces separately, followed by a combination of the two classifiers. A supervised learning method is suitable in this case. However the known interactions from public databases are still extremely small compared to the whole drug-target interaction space. Another issue is that we only have the information of the interactions, but do not know which drug target pair has no interaction, i.e., no negative samples in the training

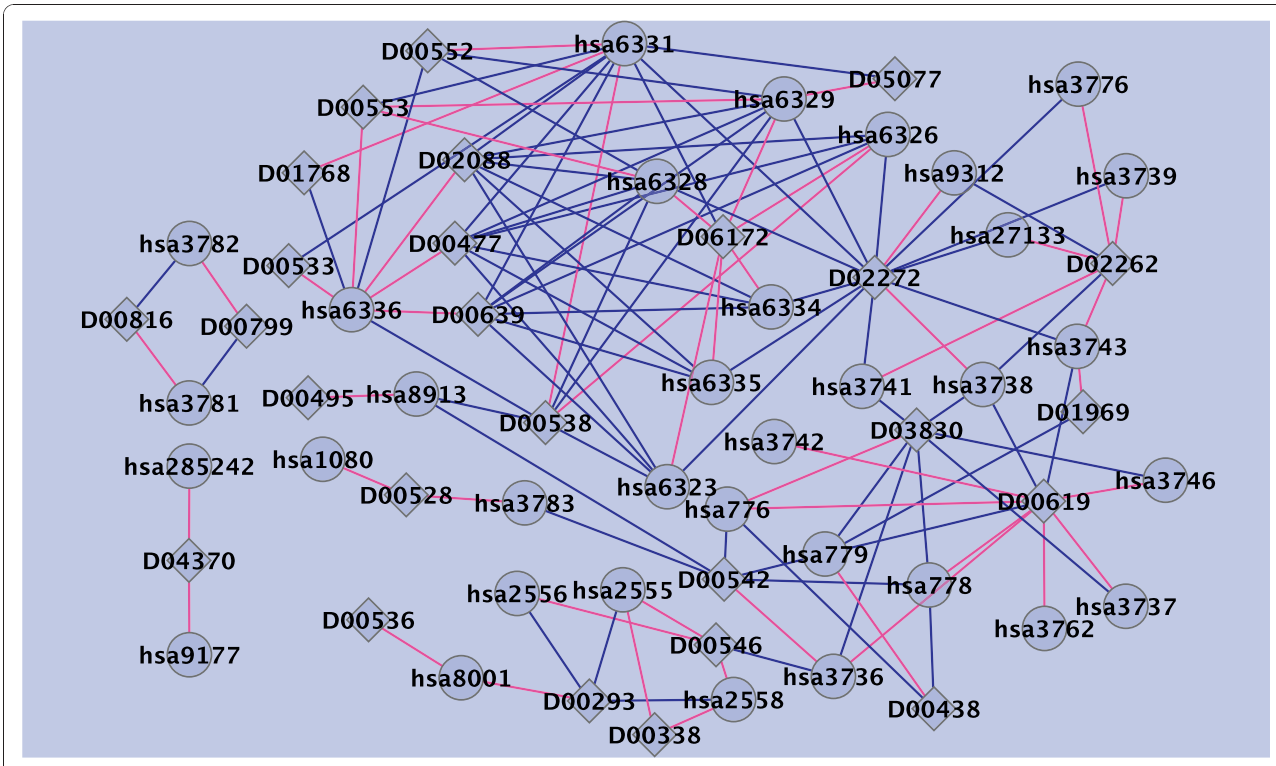

Figure 4 Predicted ion channel interaction network. Predicted ion channel interaction network. diamonds and circles represent drugs and target proteins, respectively. Bule and red lines indicate known interactions and newly predicted interactions with 50 highest scores, respectively.

process. Herein we first test a simple supervised weighted profile method. Then the standard LapRLS and drug-protein interaction network incorporated NetLapRLS are extended to predict the drug-protein interaction.

Materials

The data used here is downloaded from (http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/) [7]. Here below we provide a brief description.

