Molecular BioSystems

Dynamic Article Links

Cite this: *Mol. BioSyst.*, 2012, **8**, 1970–1978

www.rsc.org/molecularbiosystems

PAPER

Drug–target interaction prediction by random walk on the heterogeneous network†‡

Xing Chen,[abc] Ming-Xi Liu[ab] and Gui-Ying Yan*[ac]

Received 4th January 2012, Accepted 15th March 2012
DOI: 10.1039/c2mb00002d

Predicting potential drug–target interactions from heterogeneous biological data is critical not only for better understanding of the various interactions and biological processes, but also for the development of novel drugs and the improvement of human medicines. In this paper, the method of Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) is developed to predict potential drug–target interactions on a large scale under the hypothesis that similar drugs often target similar target proteins and the framework of Random Walk. Compared with traditional supervised or semi-supervised methods, NRWRH makes full use of the tool of the network for data integration to predict drug–target associations. It integrates three different networks (protein–protein similarity network, drug–drug similarity network, and known drug–target interaction networks) into a heterogeneous network by known drug–target interactions and implements the random walk on this heterogeneous network. When applied to four classes of important drug–target interactions including enzymes, ion channels, GPCRs and nuclear receptors, NRWRH significantly improves previous methods in terms of cross-validation and potential drug–target interaction prediction. Excellent performance enables us to suggest a number of new potential drug–target interactions for drug development.

Introduction

Predicting drug–target interactions from heterogeneous biological data is critical not only for better understanding of the various interactions and biological processes, but also for the development of novel drugs and improvement of human medicines.[1,2] There are about 6000–8000 targets of pharmacological interest in the human genome, but only a small number of them have been identified to be related to approved drugs so far.[3–6] As is well-known, the experimental determination of drug–target interactions are still time-consuming, expensive, and limited to small-scale research,[1,3,7–9] hence non-experimental methods are urgently needed so that the time and costs for searching for potential interactions and developing novel drugs in a genome-wide way can be decreased. Computational methods can provide new predictions for experimental scientists and narrow the scope of candidate targets to accelerate drug discovery.[10] Therefore, there is a strong incentive to develop powerful computational methods that are capable of detecting potential drug–protein interactions effectively in a genome-wide way.[9] In the past, drug research followed the one-disease, one-target, one-drug paradigm which hasn't accelerated the discovery of drugs as expected, since multiple targets are often involved in the same disease.[10] Recently, much attention has been paid to the development of multiple-target drugs in order to increase the drug efficacy and overcome drug resistance.[11–13]

In fact, the difficulty of the prediction task lies in the rarity of known drug–target interactions.[2] Molecular docking methods are invalid unless the 3D structures of target proteins are known and this problem seriously limits the wide use of this method on a genome-wide scale.[14] Campillos *et al.* (2008) identified drug–target interactions based on drug side-effect similarity, although detailed side-effect information is only available for marketed drugs.[15] Yang *et al.* made full use of disease networks to develop a novel computational method for finding multiple target optimal intervention (MTOI) solution which gives the best transformation of networks from the disease state into normal state.[11]

Recently, various statistical methods have been developed to infer potential drug–target interactions on a large scale by integrating some biological datasets, such as drug chemical structures, target protein sequences, and known drug–target interactions.

[a] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. E-mail: xingchen@amss.ac.cn, liumingxi@amss.ac.cn, yangy@amss.ac.cn*
[b] *Graduate University of Chinese Academy of Sciences, Beijing 100190, China. E-mail: xingchen@amss.ac.cn, liumingxi@amss.ac.cn*
[c] *National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, China. E-mail: xingchen@amss.ac.cn, yangy@amss.ac.cn*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb00002d

‡ **Authors' contributions.** XC and GYY conceived and developed the prediction method. XC conceived, designed and implemented the experiments. XC, MXL and GYY analyzed the results. XC and GYY wrote the paper. GYY provided guidance and supervision. All authors read and approved the final manuscript.

Keiser *et al.* (2009) proposed a computational method to predict the associations between drugs and targets based on chemical structure information,[16] however protein target information wasn't taken into consideration here. Yamanishi *et al.* (2008) integrated the information of drug–drug chemical structure similarity, protein–protein sequence similarity, and known drug–target interactions and constructed pharmacological space to predict the associations between drugs and target proteins in the framework of supervised bipartite methods.[10] Bleakley and Yamanishi (2009) used the bipartite local model (BLM) to predict drug–target interactions.[17] Under the assumption that drug–target interactions are more correlated with pharmacological effect similarity than chemical structure similarity, Yamanishi *et al.* (2010) introduced pharmacological effect information and then developed a new method for the identification of unknown drug–target interactions in the framework of supervised bipartite graph inference.[9] The common problem of the above three supervised learning methods is that they regard the unknown drug–target interactions as negative samples and they didn't utilize a wealth of unlabelled information to assist with predictions. Inaccurate negative sample selection would largely influence the predictive accuracy of the method. Xia *et al.* (2010) developed a semi-supervised method, NetLapRLS, which established two classifiers in the drug space and target space, respectively, and combined them to give the final prediction by a mean operation.[2]

The limitation of previous methods lies in two aspects. Firstly, some methods regard the unknown drug–target interactions as negative samples. Secondly, even if some semi-supervised methods make use of the unlabelled information, such as NetLapRLS, the final prediction results come from two different classifiers in the drug and target protein space, respectively. In this paper, based on the assumption that similar drugs often target similar target proteins and the framework of Random Walk with Restart (RWR),[18,19] a novel method of Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) is developed to infer potential drug–target interactions. NRWRH solves the above two problems and gives a final prediction, not two different results from two different spaces. It consists 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 nodes and seed target nodes simultaneously; then random walk on the heterogeneous network is implemented; finally, the most probable targets are selected according to the stable probability of the walk. In fact, random walk has been widely used in bioinformatics.[18–21]

