

DRUG TARGET PREDICTIONS BASED ON HETEROGENEOUS GRAPH INFERENCE

Wenhui Wang[†], Sen Yang[†], JING Li^{*}

*Department of Electrical Engineering and Computer Science, Case Western Reserve University
Cleveland, Ohio, 44106, USA*

Emails: {wxw134@case.edu, sxy221@case.edu jingli@case.edu}

A key issue in drug development is to understand the hidden relationships among drugs and targets. Computational methods for novel drug target predictions can greatly reduce time and costs compared with experimental methods. In this paper, we propose a network based computational approach for novel drug and target association predictions. More specifically, a heterogeneous drug-target graph, which incorporates known drug-target interactions as well as drug-drug and target-target similarities, is first constructed. Based on this graph, a novel graph-based inference method is introduced. Compared with two state-of-the-art methods, large-scale cross-validation results indicate that the proposed method can greatly improve novel target predictions.

1. Introduction

Drug targets are a class of molecular structures which could interact with drugs¹. Establishing new connections between existing drugs and targets or finding novel targets for a given drug plays an important role in drug development. Experimental prediction of drug-target associations is a laborious and costly task². So far there are only about a few hundreds of known targets¹. In contrast, there are many more computationally predicted targets, e.g., the so called druggable genome^{3,4}. The druggable genome denotes a set of human genes that encode proteins which might be able to bind drug-like molecules³. Though different sets of druggable genes have been predicted, the consensus on the number of druggable genes is around 3000⁴. Due to the large number of potential targets, examining each one of them with a specific drug becomes a tedious or even impossible task. From this point of view, an accurate druggable genome filtering or ranking approach becomes in urgent need.

In recent years, a large number of approaches have been proposed to address this problem. Zhu, et al. (2005)⁵ attempted to mine implicit chemical compound and gene relations from their co-occurrences in the literature. However, their results were constrained to current knowledge. Furthermore, there are many inconsistencies in target names and drug names, which may adversely affect the accuracy of the results. By using some basic biophysical principles, the structure based maximal affinity model⁶ could generate accurate predictions of druggability based solely on the crystal structure of a target's binding site. This method, however, is applicable only when the 3D structures of targets are known, which may not be available in general. More recently, several methods have combined drug-drug or target-target similarities into novel target predictions⁷⁻¹³. Phenotypic side-effect similarities were used to build a drug-drug relation network, based

[†] These authors contributed equally to this work. ^{*}To whom corresponds should be addressed.

on which novel drug-target associations were inferred⁷. Yamanishi and coauthors^{8,9} formalized the drug-target interaction inference as a supervised learning problem on a bipartite graph. The learning process was based on a unified 'pharmacological space', which was constructed by combining chemical and genomic properties. It has also been shown that chemical similarities between drugs and ligands, small molecules that bind to molecular targets, can be used to predict unanticipated associations¹⁰. Bipartite local models (BLM) used supervised methods to predict target proteins of a given drug, then to predict drugs targeting a given protein, and finally these two steps were combined to give a final prediction for each drug-target interaction¹¹. In another work, Perlman *et al.* (2011)¹² proposed a framework that combines multiple drug-drug and gene-gene similarity measures using a logistic regression model. The final classification score was used to indicate interactions between drugs and targets. Very recently, a network based inference (NBI) method was proposed to infer novel drug-target interactions¹³, which ranks drugs for a specific target based on a two-step diffusion model on the bipartite drug-target graph.

The guilt-by-association principle has been widely used in many different domains and applications (e.g., Jeh and Widom (2002)¹⁴). Chiang and Butte (2009)¹⁵ proposed a novel drug repositioning method based on the guilt-by-association principle. They claimed that suggestions for novel drug uses can be generated from the uses of drugs that cure the same diseases. This assumption was further extended by concluding that similar diseases tend to be connected with similar drugs and similar drugs tend to be connected with similar target¹⁶. Based on this assumption, the intra-similarity information can be incorporated into novel association predictions by constructing a *heterogeneous drug-target graph/network*, which includes both intra-similarity information (connections between the same kind of nodes, such as drug-drug connections and target-target connections) and interaction information (connections between different kinds of nodes, such as drug-target connections). In this paper, we propose a method, termed HGBI, for **Heterogeneous Graph Based Inference**, for novel drug target predictions based on the guilt-by-association principle and an intuitive interpretation of information flow on the drug-target heterogeneous graph. The algorithm iteratively updates the measure of strength between unlinked drug-target pairs based on all the paths in the network connecting them. We show that when properly normalized, the proposed procedure will eventually converge and stable relationships between drugs and targets can be achieved. Fig. 1 shows the framework of HGBI. Based on large scale leave-one-out cross-validation experiments, we show that HGBI exhibited superior performance and achieved much higher AUC (area under the receiver operating characteristic, i.e., ROC curve) than two existing state-of-the-art novel drug target prediction methods, BLM¹¹ and NBI¹³. In particular, when focusing on the top 1% ranked targets, HGBI successfully retrieved 1339 out of 1915 drug-target interactions, whereas BLM and NBI only retrieved 56 and 10 such interactions. Furthermore, HGBI can establish a novel interaction between a drug and a target even none of the two have directly associated targets/drugs. Some of these novel predictions are confirmed based on a new database, which is not used in this study.

