

Predicting Unknown Interactions Between Known Drugs and Targets via Matrix Completion

Qing Liao¹(✉), Naiyang Guan², Chengkun Wu², and Qian Zhang¹

¹ Department of Computer Science and Engineering,
Hong Kong University of Science and Technology, Hong Kong, Hong Kong
{qnature,qianzh}@cse.ust.hk

² College of Computer, National University of Defense Technology, Changsha, China
{chengkun_wu,ny_guan}@nudt.edu.cn

Abstract. Drug-target interactions map patterns, associations and relationships between drugs and target proteins. Identifying interactions between drug and target is critical in drug discovery, but biochemically validating these interactions are both laborious and expensive. In this paper, we propose a novel interaction profiles based method to predict potential drug-target interactions by using matrix completion. Our method first arranges the drug-target interactions in a matrix, whose entries include interaction pairs, non-interaction pairs and undetermined pairs, and finds its approximation matrix which contains the predicted values at undetermined positions. Then our method learns an approximation matrix by minimizing the distance between the drug-target interaction matrix and its approximation subject that the values in the observed positions equal to the known interactions at the corresponding positions. As a consequence, our method can directly predict new potential interactions according to the high values at the undetermined positions. We evaluated our method by comparing against five counterpart methods on “gold standard” datasets. Our method outperforms the counterparts, and achieves high AUC and F_1 -score on enzyme, ion channel, GPCR, nuclear receptor and integrated datasets, respectively. We showed the intelligibility of our method by validating some predicted interactions in both DrugBank and KEGG databases.

Keywords: Drug-target interaction · Matrix completion · Drug discovery

1 Introduction

Associations between drugs and targets are essential for understanding the pharmacology of drugs and for repositioning known drug [1–4]. Capturing associations between drugs and targets using traditional biochemical experiments is a laborious and time-consuming procedure that is also very expensive [5–7]. One alternative is to compute potential associations between drugs and targets via

in-silico way [8,9]. Molecular docking [10–15], literature mining [16] and ligand-bases [17–19] are three common computational approaches. Docking requires information about the 3D structure of a target/protein to calculate how well each drug candidate can bind with the target, but this type of information is missing for many targets, like GPCR and ion channel [20,21]. Moreover, docking is computationally expensive, which makes it difficult to process large-scale datasets. Text mining approaches is heavily relied on domain dictionaries to deal with semantic ambiguity like aliases and synonyms in the literature [16]. Ligand-based approaches such as QSAR (Quantitative Structure Activity Relationship) compare a candidate ligand with the known ligands of a target protein to predict its bindings [17,18], and the performance of ligand-bases approach decreases when the number of known ligands is limited [19].

Recently, machine learning methods has been shown to be effective in finding the drug-target interactions based on chemical properties of drug compounds, genomic properties of targets, and interaction profiles [22–27]. Those approaches share the identical assumption that similar drugs tend to interact closely with similar target proteins [28,29]. Existing studies utilized drug/target similarity information and known interactions to capture potentially novel interactions between drugs and targets. For instance, Jacob and Vert proposed a SVM-based method [30–32] called pairwise kernel method (PKM) to generate similarity kernels over drug-target pairs [27]. Yamanishi [22] proposed the kernel regression-based method (KRM) to infer unknown drug-target interaction in a unified space called “pharmacological space”. Yamanishi [23] proposed the bipartite graph inference (BLM) method, which builds a bipartite local model to predict links between drugs and targets in a bipartite graph. Gonen [24] proposed Kernelized Bayesian matrix factorization (KBMF2K) to predict interactions by projecting drug compounds and target proteins onto a unified subspace via joint Bayesian formulation. Laarhoven [25] proposed a Gaussian interaction profile (GIP) method to predict drug-target interactions by generating a Gaussian kernel from interaction profiles and similarity information among drugs and among targets. Xia [26] proposed a NetLapRLS method by incorporating Laplacian regularized least square (LapRLS) and a new kernel established from the known interaction network in a unified framework. Moreover, Wang and Zeng [33] built a restricted boltzmann machine (RBM) method to predict drug-target pairs and to describe types of predicted pairs.

