Globalized Bipartite Local Learning Model for Drug-Target Interaction Prediction

Jian-Ping Mei
School of Computer Science
Nanyang Technological University
50 Nanyang Avenue, Singapore
jpmei@ntu.edu.sg

Chee-Keong Kwoh*
School of Computer Science
Nanyang Technological University
50 Nanyang Avenue, Singapore
asckkwoh@ntu.edu.sg

Peng Yang
School of Computer Science
Nanyang Technological University
50 Nanyang Avenue, Singapore
yang0293@e.ntu.edu.sg

Xiao-Li Li
Institute for Infocomm Research
1 Fusionopolis Way #21-01
Connexis, Singapore
lleipuner@researchlabs.org

Jie Zheng
School of Computer Science
Nanyang Technological University
50 Nanyang Avenue, Singapore
zhengjie@ntu.edu.sg

ABSTRACT

Computational methods provide efficient ways to predict possible interactions between drugs and targets, which is critical in drug discovery. Supervised prediction with bipartite Local Model recently has been shown to be effective for prediction of drug-target interactions. However, this pure "local" model is unapplicable to new drug or target candidates that currently have no known interactions. In this paper, we extend the existing supervised learning approach – bipartite local model (BLM) by integrating a strategy for handling new drug and target candidates. Based on the assumption that similar drugs and targets have similar interaction profiles, we present a simple neighbor-based training data inferring procedure and integrate it into the frame work of BLM. This globalized BLM called bipartite local model with neighbor-based inferring (BLMN) then has an extended functionality for prediction interactions between new drug candidates and target candidates. Consistent good performance of BLMN has been observed in the experiment for the prediction of interaction between drugs and four important categories of targets. For the Nuclear Receptors dataset, where there are more chances for the presented strategy to be applied, 20% improvement in terms of AUPR has been achieved. This demonstrates the effectiveness of BLMN and its potential in drug-target interaction prediction.

Keywords

drug-target interaction, new candidate, neighbor-based, local model

*to whom correspondence should be addressed

1. INTRODUCTION

Identification of drug-target interaction is an important part of the drug discovery pipeline. The great advances in molecular medicine and the human genome project provide more opportunities to discover unknown associations in the drug-target interaction network. These new interactions may lead to the discovery of new drugs and also are useful for helping understand the causes of side effects of existing drugs. Since experimental way to determine drug-target interactions is costly and time-consuming, *in silico* prediction comes out to be a potential complement that provides useful information in an efficient way.

Traditional approaches for this task are generally categorized into drug-based approaches and target-based approaches. Drug-based approaches screen candidate drugs, compounds or ligands to predict whether they interact with a given target based on the assumption that similar drugs share the same target. The similarity of two drugs are measured in different ways with respect to different aspects. Other than comparing drugs according to their chemical structures [14], side-effect has also been used to measure the similarity between drugs [2]. Assuming that similar targets bind to the same ligand, target-based approaches, on the other hand, compare proteins to predict whether they bind to the given ligand, or whether they are the targets of the given drug or compound. More specifically, for a given drug, new targets are identified by comparing candidate proteins to the known targets of this drug with respect to certain descriptors such as amino acid sequence, binding sites, or ligands that bind to them. The authors of [8] review computational methods to find new targets for already approved drugs for the treatment of new diseases based on the structural similarity of their binding sites. In [11], targets are compared by the chemical similarity of ligands that bind to them. Different

from these classic drug-based or target-based approaches, chemogenomics approaches have been proposed to consider the interactions between drugs and a protein family rather than a single target [3, 12, 16, 9].

Recently, machine learning approaches have been applied to this task to explore the whole interaction space. [20] proposed a supervised bipartite graph learning approach. In this approach, the chemical space and the geometric space are mapped into a unified space so that those interacted drugs and targets are close to each other while those non-interacted drugs and targets are far away from each each other. After the mapping function to such a unified space is learned, the query pair of drug and target are also mapped in the same way to that unified space, and the probability of interaction between them is the closeness that they are in the mapped space. [1] shows that the combination of supervised learning independently based on drug and target performs very well. This approach is called the Bipartite Local Model (BLM). For a query pair of drug and target, a model of the query drug is learned with a certain classifier based on the information of its known targets. Then the probability of interaction between this drug and the query target is predicted with this model. The same procedure is applied to obtain the probability of interaction between them from the target side. Finally, an overall probability of interaction for the query pair is calculated by combing these two probabilities. It has been shown that the result based the knowledge of both directions, i.e., from the drug side and from the target side, is much better than those based on each single one. The same idea is adopted by another two following work [19] and [13]. In [19], semi-supervised approach is used instead of supervised approach to learn the local model. Laarhoven [13] found that only use the kernel based on the topology of the known interaction network is able to obtain a very good performance, although together with other types of similarities can further improve the results. Other than using one type of drug-drug similarity and one type of target-target similarity, [15] use multiple types of drug-drug similarities and target-target similarities and combine them as features to describe each drug-target pair to train the logistic regression classifier.

