Prediction of Drug-Target Interaction with Graph Regularized Non-Negative Matrix Factorization

Xiao-Ying Yan¹ ², Run-Zhou Li¹ Lei Kang¹
¹College of Computer Science, Xi'an Shiyou University, Xi'an 710065, China
²School of Automation, Northwestern Polytechnical University, Xi'an, 710072, China
xiaoying_yan@126.com

Abstract: Identification of drug-target interactions (DTIs) is very important for drug discovery, which can help to find the new uses for an old drug or to discover the off-targets for a given drug. Currently, algorithms have difficulty in finding interactions for new drugs and new targets. We proposed a novel method that uses graph regularized nonnegative matrix factorization framework to predict potential targets/drugs for new drugs/targets by using clustering approaches to construct interaction profiles for new drugs/targets. Compared with other methods, our method obtained the best performance in terms of AUPR.

1. Introduction
Drug reposition is an efficient tool to find new uses for the old drug [1]. However, designing the experiment methods to scan drug-target interactions is impossible, as they are expensive, tedious and time-consuming. This motivates the need for developing methods to predict drug-target interaction [2].

BLM [3] is a classification-based method, in which, drug-target pairs are labeled as positive or negative samples, then adopt SVM as the classifier to predict DTIs. However, we cannot obtain high quality negative samples, of which drug-target pairs are not interacted [2]. Our network-based label propagation with mutual interaction information derived from heterogeneous network (LPMIH) [4] are Network-based method, which uses the graph and network theory to infer the drug-target interactions.

It is a challenging task to predict targets/drugs for new drugs/targets. To solve these problems, Mei et al.[5] proposed BLM-NII algorithm to modify the BLM algorithm which cannot be used for new drugs/targets, by introducing the neighbor-based interaction profile inferring (NII). Van et al. [6] proposed RLS-WNN algorithm by introduced weighted nearest neighbors (WNN). Ezzat et al. [7] introduced weighted K nearest known neighbors (WKNN) into graph regularized matrix factorization model for predicting drug-target interaction. The ideas behind NII, WNN and WKNN are using weighted neighbor interaction profiles to build prior information for new drugs or targets. Recently, many research results show that drugs or targets clustered in the same module tend to function more similarly [8].

In this study, we proposed a new graph regularized nonnegative matrix factorization with clustering adjustment (namely GRNMFC-DT) method to predict potential targets/drugs for new drugs/targets, which using clustering approaches to construct the initial interaction information for new drugs/targets and adopting the graph regularized nonnegative matrix factorization (GRNMF) framework.
2. Datasets and Methods

2.1. Datasets
We used the four standard DTI datasets provided by Yamanishi et al.[1], namely enzymes (Es), ion channels (ICs), G-protein coupled receptors (GPCRs) and nuclear receptors (NRs), which can be downloaded from http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/. Some statistics of each dataset are listed in Table 1.

There are three matrices in each dataset, drug chemical structures similarity matrix \( S_{nn} \times \), target protein sequence similarity matrix \( S_{mm} \times \) and drug-target interaction matrix \( Y_{nm} \times \). For \( Y_{nm} \times \), if \( d_i \) interact with \( t_j \), \( y(i,j) = 1 \), otherwise \( y(i,j) = 0 \). The \( i \)-th row in \( Y_{nm} \times \) is the interaction profile for drug \( d_i \), denoted as \( Y(d_i) \), the \( j \)-th column in \( Y_{nm} \times \) is the interaction profile for target \( t_j \), denoted as \( Y(t_j) \).

| Dataset   | \( N_d \) | \( N_t \) | \( E_{dt} \) |
|-----------|----------|----------|-------------|
| Es        | 445      | 664      | 2926        |
| ICs       | 210      | 204      | 1476        |
| GPCRs     | 223      | 95       | 635         |
| NRs       | 54       | 26       | 90          |

2.2. Mining drug/target modules to adjust the drug-target interaction matrix
For new drug \( d_k \), the values in interaction profile \( Y(d_k) \) are all zeros. Similarly, for new target \( t_l \), the values in interaction profile \( Y(t_l) \) are all zeros too, which may cause difficulty in the prediction for new drugs/targets.

Here, we constructed new interaction profiles based on the clustering results to solve the above problem. Based on the drug similarity matrix \( S_{nn} \times \) and target similarity matrix \( S_{mm} \times \), we used the ClusterONE [9] to detect the meaningful cohesive clusters (or modules) \( M_d \) and \( M_t \), respectively.

For drug \( d_k \), we combined the interaction profile of other known drugs in the same cluster \( M_d \) to produce the following interaction profile:

\[
Y_{d_k} = \frac{1}{Q_d} \sum_{i \in M_d} w_i Y_{d_i}
\]

(1)

Here, \( d_k \) is a known drug and is at the same cluster \( M_d \) with \( d_i \); \( i \) is the ordered index of known drugs in \( M_d \); \( w_i \) is the weight coefficient and is defined as \( w_i = \mu^{-1} \ast S(d_i,d_k), \mu \in [0,1] \); \( Q_d \) is for normalization and is defined as \( Q_d = \sum_{i \in M_d} S(d_i,d_k) \).