Some methods utilize similarity information and partial drug-target interaction profiles to predict potential interactions between drugs and targets. However, similarity information might not be available in some cases. For example, it is extremely difficult to collect complete similarity information in large scale database like STITCH [34] and the 3D shape similarity of many proteins/targets, especially GPCRs are unavailable [21,35]. On the other side, drug similarity can be calculated based on different types of biological knowledge such as chemical structure (CS), and anatomical therapeutic chemical classification system (ATC) [32], and target similarity can also be calculated from genomic sequence (GS) [22,36] and gene ontology (GO) [37,38]. It is difficult to decide which types

of similarity is the most appropriate one, as each measure has its private biochemical properties. Zheng [39] showed that the effect of the same type of similarity might vary dramatically on different datasets. For example, the GS similarity over target is very critical for predicting interactions on GPCR dataset, while it is almost useless on the nuclear receptor, ion channel and enzyme datasets. Therefore, integrating different similarity metrics is still a challenging problem.

Drug-target interaction prediction problem can infer interactions with less information. For example, Cheng [40] proposed a network-based inference (NBI) method to predict interactions by using the interaction profiles only. Moreover, Cobanoglu and Bahar *etc.* [41] assumed that all the samples obey the Gaussianly distributed probability to present probabilistic matrix factorization (PMF) to predict pairs by only using the interaction profiles. Experiments demonstrate that the prediction of NBI is more reliable than drug-based similarity inference (DBSI) and target-based similarity inference (TBSI), because choosing an improper similarity data may introduce extra noise to the model building process due to inaccurate selection of similar pairs. Although above mentioned methods exhibit a satisfactory AUC value (area under ROC curve) performance, its precision and recall are still unsatisfactory.

In this paper, we proposed a novel drug-target interaction prediction method which uses the matrix completion for prediction based on only interaction information. Our method assumes that similar drugs often interact with similar proteins and converts the interaction prediction problem into a collaborative filtering problem, which infers missing entries in the interaction matrix by using known interactions. We evaluated our method by comparing with the existing methods and ten-fold cross-validation on a “gold standard” datasets including enzyme, ion channels, G-protein-coupled receptors (GPCRs), nuclear receptor and integrated datasets. Experimental results show that our method achieves high performance in both AUC and F_1 -score, and validation of predicted pairs in the latest DrugBank and KEGG databases shows that our method is intelligible.

2 Methods

2.1 Drug-Target Interaction Databases

In the drug-target interaction prediction literature, four datasets include enzyme, ion channels, GPCRs, and nuclear receptor are usually regarded as the “gold standard” dataset [22–27, 39, 40]. In this study, we also combined them together to generate an integrated dataset for further verification. Yamanishi [22] proposed a widely used benchmark for drug-target interactions which includes four subsets for different types of targets: enzyme, ion channels, GPCRs and nuclear receptor. These datasets were collected from curated databases including KEGG BRITE [42], BERENDA [43], SuperTarget [44] and DrugBank [45], respectively. The numbers of drugs, targets and drug-target interactions are listed in Table 1. We also generated an integrated dataset that combines all four subsets. The integrated data of these four datasets contains 5127 interactions between 989 target proteins and 791 drugs.

Table 1. Summary of the drug-target interaction datasets.

Datasets	# of drugs	# of targets	# of drug-target interactions
Enzyme	445	664	2926
Ion channel	210	204	1476
GPCR	223	95	635
Nuclear receptor	54	26	90
Integrated	791	989	5127

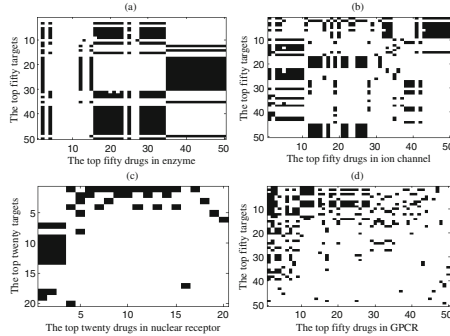


Fig. 1. Visualization of the low-rank pattern of drug-target interactions. The pattern of sorted interactions in four datasets. (a) Sorted interaction between first fifty drugs/target proteins in enzyme dataset. (b) Sorted interactions between first fifty drug/target proteins in ion channel dataset. (c) Sorted interactions between first twentieth drugs/target proteins in nuclear receptor dataset. (d) Sorted interaction between first fifty drug/target proteins in GPCR dataset.