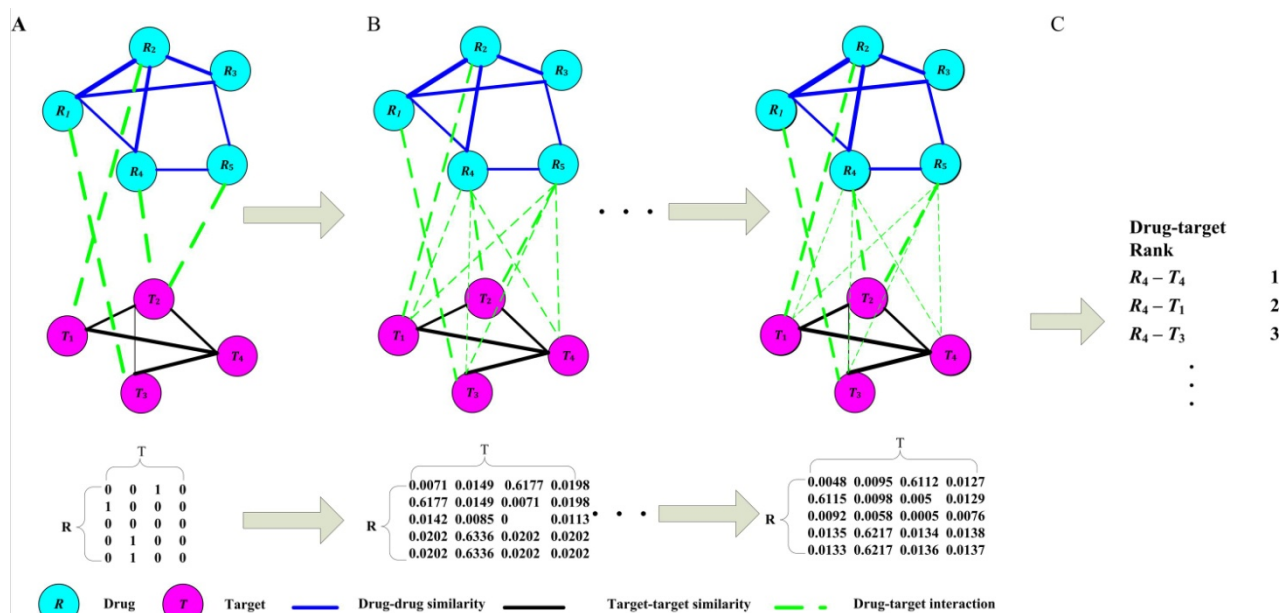


Fig. 1 The framework of HGBI. A: A heterogeneous graph is constructed based on drug-drug similarities, target-target similarities, and drug-target interactions. B: Edge weights between drugs and targets are updated iteratively by incorporating all possible paths between each drug-target pair. C: For each drug, all candidate targets are ranked according to the final weights.

2. Materials and methods

2.1 Datasets collection

There are two intra-similarity matrices which represent the drug-drug similarities and target-target similarities, respectively. In addition, there is an interaction matrix, i.e. the drug-target interaction matrix, which represents the connections between drugs and targets. The drug-drug similarity matrix includes all the FDA-approved drugs from the DrugBank database¹⁷. The similarities are calculated based on their chemical structures. First, chemical structures of all drug compounds in the Canonical SMILES format¹⁸ are downloaded from DrugBank¹⁷. Then, the Chemical Development Kit¹⁹ is used to calculate a binary fingerprint for each drug. Finally, a similarity score of two drugs is calculated using Tanimoto score²⁰ based on their fingerprints, which is in the range of [0, 1]. A druggable gene is defined as a human protein coding gene that contributes to a disease phenotype and can be modified by a small molecule drug. The term “druggable genome” has been used to denote a list of computationally predicted genes that their proteins can serve as suitable targets for developing therapeutic drugs²¹. We use the term “druggable gene” and “target” interchangeably in this study. The list of druggable genes is downloaded from the Sopic Integrated Druggable Genome Database project²¹, which includes genes from the ENSEMBL database²², the DrugBank database¹⁷ and the InterPro-BLAST database²³. The target-target similarities are calculated using the Smith-Waterman algorithm²⁴ based on the amino acid sequences of their corresponding proteins. The similarities are normalized using the same

method proposed by Bleakley and Yamanishi (2009)¹¹. Initial drug-target interactions are collected from the DrugBank database¹⁷, but limited to drugs that have associated diseases in the Online Mendelian Inheritance in Man (OMIM) database²⁵, which are the same as the one used in Ref¹⁶. For each drug-target interaction, their corresponding value in the drug-target interaction matrix is 1. All other items in the drug-target matrix are set to 0.

2.2 Basic statistics

The total number of drugs is 1409. The total number of targets is 3997. The matrix is very sparse with many isolated nodes (having no connections). The total number of connections among drugs and targets is only 2098, with 554 drugs having at least one known target and 602 targets connecting with at least one drug. Among the connected nodes, many of them have more than one connection, which means known information about drugs and targets is biased towards a very small subset of all drugs/targets. The degree distribution of each entity in the matrix is given in Fig. 2.