In the past, most of the traditional methods predicted potential drug–target interactions by constructing supervised or semi-supervised classifiers. NRWRH, making full use of the tool of the network for data integration to predict drug–target associations, is very different to traditional methods. NRWRH is different from traditional random walk with restart in two aspects. One is that the information of known drug–target interaction networks is integrated with drug chemical structure similarity and protein sequence similarity to improve the similarity measure between drugs or targets.[2] The other is that the random walk is implemented on three networks.[21] When the given drug has known targets, candidate targets can be ranked 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, potential targets of this given drug are selected based on target information of drugs which are similar to this given drug. The excellent performance of NRWRH in the four classes of drug–target interaction datasets, including enzymes, ion channels, GPCRs and nuclear receptors, is demonstrated by the leave-one-out cross-validation schema. The method is also evaluated by predicting potential drug–target interactions and finding out how many of them are confirmed by various databases. NRWRH shows superior performance in the predictive ability compared to previous methods by the confirmation from KEGG database,[22] DrugBank databases[23] and SuperTarget database.[24]

Methods

Materials

In this paper, we investigate drugs targeting four pharmaceutically useful target classes: enzymes, ion channels, GPCRs and nuclear receptors. All the data used in this paper was downloaded from http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/.[10] Here a brief description about three kinds of datasets and corresponding matrix representation used in this paper is given. Among the matrices below, (1) (2) and (5) have been constructed in previous literature.[10]

(1) The drug chemical structure similarity matrix S_d^c (here, d denotes drug and c denotes chemical structure).

The entity $S_d^c(i,j)$ in row i column j is the chemical structure similarity between drugs i and j, which is a global score found by calculating the ratio between the size of common structures and the size of union structures based on a graph alignment algorithm. Drug chemical structure similarity can be calculated by SIMCOMP[25] based on the information of chemical structure of drugs from the DRUG and COMPOUND Sections in the KEGG LIGAND database.[22] The reasons for choosing SIMCOMP for drug similarity evaluation are listed as follows. On the one hand, traditional methods for the structural similarity calculation between a set of compounds include 2D fingerprinting method, Tanimoto coefficient and so on. However, as a mathematical extension of the earlier bit-comparison method and a kind of graph comparison method, the 2D fingerprinting method has two limitations. At first, this kind of method does not include biochemical information in the representation of atoms, in other words, it does not consider distinguishing the same atoms under different environments. Secondly, as a kind of graph comparison method, its computational time would increase exponentially for larger biochemical compounds.[25] For a similarity metric such as the Tanimoto coefficient, sometimes it may yield low similarity values and it has an inherent bias towards certain similarity values.[26,27] Hence, we chose the SIMCOMP method to calculate chemical structural similarity between compounds, which could improve these limitations to some degree.[25] On the other hand, drug chemical

structure evaluation based on SIMCOMP has been widely applied to the drug-related research, especially the prediction of drug–target interactions.[1,2,9,10,17,28,29]

(2) The target protein sequence similarity matrix S_t^s (here, t denotes target and s denotes sequence information).

The entity $S_t^s(i,j)$ in row i column j is the target protein sequence similarity between targets i and j, which is calculated by normalized version of Smith–Waterman scores[30] based on the information of amino acid sequences of target proteins from the KEGG GENES database.[22]

(3) New drug–drug similarity matrix S_d^n and target–target similarity matrix S_t^n (here, d denotes drug, t denotes target and n denotes new similarity measure).

The entity $S_d^n(i,j)$ in row i column j is the number of known targets shared by drugs i and j. Correspondingly, the entity $S_t^n(i,j)$ in row i column j is the number of known drugs shared by targets i and j. Here the aim is to give another similarity measure by extracting the information from the known drug–target interactions.[2] The underlying assumption is that if two drugs have more common targets, they are more similar and if two target proteins are targeted by more common drugs, they are more similar.

(4) The integrated drug–drug similarity matrix S_d and target–target similarity matrix S_t (here, d denotes drug, t denotes target).

The entity $S_d(i,j)$ in row i column j is an integrated similarity between drug i and j based on the known chemical structure similarities and new drug–drug similarties mentioned above. Correspondingly, the entity $S_t(i,j)$ in row i column j is the integrated similarity between target i and j based on the known target sequence similarities and new target–target similarities mentioned above. New drug–drug similarities and new target–target similarities matrices 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 $\overline{S_d^n} = (D_d^n)^{-1/2} S_d^n (D_d^n)^{-1/2}$ which yields a symmetric matrix where $\overline{S_d^n}(i,j) = S_d^n(i,j)/\sqrt{D_d^n(i,i) D_d^n(j,j)}$. A similar operation is applied to S_t^n and the normalized matrix $\overline{S_t^n}$ is obtained. The drug similarity matrix can be obtained by the linear combination $S_d = w_d S_d^c + (1 - w_d)\overline{S_d^n}$. Similarly, the target similarity matrix can be obtained by $S_t = w_t S_t^s + (1 - w_t)\overline{S_t^n}$. Here parameter w_d and w_t represent the weight of traditional similarity evaluation in the integrated similarity measure.

(5) The drug target interaction adjacency matrix A.

The entity $A(i,j)$ in row i column j is 1 if protein target i is targeted by drug j based on the confirmation from the KEGG BRITE,[22] BRENDA,[31] SuperTarget[24] and DrugBank databases,[23] otherwise 0. The numbers of known interactions in the four datasets are 2926, 1476, 635 and 90, for their targets enzymes, ion channels, GPCRs and nuclear receptors, respectively. The numbers of the corresponding drugs in these classes are 445, 210, 223 and 54, respectively. The numbers of the corresponding target proteins in these classes are 664, 204, 95 and 26, respectively. These datasets are regarded as the 'gold standard' data in this study for evaluating the performance of various methods in the cross-validation schema and predicting potential drug–target interactions.

