The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

Semi-supervised Drug-Protein Interaction Prediction from Heterogeneous Spaces*

Zheng Xia[1,2] Xiaobo Zhou[2] Youxian Sun[1]
Ling-Yun Wu[3,†]

[1]State Key Lab of Industrial Control Technology, Zhejiang University, Hangzhou 310027,China
[2]Center for Biotechnology & Informatics and Department of Radiology, The Methodist Hospital Reseach Institute, Weill Medical College, Cornell University, Houston, TX 77030, USA
[3]Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080,China

Abstract 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 interaction while 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 gives better results than using the labeled data alone. Further, our semi-supervised learning method integrates known drug-protein interaction network information as well as chemical structure and genomic sequence data. We report encouraging results of our method on drug-protein interaction network reconstruction which may shed light on the molecular interaction inference and new uses of marketed drugs.

Keywords Drug-Protein Interaction Network; Semi-supervised Learning; Kernel Methods; Normalized Laplacian.

1 Introduction

Producing a new drug is an expensive and time-consuming process that is subject to a variety of regulations such as drug toxicity monitoring. Meanwhile, there have been many drugs in market approved by U.S. Food and Drug Administration (FDA). Finding the potential use in other therapeutic categories of those FDA approved drugs by predicting their targets is an efficient and time-saving method in drug discovery [12]. Additionally, predicting interactions between drugs and target proteins can help decipher many biological processes. Therefore, there is a strong incentive to develop statistical methods which is capable of detecting these potential drug-protein interactions effectively.

A variety of methods have been proposed to address this *in silico* prediction problem. One of the traditional methods is to predict the drugs interacting with a single given protein based on the chemical structure similarity in a traditional classification framework.

*Supported by National Natural Science Foundation of China (Grant No. 10631070), the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. kjcx-yw-s7), and Beijing Natural Science Foundation (Grant No. 1092011).
†Corresponding author. Email: lywu@amt.ac.cn

This kind of approach does not take advantage of the information in the protein domain. Another widely used method is molecular docking [9] which requires the 3D structure of the target protein. Unfortunately the 3D structures of many proteins are not available [1]. For example, very few G protein-coupled receptors (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 [6, 11]. In [6] the two spaces are encoded together by defining a pairwise kernel which is then fed to the support vector machine (SVM) for classification. The drawback of this kernel framework is that there will be a huge number of samples to be classified (number of drugs multiplies number of proteins) which poses much computational difficulty. Another problem is that the negative drug-protein pairs are selected randomly without experimental confirmation. Yamanishi *et al.*[11] developed a bipartite graph model where the chemical and genomic spaces as well as drug-protein interaction network are integrated into a pharmacological space. In the bipartite model, the known interactions in the training data are labeled as $+1$ and all other unknown drug-protein pairs in the training data are assumed as non-interactions with label 0. Then three different classifiers are possible: new drug candidate versus known target protein, known drugs versus new target protein and new drug candidate versus new target protein candidate. 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. And we prefer only one classifier to predict drug-protein interactions. Lastly, both 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) [2] 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, indicating 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 prediction methods based on LapRLS and NetLapRLS obtain better results than using the labeled data alone. And NetLapRLS which incorporates drug-protein network information provides superior performance than standard LapRLS.

2 Materials

The data used here is downloaded from `<http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/>`[11]. Here we give a brief description.

- Chemical data
  The chemical structure similarity between compounds are calculated by SIMCOMP [5] 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 $\mathbf{S}_d \in \mathscr{R}^{n_d \times n_d}$ which represents the chemical space information.

- 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 $\mathbf{S}_p \in \mathscr{R}^{n_p \times n_p}$ which represents the genomic space.
- Drug-protein interaction data
  At the time when the paper [11] was written, Yamanishi *et al*.[11] established four data sets, in which 445,210,223 and 54 drugs target 664 enzymes, 204 iron channels, 95 GPCRs and 26 nuclear receptors, respectively, and the numbers of known interactions of the four data sets are 2926,1476,635 and 90, respectively.

3 Methods

Semi-supervised learning (SSL) has been attracting much research attention in the machine learning community [4]. 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 [2]. One of the implementations of this framework is the Laplacian regularized least squares (LapRLS) which is very simple and has comparable performance with Laplacian regularized support vector machine.

