Prioritizing Disease Genes by Bi-Random Walk

Maoqiang Xie[1], Taehyun Hwang[2], and Rui Kuang[3,⋆]

[1] College of Software, Nankai University, Tianjin, China
[2] Masonic Cancer Center, University of Minnesota, Twin Cities, USA
[3] Department of Computer Science and Engineering,
University of Minnesota, Twin Cities, USA
kuang@cs.umn.edu

Abstract. Random walk methods have been successfully applied to prioritizing disease causal genes. In this paper, we propose a bi-random walk algorithm (BiRW) based on a regularization framework for graph matching to globally prioritize disease genes for all phenotypes simultaneously. While previous methods perform random walk either on the protein-protein interaction network or the complete phenome-genome heterogenous network, BiRW performs random walk on the Kronecker product graph between the protein-protein interaction network and the phenotype similarity network. Three variations of BiRW that perform balanced or unbalanced bi-directional random walks are analyzed and compared with other random walk methods. Experiments on analyzing the disease phenotype-gene associations in Online Mendelian Inheritance in Man (OMIM) demonstrate that BiRW effectively improved disease gene prioritization over existing methods by ranking more known associations in the top 100 out of nearly 10,000 candidate genes.

Keywords: Disease Gene Prioritization, Bi-Random Walk, Graph-based Learning.

1 Introduction

It is now well accepted that phenotypes are determined by genetic material under environmental influences. To understand the relation between disease phenotypes and genes, numerous genomic studies on large patient cohorts such as genome-wide association studies [1][2] have been conducted to identify candidate disease genes, and in the past decade, the knowledge of determined disease phenotype-gene associations has been quickly accumulated in databases such as Online Mendelian Inheritance in Man (OMIM), a database of human genes and genetic disorders. Driven by the accumulated knowledge, random walk-based algorithms, which take the advantage of the availability of large phenotypic and molecular networks (Fig. 1), were proposed to utilize the disease modules and gene modules in the networks to prioritize disease genes [3][4][5][6][7][8][9]. The human disease phenotype network [10] provides information on phenotype similarities computed

⋆ Corresponding author.

P.-N. Tan et al. (Eds.): PAKDD 2012, Part II, LNAI 7302, pp. 292–303, 2012.

by text mining of the full text and clinical synopsis of the disease phenotypes in OMIM [11]. Large molecular networks such as the human protein-protein interaction network [12] [13] or functional linkage network [6] provide functional relations among genes. Based on the observation that genes associated with the same or related diseases tend to interact with each other in the gene network and similar phenotypes tend to share the same disease genes, random walk provides an effective framework to explore the relations in the networks.

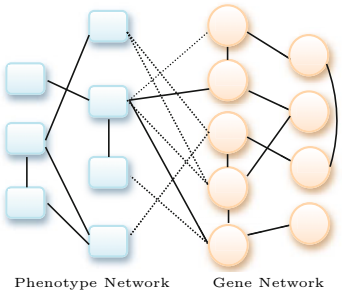

Fig. 1. Predicting missing associations in disease phenotype-gene association network. The solid and dash lines represent known and missing associations, respectively.

Motived by the graph matching problem, we postulate that phenotype-gene associations can be characterized by paired associations between close by genes in the PPI network and close by phenotypes in the phenotype similarity network. Confirmed by the high frequency of such paired associations in OMIM, we propose a bi-random walk algorithm (BiRW) to capture the patterns in the networks to unveil the association between the complete collection of disease phenotypes and genes (phenome-genome association). The key assumption is that the global structure of phenome-genome association can be represented by paired associations, and thus, the reconstruction of the complete phenome-genome association can be achieved by maximizing the number of such paired associations constrained on the known associations. BiRW algorithm iteratively adds new associations into the network by bi-random walk to evaluate the number of recovered paired associations with a decay factor penalizing the number of steps. We investigated variants of BiRW by performing bi-random walk with balanced or unbalanced steps in the the PPI network and the phenotype similarity network, and evaluated the methods by experiments on OMIM data.