For target \( t_j \), we conducted the same procedure to obtain the new interaction profile for each target:

\[
Y_{t_j} = \frac{1}{Q_t} \sum_{i \in M_t} w_j Y_{t_j}
\]

(2)

\[
w_j = \mu^{-1} \ast S(t_i,t_j), \mu \in [0,1]
\]

(3)

\[
Q_t = \sum_{j \in M_t} S(t_i,t_j)
\]

(4)
Then, we combined these two matrices $Y_d$ and $Y_t$ as following to obtain a new interaction matrix $Y_{new} = \frac{a_1 Y_d + a_2 Y_t}{a_1 + a_2}$, here we set $a_1 = a_2 = 1$. After replacing $y(i, j) = 0$ with the associated values in $Y_{new}$, we obtained the final drug-target interaction matrix $Y_{new}$.

2.3. Graph regularized nonnegative matrix factorization with clustering adjustment for prediction of drug-target interaction

2.3.1. GRNMF

For above obtained non-negative data matrix $Y'_{new}$, the standard NMF is used to decompose $Y'_{new}$ into two low rank non-negative matrices $A_{k 	imes n}$ and $B_{m 	imes l}$ ($k << \min(n, m)$), and minimize the reconstruction error between $Y'$ and $AB^T$. The most widely used error functions is defined as following:

$$\min_{A,B} \| Y' - AB \|^2_F \text{ s.t. } A \geq 0, B \geq 0 \tag{5}$$

In order to avoid over-fitting and to improve the learning performance, we adopt the graph regularized nonnegative matrix factorization framework, which has been successfully used for identifying microRNA-disease associations[10]. The modified objective function is as following [10]:

$$\min_{A,B} \| Y' - AB \|^2_F + \alpha \| A \|_F^2 + \| B \|_F^2$$

$$+ \beta \sum_{i,p=1}^{n} |a_i - a_p| \| S'_d(d_i, d_p) \|_F$$

$$+ \gamma \sum_{j,q=1}^{m} |b_j - b_q| \| S'_t(t_j, t_q) \|_F \text{ s.t. } A \geq 0, B \geq 0 \tag{6}$$

Here, $\alpha$, $\beta$, and $\gamma$ are the regularization coefficients; $a_i$ and $b_j$ are the $i$-th and $j$-th row of $A$ and $B$; $n$ and $m$ are the numbers of drugs and targets; the first part is for the NMF of $Y'$; the second part is for $A$ and $B$ smoothness[11]; the last two parts are graph regularization to minimize the distance between latent neighbors; $S'_d$ and $S'_t$ are sparse similarity matrices for drug and target.

Sparsification is needed before graph regularization[10]. In this paper, we utilize the above mentioned clustering results to produce a weight matrix $W^d$ for drugs, the element $W^d_{i,j} = 1$, if $d_i,d_j \in M_d$; otherwise, $W^d_{i,j} = 0$. At the same way, we can produce a weight matrix $W^t$ for targets.

Subsequently, sparse drug similarity matrix $S'^d_d$ and sparse target similarity matrix $S'^t_t$ are determined by $S'^d_d = W^d_d S^d_d$ and $S'^t_t = W^t_t S^t_t$, respectively.

2.3.2. Optimization

The objective function can be rewritten as[10]:

$$\min_{A,B} \| Y' - AB \|^2_F + \alpha \| A \|_F^2 + \| B \|_F^2$$

$$+ \beta Tr(A^T L_d A)$$

$$+ \gamma Tr(B^T L_t B) \text{ s.t. } A \geq 0, B \geq 0 \tag{7}$$

Here, $Tr(\cdot)$ is the trace of a matrix; $L_d = D_d - S'^d_d$ and $L_t = D_t - S'^t_t$ are the graph Laplacian matrices.
for $S_d$ and $S_t$, respectively. $D_d = \sum_i S_d^i$ and $D_t = \sum_i S_t^i$ are the diagonal matrices.

After using Lagrange function and Karush-Kuhn-Tucker (KKT) conditions [12], and iteratively updated $A$ and $B$ as follows. The more details about update computation can be find in literature [10].

\[
\begin{align*}
    d_{ik} &= d_{ik} \frac{(Y^T B + \beta S_d^i) A_d}{(AB^T + \alpha A + \beta D_d A)_d} \\
    b_{jk} &= b_{jk} \frac{(Y^T A + \gamma S_t^j) B_j}{(BA^T + \alpha B + \gamma D_t B)_j}
\end{align*}
\]

(8)

After the update process converged, we obtained the predicted drug-target interaction matrix as $Y^* = AB^T$. The top-ranked targets in each row of $Y^*$ are more likely to be interacted with the corresponding drug.