Consider the drug dataset $\mathbb{D} = \{d_1, \ldots, d_{n_d}\}$ and the target protein dataset $\mathbb{P} = \{p_1, \ldots, p_{n_p}\}$ where n_d and n_p are the numbers of the drugs and proteins in study respectively. An interaction pattern of drug d_i and target protein p_j is represented by a binary label matrix $\mathbf{Y} \in \mathscr{B}^{n_d \times n_p}$. 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 with the whole drug-target interaction space. Another problem is that we only have the information of the interaction. But we do not know which drug target pair has no interaction. That means there are no negative samples in the training process. Herein we first test a simple supervised weighted profile method. And then the standard LapRLS and drug-protein interaction network incorporated NetLapRLS are extended to predict the drug-protein interaction.

3.1 Combining weighted profiles method

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

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

where $S_d(d_i, d_k)$ is a chemical structure similarity score from $\mathbf{S}_d$ and N_{d_i} is a normalization term defined as $\mathrm{N}_{d_i} = \sum_{k=1}^{n_d} S_d(d_i, d_k)$. 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:

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

where $S_p(p_j,p_k)$ is a genomic sequence similarity score from $\mathbf{S}_p$ and N_{p_j} is a normalization term defined by $\mathrm{N}_{p_j} = \sum_{k=1}^{n_p} S_p(p_j,p_k)$. 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

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

The drug-protein pairs (d_i,p_j) in $\bar{f}(d_i,p_j)$ with high scores are predicted to interact each other. The original weighted profile method is used in [11]. However their predictions in the two spaces are not fused. Fig.1 shows that the combining weighted profile method has better performance than the predictions from the single space on the GPCRs data.

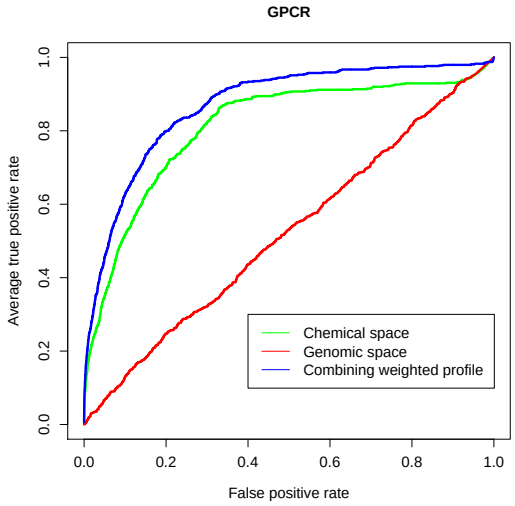

Figure 1: ROC curves of combining weighted profile, weighted profile from chemical and genomic spaces on GPCR data

3.2 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 $\mathscr{G}_d = \{\mathscr{V}_d, \mathscr{E}_d\}$ and $\mathscr{G}_p = \{\mathscr{V}_p, \mathscr{E}_p\}$,where the set of nodes or vertices is $\mathscr{V}_d = \{d_i\}$, $\mathscr{V}_p = \{p_i\}$ and the set of edges is $\mathscr{E}_d = \{e^d_{mn}\}$, $\mathscr{E}_p = \{e^p_{mn}\}$ respectively. Each drug d_i or protein p_j is treated as the node on the graph, and the weight of edge e^d_{mn} or e^p_{mn} is $W_d(m,n)$ or $W_p(m,n)$ respectively. Typically, the weight measures the similarity between two nodes. In our case, the drug domain similarity $\mathbf{W}_d$ is obtained by combining the chemical similarity $\mathbf{S}_d$ and drug-target interaction network. And the protein domain similarity $\mathbf{W}_p$ 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 $\mathbf{S}_p$ are already introduced in section 2.

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 have larger similarity. For example, in Fig. 2, the solid line means the known drug-protein interaction and the dotted line represents the interaction to be predicted. So drug D2 shares 3 same proteins with drug D1 while drug D3 shares 1 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 $\mathbf{K}_d \in \mathscr{R}^{n_d \times n_d}$ 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 $\mathbf{K}_p \in \mathscr{R}^{n_p \times n_p}$ 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 [11], our method employs a different way to extract information from the network. The shortest path concept is used in [11] while we utilize the number of common nodes shared by two proteins(drugs) to indicate a new similarity measurement.

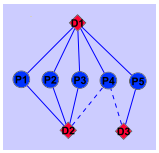

Figure 2: The example of drug-protein interaction network.

