The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010
Copyright © 2010 ORSC & APORC, pp. 219–226

NRWRH for Drug Target Prediction*

Xing Chen[1,2,†] Gui-Ying Yan[1]

[1]Academy of Mathematics and Systems Science, CAS, Beijing 100190, China
[2]Graduate University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Drug-target interaction prediction is an important problem for the development of novel drugs and human medical improvement. Many supervised and semi-supervised methods are proposed to uncover the relation between drugs and targets. Under the hypothesis that similar drugs target similar target proteins and the framework of Random Walk with Restart, the method of Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) is proposed to infer potential drug-target relationship. This method integrates three different networks (protein-protein similarity network, drug-drug similarity network, and known drug-target interaction network) into a heterogeneous network by known drug-target interactions and implements the random walk on this heterogeneous network. When applied to four classes of drug-target proteins interaction data including enzymes, ion channels, GPCRs and nuclear receptors, NRWRH significantly improves the previous methods in terms of cross validation.

Keywords Drug-target interactions; Random walk; Heterogeneous network.

1 Introduction

Identifying drug-target interactions is very meaningful not only for better understanding of the various interactions and biological processes, but also for the development of novel drugs and human medical improvement [1, 2]. There are about 6,000-8,000 targets of pharmacological interest in the human genome, but only a small number of them are identified to be related to approved drugs so far [3, 4, 5, 6]. Because the experiment for identifying drug-target interactions is time-consuming, expensive, and limited in small-scale research, computational methods are needed to decrease time and costs for drug discovery and search potential interactions in a genome-wide way [1, 3, 7, 8]. Computational methods can provide supporting evidences to the drug target experiments and accelerate the drug discovery [9]. The one-target one-drug paradigm has been denominating the drug research in the past decades, but it didn't accelerate the discovery of drugs as expected [9]. Because multiple targets are often involved in the same disease, much attention has been paid to search drugs involving many targets [10, 11, 12]. Multiple-target therapeutics can benefit the drug efficacy and are less likely to cause drug resistance [11]. Hence the need to search the targets of drug is emergent.

*This research is supported by the National Natural Science of Foundation of China under Grant Nos. 10531070, 10721101, and KJCX-YW-S7.

†Corresponding author. Email: xingchen@amss.ac.cn

Many computational methods have been developed to uncover the relation between drugs and targets. Yang et al made full use of disease network and proposed a novel computational method to search for the target of drugs and give multiple targets optimal intervention solutions restoring the disease network into normal state best [10]. Keiser et al proposed a computational method to predict the associations between drugs and targets based on the chemical structure information [13]. Thirty of predicted associations were tested by biological experiments and 23 of them were confirmed. But protein target information wasn't taken into consideration here. Yamanishi et al integrated the information of drug-drug chemical similarity, protein-protein sequence similarity, and known drug-target interactions and constructed pharmacological space [9]. A supervised learning method was proposed to associate drugs and targets. Bleakley et al used bipartite local methods to predict drug-target interactions [14]. Firstly the target of a given drug was predicted, and then the drug targeting a given target was predicted; finally the above results were combined to give the prediction of given drug-target pair. Wang et al employed support vector machine to predict drug-target interactions [1]. Gold-standard positive dataset was extracted from the database, and then gold-standard negative dataset was selected automatically to solve the sample imbalance problem. Chemical structure of the drugs and sequence information of the target proteins were used to extract the feature of the classifier and the classifier was constructed to learn the rule from data. The common problem of above three supervised learning method is that they regard the unknown drug-target interactions as negative samples. Xia et al made full use of the unlabeled information and integrated the information from chemical and genomic space [2]. A semi-supervised method, NetLapRLS, was developed. This method established classifier in the drug space and target space respectively, and combined two classifiers to give the final prediction. A good performance has been obtained because of the integration of information and the use of unlabeled data.

In the present study, based on the assumption that similar drugs target similar target proteins and the framework of Random Walk with Restart (RWR) [15, 16], we developed the method of Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) to infer potential drug-target relationship. NRWRH is composed of four steps: firstly, three networks (protein-protein similarity network, drug-drug similarity network, and known drug-target interaction network) are constructed and combined into a heterogeneous network by known drug-target interactions; secondly, the initial probability of random walk is determined to make random walk start at the given drug node and seed target nodes (these target proteins targeted by the given drug) simultaneously; then random walk on the heterogeneous network is implemented; finally we select the most probable targets according to the stable probability after the walk is stable. Random walk has been used widely in the bioinformatics [15, 16, 17, 18].