Despite the good performance of supervised local approach, it is unable to learn without any positive training data and hence is not able to provide reasonable predictions for drug or target candidates which are currently new, i.e., candidates with no existing interactions. In the existing framework of BLM, the model for the query drug-candidate or target-candidate is learned based on its own preference, i.e., each drug select new targets based on its own way of choosing those already known targets; each target select new targeting drugs according to its discovered targeting drugs. The BLM therefore does not consider information from their neighbors. In this study, we step further to present a modified BLMN to make it applicable to new drug and target candidates by incorporating some necessary globalization. Specifically, when the query involving new drug or target candidates which have no existing interactions, we first derive the initial weighted interactions of the new candidate from its neighbors, and then use them as training data to learn the model and finally give the wanted prediction. Systematic experiments are conducted to simulate the task of drug-target interaction with datasets that have been used in several previous studies. The results show that our proposed approach achieves consistent improvement in the performance in terms of AUC and AUPR scores. As these four datasets contain different portions of new drug and target candidates in our simulation, the improvements of BLMN compared to BLM are also different for the four datasets. The most significant improvement of AUPR is obtained for the Nuclear Receptor dataset, which contains the largest portion of new drug and target candidates. This shows the effectiveness of the presented strategy by inferring training data from neighbors when there is no training data readily available.

Table 1: A summary of datasets.

Dataset	Enzyme	Ion Channel	GPCR	Nuclear Receptor
n_d	445	210	223	54
n_t	664	204	95	26
$\bar{D}_d$	6.58	7.03	2.85	1.67
$\bar{D}_t$	4.41	7.24	6.68	3.46
$D_d = 1(\%)$	39.78	38.57	47.53	72.22
$D_t = 1(\%)$	43.37	11.27	35.79	30.77

2. MATERIALS

To facilitate comparison with state-of-the-art approaches, we used the same groups of four datasets which are first analyzed by [20] and then later by [1], [19], [13] and [4]. These four datasets corresponds to drug-target interactions of four important categories of protein targets, namely enzyme, ion channel, G-protein-coupled receptor and nuclear receptor, respectively [1]. Table 1 gives some statistics of each dataset including the number of drugs (n_d), the number of targets (n_t), the average number of targets ($\bar{D}_d$), the average number of targeting drugs ($\bar{D}_t$), the percentage of drugs that have one target ($D_d = 1$) and the percentage of targets that have one targeting drug ($D_t = 1$). It is indicated from this table that among the four networks, Ion Channel is the most dense while Nuclear Receptor is the most sparse.

Each dataset is described by three types of information in the form of three matrices. Together with the drug-target interaction information, the drug-drug similarity, and target-target similarity are also available. Four interaction networks were retrieved from the KEGG BRITE [10], BRENDA [17], SuperTarget [6] and DrugBank [18] these four databases. The drug-drug similarity is measured based on chemical structures from the DRUG and COMPOUND sections in the KEGG LIGAND database [10] and is calculated with SIMCOMP [7]. The target-target similarity is measured based on the amio acid sequences from the KEGG GENEsS database [10] and is calculated with a normalized version of Smith-Waterman score. More details on how the data have been collected and calculated are given in [20].

3. METHODS

3.1 Problem formalization

The problem under consideration is to predict new interactions between n_d drugs and n_t targets. We use an $n_d \times n_t$ matrix $\mathbf{A}$ to record these known interactions, i.e., $a_{ij} \in \mathbf{A} = 1$ if

[1]The datasets are download from http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/

the ith drug denoted as d_i, is known to interact with the jth target denoted as t_j. All other entries of $\mathbf{A}$ are 0. Assume n_i interactions in total involves m_d drugs and m_t targets and $m_d < n_d$ and $m_t < n_t$. This means there are some new drug and target candidates and the corresponding rows and columns of $\mathbf{A}$ are all 0. Other than the interaction network, $\mathbf{S}_d$ and $\mathbf{S}_t$ are similarity matrix of drug and target, respectively.