• ***Chemical data***

The chemical structure similarity between compounds are calculated by SIMCOMP [17] using chemical structures fetched from KEGG LIGAND database. SIMCOMP provides a global similarity score by the ratio between the size of common substructures and the size of the union structures of two compounds. Applying this operation to all compounds pairs, we constructed a similarity matrix denoted $\mathbf{S}_d \in R^{n_d \times n_d}$ which represents the chemical space information.

Table 5 Top 5 scoring predicted drug-protein interactions for the GPCRs data set

Rank	Pair	Annotation
*1	D02358	Metoprolol
	hsa154	adrenergic receptor, beta 2
*2	D00095	Adrenaline
	hsa155	beta3-adrenergic receptor agonist
*3	D00371	Theophylline
	hsa135	adenosine A2a receptor antagonist
*4	D02354	Thiethylperazine
	hsa1814	dopamine receptor D3
*5	D00283	Clozapine
	hsa1814	dopamine receptor D3(DRD3)

• ***Genomic data***

A normalized Smith-Waterman score is calculated to indicate the similarity between two amino acid sequences of target proteins which were obtained from the KEGG GENES database. All protein pairs similarities are computed to construct a similarity matrix denoted $\mathbf{S}_p \in R^{n_p \times n_p}$ which represents the genomic space.

• ***Drug-protein interaction data***

At the time of the paper [7] was written, Yamanishi *et al.* [7] found 445, 210, 223, and 54 drugs targeting 664 enzymes, 204 iron channels, 95 GPCRs, and 26 nuclear receptors, receptively, and the known interactions are 2926, 1476, 635 and 90.

Combining weighted profiles

The method of combining weighted profiles follows the idea that the label of the new sample is determined by its similarity with the training samples. For a drug d_i, its