Network-based random walk with restart on the heterogeneous network

In the present study, based on the assumption that similar drugs often target similar target proteins and the framework of Random Walk with Restart (RWR),[18,19] a novel method of Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) is developed to infer potential drug–target interactions. It consists of four steps as followed: construct the heterogeneous network; decide the initial probability; decide the transition matrix and implement random walk; obtain stable probability and rank candidate targets.

Based on the aforementioned integrated drug–drug similarity matrix, integrated target–target similarity matrix, and drug target interaction adjacency matrix, drug similarity network, target similarity network, and drug–target interaction networks are constructed, respectively. In the drug similarity network, let vertices set $D = \{d_1, d_2, \ldots, d_n\}$ denote the set of n drugs, vertex d_i and d_j are connected if the integrated similarity between drug i and j is more than 0, the integrated similarity between drug i and j is used as the weight of this edge. The target similarity network is constructed in the same way as the drug similarity network. Let vertices set $T = \{t_1, t_2, \ldots, t_m\}$ denote the set of m targets, vertex t_i and t_j are connected if the integrated similarity between target i and j is more than 0, the integrated similarity between target i and j is used as the weight of this edge. In the drug–target interactions network, if target protein j is targeted by drug i, then the edge is directed from vertex d_i to t_j. The aforementioned three kinds of network constitute the heterogeneous network, which is constructed by connecting the drug similarity network and target similarity network using the drug–target interactions network.[21] A simple example of the heterogeneous network is illustrated in Fig. 1.

NRWRH is proposed to uncover the association between drugs and targets by simulating a random walker's transition from its current nodes randomly to the neighbors in the heterogeneous network starting at some given seed nodes.[21] NRWRH allows the restart of the walk in every step at a source node with probability r. Hence the initial probability of various vertices must be determined. If we want to predict potential targets of a given drug i, d_i is denoted as the seed node in the drug network. If target protein j is targeted by drug i, then t_j is used as the seed nodes in the target network. The initial probability of the target network u_0 is formed such that equal probabilities are assigned to the seed nodes in the target network, with the sum equal to 1. In the drug network probability 1 is given to vertex d_i and probability 0 is given to other vertices, forming the initial probability of drug network v_0. 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.

To implement random walk, the transition matrix must be decided.[21] Let $M = \begin{bmatrix} M_{TT} & M_{TD} \\ M_{DT} & M_{DD} \end{bmatrix}$ be the transition matrix of the heterogeneous network, where M_{TT} and M_{DD} are inter-transition matrix indicating the probability from one target (drug) to other target (drug) in the random walk, respectively; M_{TD} is the transition matrix from target network to drug

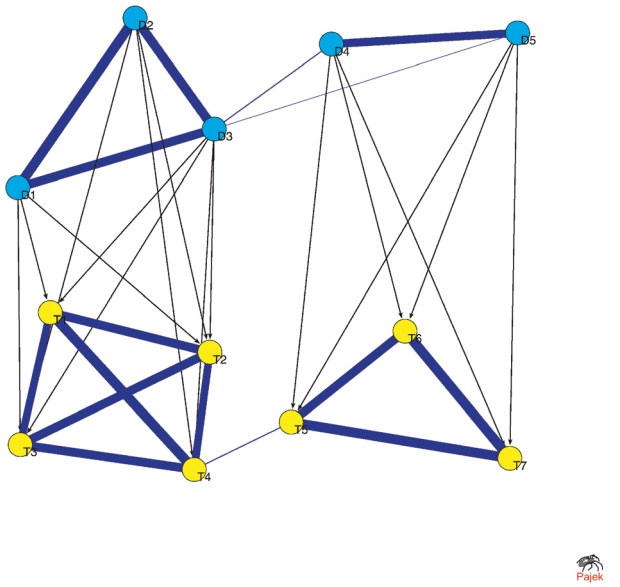

Fig. 1 A simple example of the heterogeneous network. The upper network is the drug similarity network and the lower network is the target similarity network. They are connected into a heterogeneous network by known drug–target interactions. Random walk will be implemented on this heterogeneous network to make full use of these three kinds of data.[21]

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 *vice versa*. The transition matrix is defined as follows:

The transition probability from vertex t_i to t_j is defined as

$$M_{TT}(i,j) = p(t_j|t_i) = \begin{cases} S_t(i,j)/\sum_j S_t(i,j) & \text{if } \sum_j A(i,j) = 0 \\ (1-\lambda)S_t(i,j)/\sum_j S_t(i,j) & \text{otherwise} \end{cases}.$$

The transition probability from vertex d_i to d_j is defined as

$$M_{DD}(i,j) = p(d_j|d_i) = \begin{cases} S_d(i,j)/\sum_j S_d(i,j) & \text{if } \sum_j A(j,i) = 0 \\ (1-\lambda)S_d(i,j)/\sum_j S_d(i,j) & \text{otherwise} \end{cases}$$

The transition probability from vertex t_i to d_j is defined as

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

The transition probability from vertex d_i to t_j is defined as

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

Random walk is implemented on the heterogeneous network. Let p_t be a vector in which the i-th element holds the probability of finding the random walker at node i at step t. The probability can be decided iteratively by

$$p_{t+1} = (1-r)M^T p_t + r p_0$$

The parameter r is the restart probability. At each step, the restart of the walk at the seed nodes is allowed with probability r.