3.2 Bipartite Local Learning Model (BLM)

To predict p_{ij}, the probability that a drug d_i and a target t_j interacts, the basic bipartite local model is described as follows: a local model for d_i is first learned based on the known targets of this drug $\mathbf{a}_i$ and the similarities between targets $\mathbf{S}_t$, then this model is used to predict p^d_{ij} the score between this drug to the tested protein based on the local model of d_i. The model learning and prediction process is performed independently from the query target side to get p^t_{ij}. Once both p^d_{ij} and p^t_{ij} are calculated, they are combined to get the final result $p_{ij} = f(p^d_{ij}, p^t_{ij})$.

This framework is first proposed by [1] and then used in [19] and [13]. From the above description, it is clear that the positive interactions in $\mathbf{a}_i$ and pairwise target similarity $\mathbf{S}_t$ are critical to the final prediction of p^d_{ij}. The model learned for d_i describes how this drug selects targets. Once the model is trained, the similarity between the query target and those known targets of this drug largely decides p^d_{ij}. Similarly, positive interactions in $\mathbf{a}_j$ and pairwise drug similarity $\mathbf{S}_d$ are critical to the final prediction of p^t_{ij}. Under the same BLM framework, different results are produced due to the differences in $\mathbf{S}_d$ and $\mathbf{S}_t$, the classifier used, and the merge function f to get p_{ij} based on p^d_{ij} and p^t_{ij}. In our approach, $\mathbf{S}_d$ and $\mathbf{S}_t$ are the linear combination of the similarity derived based on the network topology and the similarity derived based on other sources. Although more sophisticated ways such as Kronecker product have been used to combine two types of similarity matrices or kernel matrices, the linear combination gives comparable performance with a much lower computational complexity.

3.3 Training data inferring for new drug/target candidates

Generally, supervised learning gives better performance than unsupervised learning. However, good performance of supervised learning is largely dependent on the amount and quality of labeled training data. When the drug is new, it has no existing targets that can be used as positive labeled training data and the model for this drug thus can not be learned. Similarly, supervised local learning does not work for new target candidates. To extend the application domain of BLM to new drug and target candidates, we present a neighbor-based procedure and intergrade it into BLM. Based on the assumption that drugs which are similar to each other interact with the same targets, training data for drug/target candidates could be possibly inferred from their neighbors. The neighbors may be defined based on a particular description of drug and target, e.g. chemical similarity of drugs, sequence similarity of targets, or combinations of multiple descriptions. Assume matrices $\mathbf{S}_d$ and $\mathbf{S}_t$ record the drug-drug similarity and target-target similarity, respectively, which are obtained based on a single type

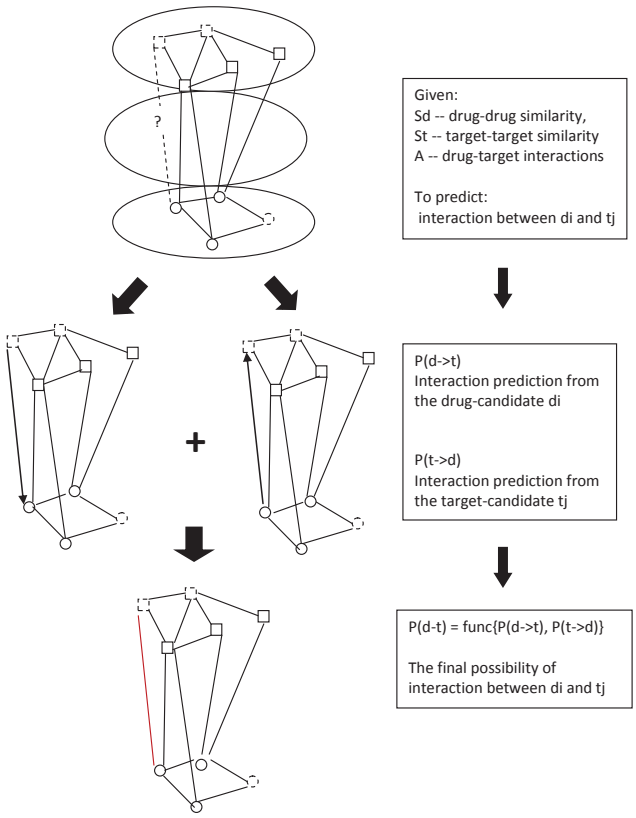


Figure 1: Drug-target interaction prediction with learning from the drug and target independently.