2.2 Motivation by Data Visualization

To predict unknown interactions from the drug-target dataset, we first analyze the “gold standard” datasets. Due to drug-target interaction matrix is too sparse to difficult find some observations, we sorted the drug and target protein in descending order by their number of interactions to make figure. Figure 1 is an example to show the interaction matrix of the first fifty or twenty drugs/target proteins based on sorted interactions in four datasets. From Fig. 1(a), we can observe that many drugs have similar interactions, such as drugs #16 to #23 [KEGG:D00538, D00224, D00377, D00391, D00401, D01069, D00136, D00569], and drugs #28 to #34 [KEGG:D03778, D03781, D00947, D00963, D01397, D02441, D03776] have very similar interactions. The same observation can also be found from the target protein side in Fig. 1(a). For instance, if a target protein has interactions from drugs #28 to #34, it may also have interaction with drug #2 [KEGG:D00521], because there exists a strong relationship between drug #28 to #34 and drug #2. From the above observation, it is evident that correlation does exist between node instances in the enzyme dataset.

For instance, if a target has interactions with drug #28 to #34, then it has a high probability to interact with #2. The same phenomenon can also be observed on the other subsets. We also analyzed the most important part of the ion channel, nuclear receptor and GPCR subset and found the same observations in Fig. 1(b), (c), and (d), respectively. In Fig. 1(c), we only select the first twenty drug/target proteins from the nuclear receptor subset, as the interaction matrix size is a 56×26 matrix. From Fig. 1(c) and (d), the observation is less evident comparing with Fig. 1(a) and (b), but the same observation can still be found from target proteins #9 to #13 [KEGG:hsa3174, hsa367, hsa4306, hsa5241, hsa5465] in Fig. 1(c) and from drugs #10 to #13 [KEGG:D00136, D00139, D00180, D00225], and from target proteins #6 to #9 [KEGG:hsa1129, hsa1131, hsa1132, hsa1133] in Fig. 1(d).

According to these observations, all four subsets have some latent factors that contribute to the prediction of interactions. In other words, from the viewpoint of drugs, all drugs interact with the target proteins in a few patterns, and thus we can leverage the known interaction information to predict unknown interactions by predicting missing values of the interaction matrix. It is also true from the viewpoint of target proteins. It therefore motivates us to apply the matrix completion technique to the drug-target interaction prediction problem.

2.3 The Drug-Target Interaction Prediction Method

The original interaction dataset needs to be pre-processed in order to apply the matrix completion. In the benchmark dataset, the value 1 denotes a confirmed/annotated interaction (positive sample), while all other unknown drug-target pairs in the training data are assumed to be non-interacted (negative sample), which is denoted by the value 0. However, our method regards 0 valued entries as the ones to be predicted. Therefore, directly performing matrix completion on original dataset may lead confusion. So we fill non-interaction entries by one value a and interaction entries by another value b , where $a \neq b$ and $\{a, b\} \neq 0$. Our method first fills missing entries with value 0 in the original matrix to be recovered by using the matrix completion, and then iteratively updates the missing entries with the predicted value.

Given a drug-target interaction matrix $M \in R^{N_d \times N_t}$ involving N_d drugs and N_t targets. The set $X_d = \{d_1, d_2, \dots, d_{N_d}\}$ is the drug set and the set $X_t = \{t_1, t_2, \dots, t_{N_t}\}$ is the target protein set. Let $M_{(i,j)} : (i, j) \in \Omega$ denote the set of the known samples, and ω the index set of the rest known samples. We formulate the interaction prediction problem as below:

$$\begin{aligned} & \min_X f_\tau(X) \\ & s.t., P_\Omega(X) = P_\Omega(M) \end{aligned} \tag{1}$$

where $f_\tau(X)$ is a nonlinear function of candidate solution matrix X . P_Ω is an orthogonal projector, and $P_\Omega(X)$ is equal to $X(i, j)$ if $(i, j) \in \Omega$, and $P_\Omega(X)$ is equal to zero otherwise. In matrix completion, the predicted matrix X is usually expected to be low-rank. We therefore rewrite (1) as the following problem to

minimize the rank of X because nuclear norm of X is a convex surrogate of its rank:

$$\begin{aligned} \min_X \tau \|X\|_* + \frac{1}{2} \|X\|_F^2 \\ \text{s.t.}, P_\Omega(X) = P_\Omega(M) \end{aligned} \tag{2}$$

where $\|X\|_*$ signifies the nuclear norm of X which is actually the sum of singular value of matrix X , and $\|X\|_F$ is the Frobernius norm of matrix X , and $\tau \geq 0$ is a thresholding which will be used in soft-thresholding operator.