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_∞. Targets with maximum in u_∞ among all the non-seed nodes is considered as the most probable target of drug i.

RWRH, NRWR and RWR for drug–target prediction

Here we also propose another three methods for predicting potential drug–target interactions: RWRH (Random Walk with Restart on the Heterogeneous network), NRWR (Network-based Random Walk with Restart), and RWR (Random Walk with Restart). The only difference between NRWRH and RWRH is that RWRH doesn't integrate the information of known drug–target interactions to improve the similarity measure. The difference between RWRH and RWR is that RWR only implements random walk on the target similarity network. Based on the framework of RWR, NRWR integrates the known drug-target interactions to improve the target similarity measurements, but random walk is still only implemented on the target network. The aim of proposing these methods is to implement performance comparison between NRWRH and these three methods to confirm the benefit from walking on the heterogeneous network and improving similarity measurements by integrating the known drug–target interactions information.

Results

The similarity nature of drug–target interactions

Here, we proposed the similarity nature of drug–target interactions: similar drugs often interact with similar target proteins. Based on this assumption, NRWRH was developed to predict the potential drug–target interactions. There are numerous examples of similar drugs that target similar target proteins. A similar nature has been illustrated in four classes of data (Fig. 2 for enzymes, Fig. S1 for ion channels, Fig. S2 for GPCRs, and Fig. S3 for nuclear receptors, see ESI†). Based on the similarity nature, the potential targets of a given drug can be identified by calculating the similarity between the candidate targets and the known seed targets.

Performance evaluation

For simplicity, we just choose $r = 0.7$, $\lambda = \eta = 0.2$, $w_d = w_t = 0.5$. These parameters can be better selected by further cross-validation. We will discuss the effect of parameters in the next section. Leave-one-out cross-validation was implemented for evaluating the performance of NRWRH in the four classes of target proteins including enzymes, ion channels, GPCRs, and nuclear receptors. Each known drug–target association was taken in turn as test datasets and other known drug–target interactions were used as training datasets. For each drug, the

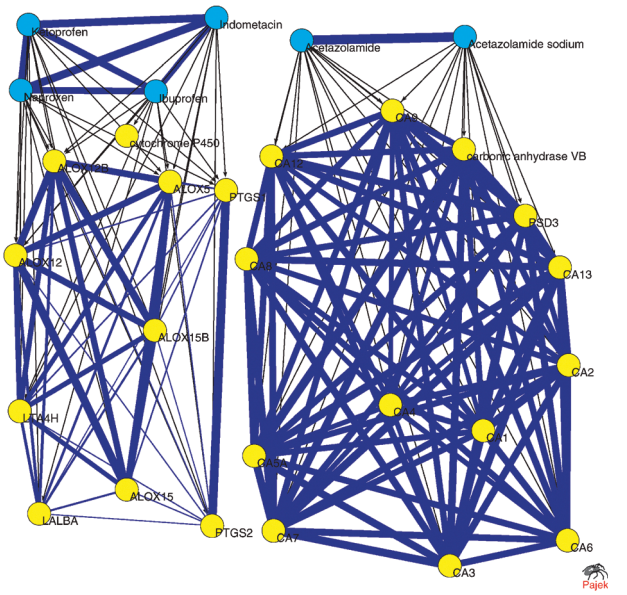

Fig. 2 The similarity nature of drug–target interactions is shown in the dataset of enzyme. Blue nodes represent the drugs and yellow nodes represent the target proteins. The black arc between the drug and the target protein means that this drug targets this protein. The blue edge between two drugs or two proteins represents the similarity between them. The thickness of the lines linking drugs or proteins indicates the degree of similarity. This figure shows similar drugs often target similar proteins. For example, the targets of similar drugs (Naproxen, Ibuprofen, Ketoprofen, and Indometacin) are similar. Also the targets of Acetazolamide and Acetazolamide sodium are significantly similar.

candidate target set was composed of all the targets that don't have any evidence to show their association with this drug. When each known drug–target association was taken as a test dataset, how well this target ranks relative to the candidate target set of this drug was assessed.

NRWRH can only prioritize the candidate targets for the given drug. That is to say, it can't prioritize all the candidate drug–target interactions for all the drugs simultaneously. Therefore, only the known targets of the given drug are used as test samples in each cross validation run. From the datasets four kinds of important protein targets, we can observe that each drug targets about 13.15, 14.06, 5.70, and 3.33 targets on average for the four kinds of datasets, respectively. This fact means that little difference between the result of leave-one-out cross validation and 10-fold cross validation would be obtained. Moreover, when the given drug has less than 10 targets, 10-fold cross validation can't be implemented. Based on the aforementioned considerations, leave-one-out cross validation was adopted in this paper. An important fact which must be pointed out is that we recalculate the new network-based drug similarity matrix and new network-based target similarity matrix in each cross validation run, without using the information about tested drug–target interaction.

The fold enrichment is a traditional measure for performance evaluation.[19] The formula is as follows: 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, then the fold enrichment is 100/2/1 = 50. Here the average fold enrichment is calculated among all the test cases. The performance of NRWRH, RWRH, RWR, and NRWR in terms of average fold enrichment is compared to show the benefits of integrating the information of known drug–target interactions to improve the drug and target similarity measure and implementing the random walk on three networks simultaneously (Fig. 3).

A second commonly used evaluation is deciding the rank of the 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.[32] Rank cutoff curves (the curve describing the relation between various cutoffs and the fraction of known drug–target interactions ranked above this cutoff) of NRWRH, RWRH, NRWR, RWR are compared in Fig. 4, which confirms the excellent performance of NRWRH.