of source or the combinations of multiple sources in certain ways. For a drug d_i that has no known targets, we obtain the inferred weighted training data with the following formula

$$\mathbf{l}(i) = \mathbf{s}^d_i \times \mathbf{A} \tag{1}$$

where each dimension

$$l_j(i) = \sum_h s^d_{ih} \times a_{hj} \tag{2}$$

This shows that the interaction weight of this drug to the $j$target is the sum of its neighbors connected to this target weighted by the similarity between this drug and its neighbors. This simple formula are two folds for a given new drug candidate: the weight to a target is high if many of its neighbors connected to it; the final weight to a target is influenced more by a neighbor with a larger similarity than those with smaller similarities. To only allow neighbors with large similarities to contribute, a threshold may be used to reduce the impact of those non-important neighbors to 0. Alternately an exponential function as below may be introduced:

$$\mathbf{l}(i) = e^{(\mathbf{s}^d_i/\beta)} \times \mathbf{A} \tag{3}$$

This procedure is applied to new target candidates in the same way.

3.4 Bipartite Local Learning Model with Neighbor-based Inferring (BLMN)

Integrating the strategy of inferring labels from neighbors for drug/target candidates into the BLM framework forms the Bipartite Local Learning Model with Neighbor-based Inferring (BLMN). Figure 1 illustrates the idea of drug-target interaction prediction with learning from the drug and target independently.

Since learning from the drug-candidate side and target-candidate side follow the similar procedures, here only the steps of neighbor-based local model learning and prediction from the drug-candidate is given as below:

1. calculate network-based similarity $\mathbf{S}_d^n$
2. combine $\mathbf{S}_d^n$ with $\mathbf{S}_d^0$ to get $\mathbf{S}_d$
3. if d_i is new, inferred labels $\mathbf{l}_i$ from its neighbors
4. Learn a local model for d_i with existing training data or derived training data.
5. Predict the score between this drug to the query protein target p_{ij}^d with the local model and $\mathbf{S}_t$.

Learning from the neighbors allows drugs and targets to obtain training data when themselves do not have any known interactions. This procedure actually introduced some degree of globalization of the original local model to give more chances or an enlarged scope for the learning process. However, too much globalization is not desired as it will decrease the local characteristics and make the models for each drug or target less discriminative. Moreover, the low quality of neighbors or similarity may cause negative impact when neighbors preferences are too much relied upon. Therefore, we only activate the learning from the neighbors for totally new candidates. For other cases, we still train the model locally on its own preference, i.e., the known interaction profile.

3.5 Compared approaches

The proposed work is closely related two existing BLM based approaches [1] and [13]. The main differences of these two approaches are summarized as below:

- BY(2009) [1]: only use chemical similarity and sequence similarity. SVM is used as the classifier. The final probability is the maximum of two independently obtained ones, i.e., $p_{ij} = \max\{p_{ij}^d, p_{ij}^t\}$.
- Laarhoven et al (2011) [13]: chemical similarity is combined with network topology based similarity to get drug-drug similarity; sequence similarity is combined with network topology based to get the target-target similarity. Regularized Least Squares (RSL) classifier is used. The final probability is the mean of two independently obtained ones,i.e., $p_{ij} = 0.5 * (p_{ij}^d + p_{ij}^t)$. Other than this simple linear combination, the version with Kronecker product is shown to be slightly better at the cost of high complexity.

Next we compare the proposed BLMN with these two approaches. We use *func=max* in our experiment as we find *max* is overall slightly better than *mean*. To compare with the state-of-the-art approach, we use the same classifier RSL (with parameter $\delta = 1$) as in [13]. To directly show the improvements attributed to the neighbor-based training data inferring strategy, we also report the results of BLM, i.e., the results of BLMN without using neighbor-based interring.

4. EVALUATION

In order to evaluate the performance of the presented approach, systematic experiments are carried out to simulate the process of bipartite network inference from biological data of four drug-target interaction networks as summarized in Table 1. We use the same leave-one-out cross validation (LOOCV) as in [13]. In each run of the method, one drug-target pair is left out by setting its entry in matrix $\mathbf{A}$ to 0. Then we try to recover its true value using the remaining data.

We measure the quality of the predicted $\mathbf{P}$ compared to $\mathbf{A}$ in terms of the area under ROC curve or true positive vs. false positive curve (AUC) and the area under the precision vs. recall curve (AUPR). It has been discussed in the literature that for skewed datasets where the number of true positives are much less than false positives, the precision-recall curve gives a more informative picture of the performance of the algorithm than the ROC curve [5]. Since in the current task, the known interactions are much less than those unknown ones, or the number of nonzero entries in $\mathbf{A}$ is much less than that of zero entries, the precision-recall curve should be a better measurement than the ROC curve here.