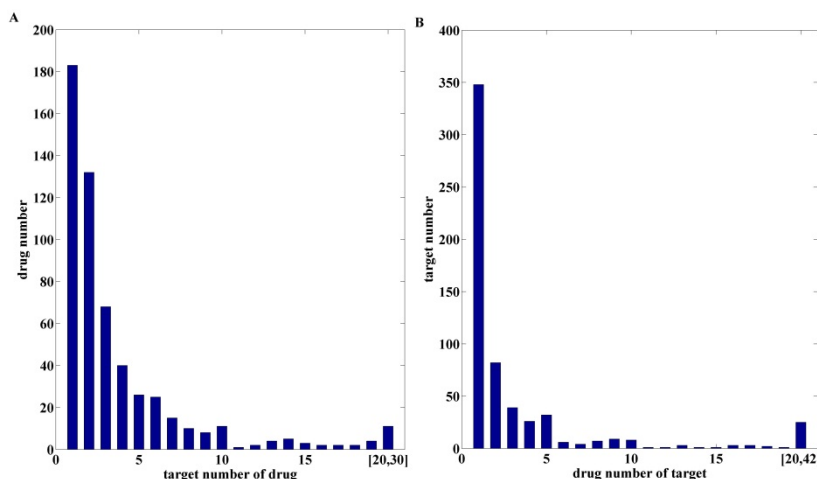


Fig. 2. Degree distributions of drugs (A) and targets (B) among the initial drug-target interactions.

2.3 Intra-similarity analysis

Before performing the proposed approach, we first study the statistical characteristics of these datasets. The distributions of the two intra-similarity matrices, i.e. the drug-drug and the target-target similarity matrices, are presented in Fig. 3A&C, which show the majority of the similarities are quite small. According to previous studies^{26,27}, low level similarity values provide little information for interaction inference. Furthermore, including the mess of low values could adversely affect prediction performance. Therefore, for the constructed heterogeneous graph, two nodes of the same type are connected only if their similarity score ≥ 0.3 . It is also noted that although the entries between a node to itself have already been excluded, there are still some entries with value 1 (Fig. 3B&D), which is mainly due to the representation issue. To ensure that a node can only have a similarity score of 1 to itself, we replace those 1s with 0.99, which should not affect the final results much.

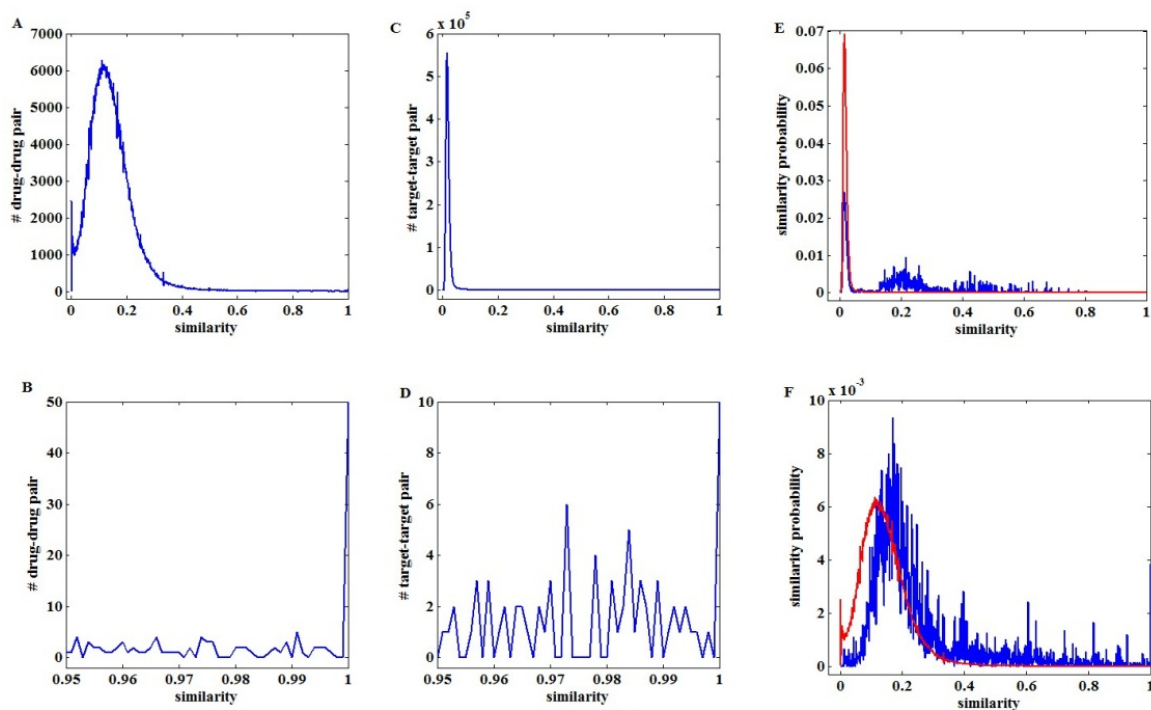


Fig. 3. Intra-similarity distributions. A: the distribution of drug-drug similarities; B: the tail distribution of drug-drug similarities in the range of (0.95, 1); C: the distribution of target-target similarities; D: the tail distribution of target-target similarities in the range of (0.95, 1); E: similarity distributions of targets sharing the same drugs (blue curve), and from different drugs (red curve); F: similarity distributions of drugs sharing the same targets (blue curve), and from different targets (red curve).