According to [46], the problem (2) can be optimized by using the Lagrangian multiplier method. Specially, we introduce a Lagrangian multiplier Y and get the Lagrangian function of (2) as below:

$$L(X, Y) = f_\tau(X) + \langle Y, P_\Omega(M) - P_\Omega(X) \rangle \tag{3}$$

Applying the Uzawas algorithm [47] to find a saddle point of (3) until convergence.

The Uzawas algorithm first updates X with Y fixed as

$$X^k = D_\tau(Y^{k-1}) \tag{4}$$

followed by updating Y with X fixed as

$$Y^k = Y^{k-1} + \delta_k P_\Omega(M - X^k) \tag{5}$$

where $\{\delta_k\}_{k \geq 1}$ is a sequence of step size, and the soft-thresholding operator D_τ is defined as follows:

$$\begin{aligned} D_\tau(X) &:= U D_\tau(\Sigma) V \\ D_\tau(\Sigma) &= \text{diag}(\{\delta_i - \tau\}_+), \end{aligned} \tag{6}$$

Singular value decomposition (SVD) of a matrix X can obtain a sequence of positive singular values σ_i . And $\text{diag}(\{\sigma_i - \tau\}_+)$ is the positive part of $\sigma_i - \tau$, and $\sigma_i - \tau$ is equal to zero, if $\sigma_i - \tau < 0$. The physical meaning of soft-thresholding operator can be understood that it filters the data and only leaves the most important part of the dataset. Therefore, noise or redundancy information can be eliminated via the soft-thresholding operator. According to [42], the iterative method can converge to an unique solution when $0 < \delta < 2$.

The matrix completion method iteratively updates (4) and (5) until the stopping criteria is met. In this study, we choose the well-known K.K.T. conditions [48] as the stopping criteria:

$$P_\Omega(M - X^k)_F \leq \epsilon P_\Omega(M)_F \tag{7}$$

where ϵ is the predefined tolerance, e.g., 10^{-4} .

2.4 Performance Metrics

We choose two metrics including AUC (Area under the Receiver Operating Characteristic Curve) and F_1 -score to evaluate the performance of our method.

The ROC curve plots the values of TPR (true positive rate) versus FPR (false positive rate), and it is one of the well-known metrics to evaluate the performance of existing interaction prediction method in the current study. TPR and FPR are defined as:

$$TPR = \frac{TP}{TP + FN} \quad (8)$$

$$FPR = \frac{FP}{FP + TN} \quad (9)$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, false negative, respectively. However, AUC is insufficient to evaluate the performance of methods on bio-dataset because most bio-datasets have a highly imbalanced class distribution between positive samples and negative samples [49]. For example, the enzyme dataset contains less than 1% interaction entries (positive samples) in the whole dataset and the remaining 99% elements are non-interaction entries (negative samples). A naive prediction method that randomly predicts all elements as non-interaction entries can achieve a small false positive rate because most elements in original dataset are non-interaction entries. On the other hand, the true positive rate can be high even though we only find one correct interaction entry, because the number of real interaction is small. That is why existing prediction methods can easily achieve a decent of AUC about 80%–99%. Practically speaking, the ability of predicting as many as potentially correct interactions can maximize the probability of validating the unverified interaction pairs via biochemical experiments. Since AUC is insufficient to evaluate performance of model, F_1 -score is used as a standard information retrieval metric, to evaluate the performance of our method and its counterparts. The F_1 -score is defined based on two critical metrics including Precision and Recall, i.e.

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (12)$$

The difference between the F_1 -score and AUC is the precision term and the utilization of FPR. FPR mainly focus on the non-interaction prediction performance while the F_1 -score mainly focus on the overall interaction prediction performance. In general, precision represents the interaction prediction success rate.

3 Result

This section evaluates the effectiveness of our method in terms of both AUC and F_1 -score with 10-fold cross validation CV on “gold standard datasets”. In this experiment, we empirically set the step size = 1.5 and the stopping tolerance = 10^{-4} . We compared our method to five representative methods including

Table 2. AUC value of 10-fold cross validation of six methods sample.