We applied the algorithm with three different groups of inputs: $\mathbf{S}^0$ for drug is the chemical similarity and for target is the sequence similarity; $\mathbf{S}^n$ denotes that the drug-drug similarity and target-target similarity are derived from the existing network; $\mathbf{S}^n + \mathbf{S}^0$ denotes that the drug-drug similarity and target-target similarity are combinations of network derived and biological data derived information. As suggested in [13], gaussian kernel is used to calcualte the network-based similarity with bandwidths $\gamma = \frac{1}{n}\sum_{i=1}^{n} a_{ij}^2$.

4.1 Performance comparison

Table 2 gives the AUC and AUPR scores of different approaches on the four datasets with three different types of similarities. The results of BY (2009) and Laarhoven et al (2011) are the best results quoted from [1] and [13], respectively. From this table, it is shown that in all the cases, BLMN outperforms the other three approaches. The results of BLM and BY(2009) with $\mathbf{S}^0$ are similar as the only difference between them is the former use RSL as the classifier while the later use SVM. The results of BLM and Laarhoven et al (2011) are also close in most of the cases although the later used Kronecker product, which is a more complicated way to combine two types of similarities.

These results clearly show that neighbor-based training data inferring is very useful for improving the final result when the dataset contains new drug/target candidates. Despite the consistent improvements of BLMN compared to the other three on all the four datasets, the amounts of improvements differ for different datasets. If we compare the improvements of the proposed approaches over the four datasets, it is seen that the improvement with respect to BLM on Nu-

clear Receptor is the most significant, i.e., 18% in AUPR with $\mathbf{S}^0 + \mathbf{S}^n$. The improvement on GPCR also achieved 10% in AUPR. Compared to these two datasets, the improvement on Enzyme and Ion Channel, which are 4% and 2%, respectively, are not so significant. Such kind of differences in performance of the proposed approach are consistent with our expectation according to the differences in the structure of the datasets. Although all the datasets do not contain new drug/target candidates, in our experiment, the real interaction to be predicted is leave out. This means drugs and targets with degree equal to 1 turn out to have no positive training data and thus they are simulated to be "new" in the experiments. As shown in Table 1, the percentage of drugs that only have one target is the highest for Nuclear Receptor (72.22 %), and then GPCR (47.53 %). The percentage of targets that only interact with one drug for Enzyme is the highest of the four which is 43.37%, while this number for Ion Channel is only 11%, which is much lower than the other three. This means Nuclear Receptor has a much larger portion of "new" drugs and targets than Ion Channel. Therefore, it has more chances for BLMN to improve the results for Nuclear Receptor where the training data inferring is applied more frequently.

It is also observed that although network-derived similarity alone provides good information, combining biological information can further improves the result especially when the network is sparse, e.g., the results of both BLM and BLMN for Ion Channel with only $\mathbf{S}^n$ is very close to those with $\mathbf{S}^n + \mathbf{S}^0$ while significant improvements are achieved for both approaches on Nuclear Receptor when $\mathbf{S}^0$ is further combined with $\mathbf{S}^n$. This shows that combining multiple types of similarities usually gives better results when no single type of similarity is good enough.