This study is based on the assumption that similar drugs tend to be associated with similar targets and dissimilar drugs are prone to be associated with dissimilar targets. To study the validation of the assumption on the collected real datasets, similarities of drugs for the same targets and similarities of drugs from different targets are compared. The average similarity of drugs for the same targets is calculated by averaging the similarities of all drug pairs that belong to the same target. To determine the average similarity of drugs from different targets, the similarity values of all drug pairs that are across different targets are averaged. Similarly, we examine the similarities among targets that share the same drugs and similarities among targets that do not share any drugs. The results are given in Fig. 3E&F. The average similarity of drug pairs from the same targets and the average similarity of target pairs from the same drugs are 0.2445 and 0.1836, respectively. In contrast, the average similarity of drug pairs from different targets and the average similarity of target pairs from different drugs are 0.1429 and 0.0231, respectively. We further test the differences of the corresponding distributions using the Wilcoxon rank sum test. Both tests reject the null hypothesis that the distributions are the same at the 5% significance level. Based on these results, it can be concluded that drugs (targets) associated with the same targets (drugs) possess higher similarity values than those associated with different targets (drugs). The guilt-by-association principle can be utilized in this study.

2.4 Drug-target association predictions

The drug-target heterogeneous graph has two kinds of nodes: drug nodes and target nodes. Let $R = \{R_1, R_2, \dots, R_n\}$ denote the n drug nodes, and $T = \{T_1, T_2, \dots, T_m\}$ denote the m target nodes. A drug is connected with another drug if and only if their similarity is greater than a pre-defined threshold (0.3 in this study), which is assigned as the weight of the edge. Edges and weights for target pairs are constructed similarly. Finally, a drug and a target are connected if they interact in the original drug-target interaction dataset. The weights of all drug-target edges are originally assigned 1. Let E_{rr} , E_{tt} , and E_{rt} represent drug-drug, target-target and drug-target edges, respectively, and W_{rr} , W_{tt} , and W_{rt} represent the weights on these three kinds of edges. The heterogeneous drug-target graph can be represented as $G_{RT} = \{\{R, T\}, \{E_{rr}, E_{tt}, E_{rt}\}, \{W_{rr}, W_{tt}, W_{rt}\}\}$. Based on this graph, the novel target prediction problem can be transformed into a novel drug-target edge prediction problem on the constructed drug-target graph. This means that the original heterogeneous graph G_{RT} can be considered as an incomplete graph with missing edges between R (drug) nodes and T (target) nodes. The objective is to capture hidden interactions between drugs and targets based on the drug-drug similarities, target-target similarities, and known drug-target interactions. The novel drug-target edge prediction problem can be formalized as follows:

Input: $G_{RT} = \{\{R, T\}, \{E_{rr}, E_{tt}, E_{rt}\}, \{W_{rr}, W_{tt}, W_{rt}\}\}$

Output: $G_{RT}^{new} = \{\{R, T\}, \{E_{rr}, E_{tt}, E_{rt}^{new}\}, \{W_{rr}, W_{tt}, W_{rt}^{new}\}\}$

where E_{rt}^{new} and W_{rt}^{new} represent the newly calculated edges and their weights respectively.

Based on the guilt-by-association assumption, the intra-similarities and drug-target associations can be combined together to predict novel interactions between drugs and targets. For example, given the graph G_{RT} , one way to calculate the association coefficient (i.e., weight) between each initially unconnected drug-target pair is based on the following equation,

$$w(r, t) = \sum_{r_i \in R} \sum_{t_j \in T} w(r, r_i) \times w(r_i, t_j) \times w(t, t_j) \quad (1)$$

Here r is a drug and t is a target and they are not connected in the original graph. r_i 's and t_j 's are the neighbors of r and t that are connected with each other in G_{RT} . $w(r, r_i)$ is the weight between r and r_i , and $w(t, t_j)$ is the weight between t and t_j . Equation 1 basically means that one can establish a new weight between a drug and a target by summarizing all paths of length three, consisting one edge in each of E_{rr}, E_{tt}, E_{rt} . This is essentially the same idea adopted by NBI¹³.

Naturally, once new relationships/weights between drugs and targets being established based on equation 1, they themselves can be utilized again to generate more relationships. An iterative procedure can be constructed, which can be represented as matrix multiplications: $W_{rt}^{i+1} = W_{rr} \times W_{rt}^i \times W_{tt}$. In general, there are two related issues that need to be resolved in order to make the proposed iterative approach to work. First, one may want to treat the initial links between drugs and targets differently from the inferred links because the initial links deserve more credibility. Second, it is desirable if the matrix W_{rt} will converge, which means that the information propagation is stabilized at the end. In this study, we propose an iterative approach based on equation 2, which naturally solves the first

problem based on its formulation. Furthermore, we show that with proper normalization, it also solves the second problem.

$$W_{rt}^{i+1} = \alpha W_{rr} \times W_{rt}^i \times W_{tt} + (1 - \alpha) W_{rt}^0 \quad (2)$$

In this formula, $W_{rt}^0 = W_{rt}$, represents the initial interactions between drugs and targets, α is a decay factor with its value between 0 and 1. In each iteration, the original drug-target interactions will contribute to the newly constructed interactions, and the contribution is controlled by the scale factor $1 - \alpha$. Theoretically, one can optimize α based on results from cross validations. In this study, we fix $\alpha = 0.4$. By iteratively using this formula, the strength between a drug and a target will eventually include all the possible paths connecting them in the heterogeneous graph. We prove that when W_{rr} and W_{tt} are properly normalized, it is guaranteed that equation 2 will converge. The result is summarized as Theorem 1 and the details of the proof can be found in the appendix. To obtain the final solution based on equation 2, we use an iterative propagation-based algorithm²⁸. Once the final result is given, for each drug, all the targets will be ranked according to the strength of their links to the drug.