When leave-one-out cross-validation is implemented, the ROC curve of each drug can be obtained to assess how well the known targets of this drug rank relative to the candidate targets. ROC curve plots the true positive rate (sensitivity) *versus* false positive rate (1−specificity) at different cutoffs. Finally, an AUC is calculated for each ROC curve. AUC = 1 indicates perfect performance and 0.5 shows random performance. The performance of the method is evaluated by counting how many drugs have an average AUC larger than different cutoffs. The AUC comparison in each dataset is shown (Fig. 5 for enzyme, Fig. S4 for ion channel, Fig. S5 for GPCR, Fig. S6 for the nuclear receptor, see ESI†). The ROC curve of all the drug–target interactions can also be obtained and the comparison between various methods in term of AUC is shown in Fig. 6. It has been observed that NRWRH has obtained a very good performance.

To further confirm the superior performance of NRWRH to previous methods and consider the fact that high-confidence prediction results are interesting in practical applications, we also provide the cross validation result based on sensitivity and specificity when the upper one percentile and top one target in

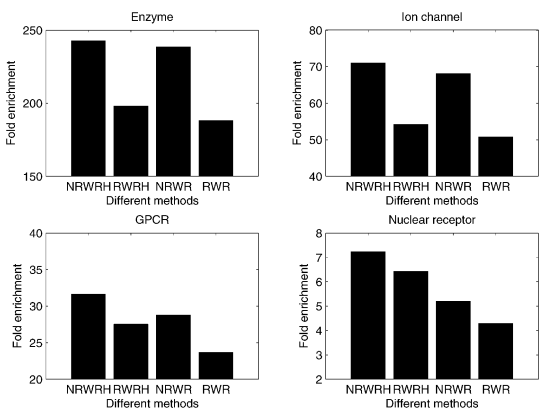

Fig. 3 Performance evaluation in terms of fold enrichment. The performance of NRWRH, RWRH, RWR, and NRWR are compared in terms of fold enrichment to show the benefits of integrating the information of known drug–target interactions to improve the drug and target similarity measure and implementing the random walk on three networks.

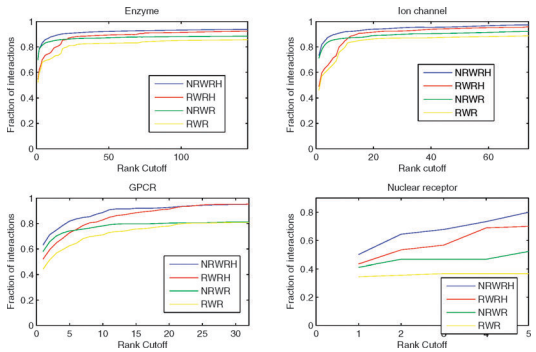

Fig. 4 Performance evaluation in terms of rank cutoff curves. Rank cutoff curves (the curve describing the relation between various cutoffs and the fraction of known drug–target interactions ranked above this cutoff) of NRWRH, RWRH, NRWR, RWR are compared here to confirm the correctness of integrating the information of known drug–target interactions to improve the drug and target similarity measure and implementing the random walk on three networks.

the prediction list is chosen as a threshold. The comparisons between NRWRH, RWRH, RWR, and NRWR in terms of sensitivity and specificity in various datasets are shown, respectively (Table 1: enzyme; Table S1: ion channel; Table S2: GPCR; Table S3: nuclear receptors, see ESI†). The results demonstrate the superior performance of NRWRH to other methods.

Effects of parameters

There are five parameters in NRWRH: restart probability r, jumping probability λ, η controlling the impact of two kinds of seed nodes, w_d and w_t controlling the impact of traditional similarity evaluation and new network-based similarity measure about the drugs and target proteins, respectively. It has been demonstrated that the predictive result is robust to the restart probability.[19–21] Therefore we choose a restart probability of 0.7 according to the selection in a previous paper about disease gene identification.[21]

To confirm that NRWRH is robust to the selection of λ, we set various values of λ ranging from 0.1 to 0.9 and calculated overall AUC in the framework of the leave-one-out cross validation. When jumping probability is equal to 0 or 1, then the random walk will be implemented only on the single

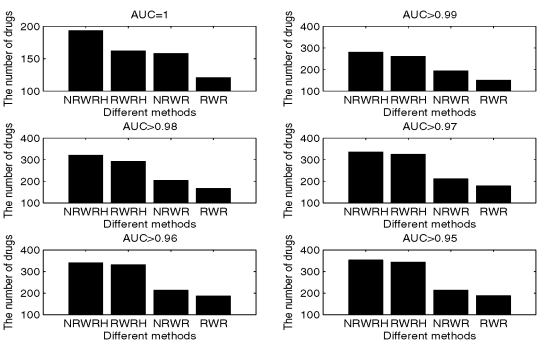

Fig. 5 Performance evaluation in terms of AUC distribution in the enzyme dataset. The performance of the methods is evaluated by counting how many drugs have an AUC larger than different cutoffs. The AUC comparison in the enzyme dataset is shown here.

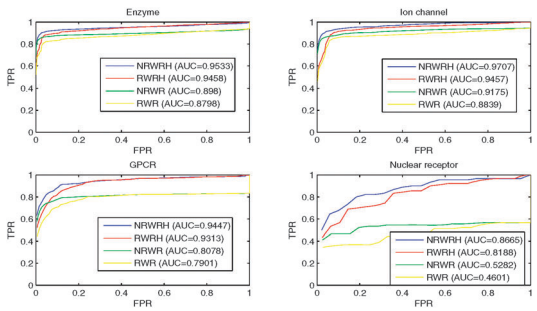

Fig. 6 Performance evaluation in terms of overall AUC. Comparison between NRWRH, RWRH, NRWR and RWR in terms of ROC curve and AUC of all the drug–target interactions is shown to confirm the benefits of integrating the information of known drug–target interactions to improve the drug and target similarity measure and implementing the random walk on three networks.