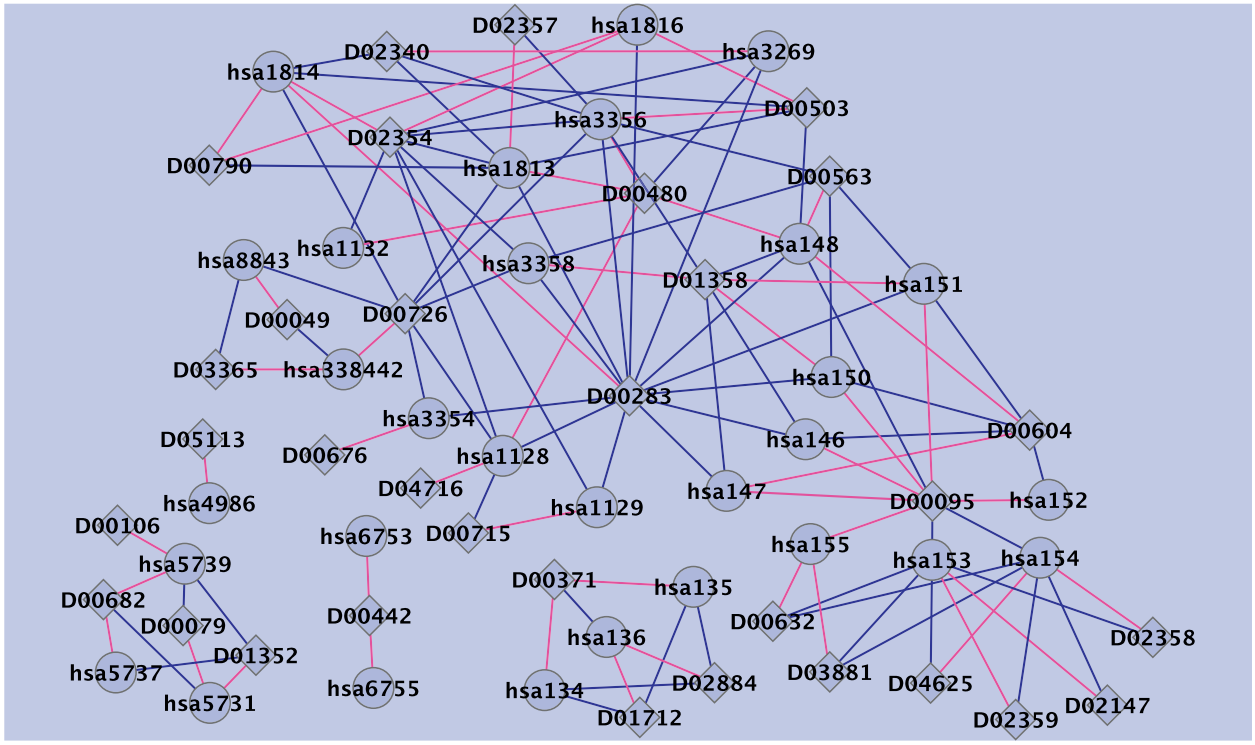

Figure 5 Predicted GPCRs interaction network. Predicted GPCRs interaction network. diamonds and circles represent drugs and target proteins, respectively. Blue and red lines indicate known interactions and newly predicted interactions with 50 highest scores, respectively.

interaction $f(d_i, p_j)$ with a protein p_j in $\mathbb{P}$ is predicted with the following formulation:

$$f(d_i, p_j) = \frac{1}{\mathbf{N}_{d_i}} \sum_{k=1}^{n_d} s_d(d_i, d_k)\mathbf{Y}_{kj} \quad (1)$$

where $s_d(d_i, d_k)$ is a chemical structure similarity score from S_d and $\mathbf{N}_{d_i}$ is a normalization term defined as $\mathbf{N}_{d_i} = \sum_{k=1}^{n_d} s_d(d_i, d_k)$. Meanwhile, for a protein p_j, its interaction $f(p_j, d_i)$ with a drug d_i can also be calculated in the genomic space by:

$$f(p_j, d_i) = \frac{1}{\mathbf{N}_{p_j}} \sum_{k=1}^{n_p} s_p(p_j, p_k)\mathbf{Y}_{ik} \quad (2)$$

where $s_p(p_j, p_k)$ is a genomic sequence similarity score from $\mathbf{S}_p$ and $\mathbf{N}_{p_j}$ is a normalization term defined by $\mathbf{N}_{p_j} = \sum_{k=1}^{n_p} s_p(p_j, p_k)$ Note that Equations (1) and (2) are estimating the interaction of the same drug-protein pair ($d_i \sim p_j$) from different data sources. The two predictions should be combined to give the final prediction by

$$\bar{f}(d_i, p_j) = \frac{f(d_i, p_j) + f(p_j, d_i)}{2} \quad (3)$$

The drug-protein pairs (d_i, p_j) in $\bar{f}(d_i, p_j)$ (d_i, p_j) with high scores are predicted to interact each other. The original weighted profile method is used in [7]. However their predictions in the two spaces are not fused. Figure 7 shows the method of combining weighted profiles provides better prediction than methods using the single space on the four data sets.

Table 6 Top 5 scoring predicted drug-protein interactions for the nuclear receptor data set

Rank	Pair	Annotation
*1	D00348	Isotretinoin
	hsa5915	retinoic acid receptor, beta
*2	D00348	Isotretinoin
	hsa5916	retinoic acid receptor, gamma
*3	D00182	Norethindrone
	hsa2099	estrogen receptor 1
*4	D00348	Isotretinoin
	hsa6256	retinoid X receptor, alpha
*5	D00348	Isotretinoin
	hsa6257	retinoid X receptor, beta