2 Methods

The disease phenotype-gene association network (or phenome-genome association network) is a heterogeneous network composed of a phenotype network, a gene network and the known phenotype-gene associations modeled by a bipartite graph (Fig. 1). Let $P_{(m\times m)}$, $G_{(n\times n)}$ and $A_{(m\times n)}$ be the adjacency matrix of the phenotype network, the gene network and the association bipartite graph respectively, where m is the number of phenotypes and n is the number of genes. The objective is to predict the missing associations based on the heterogenous disease phenotype-gene association network by reconstructing an association matrix $R_{(m\times n)}$. The magnitude of each R_{ij} provides the degree of association between phenotype i and gene j. In the following, we first introduce the loss function for the learning problem and then the Bi-Random Walk algorithm (BiRW) that minimizes the cost function for learning R.

2.1 Loss Function

Our assumption is that similar (or the same) phenotypes are more likely to share the same causal gene or causal genes that interact with each other. More specifically, we assume that the predicted paired associations should form the following subgraph patterns: 1) the triangle with two phenotype nodes and one gene node following the assumption "similar phenotypes may share the same causal gene", 2) the triangle with one phenotype node and two gene nodes following the assumption "causal genes of the same disease phenotype tend to interact", and 3) the rectangle with two phenotype nodes and two gene nodes following the assumption "genes associated with similar phenotypes tend to interact". Based on the assumptions, we define the following loss function over R,

$$L(R) = \alpha \sum_{u,v,i,j} (P \otimes G)_{(i,u),(j,v)} (R_{i,u} - R_{j,v})^2 + (1-\alpha) \sum_{i,u} (R_{i,u} - A_{i,u})^2,$$

where $P \otimes G$ is the Kronecker product of P and G. Each $P \otimes G_{(i,u),(j,v)}$ is 1 if $P_{i,j} = 1$ and $G_{u,v} = 1$, in other words phenotype i and j are neighbors and gene u and v are also neighbors, and otherwise 0. In this loss function, the first term enforces a smoothness on R where phenotypes (i, j) and gene (u, v) should form paired associations with phenotype i aligned with gene u and phenotype j aligned with gene v when (i, j) are neighbors and (u, v) are also neighbors. The second term uses prior knowledge A as a regularization term. The trade-off between these two competing constraints is controlled by a positive parameter $\alpha \in (0, 1]$. Intuitively, the cost function in equation (1) evaluates that by associating a phenotype and a gene in R, how many paired associations are curated. The interpretation is closely related to global network alignment algorithms that were applied to align protein-protein interaction networks across species [14][15][16][17][18]. Since the first term is actually a quadratic term of the elements in R with Hessian $D - (P \otimes G)$, the Laplacian of graph $P \otimes G$, the loss function can be rewritten as the following quadratic function,

$$\min_{R} \quad \alpha \overrightarrow{R}^T (D - (P \otimes G)) \overrightarrow{R} + (1-\alpha) || \overrightarrow{R} - \overrightarrow{A} ||^2, \tag{1}$$

where $\overrightarrow{R}$ is the vector concatenated from the rows in R and D is the diagonal matrix with the row sum of $P \otimes G$ as the diagonal entries.

2.2 Bi-Random Walk

To minimize the loss function in equation (1), a straightforward method is to apply random walk with restart on the Kronecker product matrix $P \otimes G$. Since $P \otimes G$ is $(m \times n)$ by $(m \times n)$, this approach does not scale to the large network. We propose a bi-random walk strategy (BiRW), which performs random walk on the phenotype network and the gene network simultaneously. BiRW aims to maximize the number of paired associations by bi-random walk on both phenotype network and gene network to evaluate potential candidate associations