Table 2: Comparison of AUC and AUPR for the four datasets

Dataset	Data	Method	AUC	AUPR
Enzyme	$\mathbf{S}^0$	BY(2009)	97.6	83.3
		BLM	96.6	84.6
		BLMN	98.4	87.9
	$\mathbf{S}^n$	Laarhoven et al (2011)	98.3	88.5
		BLM	98.2	88.0
		BLMN	99.2	91.7
	$\mathbf{S}^n + \mathbf{S}^0$	Laarhoven et al (2011)	97.8	91.5
		BLM	97.9	90.6
		BLMN	**99.2**	**95.0**
Ion Channel	$\mathbf{S}^0$	BY(2009)	97.3	78.1
		BLM	97.3	81.7
		BLMN	98.0	84.6
	$\mathbf{S}^n$	Laarhoven et al (2011)	98.6	92.7
		BLM	98.5	92.5
		BLMN	99.0	95.6
	$\mathbf{S}^n + \mathbf{S}^0$	Laarhoven et al (2011)	98.4	94.3
		BLM	98.2	93.5
		BLMN	**99.1**	**95.8**
GPCR	$\mathbf{S}^0$	BY(2009)	95.5	66.7
		BLM	94.8	64.4
		BLMN	98.0	75.0
	$\mathbf{S}^n$	Laarhoven et al (2011)	94.7	71.3
		BLM	94.4	70.6
		BLMN	97.5	83.9
	$\mathbf{S}^n + \mathbf{S}^0$	Laarhoven et al (2011)	95.4	79.0
		BLM	95.0	76.5
		BLMN	**98.5**	**86.9**
Nuclear Receptor	$\mathbf{S}^0$	BY(2009)	88.1	61.2
		BLM	86.6	58.2
		BLMN	97.3	77.2
	$\mathbf{S}^n$	Laarhoven et al (2011)	90.6	61.0
		BLM	90.9	62.9
		BLMN	95.7	77.0
	$\mathbf{S}^n + \mathbf{S}^0$	Laarhoven et al (2011)	92.2	68.4
		BLM	92.8	70.3
		BLMN	**98.5**	**88.8**

4.2 Analysis

To take a close look at the difference in the results due to the neighbor-based inferring strategy, we compare these top ranked interactions of the Nuclear Receptor dataset produced by BLMN and BLM. Since this dataset has 90 known interactions, we inspect the top 90 interactions predicted by each algorithm.

As illustrated in Fig 3, among the top 90 predicted interactions, BLM correctly detected 56 known interactions while BLMN detected 73. 52 known interactions are ranked within 90 by both. Although compared to BLM, 4 are missed by BLMN, this four interactions get an average rank 106, which indicates that these 4 are still recognized as high probable interactions by BLMN. Nevertheless, 21 interactions detected by BLMN are missed by BLM. The average rank of these 21 produced by BLM is 335 as some of them ranks very low. Among these 21, four pairs namely D00163 (Chenodeoxycholic acid) – hsa9971 (nuclear receptor subfamily 1, group H, member 4), D00506 (Phenobarbital) – hsa9970 (nuclear receptor subfamily 1, group I, member 3) , D01441 (Imatinib mesilate) – hsa6095 (RAR-related orphan receptor A), and D00040 (Cholesterol) – hsa6095 (RAR-related orphan receptor A), which are assigned extremely low ranks by BLM are successfully detected by BLMN. After checking, we find that the query drug D00506 of the first pair only has one target which happens to be the query target has9970, and the query target is known to be only interacted with the query drug. The second pair has the same situation as the first one. Since we leave out the true interaction in our simulation, the testing for these two pairs becomes to predict interaction between new drug candidate and new target candidate. Since no training data is available from both the drug side and target side, BLM is unable to learn a effective model and simply assigns a very small possibility for them. For the other two pairs, the query drug has the only one target which is the query target while the query target has two known interactions including the one with the query drug. In our leave-one-out simulation, such kind of pairs consists of a new drug candidate and a existing target candidate. For BLM, it is unable to get reasonable result from the drug side due to the lack of training data, while it may still detect the interaction if the query drug is similar enough to the drug that known to interact with the query target. However, if the query drug is not similar to that drug, the overall possibility of the interaction between this pair is still considered to be low. Although difficulty is presented for cases when training data is absent for both or one of the query drug and query target, BLMN successfully detected these four pairs to be interact. This clearly shows the effectiveness of the presented approach to infer training data from the interactions of neighbors.

4.3 New interactions predicted

Table 3 gives the top 10 new interactions predicted by BLMN. Figure 4a plots the subnetwork containing existing links and