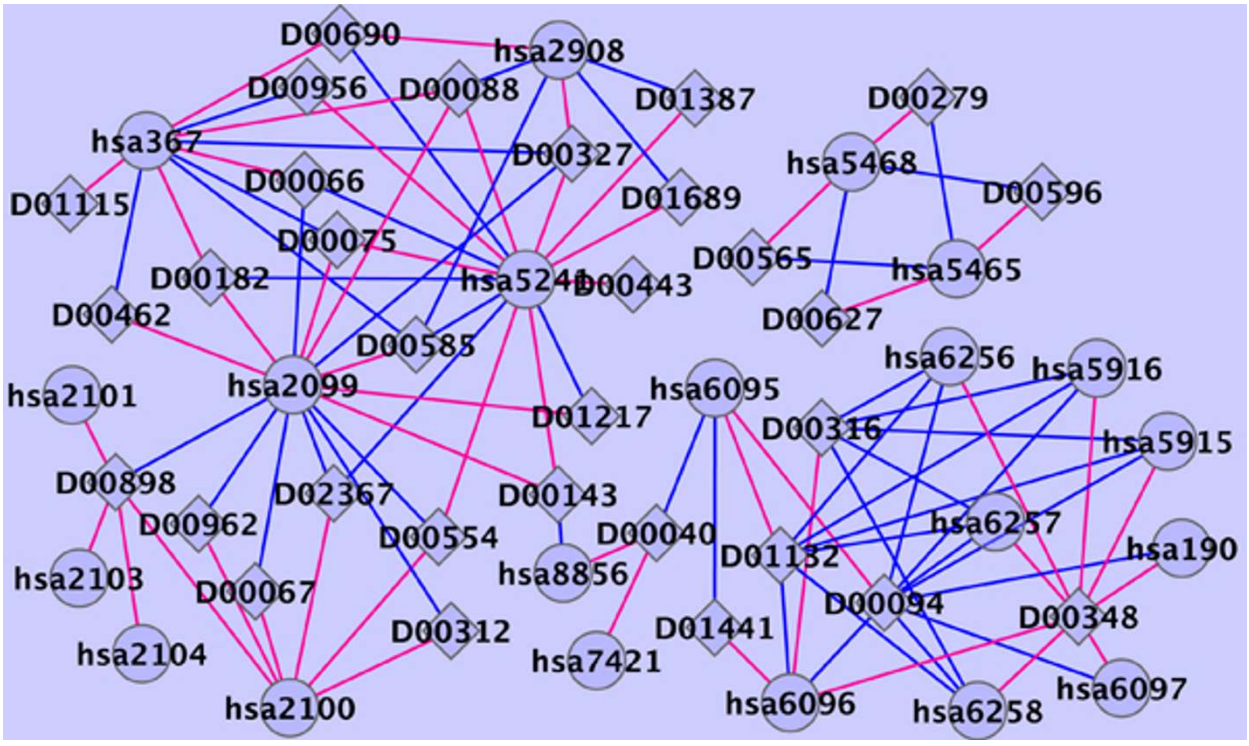

Figure 6 Predicted nuclear receptor interaction network. Predicted nuclear receptor interaction network. diamonds and circles represent drugs and target proteins, respectively. Blue and red lines indicate known interactions and newly predicted interactions with 50 highest scores, respectively.

LapRLS and NetLapRLS for drug-protein interaction prediction

In LapRLS and NetLapRLS, the data-dependent regularization terms are normalized Laplacian operation on graphs. Herein two undirected graphs of drug domain and protein domain including both labeled and unlabeled samples are represented by $\mathbf{K}_d \in R^{n_d \times n_d}$ and $\mathbf{K}_p \in R^{n_p \times n_p}$, where the set of nodes or vertices is $\mathbf{W}_d = \frac{\gamma_{d1}\mathbf{S}_d + \gamma_{d2}\mathbf{K}_d}{\gamma_{d1}+\gamma_{d2}}$. and the set of edges is $\mathcal{E}_d = \{ed_{mn}\}$, $\mathcal{E}_p = \{ed_{mn}\}$ respectively. Each drug d_i or protein p_j is treated as the node on the graph and the weight of edge $ed_{mn}\{ep_{mn}\}$ is $wd_{mn}(wp_{mn})$.

Typically, the weight measures the similarity between two nodes. In our case, the drug domain similarity $\mathbf{W}_d = \{wd_{mn}\}$ is obtained by combining the chemical similarity $\mathbf{S}_d$ and drug-target interaction network. The protein domain similarity $\mathbf{W}_p = \{wp_{mn}\}$ is derived by combining the genomic similarity $\mathbf{S}_p$ and drug-protein interaction network spaces. The chemical similarity $\mathbf{S}_d$ and genomic similarity $\mathbf{S}_p$ have already been introduced in Section Materials.