THEOREM 1. When W_{rr} and W_{tt} are properly normalized utilizing equation 3, it is guaranteed that formula (2) will converge.

$$w(r_i, r_j) = \frac{w(r_i, r_j)}{\sqrt{\sum_{k=1}^n w(r_i, r_k) \sum_{k=1}^n w(r_k, r_j)}}, w(t_i, t_j) = \frac{w(t_i, t_j)}{\sqrt{\sum_{k=1}^m w(t_i, t_k) \sum_{k=1}^m w(t_k, t_j)}} \quad (3)$$

3 Experiments

3.1 Evaluation metrics

In order to systematically evaluate the proposed approach on the collected datasets, we adopt a leave-one-out cross-validation (LOOCV) strategy in our experiments. For each drug, one of its connections to a target is treated as the test data, and it is ranked with all other targets in descending order according to the calculated drug-target association coefficients using the remaining connections as training data. For each specific ranking threshold, if the rank of the testing connection is above the threshold, it is regarded as a true positive. On the other hand, if the rank of an unknown connection is above the threshold, it is regarded as a false positive. True positive rate (TPR) and false positive rate (FPR) are calculated by varying thresholds to construct the ROC curve²⁹. The area under the curve (AUC) value represents the overall performance of the algorithm. In addition to LOOCV, we also perform 10-fold cross-validation, where all the drug-target connections are randomly partitioned into 10 subsets and each subset is treated as the test set in each iteration. Furthermore, in practice, it is natural that most researchers only focus on top ranked targets. Therefore, we also examine the performance of the algorithm on the top ranked results, i.e., the number of

correctly retrieved connections based on various top percentiles (the most left side of the ROC curve). In addition, to test the capacity of the algorithm in detecting novel interactions for drugs with no known targets, we collect all drugs that only have a single known target and perform the experiment by removing the only interaction. Finally, using all the data as training data, we test our algorithm again and compare the top ranked targets with those in another database³⁰ that is not used in training.

3.2 Comparison with existing methods

To evaluate the proposed approach, we choose to compare its performance with BLM¹¹ and NBI¹³. BLM is considered one of the state-of-the-art approaches in drug-target interaction predictions. In this study, BLM is implemented the same way as the one in the original paper¹¹. The predicted scores generated from SVM are used as the ranking criterion, which means that larger predicted scores yield higher ranks. In order to choose a proper number of negative samples for SVM training of BLM, we perform cross-validation. Based on the results (Fig S1 in appendix), the number of negative training samples was set to be $\max\{20, 2 \times \text{num_positive_samples}\}$. The result of BLM is obtained by averaging five runs with the same configuration but different negative training samples. We choose to compare with NBI¹³ because it can be viewed as a simplified version of the proposed approach, in the sense that only a two-step diffusion of the matrices (similar to equation 1) is used in NBI, while our approach uses the converged matrix. NBI¹³ is implemented according to the original paper.

3.3 Experimental results

3.3.1 Predictions for drugs with known connected targets

The cross-validation experiments for target predictions were conducted using all drugs with at least two known targets. In total, 371 such drugs and 1915 initial drug-target edges were considered. The ROC curves and AUC values of NBI, BLM, HGBI (LOOCV and 10-fold) are given in Fig.4A. It shows that HGBI (AUC:0.93) significantly outperforms both BLM (AUC:0.89) and NBI (AUC:0.73) for LOOCV. Furthermore, HGBI almost has the same performance when using 10-fold cross-validation. The numbers of correctly retrieved drug-target interactions according to different ranking thresholds are also given in Fig.4B. Results show that when focusing on the top ranked results, the performance of HGBI is much better compared with NBI and BLM, especially for the top 1% ranked targets, in which case HGBI correctly retrieved 1339 drug-target interactions, whereas BLM and NBI only retrieved 56 and 10 such interactions.

3.3.2 Predictions for drugs without known connected targets

To demonstrate the effectiveness of the proposed approach in detecting novel targets for drugs without known targets, only drugs with exactly one connected target in the dataset were collected in this experiment. There are in total 183 such drugs. Because BLM cannot predict novel targets for drugs without known targets, we only compare HGBI with NBI here. The ROC curves of NBI and HGBI are given in Figure 5. Again, HGBI (AUC:0.93) achieves much better performance than NBI (AUC:0.72).