### 3. Test Results and Discussions

To evaluate the GRNMFC-DT performance, we compared with the existing methods: BLM-NII [5], RLS-WNN [6] and WKNKN+WGRMF [7] on the four datasets. Since we only focus on the ability to predict targets for new drugs as well as to predict drugs for new targets, we applied two kinds of 10-CVs, that is 10-CV$_d$ and 10-CV$_t$. For new drugs prediction, all drugs were randomly divided into ten subsets with equal size. Each in turn as the test set, and all the interactions of these drugs should be deleted (CV$_d$), the other nine subsets are training set. In the same way, for new targets prediction, all the targets are divided into test set and train set, in which all the interaction in test set are deleted (CV$_t$). For parameters, based on the results of CV experiments on the training dataset, we selected $k=25$ and $\alpha = \beta = \gamma = 1$.

#### Table 2 AUPR results of BLM-NII, RLS-WNN, WKNKN+WGRMF and GRNMFC-DT for new drugs in 10-CV$_d$ test

| Methods          | Es (0.011) | ICs (0.009) | GPCRs (0.010) | NRs (0.043) |
|------------------|------------|-------------|---------------|-------------|
| BLM-NII          | 0.167      | 0.201       | 0.233         | 0.410       |
| RLS-WNN          | 0.386      | 0.319       | 0.363         | 0.519       |
| WKNKN+WGRMF      | 0.401      | 0.369       | 0.401         | 0.528       |
| GRNMFC-DT        | 0.420      | 0.371       | 0.432         | 0.541       |

The performance is measured by using AUPR (area under the PR curve), as AUPR heavily punishes highly ranked false positives [4]. The results of our GRNMFC-DT method and other three methods on CV$_d$ and CV$_t$ are listed in Table 2 and Table 3. From those, we can see that GRNMFC-DT achieved higher performance than other methods, indicating that our GRNMFC-DT method can be used for predicting targets for new drugs and predicting drugs for new targets.

#### Table 3 AUPR results of BLM-NII, RLS-WNN, WKNKN+WGRMF and GRNMFC-DT for new targets in 10-CV$_t$ test

| Methods          | Es (0.011) | ICs (0.009) | GPCRs (0.010) | NRs (0.043) |
|------------------|------------|-------------|---------------|-------------|
| BLM-NII          | 0.583      | 0.634       | 0.447         | 0.418       |
| RLS-WNN          | 0.761      | 0.746       | 0.547         | 0.468       |
| WKNKN+WGRMF      | 0.798      | 0.799       | 0.585         | 0.446       |
| GRNMFC-DT        | 0.831      | 0.800       | 0.601         | 0.459       |

#### 4. Conclusions

In this article, we presented a novel method GRNMFC-DT to predict targets/drugs for new
drugs/targets by using clustering approaches to construct the initial interaction information for new drugs/targets and based on graph regularized nonnegative matrix factorization framework to predict. Compared with other recent state-of-the-art methods, those are BLM-NII, RLS-WNN and WKNKN+WGRMF, our algorithm GRNMFC-DT obtains the best performance in terms of AUPR.

Acknowledgements
This work was supported by National Natural Science Foundation of China under grant 51707158 and Special scientific research project of Education Department of Shaanxi Province of China under grant 17JK0603.

References
[1] Ashburn T T, Thor K B. Drug repositioning: identifying and developing new uses for existing drugs[J]. Nature Reviews Drug Discovery, 2004, 3(8): 673-683.
[2] Zhang S W, Yan X Y. Some Remarks on Prediction of Drug-Target Interaction with Network Models[J]. Current Topics in Medicinal Chemistry, 2017, 17(999).
[3] Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models[J]. Bioinformatics, 2009, 25(18): 2397-2403.
[4] Yan X-Y, Zhang S-W, Zhang S-Y. Prediction of drug–target interaction by label propagation with mutual interaction information derived from heterogeneous network[J]. Molecular BioSystems, 2016.
[5] Mei J-P, Kwoh C-K, Yang P, et al. Drug–target interaction prediction by learning from local information and neighbors[J]. Bioinformatics, 2013, 29(2): 238-245.
[6] Van L T, Marchiori E. Predicting Drug-Target Interactions for New Drug Compounds Using a Weighted Nearest Neighbor Profile[J]. PloS one, 2013, 8(6): e66952.
[7] Ezzat A , Zhao P , Wu M , et al. Drug-Target Interaction Prediction with Graph Regularized Matrix Factorization[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2017, PP(99):646-656.
[8] Zhao S, Shao L. A co-module approach for elucidating drug–disease associations and revealing their molecular basis[J]. Bioinformatics, 2012, 28(7): 955.
[9] Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks[J]. Bioinformatics, 2015, 9(5): 471-472.
[10] Xiao Q, Luo J, Liang C, et al. A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations[J]. Bioinformatics, 2017, 34(2): 239–248.
[11] Guan N , Tao D , Luo Z , et al. Manifold Regularized Discriminative Nonnegative Matrix Factorization With Fast Gradient Descent[J]. IEEE Transactions on Image Processing, 2011, 20(7):2030-2048.
[12] Facchinei F , Kanzow C , Sagratella S . Solving quasi-variational inequalities via their KKT conditions[J]. Mathematical Programming, 2014, 144(1-2):369-412.