Table 1 Performance evaluation in terms of sensitivity and specificity in the dataset of enzyme

Enzyme	Sensitivity (Top 1%)	Specificity (Top 1%)	Sensitivity (Top 1)	Specificity (Top 1)
NRWRH	0.8592	0.9913	0.7016	0.9995
NRWR	0.8329	0.9913	0.6965	0.9995
RWRH	0.7351	0.9911	0.5437	0.9993
RWR	0.6910	0.9910	0.5195	0.9992

network, not the heterogeneous network. Hence, we restrict the value of jumping probability to the interval from 0.1 to 0.9. Table 2 shows the effects of jumping probability on the cross validation result in the four kinds of datasets.

Parameter η controls the impact of two kinds of seed nodes, *i.e.* drug node and target node. Parameters w_d and w_t denote network-based similarity weights for the drugs and targets, respectively. The effects of these three parameter values on cross validation result in the term of overall AUC for four kinds of datasets, see Table S4 (η), Table S5 (w_d), and Table S6 (w_t), ESI†. As seen from the above four tables for parameter effect discussion, we can conclude that NRWRH is robust to the selection of parameter values.

Comparison with other methods

Drug target interaction prediction methods based on supervised classifiers evaluate the score of both positive sample and negative sample in cross validation schema, while NRWRH only evaluates the score of positive samples. Considering the difference of test samples between NRWRH and the supervised classifier methods for drug–target interaction prediction, NRWRH can't be directly compared with supervised methods. Hence, NRWRH was compared with the three methods below: NetLapRLS, LapRLS and the Weighted profile method (WPM).[2] Performance comparison was implemented in the same way as the above section. Leave-one-out cross-validation was implemented and three performance measures (fold enrichment, rank cutoff curve, and AUC) were calculated. The fold enrichment comparison is shown in Fig. 7 and the rank cutoff curve is shown in Fig. 8. AUC comparison of each drug is also shown (Fig. 9 for enzyme, Fig. S7 for ion channel, Fig. S8 for GPCR, Fig. S9 for nuclear receptors, see ESI†). The ROC

Table 2 The effect of jumping probability value on the cross validation result of NRWRH

λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Enzyme	0.9520	0.9533	0.9543	0.9550	0.9553	0.9554	0.9553	0.9550	0.9541
GPCR	0.9436	0.9447	0.9464	0.9476	0.9483	0.9486	0.9487	0.9485	0.9479
Ion channel	0.9692	0.9707	0.9722	0.9740	0.9758	0.9768	0.9764	0.9752	0.9715
Nuclear receptor	0.8653	0.8665	0.8699	0.8757	0.8767	0.8762	0.8755	0.8740	0.8723

curves of all the drug–target interactions predictions by these methods are shown in Fig. 10. Comparisons in terms of sensitivity and specificity in various datasets are shown, respectively (Table 3: enzyme; Table S7: ion channel; Table S8: GPCR; Table S9: nuclear receptors, see ESI†). In almost all the datasets, NRWRH is significantly improved over previous methods and shows a very good performance.

Predicting potential drug–target interactions

After confirming the usefulness of NRWRH by cross validation, predicting potential drug–target interactions for the four classes of target proteins was conducted. Here all the known drug–target interactions in the gold standard data were used as training data. We focused on the most probable target of each drug predicted by NRWRH and NetLapRLS in different datasets. The validity of the predicted drug–target pairs was investigated based on the KEGG database, DrugBank database, and SuperTarget database, because the gold standard data in this study was collected before

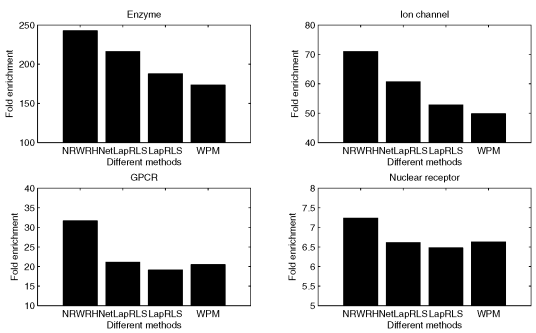

Fig. 7 Performance improvement in terms of fold enrichment. The performance of NRWRH, NetLapRLS, LapRLS and WPM are compared in terms of fold enrichment to show the superior performance of NRWRH over previous methods.

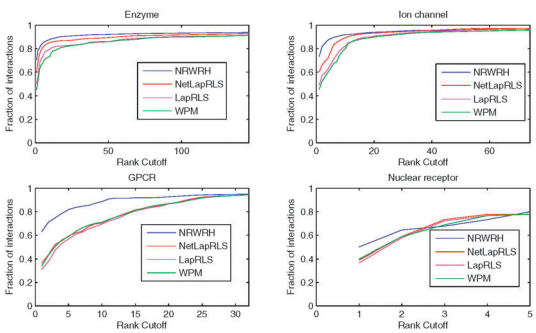

Fig. 8 Performance improvement in terms of rank cutoff curves. Rank cutoff curves (the curve describing the relation between various cutoffs and the fraction of known drug–target interactions ranked above this cutoff) of NRWRH, NetLapRLS, LapRLS and WPM are compared here to confirm the excellent performance of NRWRH.