AUC	Enzyme	Ion channel	GPCR	Nuclear receptor	Integrated data
Our method	0.9708	0.9778	0.9123	0.6640	0.9659
NBI	0.8941	0.9284	0.8357	0.6653	0.9087
PMF	0.9109	0.9575	0.9311	0.8245	0.8314
GIP	0.9516	0.9761	0.9272	0.8609	NA
KBMF2K	0.8475	0.9111	0.8741	0.8490	NA

Table 3. F_1 -score of 10-fold cross validation of six methods.

F_1 score	Enzyme	Ion channel	GPCR	Nuclear receptor	Integrated data
Our method	0.8437	0.8975	0.7083	0.5204	0.8281
NBI	0.8325	0.8233	0.6663	0.4960	0.7925
PMF	0.6556	0.8351	0.6940	0.5295	0.5536
GIP	0.7150	0.8273	0.6730	0.6021	NA
KBMF2K	0.6889	0.6764	0.5580	0.5381	NA

NBI [40], GIP [25], KBMF2K [24], PMF [41] and NetLapRLS [26] according to Hao’s review [49]. For GIP, KBMF2K, and NetLapRLS, we used the source codes provided by the authors; for NBI and PMF, we implemented the algorithm described in their paper. Note that GIP, KBMF2K and NetLapRLS need to exploit both similarity information and interaction profiles as input to predict interactions. As the performance of NetLapRLS was verified in a slightly different way from the remaining methods, we first compared our method to NBI, GIP, KBMF2K, and then compared it against NetLapRLS separately in the next subsection.

3.1 Performance Comparison of NBI, GIP, KBMF2K, PMF and Our Method on Gold Standard Datasets

One reason to use F_1 -score as the performance metric is that existing interaction datasets tend to exhibit an imbalanced distribution of positive and negative samples, as illustrated in Table 1. A large number of non-interaction entries make AUC insufficient for measuring the performance. On the contrary, the F_1 -score penalizes false positives much more than ROC [49, 50], and thus it can characterize the performance better in interaction prediction. Tables 2 and 3 list the AUC and F_1 -scores of NBI, PMF, GIP, KBMF2K and our methods on both “gold standard” and integrated datasets, respectively. Table 2 shows that our method achieves the highest AUC value (around 0.97) on the enzyme, ion channel and the integrated datasets; and its AUC value is also good on the GPCR dataset (above 0.9). Table 3 demonstrates that our method has the highest F_1 -scores on all test sets except for the nuclear receptor subsets.

In summary, our method can achieve the best performance on the Enzyme, Ion channel and integrate databases while it requires much less information comparing with the similarity-bases methods. Moreover our method has the higher AUC and F_1 value on most datasets comparing with the NBI and PMF methods.

3.2 Comparison with the NetLapRLs Method

In this experiment, we compared the performance of our method against that of the NetLapRLS method. The main difference between NetLapRLS method and other methods is that the NetLapRLS method only utilizes known interactions entries (positive instances) for prediction, while the remaining methods treat unknown interaction as non-interactions (negative instances) in the training data. We investigate the effects of introducing negative instances into the training data on the performance of our method and NetLapRLS methods. To do this, we performed a series of performance test by randomly adding 0%, 10%, ..., 90% negative samples into the training dataset. In each test, we performed 10-fold cross validation.

When the percentage of negative samples equals to 0%, our method only uses positive samples to predict all unknown interaction entries like NetLapRLs.

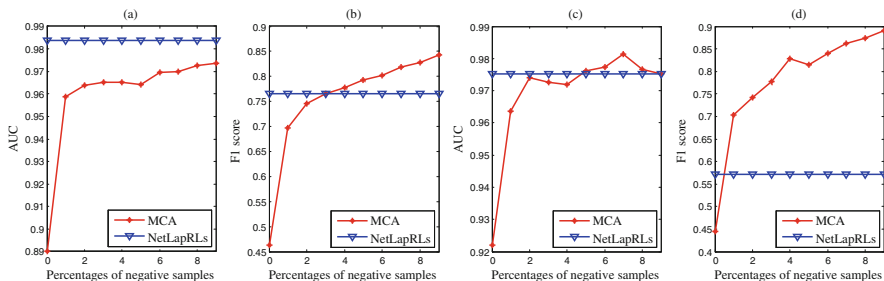


Fig. 2. AUC and F_1 of our method versus NetLapRLs on Enzyme (a), (b) and on Ion channel datasets (c), (d), respectively.