To our knowledge, there are no computational methods by using the idea of random walk to predict potential drug-target interactions before this work. NRWRH is different from traditional random walk with restart in two aspects. The first is that we use the information of known drug-target interaction network to improve drug chemical structure similarity and protein sequence similarity (motivated by [2]). The other difference is that we implement the random walk on three networks (motivated by [18]). When searching for targets for a drug having known targets, we can rank candidate targets by calculating the similarity between candidate targets and known targets. However, if the drug has no known target, only using target similarity will be insufficient and hence drug similarity must be used. In this case, we should select potential targets of this given drug based on target information of drugs which are similar to the given drug. NRWRH is applied to four classes of drug-target proteins interaction data including enzymes, ion channels, GPCRs and nuclear receptors. Leave-one-out cross-validation is implemented and a significant performance improvement over NetLapRLS has been obtained. So we can reach a conclusion that NRWRH successfully makes full use of the information of drug similarity, protein similarity and known drug-target interactions.

2 Materials

All the data used in this paper is downloaded from http://web.kuicr.kyoto-u.ac.jp/supp /yoshi/drugtarget/ [9]. Here we give a brief description.

2.1 The drug chemical structure similarity

The drug chemical structure similarity is calculated by SIMCOMP [19] based on the information of chemical structure from the DRUG and COMPOUND Sections in the KEGG LIGAND database [20]. The similarity is a global score based on the ratio between the size of common structures and the size of union structures. The drug chemical similarity matrix is denoted as S_d^c.

2.2 The target protein sequence similarity

The target protein sequence similarity is calculated by normalized Smith-Waterman score [21] based on the information of amino acid sequences from the KEGG GENES database [20]. The target protein similarity matrix is denoted as S_t^s.

2.3 Drug-target interaction data

Yamanishi et al [9] collected four classes of drug-target proteins interaction data including enzymes, ion channels, GPCRs and nuclear receptors from the KEGG BRITE [20], BRENDA [22], SuperTarget [23], and DrugBank databases [24]. The number of known interactions of the four datasets are 2926, 1476, 635, and 90. These datasets are used as gold standard data in the study. Further detail about data collection has been given in [9].

3 Methods

3.1 Construction of the heterogeneous network

The chemical structure of the drug and the sequence similarity of the protein have been introduced in Section Materials. Here the aim is to extract the information from the known drug-target interaction network (motivated by [2]). The underlying assumption is that if two drugs have more common targets, they are more similar. So another drug similarity matrix S_d^n can be obtained. The entry of the matrix is the number of targets shared by two drugs. Similarly, another target similarity matrix S_t^n can be established. The entry of this matrix is the number of drugs shared by two targets.

Drug target network similarity must be normalized. For S_d^n, a diagonal matrix D_d^n is defined such that $D_d^n(i,i)$ is the sum of row i of S_d^n. We set normalized matrix $\widehat{S_d^n} =$

$(D_d^n)^{-1/2} S_d^n (D_d^n)^{-1/2}$ which yields a symmetric matrix where $\widehat{S_d^n}(i,j) = S_d^n(i,j) / \sqrt{D_d^n(i,i) D_d^n(j,j)}$. Similar operation is applied to S_t^n and normalized matrix $\widehat{S_t^n}$ is obtained. The drug similarity can be obtained by the linear combination $S_D = w_d S_d^c + (1 - w_d)\widehat{S_d^n}$. Similarly, the target similarity can be obtained by $S_T = w_t S_t^s + (1 - w_t)\widehat{S_t^n}$.

Three kinds of data have been obtained: drug similarity, target similarity, and known drug-target interactions. These data can be represented by three networks, namely drug similarity network, target similarity network, and drug-target interactions network. In the drug similarity network, two drugs are connected if the similarity between them is more than 0. Drug similarity network is weighted and the weight of each edge is the similarity score between two corresponding drugs. Target similarity network is constructed in the same way as drug similarity network. In the drug-target interactions network, drug and target are connected if this drug targets the target protein. $(B)_{n \times m}$ is denoted as the adjacency matrix of the drug-target interaction network, where n and m represent the number of targets and drugs. The heterogeneous network is constructed by connecting drug similarity network and target similarity network using drug-target interactions network (motivated by [18]).