Next we need to extract the information from the drug-protein interaction network space. The underlying assumption made here is that if two drugs share more target proteins, they are more similar. For example, in Figure 8, the blue line means the known drug-protein interaction while the red line represents the interaction to be predicted. So drug D2 shares 3 same proteins with drug D1 while drug D3 shares a common protein with drug D1. Drug D1 interacts with Protein P4. Based on the assumption here, we can infer that it is more probable that drug D2 interacts with protein P4 than drug D3 does. So another similarity matrix for drug domain from drug-protein interaction network $\mathbf{W}_p = \frac{\gamma_{p1}\mathbf{S}_p + \gamma_{p2}\mathbf{K}_p}{\gamma_{p1}+\gamma_{p2}}$. can be established whose each entry is the number of proteins shared by drug d_i and d_j. Similarly, we can also derive the network similarity matrix $\mathbf{D}_d(k,k) = \sum_{m=1}^{n_d} wd_{k,m}$ whose each entry is the number of drugs shared by protein p_j and p_i. Though drug-protein interaction network was also used in [7], our method employs a different way to extract information from the network. The shortest path concept is

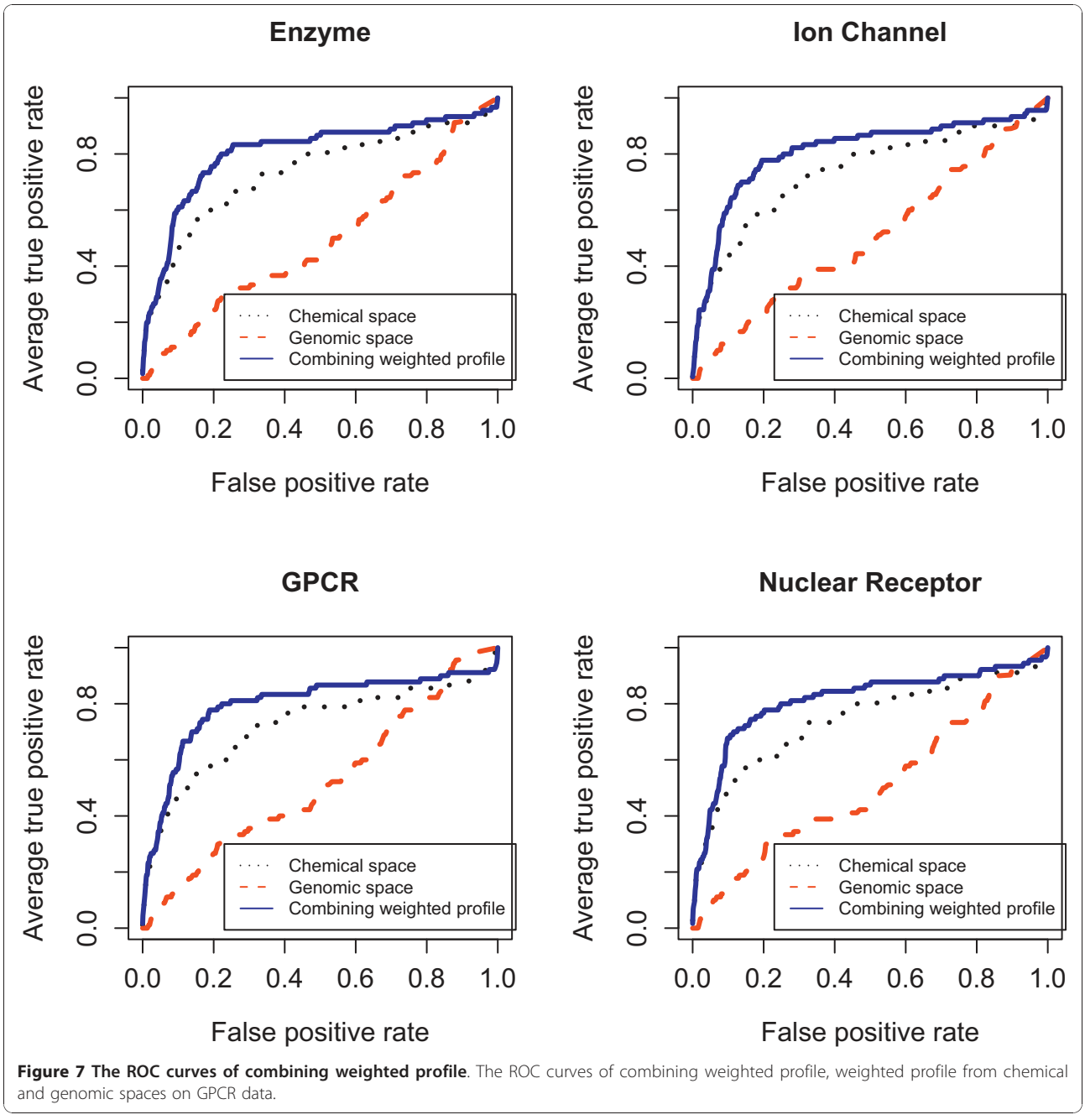

Figure 7 The ROC curves of combining weighted profile. The ROC curves of combining weighted profile, weighted profile from chemical and genomic spaces on GPCR data.