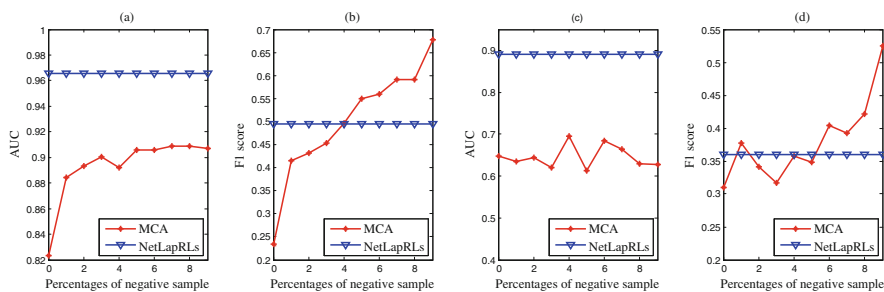


Fig. 3. AUC and F_1 of our method versus NetLapRLs on GPCRs (a), (b) and on Nuclear receptor datasets (c), (d), respectively.

We can see from Figs. 2 and 3 that when the percentage of negative samples increases, both AUC and F_1 -score of our method can be gradually improved (especially for F_1 -scores). More importantly, we can clearly see that even if the AUC values are stable, the F_1 -score significantly improves. This also reflects the importance of including F_1 -score for performance evaluation.

4 Discussion

4.1 Validated New Pairs in the Latest Databases

In order to illustrate the capability of our method in the real case, we developed a small tool [51] by Python that can automatically validate predicted links using the knowledge from DrugBank and KEGG. We decide the value of threshold when the F_1 score is the highest. If the prediction value is larger than the threshold, we regard it as candidate interaction, otherwise, non-interaction. Table 4 summaries the validation links in four datasets. According to Table 4, the percentages of new validated interactions are 16.90 %, 17.24 %, 40.78 % and 22.22 % for enzyme, ion channel, GPCR and nuclear subsets datasets, respectively. Figure 4 shows two instances of new validated drug-target interactions we truly find them in the latest database. Both drugs D00691 and D00528 interact with the same target hsa5150 and both drugs D00563 and D00283 interact with the same target hsa152. We can find that D00691 has a similar chemical structure with D00528, and D00563 has a similar chemical structure with D00283. Although our method does not apply any similarity information in the model, the result also reflects that the similar drugs tend to interact closely with similar target proteins. More materials are available at [51].

Table 4. Summary of validated interactions of four datasets.

	#Predicted interactions	#Validated interactions
Enzyme	71	12
Ion channel	58	10
GPCR	76	31
Nuclear receptor	18	4

4.2 Limitation

Due to our method only utilized interaction profiles to mine potential interaction, it requires only existing correlation between samples. In this work, we analysis four datasets and find some drugs have very similar interactions between targets. Therefore, we can fill the missing interaction entries based on observed interactions. Some biological missing value cannot be predicted by this method such as IC_{50} , EC_{50} , K_i and K_d value [52] in structure activity relationship (SAR) dataset because it does not have clear correlations between samples.

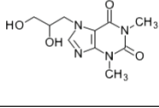
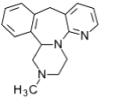
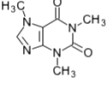
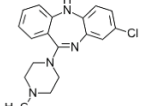
	Drug	Drug Chemical Structure	Target		Drug	Drug Chemical Structure	Target
1	D00691		hsa5150	2	D00563		hsa152
	D00528				D00283		

Fig. 4. Examples of drug chemical structures of the same target.

5 Conclusions

Our method is the first work to predict drug-target interactions by using the matrix completion technique [2] based on the observation that most drug-target interaction matrices are low-rank. Our method first fills missing entries with 0 in the original matrix to be recovered, and it iteratively updates the missing entries with predictive value. Moreover, in extreme case, a strong drug-drug and target-target correlation makes the interaction matrix to a low-rank one. Therefore, our method tends to adopt the lowest-rank approximate matrix as the correct solution during the iterative process.