3.2 Initial probability

NRWRH simulates a random walker's transition from its current nodes randomly to the neighbors in the heterogeneous network starting at some given seed nodes (motivated by [18]). NRWRH allows the restart of the walk in every step at source node with probability r. If we want to predict some potential targets of a given drug, this drug is denoted seed nodes in the drug network and targets which are targeted by this drug are used as seed nodes in the target network. The initial probability u_0 of the target network is formed such that equal probabilities are assigned to the seed nodes in the target network, with the sum equal to 1. The initial probability of the drug network v_0 is obtained similarly. Hence, the initial probability of the heterogeneous network is $p_0 = \begin{bmatrix} (1-\eta)u_0 \\ \eta v_0 \end{bmatrix}$. The parameter $\eta \in [0,1]$ weights the importance of drug network and target network.

3.3 Transition matrix and random walk

To implement random walk, transition matrix must be decided (motivated by [18]). Let $M = \begin{bmatrix} M_T & M_{TD} \\ M_{DT} & M_D \end{bmatrix}$ be the transition matrix of the heterogeneous network, where M_T and M_D are inter-transition matrix, $M_{TD}$is the transition matrix from target network to drug network, and M_{DT} is the transition matrix from drug network to target network.Let λ be the probability of jumping from target network to drug network or vise versa. Transition matrix is defined as follows.

The transition probability from target t_i to target t_j is defined as

$$(M_T)_{i,j} = p(t_j|t_i) = \begin{cases} (S_T)_{ij} / \sum_j (S_T)_{ij} & if \sum_j B_{ij} = 0 \\ (1-\lambda)(S_T)_{ij} / \sum_j (S_T)_{ij} & otherwise \end{cases}$$

The transition probability from drug d_i to drug d_j is defined as

$$(M_D)_{i,j} = p(d_j|d_i) = \begin{cases} (S_D)_{ij} / \sum_j (S_D)_{ij} & if \sum_j B_{ji} = 0 \\ (1-\lambda)(S_D)_{ij} / \sum_j (S_D)_{ij} & otherwise \end{cases}$$

Table 1: Average fold enrichment comparison between NRWRH and NetLapRLS is shown to confirm that NRWRH has an excellent performance.

Method	Enzyme	Ion channel	GPCR	Nuclear receptor
NRWRH	242.6985	70.9746	31.6537	7.2362
NetLapRLS	216.2111	60.7005	21.0765	6.6087

The transition probability from target t_i to drug d_j is defined as

$$(M_{TD})_{i,j} = p(d_j|t_i) = \begin{cases} \lambda B_{ij}/\sum_j B_{ij} & if \sum_j B_{ij} \neq 0 \\ 0 & otherwise \end{cases}$$

The transition probability from drug d_i to target t_j is defined as

$$(M_{DT})_{i,j} = p(t_j|d_i) = \begin{cases} \lambda B_{ji}/\sum_j B_{ji} & if \sum_j B_{ji} \neq 0 \\ 0 & otherwise \end{cases}$$

Let p_t be a vector in which the i-th element holds the probability of being at node i at step t. The probability can be decided iteratively by $p_{t+1} = (1-r)M^T p_s + rp_0$. The parameter r is the restart probability. In each step, the restart of the walk at the seed nodes is allowed with probability r.

3.4 Stable probability and target ranking

After some steps, stable probability $p_\infty = \begin{bmatrix} (1-\eta)u_\infty \\ \eta v_\infty \end{bmatrix}$ is obtained by implementing the iteration until the change between p_t and p_{t+1} (measure by the L_1 norm) is less than 10^{-10} . Targets are ranked based on u_∞.

4 Results

For simplicity, we just choose r=0.7 ([18]), $\lambda = \eta = 0.2$, $w_d = w_t = 0.5$. These parameters can be better selected by further cross validation. Leave-one-out cross validation is implemented for evaluating the performance of method in the four classes of target proteins including enzymes, ion channels, GPCRs and nuclear receptors. Each known drug-target association is taken in turn as test dataset and other known drug-target interactions are used as training datasets.NRWRH is compared with NetLapRLS to show its predictive ability.