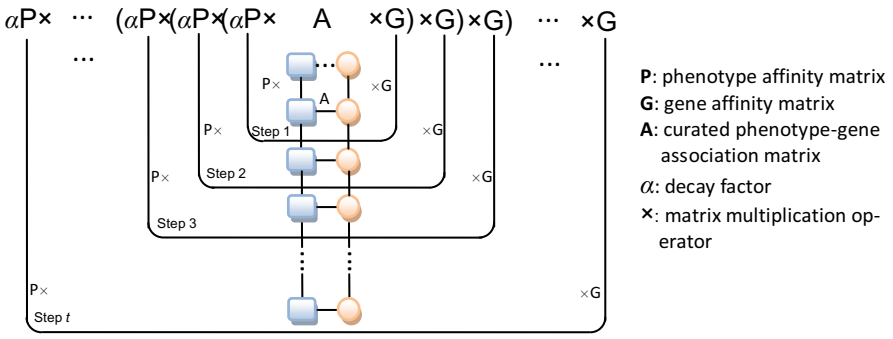

Fig. 2. Illustration of bi-random walk algorithm. P and G are the affinity matrices of the phenotype network and the gene network, respectively. A is the bipartite graph of the known phenotype-gene association from OMIM. By iteratively extending the phenotype path and the gene path (achieved by multiplying P on the left or G on the right in each step), the algorithm maximizes the number of paired associations (loops between phenotypes and genes) with the steps weighted by a decay factor $\alpha \in (0, 1)$. The dashed edge indicates a potential association to add into the network. The iterative algorithm finds the number of new paired associations formed by introducing this additional connection.

(Fig. 2). By iteratively extending the phenotype path and the gene path (achieved by multiplying P on the left and G on the right in each step), the algorithm evaluates each candidate association by the number of closed loops weighted by a decay factor $\alpha \in (0, 1)$. The decay factor down-weights the importance of newly formed loops as the number of random walk steps is getting larger. Here, the matrix multiplications $(P_{(m\times m)} \cdot A_{(m\times n)} \cdot G_{(n\times n)})$ mimic jumps on the phenotype network, the gene network and the association network. In the first step, each element $(P \cdot A \cdot G)_{(i,j)}$ represents the number of paired associations obtained by connecting a target phenotype i to a candidate gene j with phenotype or gene paths length 1. If we ignore the decay factor for now, more generally, after t steps of multiplication $P \ldots (P \cdot (P \cdot A \cdot G) \cdot G) \ldots G = P^t \cdot A \cdot G^t$, the loop patterns curated with up to t steps of random walks can be evaluated. To achieve the best solution $R_{(m\times n)}$, we formulated the problem as $R = P \cdot R \cdot G$, assuming P is column-normalized, G is row-normalized, and the elements in R add to 1. $P \cdot R \cdot G$ can be rewritten in a vector form $P \otimes G\overrightarrow{R}$. Each bi-random walk is the same as a random walk on the Markov matrix $P \otimes G$. Thus, applying bi-random walk is identical to using power method to find the stationary distribution of $P \otimes G$. Note that the idea is also similar to a normalized and relaxed version of regular graph-matching methods [17], which maximize the number of matched edges in two graphs (the phenotype network and the gene network). In addition, the known OMIM associations A normalized the same as R is introduced as priori knowledge. The complete form of the model is as follows,

$$R = \alpha P \cdot R \cdot G + (1 - \alpha)A, \tag{2}$$

The decay factor α also plays the role to balance the objective of closed loops for evaluating candidate associations and the consistence with the known associations in A. This equation can be solved by iteratively updating R by calculating the right side of the equation (2) with the current R. The process also converges to a unique solution [18]. Candidate associations can then be selected by the magnitude of the scores in R. Essentially, this algorithm is mathematically equivalent to the label propagation algorithm in [19], and it was shown that the algorithm minimizes the cost function in equation (1).

2.3 Unbalanced Bi-Random Walk

As illustrated in Fig. 2, the steps to walk on the phenotype network and the gene network explicitly summarize the closed loops in the previous step. Theoretically, the random walk in the two directions will eventually converge to a stationary distribution as the unique solution. However, since only the closed loops of smaller path lengths are informative for predicting associations, excessively counting loops obtained by a large number of random walk steps could introduce false positives. Moreover, the phenotype similarity network and the gene network contain different topologies and structures, and thus, the optimal number of random walk steps might be different on the two networks. To address the problem, we restrict the number of random walk steps on the two sides by introducing two additional parameters l and r as the numbers of maximal iterations in the following left/right random walk on the networks,

$$\begin{aligned} \text{Left Walk:} \quad & R_t = \alpha P \cdot R_{t-1} + (1-\alpha)A \\ \text{Right Walk:} \quad & R_t = \alpha R_{t-1} \cdot G + (1-\alpha)A \end{aligned} \tag{3}$$

Left Walk and Right Walk could be applied alternatively to introduce additional steps in either phenotype network or gene network. The new formula does not converge as equation 2 to a closed-form but it carries the same interpretation that each left or right walk extends either the phenotype path length or the gene path length. Empirically, l, r and α are the parameters tuned by cross-validation on the training data.