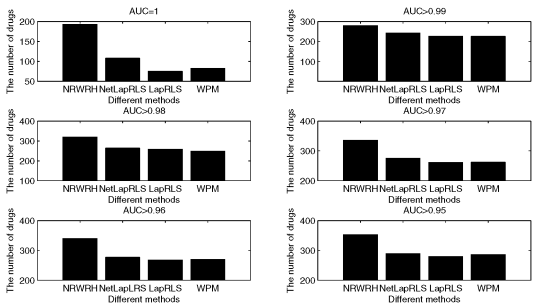

Fig. 9 Performance improvement in terms of AUC distribution. The performance of the methods is evaluated by counting how many drugs have an AUC larger than different cutoffs. The AUC comparison between NRWRH, NetLapRLS, LapRLS and WPM in the enzyme dataset is shown to confirm the good performance of NRWRH.

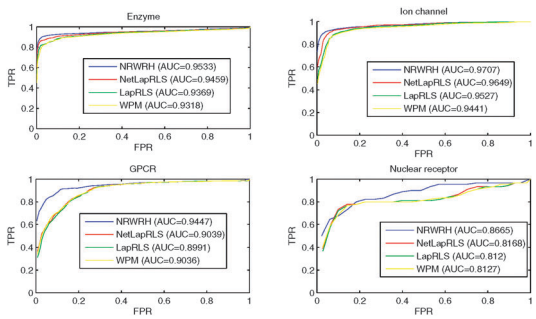

Fig. 10 Performance improvement in terms of overall AUC. Comparison between NRWRH, NetLapRLS, LapRLS and WPM in terms of ROC curve and AUC of all the drug-target interactions is shown to confirm the performance advantage of NRWRH compared to previous methods.

Table 3 Performance improvement over previous methods in terms of sensitivity and specificity in the datasets of enzymes

Enzyme	Sensitivity (Top 1%)	Specificity (Top 1%)	Sensitivity (Top 1)	Specificity (Top 1)
NRWRH	0.8592	0.9913	0.7016	0.9995
NetLapRLS	0.8291	0.9913	0.6025	0.9994
LapRLS	0.7628	0.9912	0.4993	0.9992
WPM	0.7075	0.9911	0.4477	0.9991

2008 and some new drug–target interactions have also been introduced into these databases from then on. All the prediction results of the most probable target for each drug can be obtained in the ESI† (Table S10 for the enzyme, Table S11 for the ion channel, Table S12 for GPCR, Table S13 for nuclear receptors). The drug–target interactions predicted by NRWRH and NetLapRLS and also confirmed by at least one database have also been shown in the Table S14, see ESI.†

Because of space limitations, we focused on the results for the GPCR dataset below (Table 4). We confirmed that

Table 4 The drug–target interactions predicted by NRWRH and NetLapRLS in the dataset of GPCR

Drug ID	Target ID	Evidence	Method	Drug ID	Target ID	Evidence	Method
D00049	hsa:8843	DrugBank	NRWRH	D01692	hsa:147	DrugBank	NRWRH
D00059	hsa:1816	KEGG	NRWRH	D01891	hsa:5732	KEGG	NRWRH
D00079	hsa:5731	DrugBank	NRWRH	D02250	hsa:6751	KEGG	NRWRH
D00270	hsa:3358	KEGG	NRWRH	D02340	hsa:1812	DrugBank	NRWRH
D00397	hsa:1133	KEGG	NRWRH	D02349	hsa:151	KEGG	NRWRH
D00371	hsa:135	KEGG, DrugBank	NRWRH	D02357	hsa:3358	KEGG, DrugBank	NRWRH
D00415	hsa:3355	Supertarget, DrugBank	NRWRH	D00442	hsa:6755	KEGG, DrugBank	NRWRH
D00419	hsa:5731	KEGG	NRWRH	D03490	hsa:155	KEGG	NRWRH
D02725	hsa:5732	KEGG	NRWRH	D05113	hsa:4986	DrugBank	NRWRH
D00498	hsa:4986	KEGG, DrugBank	NRWRH	D01118	hsa:1129	KEGG	NRWRH
D00514	hsa:151	KEGG	NRWRH	D00540	hsa:1132	KEGG	NRWRH
D00524	hsa:1132	KEGG	NRWRH	D00559	hsa:1813	KEGG	NRWRH
D00525	hsa:1129	KEGG	NRWRH	D00563	hsa:152	KEGG	NRWRH
D00779	hsa:1132	KEGG	NRWRH	D00837	hsa:4985	DrugBank	NRWRH
D01358	hsa:151	KEGG	NRWRH	D02349	hsa:154	KEGG	NetLapRLS
D00560	hsa:148	KEGG	NetLapRLS	D01358	hsa:150	KEGG	NetLapRLS
D00095	hsa:155	KEGG, Supertarget	NRWRH, NetLapRLS	D01103	hsa:1129	KEGG	NRWRH, NetLapRLS
D01386	hsa:153	KEGG	NRWRH, NetLapRLS	D04625	hsa:154	KEGG	NRWRH, NetLapRLS
D03415	hsa:148	KEGG	NRWRH, NetLapRLS	D00232	hsa:1132	KEGG	NRWRH, NetLapRLS
D04375	hsa:151	KEGG	NRWRH, NetLapRLS	D00715	hsa:1129	KEGG	NRWRH, NetLapRLS

37 predictions by the NRWRH in the GPCR dataset are now annotated in at least one database. By contrast, NetLapRLS can only find 11 drug–target interactions. We take this result as strong evidence to support the practical relevance of our approach. Twenty-nine drug–target interactions predicted by NRWRH can't be obtained by NetLapRLS, while only three interactions predicted by NetLapRLS can't be obtained by NRWRH. Similar conclusions can also be reached in the other datasets.