used in [7] while we utilize the number of common nodes shared by two proteins(drugs) to indicate a new similarity measurement.

Now the drug domain similarity $\mathbf{W}_d$ can be derived from the chemical similarity and drug-protein network similarity by linear combination $\mathbf{D}_p(k,k)=\sum_{m=1}^{n_p} wp_{k,m}$. Similarly, the protein domain similarity $\mathbf{W}_p$ can be obtained by $\mathbf{L}_d=\mathbf{D}_d^{-1/2}\Delta_d\mathbf{D}_d^{-1/2}=\mathbf{I}_{n_d\times n_d}-\mathbf{D}_d^{-1/2}\mathbf{W}_d\mathbf{D}_d^{-1/2}$. Compared with the standard LapRLS, our NetLapRLS incorporates drug-protein network information into the prediction model. In the following paragraph, we just describe the method NetLapRLS from which the standard LapRLS can be deduced by setting $\gamma_{Md2} = \gamma_{p2} = 0$.

Given the similarity matrices of drug domain and protein domain, we first perform Laplacian operation on the two graphs which is required by our semi-supervised learning method. The node degree matrices $\mathbf{D}_d$ and $\mathbf{D}_p$ are two diagonal matrices with their (k, k)-element defined as $\mathbf{L}_p=\mathbf{D}_p^{-1/2}\Delta_p\mathbf{D}_p^{-1/2}=\mathbf{I}_{n_p\times n_p}-\mathbf{D}_p^{-1/2}\mathbf{W}_p\mathbf{D}_p^{-1/2}$ and

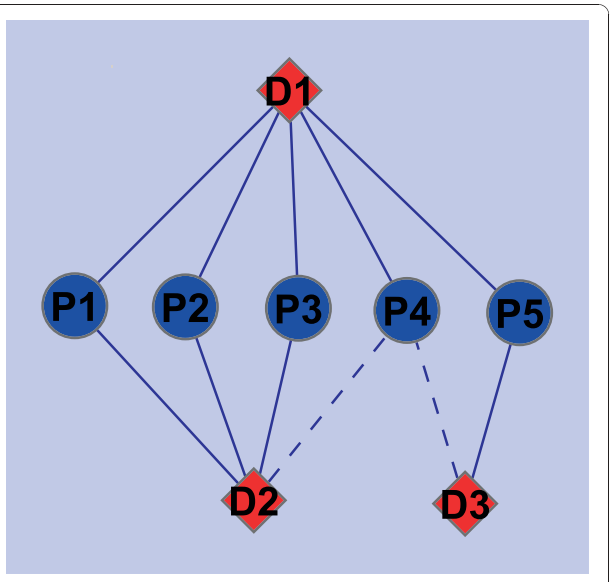

Figure 8 The example of drug-protein interaction network. The example of drug-protein interaction network.