2.4 BiRW Algorithms

Given phenotype network P, gene network G, and the phenotype-gene associations A, we first normalize the matrices $\bar{P} = D_P^{-\frac{1}{2}} \cdot P \cdot D_P^{-\frac{1}{2}}$ and $\bar{G} = D_G^{-\frac{1}{2}} \cdot G \cdot D_G^{-\frac{1}{2}}$, where D_P is a diagonal matrix with diagonal elements $D_{Pii} = \sum_j P_{ij}$, and $\bar{G}$ is the same normalized from G. Depending on the arrangement of the left/right walk, we consider three variations of BiRW.

BiRW_bl: This algorithm exactly implements the balanced BiRW given in equation (2), and computes the closed-form solution of equation (1).

BiRW_bl($\bar{P}, \bar{G}, A, \alpha$)
1 $R_0 = \frac{A}{sum(A)}, t = 1$
2 **Do until convege**
3 $R_t = \alpha \bar{P} \cdot R_{t-1} \cdot \bar{G} + (1-\alpha)A$
4 $t = t + 1$
5 **return** (R)

BiRW_avg: This algorithm implements the unbalanced BiRW with the averaged output from the left walk and the right walk in each step.

BiRW_avg($\bar{P}, \bar{G}, A, \alpha, l, r$)
1 $R_0 = \frac{A}{sum(A)}$
2 **for** $t = 1$ **to** $max(l, r)$
3 **if** $t <= l$
4 $R_{t\_left} = \alpha \bar{P} \cdot R_{t-1} + (1-\alpha)A$
5 **if** $t <= r$
6 $R_{t\_right} = \alpha R_{t-1} \cdot \bar{G} + (1-\alpha)A$
7 $R_t = (\delta_{t \le r} \cdot R_{t\_left} + \delta_{t \le l} \cdot R_{t\_right})/(\delta_{t \le l} + \delta_{t \le r})$
8 **return** (R)

In the algorithm, $\delta_{t \le x}$ is 1 if $t \le x$ and 0 otherwise.

BiRW_seq: This algorithm implements the unbalanced BiRW with sequential walk with left walk followed by right walk in each step.

BiRW_seq($\bar{P}, \bar{G}, A, \alpha, l, r$)
1 $R_0 = A = \frac{A}{sum(A)}$
2 **for** $t = 1$ **to** $max(l, r)$
3 **if** $t <= l$
4 $R_{t\_left} = \alpha \bar{P} \cdot R_{t-1} + (1-\alpha)A$
5 **if** $t <= r$
6 $R_t = \alpha R_{t\_left} \cdot \bar{G} + (1-\alpha)A$
7 **return** (R)

3 Comparison of Random Walk Algorithms

In this section, we compare BiRW with the other random walk or label propagation algorithms for disease gene prioritization [4][6][7][8][9]. For example, PRINCE performs label propagation on the PPI network to prioritize disease genes [8]. The initial probabilities on the gene nodes are normalized from the causative genes of the nearest neighbors of the query phenotype p chosen by a logistic function. The initial scores are propagated in the stochastic matrix normalized from the PPI network. After convergence, the unique solution of label propagation is used to rank the genes. RWRH [9] runs the same label propagation algorithm on the combined heterogeneous network of all the three networks to rank genes for a query phenotype. MINProp [7] is based on a principled way to integrate three networks in an optimization framework and performs iterative label propagation on each individual subnetwork. These disease gene prioritization

algorithms rank genes based on their predicted association against a particular query phenotype while BiRW is a global approach which identifies the missing associations of all the phenotypes simultaneously. Thus, conceptually, BiRW is a phenome-genome approach while the other algorithms are phenotype-wise approaches, none of which explores the relation between the predicted associations across the phenotypes. To illustrate the difference between BiRW and the other methods, we compared the initialization and the random walk steps of the algorithms in Table 1. The first difference is that these methods learn with the structure of different networks. Random Walk, Diffusion Kernel and PRINCE perform random walk only on the PPI network combined with the direct neighbors inferred from the phenotype network and the known associations. RWRH and MINProp perform random walk on the complete heterogenous phenome-genome association network. BiRW performs random walk on the Kronecker product graph of the phenotype network and the gene network in the balanced case or on the phenotype network and the gene network separately in the unbalanced case.