We chose both AUC and F_1 -score to evaluate the prediction performance. Five representative models: NBI, PMF, GIP, KBMF2K, NetLapRLS are used for comparative study. Among them, GIP, KBMF2K and NetLapRLS apply interaction profiles as well as similarity information of drugs/target proteins. NBI, PMF and our method predict the interaction pairs only based on interaction information. Our method outperforms other methods in terms of both AUC and F_1 -score. As there is no standard similarity strategy of bio-data, our method only applies interaction profiles to predict drug-target interactions. For future work, we will extend our method by introducing similarity learning on bio datasets.

Acknowledgments. This work was supported by The National Natural Science Foundation of China (under grant No. U1435222 and No. 61502515). And this work was also supported in part by grants from 973 project 2013CB329006, RGC under the contract CERG 16212714.

References

1. Hopkins, A.L.: Drug discovery: predicting promiscuity. *Nature* **462**(7270), 167–168 (2009)
2. Cai, J.-F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **20**(4), 1956–1982 (2010)

- Ashburn, T.T., Thor, K.B.: Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **3**(8), 673–683 (2004)
- Dudley, J.T., Deshpande, T., Butte, A.J.: Exploiting drug-disease relationships for computational drug repositioning. *Briefings Bioinform.* **12**(4), 303–311 (2011)
- Swamidass, S.J.: Mining small-molecule screens to repurpose drugs. *Briefings Bioinform.* **12**(4), 327–335 (2011)
- Moriaud, F., Richard, S.B., Adcock, S.A., Chanas-Martin, L., Surgand, J.-S., Jeloul, M.B., Delfaud, F.: Identify drug repurposing candidates by mining the protein data bank. *Briefings Bioinform.* **12**(4), 336–340 (2011)
- Whitebread, S., Hamon, J., Bojanic, D., Urban, L.: Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov. Today* **10**(21), 1421–1433 (2005)
- Haggarty, S.J., Koeller, K.M., Wong, J.C., Butcher, R.A., Schreiber, S.L.: Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays. *Chem. Biol.* **10**(5), 383–396 (2003)
- Kuruvilla, F.G., Shamji, A.F., Sternson, S.M., Hergenrother, P.J., Schreiber, S.L.: Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. *Nature* **416**(6881), 653–657 (2002)
- Manly, C.J., Louise-May, S., Hammer, J.D.: The impact of informatics and computational chemistry on synthesis and screening. *Drug Discov. Today* **6**(21), 1101–1110 (2001)
- Cheng, A.C., Coleman, R.G., Smyth, K.T., Cao, Q., Soulard, P., Caffrey, D.R., Salzberg, A.C., Huang, E.S.: Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **25**(1), 71–75 (2007)
- Rarey, M., Kramer, B., Lengauer, T., Klebe, G.: A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **261**(3), 470–489 (1996)
- Shoichet, B.K., Kuntz, I.D., Bodian, D.L.: Molecular docking using shape descriptors. *J. Comput. Chem.* **13**(3), 380–397 (1992)
- Halperin, I., Ma, B., Wolfson, H., Nussinov, R.: Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins Struct. Funct. Bioinform.* **47**(4), 409–443 (2002)
- Shoichet, B.K., McGovern, S.L., Wei, B., Irwin, J.J.: Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **6**(4), 439–446 (2002)
- Kolb, P., Ferreira, R.S., Irwin, J.J., Shoichet, B.K.: Docking and chemoinformatic screens for new ligands and targets. *Curr. Opin. Biotechnol.* **20**(4), 429–436 (2009)
- Zhu, S., Okuno, Y., Tsujimoto, G., Mamitsuka, H.: A probabilistic model for mining implicit chemical compound–generations from literature. *Bioinformatics* **21**(Suppl. 2), ii245–ii251 (2005)
- Butina, D., Segall, M.D., Frankcombe, K.: Predicting adme properties in silico: methods and models. *Drug Discov. Today* **7**(11), S83–S88 (2002)
- Byvatov, E., Fechner, U., Sadowski, J., Schneider, G.: Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **43**(6), 1882–1889 (2003)
- Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J., Shoichet, B.K.: Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**(2), 197–206 (2007)
- Klabunde, T., Hessler, G.: Drug design strategies for targeting g-protein-coupled receptors. *ChemBiochem* **3**(10), 928–944 (2002)
- 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* **24**(13), i232–i240 (2008)