$\mathbf{F}_d \in R^{n_d \times n_p}$. The Laplacian operation of the two graphs is defined as $\Delta_d = \mathbf{D}_d - \mathbf{W}_d$ and $\Delta_p = \mathbf{D}_p - \mathbf{W}_p$ respectively. The normalized graph Laplacians are $\mathbf{F}_p \in R^{n_p \times n_d}$. respectively.

NetLapRLS defines a continuous classification function $\mathbf{F}$ that is estimated on the graph to minimize a cost function. The cost function typically enforces a trade-o between the smoothness of the function on the graph of both labeled and unlabeled data and the accuracy of the function at fitting the label information for the labeled nodes. Herein we extend NetLapRLS to the matrix form. The two continuous classification functions are defined by $\mathbf{F}_d^* = \min_{\mathbf{F}_d} J(\mathbf{F}_d) = \|\mathbf{Y} - \mathbf{F}_d\|_{\mathcal{F}}^2 + \beta_d Trace\left(\mathbf{F}_d^T \mathbf{L}_d \mathbf{F}_d\right)$ and $\mathbf{F}_d^* = \mathbf{W}_d \alpha_d^*$. Let's first address the prediction $\mathbf{F}_d$ on the drug domain. The cost function of NetLapRLS is defined as follows

$$\alpha_d \in R^{n_d \times n_p} \tag{4}$$

where $\alpha_d^* = \arg \min_{\alpha_d \in R^{n_d \times n_p}} \{\|\mathbf{Y} - \mathbf{W}_d \alpha_d\|_{\mathcal{F}}^2 + \beta_d Trace(\alpha_d^T \mathbf{W}_d \mathbf{L}_d \mathbf{W}_d \alpha_d)\}$ is Frobenius norm and *Trace* is the trace of a matrix. Representer theorem [18] shows that the solution is a linear combination

$$-\mathbf{W}_d(\mathbf{Y} - \mathbf{W}_d \alpha_d) + \beta_d \mathbf{W}_d \mathbf{L}_d \mathbf{K}_d \alpha_d = 0$$

Substituting this form into equation (4), we arrive at a convex differentiable objective function with respect to variable $\alpha_d^* = (\mathbf{W}_d + \beta_d \mathbf{L}_d \mathbf{W}_d)^{-1} \mathbf{Y}$

$$\mathbf{F}_d^* = \mathbf{W}_d(\mathbf{W}_d + \beta_d \mathbf{L}_d \mathbf{W}_d)^{-1} \mathbf{Y} \tag{5}$$

The derivative of the objective function vanishes at the minimizer:

$$\mathbf{F}_p^* = \mathbf{W}_p(\mathbf{W}_p + \beta_p \mathbf{L}_p \mathbf{W}_p)^{-1} \mathbf{Y}^T \tag{6}$$

which leads to the following solution:

$$\mathbf{F}^* = \frac{\mathbf{F}_d^* + (\mathbf{F}_p^*)^T}{2} \tag{7}$$

Then we get the prediction from the drug domain in the following form:

$$\mathbf{F}_d^* = \mathbf{W}_d (\mathbf{W}_d + \beta_d \mathbf{L}_d \mathbf{W}_d)^{-1} \mathbf{Y} \tag{8}$$

Similarly, we can also derive the prediction in the protein domain by

$$\mathbf{F}_p^* = \mathbf{W}_p (\mathbf{W}_p + \beta_p \mathbf{L}_p \mathbf{W}_p)^{-1} \mathbf{Y}^T \tag{9}$$