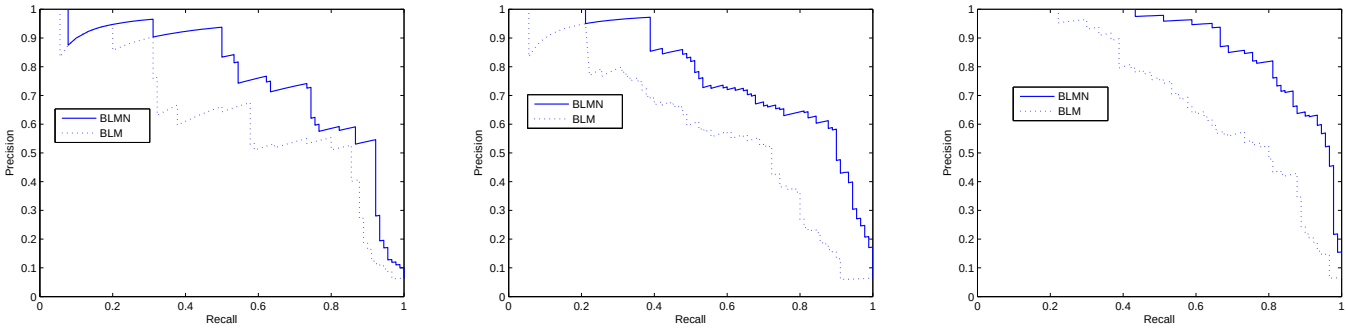


Figure 2: Comparison of Precision-Recall curve between BLM and BLMN on Nuclear Receptor with different similarities. (a) S^0, (b) S^n, (c) $S^0 + S^n$.

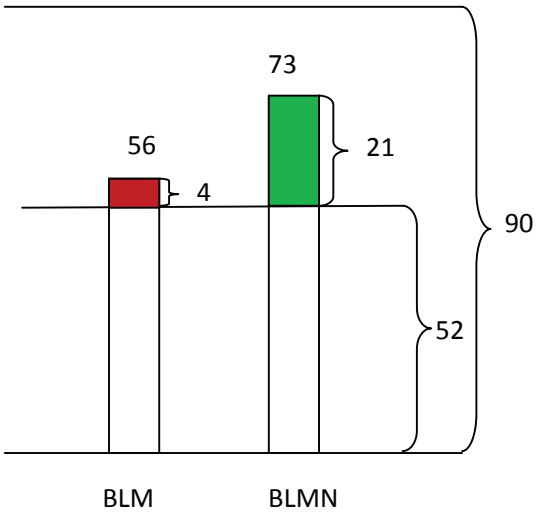


Figure 3: Performance of BLM and BLMN on Nuclear Receptor.

Table 3: Top 10 predicted interactions

Rank	Pair	Description
1	D00316	Etretinate
	hsa6096	RORB; RAR-related orphan receptor beta
2	D01132	Tazarotene
	hsa6097	RORC; RAR-related orphan receptor gamma
3	D00182	Norethisterone
	hsa2099	ESR1; Estrogen receptor 1
4	D00094	Tretinoin
	hsa6095	RORA; RAR-related orphan receptor A
5	D00348	Isotretinoin
	hsa5915	RARB; retinoic acid receptor, beta
6	D00348	Isotretinoin
	hsa5916	RARG; retinoic acid receptor, gamma
7	D00348	Isotretinoin
	hsa6256	RXRA; retinoid X receptor, alpha
8	D00348	Isotretinoin
	hsa6257	RXRB; retinoid X receptor, beta
9	D00348	Isotretinoin
	hsa6258	RXRG; retinoid X receptor, gamma
10	D01132	Tazarotene
	hsa190	NR0B1; nuclear receptor subfamily 0, group B, member 1

9 of those predicted links.

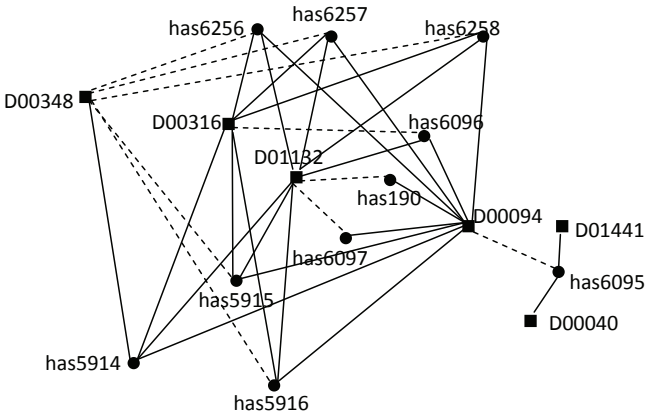
D00316 is predicted to interact with has9096 with a high possibility because D00316 shares six targets with D01132 and D00094, which interact with has9096. Similarly, D01132 shares seven targets with D00094, which indicates these two drugs are similar in terms of interaction profile. Therefore, D01132 is expected to also interact with has6097 and has190, another two targets of D00094. Since has6097 also has a similar sequence with has6096, the overall possibility that D01132 interacts with has6097 is larger than that with has190. D00094 is predicted to interact with has6095, which has large similarities to two known targets has6096 and has6097. Since the chemical similarity between D00348 and D00094 calculated with SIMCOMP is 1, D00348 is expected to share the same targets with D00094.