4.1 Fold enrichment

For each drug, the candidate target set is composed of all the targets that don't have evidences to show their association with this drug. When each known drug-target association is taken as test dataset, how well this target ranks relative to the candidate target set of this drug is assessed by fold enrichment. The formula is fold enrichment=the number of candidate targets/2/the rank of left out target. Fold enrichment actually represents the average rank of a target before prioritization divided by the rank after prioritization. For example, if the test target is ranked 1st in the candidate target set of 100 targets,

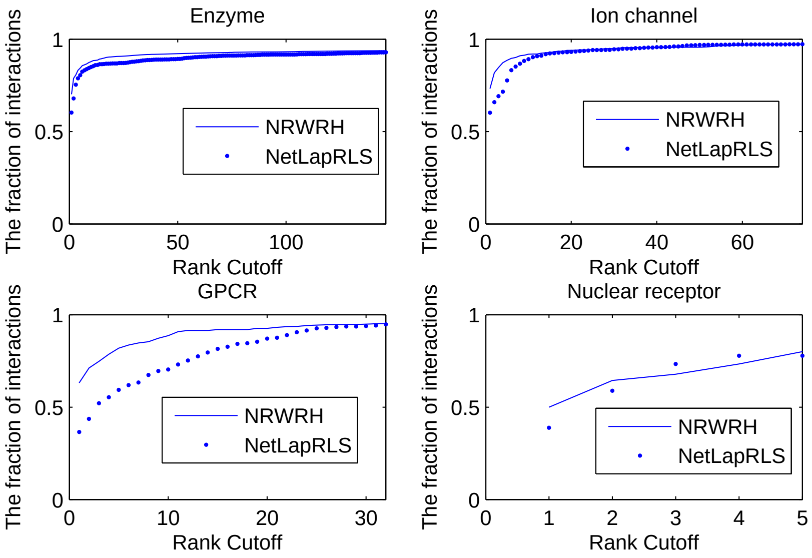

Figure 1: Drug target prediction performance based on leave-one-out cross validation. Each known drug-target interaction is treated as a test case. All the test cases are pooled together and the fraction of the tested interactions ranked above various rank cutoffs is calculated. NRWRH and NetLapRLS are applied to four classes of target proteins. The maximum rank cutoff for each dataset is five percent of the number of drug-target interactions in the dataset.

then the fold enrichment is 100/2/1=50. Average fold enrichment comparison between NRWRH and NetLapRLS is shown in Table 1 to confirm that NRWRH has an excellent performance.

4.2 Rank cutoff curve

A second commonly used evaluation is deciding the rank of test target among the candidate target set in each test case and calculating the fraction of test targets ranked above various cutoffs by considering all the test cases [25]. For the comparison with NetLapRLS, rank cutoff curve is shown in Figure 1 (the curve describing the relation between various cutoffs and the fraction of known drug-target interactions ranked above this cutoff), which still confirms the performance advantage of NRWRH compared to NetLapRLS.

5 Conclusion

In this work, we propose NRWRH to predict potential drug-target interactions by integrating the drug chemical structure, protein sequence, and known drug-target interactions. Methods are applied to four classes of target proteins including enzymes, ion channels, GPCRs and nuclear receptors. Cross validation is implemented to demonstrate the superior performance of NRWRH. We conclude that NRWRH successfully makes full use of the information of drug similarity and known drug-target interactions. The success of our method can be attributed to a combination of several factors. First we combine three different networks into a heterogeneous network and implement random walk on this heterogeneous network. Also we use the information of known drug-target interactions to improve the drug similarity and protein similarity. Finally when the drug has no known target, we can predict potential targets of this given drug based on the target information of drugs which are similar to the given drug. If more known drug-target interactions can be obtained, the performance of the method will be further improved. In the future work, we plan to integrate more biological relevant information to define drug-drug similarity and target-target similarity.

Acknowledges

We thank Prof. Lingyun Wu and Yong Wang for helpful discussion and Yoshihiro Yamanishi for making their data publicly available.