Now the drug domain similarity $\mathbf{W}_d$ can be derived from the chemical similarity and drug-protein network similarity by linear combination $\mathbf{W}_d = \frac{\gamma_{d1}\mathbf{S}_d + \gamma_{d2}\mathbf{K}_d}{\gamma_{d1}+\gamma_{d2}}$. Similarly, the protein domain similarity $\mathbf{W}_p$ can be obtained by $\mathbf{W}_p = \frac{\gamma_{p1}\mathbf{S}_p + \gamma_{p2}\mathbf{K}_p}{\gamma_{p1}+\gamma_{p2}}$. 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 $\gamma_{d2} = \gamma_{p2} = 0$.

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 matrixes $\mathbf{D}_d$ and $\mathbf{D}_p$ are two diagonal matrixes with their (k,k)-element defined as $D_d(k,k) = \sum_{m=1}^{n_d} W_d(k,m)$ and $D_p(k,k) = \sum_{m=1}^{n_p} W_p(k,m)$. The Laplacian operation of the two graphs are defined as $\Delta_d = \mathbf{D}_d - \mathbf{W}_d$ and $\Delta_p = \mathbf{D}_p - \mathbf{W}_p$ respectively. The normalized graph Laplacian are $\mathbf{L}_d = \mathbf{D}_d^{-1/2}\Delta_d\mathbf{D}_d^{-1/2} = \mathbf{I}_{n_d \times n_d} - \mathbf{D}_d^{-1/2}\mathbf{W}_d\mathbf{D}_d^{-1/2}$ and $\mathbf{L}_p = \mathbf{D}_p^{-1/2}\Delta_p\mathbf{D}_p^{-1/2} = \mathbf{I}_{n_p \times n_p} - \mathbf{D}_p^{-1/2}\mathbf{W}_p\mathbf{D}_p^{-1/2}$ 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 tradeoff 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 $\mathbf{F}_d \in \mathscr{R}^{n_d \times n_p}$ and $\mathbf{F}_p \in \mathscr{R}^{n_p \times n_d}$. Let's first address the prediction $\mathbf{F}_d$ on the drug domain. The cost function of NetLapRLS is defined as follows

$$\mathbf{F}_d^* = \min_{\mathbf{F}_d} J(\mathbf{F}_d) = \|\mathbf{Y} - \mathbf{F}_d\|_{\mathscr{F}}^2 + \beta_d \|\mathbf{F}_d^T \mathbf{L}_d \mathbf{F}_d\|_{\mathscr{F}}^2 \tag{3}$$

where $\|\cdot\|_{\mathscr{F}}$ is Frobenius norm. Representer theorem [8] shows that the solution is a linear combination

$$\mathbf{F}_d^* = \mathbf{W}_d \alpha_d^*$$

Substituting this form into equation (3), we arrive at a convex differentiable objective function with respect to variable $\alpha_d \in \mathscr{R}^{n_d \times n_p}$

$$\alpha_d^* = \arg \min_{\alpha_d \in \mathscr{R}^{n_d \times n_p}} \{\|\mathbf{Y} - \mathbf{W}_d \alpha_d\|_{\mathscr{F}}^2 + \beta_d \|\alpha_d^T \mathbf{W}_d \mathbf{L}_d \mathbf{W}_d \alpha_d\|_{\mathscr{F}}^2\}$$

The derivative of the objective function vanishes at the minimizer:

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

which leads to the following solution:

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

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

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

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

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

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

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

4 Results

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 carry out a ten-fold cross-validation by splitting the gold 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. The performance is evaluated by using a receiver operating curve(ROC)[10]. For simplicity, we just 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 as 0, the 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, we can see LapRLS and NetLapRLS methods which use unlabeled information provide better performance with respect to AUC score and sensitivity. And the proposed NetLapRLS which incorporates the drug-protein interaction network information gets better result than the standard LapRLS, especially with respective to the sensitivity which is dramatically improved.

Table 1: Statistics of the prediction performance. The AUC is the area under the ROC curve, normalized to 1. The cutoff for sensitivity and specificity is set to select the top τ predictions, where τ is the number of the interactions in the testing data.