It is known that has2099 is the target of 10 drugs. Among these drugs, D00066 is the one which is assigned a largest weight during model learning. D00182 shares a target has5241 with D00066, and these two drugs also have a relative large chemical similarity. Combine both, D00182 is very similar with D00066 and hence is predicted to interact with has2099.

According to these results, it is observed that to be assigned with a high possibility of interaction for a given pair of drug and target, they usually satisfy one or more of the following conditions: the query drug has similar chemical structure with drugs that known to interact with the query target; the query drug shares targets with some drugs which interact with the query target; the query target has similar sequence with the known targets of the query drug; the query target and known targets of the query drug interact with the common drugs.

5. CONCLUSION AND DISCUSSION

We have proposed to integrate a neighbor-based inferring procedure into local learning model. This improved model is applicable to new drug/target candidates which have no interactions. From our experimental results, it has been demonstrated that the proposed strategy of inferring training data from neighbors is quite useful for producing effective prediction of interaction involving new candidates. With the presented strategy, the performance of prediction model is



(a)

Figure 4: Existing interactions (solid line) and predicted interactions (dashed line) between drugs (dot) and protein targets (square). (a) 1,3,5,8, and 9 (b) 6,7 and 10 (c) 2 (d) 4.

consistently improved when applied to datasets containing new drug/target candidates. As expected, the improvement is more significant when a larger portion of drug/target candidates are contained which means the proposed strategy is activated more frequently.

In the current work, we only apply neighbor-based inferring for drugs and targets that are totally new, i.e., have no existing interactions, and we find the results are already good enough to show the usefulness of this strategy. It may be applied to a larger portion of drugs and targets which do not have sufficient training data. However, too much emphasis on neighbors tends to eliminate the local characteristics of each drug and target and could cause degeneration of the results. It would be an interesting future work to explore the balance between local information and global information that used in model learning.

Although network-based similarity is reliable and robust, further incorporating other types of data sources is still beneficial to improve the prediction. It provides chances to find new interactions with additional information that enriches the knowledge contained in the interaction network, especially when the network is too sparse to provide enough information.

6. REFERENCES

[1] K. Bleakley and Y. Yamanishi. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, 2009.

[2] M. Campillos *et al.* Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008.

[3] P. R. Caron *et al.* Chemogeominc approaches to drug discovery. *Curr. Opin. Chem. Biol.*, 5:464–470, 2001.

[4] X. Chen *et al.* Drug-target interaction prediction by random walk on the heterogeneous network. *Molecular BioSystems*, DOI: 10.1039/c0xx00000x, 2012.

[5] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proc. 23rd International Conference on Machine Learning*, pages 233–240, 2006.

[6] S. Gnther *et al.* Supertarget and matador: resources for exploring drug-target relationships. *Nucleic acids res.*, 36(Database issue):D919–D922, 2008.

[7] M. Hattori *et al.* 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):11853–11865, 2003.

[8] V. J. Haupt and M. Schroeder. Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Breifings in Bioinformatics*, 2011.

[9] L. Jacob and J.-P. Vert. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, 24(19):2149–2156, 2008.

[10] M. Kanehisa *et al.* From genomics to chemical genomics: new developments in kegg. *Nucleic acids res.*, 34(Database):D354–357, 2006.

[11] M. J. Keiser *et al.* Predicting new molecular targets for known drugs. *Nature*, 462:175–181, 2009.

[12] H. Kubinyi and G. Müller. *Chemogenomics in Drug Discovery.* Wiley-VCH, Weinheim, 2004.

[13] T. V. Laarhoven, S. B. Nabuurs, and E. Marchiori. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics*, 2011.

[14] Y. C. Martin *et al.* Do structurally similar molecules have similar biological activity? *J. Med. Chem*, 45:4350–4358, 2002.

[15] L. Perlman *et al.* Combining durg and gene similarity measures for drug-target elucidation. *Journal of computational biology*, 18:133–145, 2011.

[16] D. Rognan. Chemogenomic approaches to rational drug design. *British Journal of Pharmacology*, 152:38–52, 2007.

[17] I. Schomburg *et al.* Brenda, the enzyme database: updates and major new developments. *Nucleic Asids Res.*, 32(supl-1):D431–433, 2004.

[18] D. S. Wishart *et al.* Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids res.*, 36(Database issue):D901–906, 2008.

[19] Z. Xia *et al.* Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Systems Biology*, 4 (Suppl 2):S6, 2010.

[20] Y. Yamanishi *et al.* Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24:i232–i240, 2008.