23. Bleakley, K., Yamanishi, Y.: Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics* **25**(18), 2397–2403 (2009)
24. Gönen, M.: Predicting drug-target interactions from chemical and genomic kernels using bayesian matrix factorization. *Bioinformatics* **28**(18), 2304–2310 (2012)
25. van Laarhoven, T., Nabuurs, S.B., Marchiori, E.: Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* **27**(21), 3036–3043 (2011)
26. Xia, Z., Wu, L.-Y., Zhou, X., Wong, S.T.: Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.* **4**(Suppl. 2), S6 (2010)
27. Jacob, L., Vert, J.-P.: Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **24**(19), 2149–2156 (2008)
28. Klabunde, T.: Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.* **152**(1), 5–7 (2007)
29. Schuffenhauer, A., Floersheim, P., Acklin, P., Jacoby, E.: Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **43**(2), 391–405 (2003)
30. Nagamine, N., Sakakibara, Y.: Statistical prediction of protein-chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics* **23**(15), 2004–2012 (2007)
31. Nagamine, N., Shirakawa, T., Minato, Y., Torii, K., Kobayashi, H., Imoto, M., Sakakibara, Y.: Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening. *PLoS Comput. Biol.* **5**(6), e1000397–e1000397 (2009)
32. Yabuuchi, H., Nijijima, S., Takematsu, H., Ida, T., Hirokawa, T., Hara, T., Ogawa, T., Minowa, Y., Tsujimoto, G., Okuno, Y.: Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.* **7**(1), 472 (2011)
33. Wang, Y., Zeng, J.: Predicting drug-target interactions using restricted boltzmann machines. *Bioinformatics* **29**(13), i126–i134 (2013)
34. Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L.J., Bork, P.: Stitch 3: zooming in on protein-chemical interactions. *Nucleic Acids Res.* **40**(D1), D876–D880 (2012)
35. Ballesteros, J., Palczewski, K.: G protein-coupled receptor drug discovery: implications from the crystal structure of rhodopsin. *Curr. Opin. Drug Discov. Dev.* **4**(5), 561 (2001)
36. Yamanishi, Y., Kotera, M., Kanehisa, M., Goto, S.: Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* **26**(12), i246–i254 (2010)
37. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**(1), 25–29 (2000)
38. Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., Wang, S.: Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics* **26**(7), 976–978 (2010)
39. Zheng, X., Ding, H., Mamitsuka, H., Zhu, S.: Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1025–1033. ACM (2013)
40. Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., Tang, Y.: Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* **8**(5), e1002503 (2012)

41. Cobanoglu, M.C., Liu, C., Hu, F., Oltvai, Z.N., Bahar, I.: Predicting drug-target interactions using probabilistic matrix factorization. *J. Chem. Inf. Model.* **53**(12), 3399–3409 (2013)
42. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M.: From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**(Suppl. 1), D354–D357 (2006)
43. Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., Schomburg, D.: Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Res.* **32**(Suppl. 1), D431–D433 (2004)
44. Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E.G., Gewiess, A., Jensen, L.J., et al.: Supertarget and matorator: resources for exploring drug-target relationships. *Nucleic Acids Res.* **36**(Suppl. 1), D919–D922 (2008)
45. Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M.: Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **36**(Suppl. 1), D901–D906 (2008)
46. Bertsekas, D.P.: *Nonlinear programming* (1999)
47. Elman, H.C., Golub, G.H.: Inexact and preconditioned uzawa algorithms for saddle point problems. *SIAM J. Numer. Anal.* **31**(6), 1645–1661 (1994)
48. Boyd, S., Vandenberghe, L.: *Convex Optimization*. Cambridge University Press, Cambridge (2009)
49. Ding, H., Takigawa, I., Mamitsuka, H., Zhu, S.: Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Briefings in Bioinform.*, bbt056 (2013)
50. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240. ACM (2006)
51. Wu, C., Liao, Q.: An useful tool for finding drug-target interaction in drugbank and KEGG. <http://www.cse.ust.hk/~qnature/>
52. Hu, Y., Bajorath, J.: Compound promiscuity: what can we learn from current data? *Drug Discov. Today* **18**(13), 644–650 (2013)