Data	Methods	AUC	Sensitivity	Specificity
Enzyme	Weighted profile	0.922	0.06	0.999
	LapRLS	0.950	0.53	0.999
	NetLapRLS	0.983	0.75	0.999
Ion channel	Weighted profile	0.907	0.17	0.997
	LapRLS	0.961	0.36	0.998
	NetLapRLS	0.986	0.72	0.999
GPCR	Weighted profile	0.869	0.13	0.997
	LapRLS	0.934	0.24	0.998
	NetLapRLS	0.971	0.50	0.998
Nuclear receptor	Weighted profile	0.810	0.11	0.994
	LapRLS	0.850	0.16	0.994
	NetLapRLS	0.888	0.21	0.995

Due to the limitation of space, we just focus on the result analysis on GPCRs using our NetLapRLS. Figure 3 shows the predicted top 50 scoring predictions drug-protein interaction network on the GPCRs data using the all known interactions as training data set. Table 2 shows the list of the top 3 predicted drug-protein pairs, with annotation as given in the KEGG database [7]. Searching the latest version of KEGG drug database, we found all the three prediction are now annotated as interacting drug-target pairs. Additionally, 6 predicted new targets (adrenergic receptor class) of drug adrenaline (D00095) are also confirmed in the latest KEGG drug database.

Table 2: Top 3 scoring predicted drug-protein interactions for the GPCRs data.

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

5 Conclusion

In this work, we presented a semi-supervised learning method NetLapRLS for drug-protein interaction prediction by integrating chemical space,genomic space and drug-protein interaction network space. Our method did not need the negative samples and gave a prediction for 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. Further-

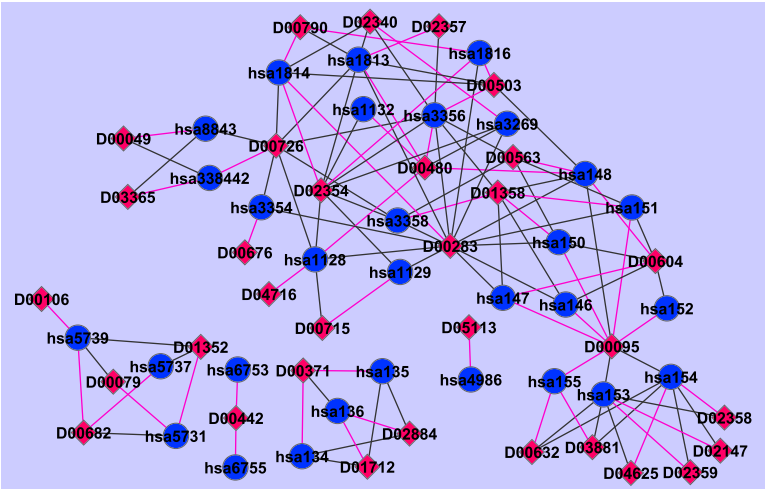

Figure 3: Predicted GPCRs interaction network. Red diamonds and blue circles represent drugs and target proteins, respectively. Gray and red edges indicate known interactions and newly predicted interactions with 50 highest scores, respectively.

more, recently added drug-target interactions to the KEGG immediately allowed us to confirm the 3 most strongly-predicted drug-target interactions and 6 targets of D00095 on GPCRs dataset obtained using our method. This enhanced the strength of our proposed method for real drug-target prediction problems.

The ideal way to use semi-supervised learning for predicting compound-protein interactions is to incorporate different spaces by a multi-task kernel and is fed to typical semi-supervised learning. However, implementation of such large scale semi-supervised learning method will pose considerable computational problems. In the future, we want to incorporate more sophisticated or biologically relevant information into the kernel similarity such as side effect [3] to improve the prediction accuracy.

Acknowledgements

We thank the Bioinformatics Core in The Methodist Hospital Research Institute for their support and Dr.Yamanishi *et al*. for making their data publicly available.

References

[1] J. Ballesteros and K. Palczewski. G protein-coupled receptor drug discovery: implications from the crystal structure of rhodopsin. *Curr Opin Drug Discov Devel*, 4(5):561–74, 2001.

[2] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[3] M. Campillos, M. Kuhn, A. C. Gavin, L. J. Jensen, and P. Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–6, 2008.

[4] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-supervised learning*. MIT press, 2006.

[5] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc*, 125(39):11853–65, 2003.

[6] L. Jacob and J. P. Vert. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, 24(19):2149–56, 2008.

[7] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res*, 34(Database issue):D354–7, 2006.

[8] B. Schölkopf and A. J. Smola. *Learning with kernels*. MIT press Cambridge, Mass, 2002.

[9] B. K. Shoichet, S. L. McGovern, B. Wei, and J. J. Irwin. Lead discovery using molecular docking. *Curr Opin Chem Biol*, 6(4):439–46, 2002.

[10] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–1, 2005.

[11] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–40, 2008.

[12] L. Yao and A. Rzhetsky. Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res*, 18(2):206–13, 2008.