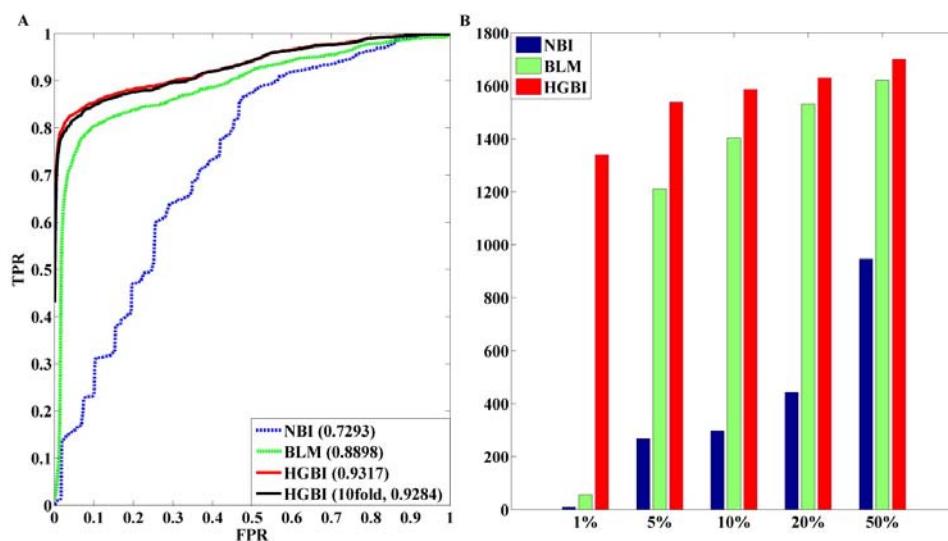


Fig.4 A: ROC curves of drug-target association predictions. AUC of each curve is indicated in the parentheses; B: The number of retrieved drug-target interactions using different thresholds. *x-axis* is the ranking thresholds in percentile.

3.3.3 Case studies on drug-target association predictions

Finally, using all the data as the training data, HBGI can make new predictions for all the drugs in the database. To further analyze its performance for practical usage, six drugs, i.e. Citalopram (Drugbank ID: DB00215), Escitalopram (DB01175), Terfenadine (DB00342), Diphenidol (DB01231), Fexofenadine (DB00950), and Naltrexone (DB00704), were randomly chosen for the case studies. For each drug, all their initial targets and the top 10 predicted targets were collected. A subset of these drugs, targets, and their connections are also illustrated in Fig 5B, which only shows upto 3 known targets and the top 3 predicted targets for clarity. Several observations can be made based on Fig 5B. First, similar drugs tend to share similar predicted targets, such as the drugs Diphenidol and Terfenadine. Second, predictions for drugs without known connected targets, such as the drug Fexofenadine, can be performed using HBGI. Because it is connected with the drug Diphenidol, one of Diphenidol's targets (target Entrez_ID: 1128 in Fig 5B) is predicted to be associated with it.

We further searched the Supertarget database³⁰, which is an extensive web resource for analyzing drug-target interactions. Some of the top ranked predictions by HBGI are supported by newly reported discoveries in the database. For example, in the Supertarget database, Citalopram (DB00215) has two new targets SLC6A3 (6531) and SLC6A2 (6530) that were not in the DrugBank database. They were ranked as the 2nd and the 17th among all target candidates by HBGI. Similarly, ADRA1D (146) & CHRM1 (1128) were not associated with Terfenadine (DB00342) in the DrugBank database. They were ranked as the 3rd & the 8th, respectively.

In a very recent study³¹, the authors experimentally validated 123 unique drug-target relationships. In comparison of our prediction with these newly validated relationships, our median rank for this data set is 16 (out of 3997 targets) and 43 of them are ranked in top 10.

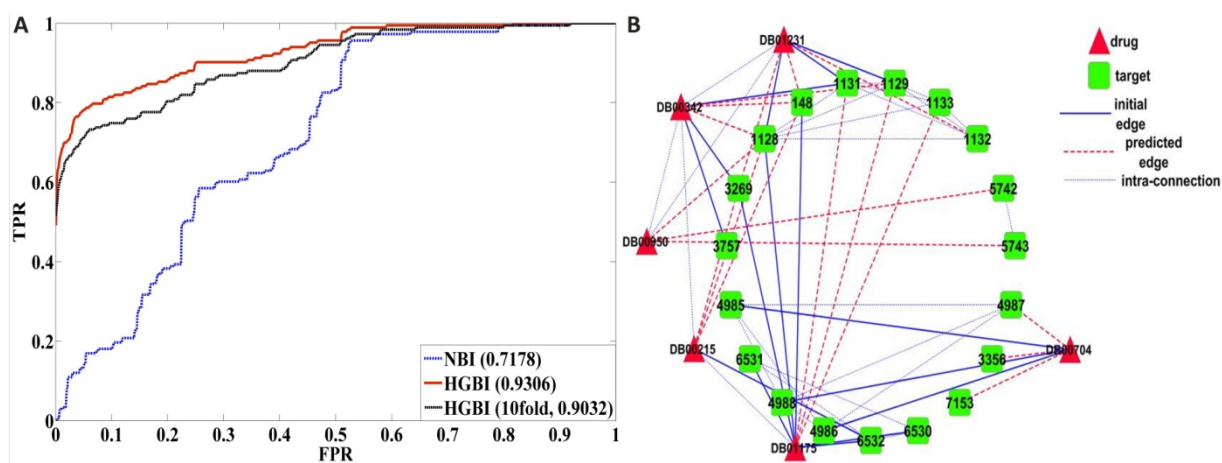


Fig.5 A. ROC curves of novel target predictions for drugs without known targets. B. The subnetwork for the case study.