In the end, the predictions from drug and protein domains are combined into

$$\mathbf{F}^* = \frac{\mathbf{F}_d^* + (\mathbf{F}_p^*)^T}{2} \tag{10}$$

Acknowledgements
We thank the colleagues of the Bioinformatics and bioengineering programmatic core, TMHRI for their support and discussion, and Dr. Yamanishi *et al.* for making their data publicly available. Ling-Yun Wu is supported by National Natural Science Foundation of China under grant number 60970091.
This article has been published as part of BMC Systems Biology Volume 4 Supplement 2, 2010: Selected articles from the Third International Symposium on Optimization and Systems Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/1752-0509/4?issue=S2

Author details
[1]Bioinformatics and Bioengineering Program, The Methodist Hospital Research Institute, Weill Medical College, Cornell University, Houston, TX 77030, USA. [2]Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China.

Authors' contributions
Zheng Xia and Ling-Yun Wu co-developed the method, implemented the code and drafted the manuscript. Xiaobo Zhou and Stephen T.C. Wong supervised this project and gave critical revision of the manuscript. Stephen Wong provided the financial support for the work from his bioinformatics and bioengineering program grant from The Methodist Hospital Research Institute (TMHRI), Houston, Texas. All authors read and approved the manuscript.

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

Published: 10 September 2010

References

1. Yao L, Rzhetsky A: **Quantitative systems-level determinants of human genes targeted by successful drugs.** *Genome Res* 2008, **18(2)**:206-13.
2. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK: **Relating protein pharmacology by ligand chemistry.** *Nat Biotechnol* 2007, **25(2)**:197-206.
3. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KLH, Edwards DD, Shoichet BK, Roth BL: **Predicting new molecular targets for known drugs.** *Nature* 2009, **462(7270)**:175-181.
4. Shoichet BK, McGovern SL, Wei B, Irwin JJ: **Lead discovery using molecular docking.** *Curr Opin Chem Biol* 2002, **6(4)**:439-46.
5. Ballesteros J, Palczewski K: **G protein-coupled receptor drug discovery: implications from the crystal structure of rhodopsin.** *Curr Opin Drug Discov Devel* 2001, **4(5)**:561-74.
6. Jacob L, Vert JP: **Protein-ligand interaction prediction: an improved chemogenomics approach.** *Bioinformatics* 2008, **24(19)**:2149-56.
7. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M: **Prediction of drug-target interaction networks from the integration of chemical and genomic spaces.** *Bioinformatics* 2008, **24(13)**:i232-40.
8. Bleakley K, Yamanishi Y: **Supervised prediction of drug-target interactions using bipartite local models.** *Bioinformatics* 2009, **25(18)**:2397.
9. Belkin M, Niyogi P, Sindhwani V: **Manifold regularization: A geometric framework for learning from labeled and unlabeled examples.** *Journal of Machine Learning Research* 2006, **7**:2399-2434.
10. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R.** *Bioinformatics* 2005, **21(20)**:3940-1.
11. Yildirim MA, Goh KI, Cusick ME, Barabsi AL, Vidal M: **Drug-target network.** *Nat Biotechnol* 2007, **25(10)**:1119-1126[http://dx.doi.org/10.1038/nbt1338].
12. Yamanishi Y: **Supervised bipartite graph inference.** *Proceedings of the Conference on Advances in Neural Information and Processing System* **21**:1433-1440.
13. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006, , **34 Database:** D354-7.
14. Wishart D, Knox C, Guo A, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J: **DrugBank: a comprehensive resource for in silico drug discovery and exploration.** *Nucleic Acids Research* 2006, , **34 Database:** D668.
15. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P: **Drug target identification using side-effect similarity.** *Science* 2008, **321(5886)**:263-6.
16. Chapelle O, Schölkopf B, Zien A: *Semi-supervised learning* MIT press 2006.
17. Hattori M, Okuno Y, Goto S, Kanehisa M: **Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways.** *J Am Chem Soc* 2003, **125(39)**:11853-65.
18. Schölkopf B, Smola AJ: *Learning with kernels* MIT press Cambridge, Mass 2002.

doi:10.1186/1752-0509-4-S2-S6
Cite this article as: Xia *et al.*: Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Systems Biology* 2010 **4**(Suppl 2):S6.