Discussion

The success of NRWRH can be attributed to a combination of several factors. Firstly, three different networks are combined into a heterogeneous network and random walk is implemented on this heterogeneous network. Moreover, known drug–target interactions are used to improve the drug similarity and protein similarity. Finally, when the drug has no known target, potential targets of this given drug can be predicted based on the target information of drugs which are similar to the given drug. In a word, the superior performance of NRWRH to other traditional methods is attributed to the fact that NRWRH makes full use of the tool of the network for the data integration to predict drug–target associations, while traditional methods always constructed supervised or semi-supervised classifiers to infer drug–target interactions. NRWRH is expected to be useful for drug development.

When predicting potential target proteins for new drugs which do not have any known target information, network-based drug (target) similarity matrix is zero matrices, limiting the wide application of NRWRH. Therefore, the performance of NRWRH could be further improved by obtaining more known drug–target interactions. In future work, we plan to integrate more biologically relevant information to define drug–drug similarity and target–target similarity for further improvement on the performance of the drug–target prediction methods.

Conclusions

In this work, NRWRH is developed to predict potential drug–target interactions by integrating the drug chemical structure information, protein sequence information, and known drug–target interactions information on a large scale. The originality of the proposed method lies in the integration of three different networks (drug similarity network, target similarity network and known drug–target interaction networks) into a heterogeneous network. NRWRH is applied to four classes of target proteins including enzymes, ion channels, GPCRs and nuclear receptors. Cross-validation and potential drug–target interaction predictions are implemented to demonstrate the superior performance of NRWRH to previous methods.

Acknowledgements

The financial support from the National Natural Science of Foundation of China under Grant Nos. 10531070, 10721101, KJCX-YW-S7 and NCMIS is highly appreciated. We thank Prof. Lingyun Wu and Yong Wang for helpful discussion and Yoshihiro Yamanishi for making their data publicly available. We also thank anonymous reviewers for valuable suggestions.

Notes and references

1 Y. C. Wang, Z. X. Yang, Y. Wang and N. Y. Deng, *Lett. Drug Des. Discovery*, 2010, **7**, 370–378.
2 Z. Xia, L. Y. Wu, X. Zhou and S. T. Wong, *BMC Syst. Biol.*, 2010, **4**(Suppl 2), S6.
3 Q. Li and L. Lai, *BMC Bioinformatics*, 2007, **8**, 353.
4 J. Drews, *Science*, 2000, **287**, 1960–1964.
5 J. P. Overington, B. Al-Lazikani and A. L. Hopkins, *Nat. Rev. Drug Discovery*, 2006, **5**, 993–996.
6 Y. Landry and J. P. Gies, *Fundam. Clin. Pharmacol.*, 2008, **22**, 1–18.
7 S. J. Haggarty, K. M. Koeller, J. C. Wong, R. A. Butcher and S. L. Schreiber, *Chem. Biol.*, 2003, **10**, 383–396.
8 F. G. Kuruvilla, A. F. Shamji, S. M. Sternson, P. J. Hergenrother and S. L. Schreiber, *Nature*, 2002, **416**, 653–657.
9 Y. Yamanishi, M. Kotera, M. Kanehisa and S. Goto, *Bioinformatics*, 2010, **26**, i246–254.
10 Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda and M. Kanehisa, *Bioinformatics*, 2008, **24**, i232–240.
11 K. Yang, H. Bai, Q. Ouyang, L. Lai and C. Tang, *Mol. Syst. Biol.*, 2008, **4**, 228.

12 G. R. Zimmermann, J. Lehar and C. T. Keith, *Drug Discovery Today*, 2007, **12**, 34–42.
13 S. Frantz, *Nature*, 2005, **437**, 942–943.
14 M. Rarey, B. Kramer, T. Lengauer and G. Klebe, *J. Mol. Biol.*, 1996, **261**, 470–489.
15 M. Campillos, M. Kuhn, A. C. Gavin, L. J. Jensen and P. Bork, *Science*, 2008, **321**, 263–266.
16 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. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth, *Nature*, 2009, **462**, 175–181.
17 K. Bleakley and Y. Yamanishi, *Bioinformatics*, 2009, **25**, 2397–2403.
18 T. Can, O. Camoglu and A. K. Singh, 2005.
19 S. Kohler, S. Bauer, D. Horn and P. N. Robinson, *Am. J. Hum. Genet.*, 2008, **82**, 949–958.
20 X. Chen, G. Y. Yan and X. P. Liao, *OMICS*, 2010, **14**, 337–356.
21 Y. Li and J. C. Patra, *Bioinformatics*, 2010, **26**, 1219–1224.
22 M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki and M. Hirakawa, *Nucleic Acids Res.*, 2006, **34**, D354–357.
23 D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam and M. Hassanali, *Nucleic Acids Res.*, 2007, **36**, D901–906.
24 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, *Nucleic Acids Res.*, 2008, **36**, D919–922.
25 M. Hattori, Y. Okuno, S. Goto and M. Kanehisa, *J. Am. Chem. Soc.*, 2003, **125**, 11853–11865.
26 D. R. Flower, *J. Chem. Inf. Model.*, 1998, **38**, 379–386.
27 J. W. Godden, L. Xue and J. Bajorath, *J. Chem. Inf. Model.*, 2000, **40**, 163–166.
28 W. M. Yu, X. A. Cheng, Z. B. Li and Z. R. Jiang, *Drug Dev. Res.*, 2011, **72**, 219–224.
29 T. van Laarhoven, S. B. Nabuurs and E. Marchiori, *Bioinformatics*, 2011, **27**, 3036–3043.
30 T. F. Smith and M. S. Waterman, *J. Mol. Biol.*, 1981, **147**, 195–197.
31 I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn and D. Schomburg, *Nucleic Acids Res.*, 2004, **32**, D431–433.
32 B. Linghu, E. S. Snitkin, Z. Hu, Y. Xia and C. Delisi, *GenomeBiology*, 2009, **10**, R91.