4 Conclusion

In this paper, we have proposed a drug target prediction approach, HGBI, which integrates drug-drug similarities, target-target similarities and drug-target interactions into a heterogeneous graph and models the drug-target interactions as the stabilized information flow between them across the heterogeneous graph. Experiments have shown that HGBI significantly outperforms two existing methods in predicting novel targets for drugs with and without known targets. A case study has illustrated that HGBI can be used in practice to rank candidate targets for drugs and many top ranked ones can be utilized for further investigations.

Although equation 2 is similar to the framework of random walk with restart (RWWR)³², it is different in the sense that 1) equation 2 is defined on a heterogeneous graph and only connections between nodes of different types need to be derived; 2) because of the heterogeneous graph, all the information on similar drugs, similar targets and drug-target interactions has been used in predicting new drug-target associations.

For future directions, first, instead of using top ranked targets, it is possible to adopt an automatic threshold to declare significant predictions using the same idea in Ref³³. Furthermore, many approaches have been developed to identify disease genes. However, the relationships between disease genes and drug targets are not totally characterised. It would be interesting to include disease gene information in drug target predictions.

5. Acknowledgements

This research is supported by an NIH/NLM grant (LM008991).

References

1. P. Imming, *et al.*, *Nat Rev Drug Discov.* **10**, 821 (2006).
2. S. Haggarty, *et al.*, *Chem Biol.* **5**, 383 (2003).
3. A. Hopkins and C. Groom, *Nat Rev Drug Discov.* **9**, 727 (2002).
4. A. Russ and S. Lampel, *Drug Discov Today.* **23-24**, 1607 (2005).
5. S. Zhu, *et al.*, *Bioinformatics*, **suppl 2**, ii245 (2005).
6. A. Cheng, *et al.*, *Nat Biotechnol*, **1**, 71 (2007).

7. M. Campillos, *et al.*, *Science*, **5886**, 263 (2008).
8. Y. Yamanishi and M. Araki, *et al.*, *Bioinformatics*, **13**, i232 (2008).
9. Y. Yamanishi, *Proceedings of NIPS*, **21**, 1433 (2008)..
10. M. Keiser, *et al.*, *Nature*, **7270**, 175 (2009).
11. K. Bleakley and Y. Yamanishi, *Bioinformatics*, **18**, 2397 (2009).
12. L. Perlman, *et al.*, *J Comput Bio*, **2**, 133 (2011).
13. F. Cheng, *et al.*, *PLoS Comput Biol*, **5**, e1002503 (2012).
14. G. Jeh and J. Widom, *KDD*, 538 (2002).
15. A. Chiang, and A. Butte, *Clin Pharmacol Ther*, **5**, 507 (2009).
16. A. Gottlieb, *et al.*, *Mol Syst Biol*, **496**, 496 (2011).
17. C. Knox, *et al.*, *Nucleic Acids Res*, **Database issue**, 1035 (2011).
18. D. Weininger, *Journal of Chemical Information and Modeling*, **1**, 31 (1988).
19. C. Steinbeck, *et al.*, *Curr Pharm Des*, **17**, 2111 (2006).
20. T. Tanimoto, *IBM Internal Report 17th Nov*, (1957).
21. Sophic, <http://www.sophicalliance.com/documents/sophicdocs/White%20Paper%20Update%2001-27-11/The%20Druggable%20Genome012511.pdf>, (2012).
22. P. Flicek, *et al.*, *Nucleic Acids Res*, **Database issue**, 800 (2011).
23. S. Hunter, *et al.*, *Nucleic Acids Res*, **Database issue**, 211 (2009).
24. T. Smith and M. Waterman, *J Mol Biol*, **1**, 195 (1981).
25. A. Hamosh *et al.*, *Nucleic Acids Res*, **30**, 52 (2002).
26. Y. Chen, *et al.*, *Bioinformatics*, **13**, i167 (2011).
27. M. van Driel, *et al.*, *Eur J Hum Genet*, **5**, 535 (2006).
28. V. Oron, *et al.*, *PLoS Comput Biol*, **1**, e1000641 (2010).
29. T. Sing, *et al.*, *Bioinformatics*, **20**, 3940 (2005).
30. N. Hecker, *et al.*, *Nucleic Acids Res*, **Database issue**, D1113 (2012).
31. E. Lounkine, *et al.*, *Nature*, **486**, 361 (2012).
32. Tong, *et al.*, *Proceedings of ICDM*, 613-622(2005).
33. Chen, *et al.*, *PLoS One*, **6**, e21137 (2011).

6. Appendix

PROOF of Theorem 1: To make the proof process clear, Let A , B and X denote W_{rr} , W_{tb} and W_{rt} respectively. A , B and X are $n \times n$, $m \times m$ and $n \times m$ matrices respectively. A_i and A^j denote the i -th row of A and j -th column of A respectively. a_{ij} is used to represent the value at the i -th row and j -th column of matrix A . These conventions are also used for matrix B and X .