References

[1] Y.C. Wang, Z.X. Yang, Y. Wang, and N.Y. Deng. Computationally probing drug-protein interactions via support vector machine. Lett Drug Des Discov, 7(5), 370-378, 2010.

[2] Z. Xia, L.Y. Wu, X.B. Zhou, and S.T.C. Wong. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst Biol, 2010, in press.

[3] Q.L. Li, and L.H. Lai. Prediction of potential drug targets based on simple sequence properties. BMC Bioinformatics, 8, 353-364, 2007.

[4] J. Drews. Drug discovery. A historical perspective. Science, 287(5460), 1960-1964, 2000.

[5] J.P. Overington, B. Al-Lazikani, and AL Hopkins. How many drug targets are there? Nat Rev Drug Discov, 5(12):993-996, 2006.

[6] Y. Landry, and J.P. Gies. Drugs and their molecular targets: an updated overview. Fundam Clin Pharmacol.22(1), 1-18, 2008.

[7] S.J. Haggarty, K.M. Koeller, J.C. Wong, R.A. Butcher, S.L. Schreiber. Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays. Chem Biol. 10(5), 383-396, 2003.

[8] F.G. Kuruvilla, A.F. Shamji, S.M. Sternson, P.J. Hergenrother, S.L. Schreiber. Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. Nature. 416(6881), 653-657, 2002.

[9] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics, 24(13), 232-240, 2008.

[10] K. Yang, H. Bai, Q. Ouyang, L. Lai, and C. Tang. Finding multiple target optimal intervention in disease-related molecular network. Mol Syst Biol. 4, 228, 2008.

[11] G.R. Zimmermann, J. LehÍćr, and C.T. Keith. Multi-target therapeutics: when the whole is greater than the sum of the parts. Drug Discov Today. 12(1-2), 34-42, 2007.

[12] S. Frantz. Drug discovery: playing dirty. Nature. 13;437(7061), 942-3, 2005.

[13] M.J. Keiser, V. Setola, J.J. Irwin, C. Laggner, A.I. Abbas, S.J. Hufeisen, N.H. Jensen, M.B. Kuijer, R.C. Matos, T.B. Tran, R. Whaley, R.A. Glennon, J. Hert, K.L.H. Thomas, D.D. Edwards, B.K. Shoichet, B.L. Roth: Predicting new molecular targets for known drugs. Nature, 462(7270), 175-181, 2009.

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

[15] T. Can, O. Camoglu, and A.K. Singh. Analysis of protein-protein interaction networks using random walks. In BIOKDD '05: Proceedings of the 5th international workshop on Bioinformatics (New York, USA: Association for Computing Machinery), 61-68, 2005.

[16] S. Kohler, S. Bauer, D. Horn, and P.N. Robinson. Walking the Interactome for Prioritization of Candidate Disease Genes. Am J Hum Genet. 82(4), 949-958, 2008.

[17] X. Chen, G.Y. Yan, and X.P. Liao. A Novel Candidate Disease Genes Prioritization Method Based on Module Partition and Rank Fusion. OMICS, 2010, in press.

[18] Y. Li, and J.C. Patra. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics. 26(9), 1219-1224, 2010.

[19] M. Hattori, Y. Okuno, S. Goto, M. Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J Am Chem Soc. 125(39):11853-11865, 2003.

[20] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh S. Kawashima, T. Katayama, M. Araki, M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res, 34(Database issue), D354-357, 2006.

[21] T.F. Smith, and M. Waterman. Identification of common molecular subsequences. J. Mol. Biol., 147, 195-197, 1981.

[22] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg. Brenda, the enzyme database: updates and major new developments. Nucleic Acids Res., 32 (Database issue), D431-D433, 2004.

[23] S. Gunther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E.G. Urdiales, A. Gewiess, L.J. Jensen, R. Schneider, R. Skoblo, R.B. Russell, P. E. Bourne, P. Bork, and R. Preissner. Supertarget and matador: resources for exploring drug-target relationships. Nucleic Acids Res., 36 (Database issue), D919-D922, 2008.

[24] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res., 36 (Database issue), D901-D906, 2008.

[25] B. Linghu, E.S. Snitkin, Z.J. Hu, Y. Xia, C. DeLisi. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol.10(9):R91, 2009.