Then according to equation (2), we have $x_{ij} = \alpha A_i X B^j + (1 - \alpha) x_{ij}^0$. For X^1 , we can also get

$$\begin{bmatrix} x_{1,1} \\ \vdots \\ x_{n,1} \end{bmatrix} = \alpha \begin{bmatrix} a_{1,1}b_{1,1}, \dots, a_{1,n}b_{1,1}, \dots, a_{1,1}b_{m,1}, \dots, a_{1,n}b_{m,1} \\ \vdots \\ a_{n,1}b_{1,1}, \dots, a_{n,n}b_{1,1}, \dots, a_{n,1}b_{m,1}, \dots, a_{n,n}b_{m,1} \end{bmatrix} \begin{bmatrix} x_{1,1} \\ \vdots \\ x_{n,1} \\ \vdots \\ x_{1,m} \\ \vdots \\ x_{n,m} \end{bmatrix} + (1 - \alpha) \begin{bmatrix} x_{1,1}^0 \\ \vdots \\ x_{n,1}^0 \end{bmatrix}$$

If we use $A_i \times B^j$ to denote $[a_{i,1}b_{1,j} \ \dots \ a_{i,n}b_{1,j} \ \dots \ a_{i,1}b_{m,j} \ \dots \ a_{i,n}b_{m,j}]$, then equation (2) can be written as

$$\begin{bmatrix} X^1 \\ \vdots \\ X^m \end{bmatrix} = \alpha \begin{bmatrix} A_1 \times B^1 \\ \vdots \\ A_n \times B^1 \\ \vdots \\ A_1 \times B^m \\ \vdots \\ A_n \times B^m \end{bmatrix} \begin{bmatrix} X^1 \\ \vdots \\ X^m \end{bmatrix} + (1-\alpha) \begin{bmatrix} X^{10} \\ \vdots \\ X^{m0} \end{bmatrix} \quad (4)$$

Let C denote $\begin{bmatrix} A_1 \times B^1 & \cdots & A_n \times B^1 & \cdots & A_1 \times B^n & \cdots & A_n \times B^m \end{bmatrix}^T$ and $i = sn + t$, $j = rn + \theta$, $s = sI\{t > 0\} + (s-1)I\{t = 0\}$, $t = tI\{t > 0\} + nI\{t = 0\}$, $r = rI\{\theta > 0\} + (r-1)I\{\theta = 0\}$, $\theta = \theta I\{\theta > 0\} + nI\{\theta = 0\}$, $0 \leq t, \theta < n$.

Then we get: $c_{i,j} = a_{t,\theta} b_{r+1,s+1}$ and $c_{j,i} = a_{\theta,t} b_{s+1,r+1}$.

By comparing the above two equations, we can easily find that C is a symmetrical matrix with row and column number $n \times m$. If we use X^* to represents $\begin{bmatrix} X^1 & \cdots & X^n \end{bmatrix}^T$, then equation (4) can also be written as:

$$X^* = \alpha C X^* + (1-\alpha) X^{*0} \quad (5)$$

According to (Vanunu, et al. 2010)²⁹, in order to get a converged solution for equation (5), C is normalized as: $C^{norm} = D^{-1/2} C D^{-1/2}$, where D is diagonal matrix with $d_{i,i}$ equals to the sum of the i -th row of C . Therefore, we

$$\text{also have } c_{i,j}^{norm} = \frac{c_{i,j}}{\sqrt{d_{i,i} d_{j,j}}} \text{ and } d_{i,i} = \sum_{u=1}^{nm} c_{i,u} = \sum_{u=1}^{nm} a_{t,\theta_u} b_{r_u+1,s+1} = \sum_{p=1}^n a_{t,p} \sum_{q=1}^m b_{q,s+1}$$

where $u = r_u n + \theta_u$. By incorporating the above equation into $c_{i,j}^{norm}$, we can get

$$c_{i,j}^{norm} = \frac{a_{t,\theta} b_{r+1,s+1}}{\sqrt{\sum_{p=1}^n a_{t,p} \sum_{q=1}^m b_{q,s+1}} \sqrt{\sum_{p=1}^n a_{\theta,p} \sum_{q=1}^m b_{q,r+1}}} = \frac{a_{t,\theta}}{\sqrt{\sum_{p=1}^n a_{t,p} \sum_{p=1}^n a_{\theta,p}}} \frac{b_{r+1,s+1}}{\sqrt{\sum_{q=1}^m b_{q,s+1} \sum_{q=1}^m b_{q,r+1}}}$$

$$\text{Therefore, if we normalize } A \text{ and } B \text{ as } a_{i,j}^{norm} = \frac{a_{i,j}}{\sqrt{\sum_{p=1}^n a_{i,p} \sum_{p=1}^n a_{j,p}}} \text{ and } b_{i,j}^{norm} = \frac{b_{i,j}}{\sqrt{\sum_{q=1}^m b_{q,i} \sum_{q=1}^m b_{q,j}}}$$

We can get $c_{i,j}^{norm} = a_{t,\theta}^{norm} b_{r+1,s+1}^{norm}$.

With this equation, we can rewrite equation (5) as $X^* = \alpha C^{norm} X^* + (1-\alpha) X^{*0}$

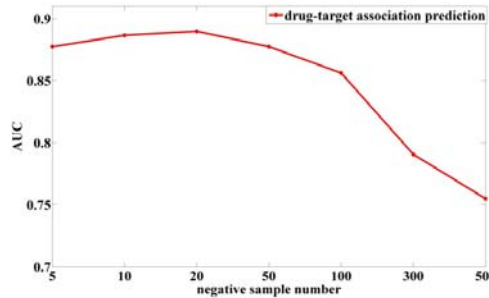


Figure S1 Cross-validation results of BLM using different numbers of negative samples.