Table 1. Comparison of random walk algorithms for disease gene prioritization. We denote the target variables for assigning prediction scores on the phenotype nodes and the gene nodes $p_{(m\times 1)}$ and $g_{(n\times 1)}$, respectively. q is the index of the query phenotype. For any matrix X, $\hat{X}$ represents the row normalized stochastic matrix from X. α, β and λ are positive parameters $\in (0,1)$.

Algorithm	Initialization and Random walk step(s)
Random Walk [4][6]	$g^0 = (A_{q*})'$ $g^t = \alpha G g^{t-1} + (1-\alpha) g^0$
Diffusion Kernel [4]	$g^0 = (A_{q*})'$ $g = (e^{-\beta(D_G - G)}) * g^0$
PRINCE [8]	$g^0(i) = logit(max_l(P_{ql} * A_{li}))$ $g^t = \alpha \bar{G} g^{t-1} + (1-\alpha) g^0$
RWRH [9]	$g^0 = 0, \begin{cases} p^0(i) = 0, \forall i \neq q \\ p^0(q) = 1 \end{cases}$ $\begin{pmatrix} p^t \\ g^t \end{pmatrix} = \alpha \begin{pmatrix} (1-\lambda)\hat{P} & \lambda\hat{A} \\ \lambda\hat{A^T} & (1-\lambda)\hat{G} \end{pmatrix} \begin{pmatrix} p^{t-1} \\ g^{t-1} \end{pmatrix} + (1-\alpha) \begin{pmatrix} p^0 \\ g^0 \end{pmatrix}$
MINProp [7]	$g^0 = 0, \begin{cases} p^0(i) = 0, \ \forall i \neq q \\ p^0(q) = 1, \end{cases}$ Repeat to solve two random walk problems until converge 1) $p^t = \beta \bar{P} p^{t-1} + (1-\beta)(\frac{1-2\beta}{1-\beta} p^0 + \frac{\beta}{1-\beta} \bar{A} g)$ 2) $g^t = \alpha \bar{G} g^{t-1} + (1-\alpha)(\frac{1-2\alpha}{1-\alpha} g^0 + \frac{\alpha}{1-\alpha} \bar{A}' p)$
BiRW	$R = 0$ $R^t = \alpha \bar{P} R^{t-1} \bar{G} + (1-\alpha) * \bar{A}$

Another mathematical difference between BiRW and the other algorithms lies in the formulation of using the known associations in A. PRINCE uses the known associations to decide an initial set of genes that are associated with a query phenotype. RWRH and MINProp directly use A as part of the large network for

random walk. BiRW treats R as the target variable and the known association A as a regularization of R, intuitively, because A is only partially known and most of the zero entries of A are "unknown" instead of "no association". Thus, using A as a regularization instead of directly as part of the network for graph structure-based learning is probably a more rigorous modeling because the incompleteness of the bipartite network might mislead the random walk.

4 Experiments and Discussions

BiRW was compared to CIPHER [5], PRINCE [8] and RWRH [9], three of the best performing algorithms for disease gene prioritization, by 100-fold cross-validation and testing of an independent holdout set with OMIM data. We also compared the three variants of BiRW, BiRW_avg (default for BiRW), BiRW_seq and BiRW_bl, with similar experiments.

4.1 Data Preparation

The disease phenotype network is an undirected graph with 5080 vertices representing OMIM disease phenotypes, and edges weighted in $[0, 1]$. The edge weights measure the similarity between two phenotypes by their overlap in the text and the clinical synopsis in OMIM records, calculated by text mining [10]. The disease-gene associations are represented by an undirected bipartite graph with edges connecting phenotype nodes with their causative gene nodes. Two versions (May-2007 Version and May-2010 Version) of OMIM associations were used in the experiments. May-2007 Version contains 1393 associations between 1126 disease phenotypes and 916 genes, and May-2010 Version contains 2469 associations between 1786 disease phenotypes and 1636 genes. Human protein-protein interaction (PPI) network was obtained from HPRD [12]. The PPI network contains 34,364 curated binary interactions between 8919 genes.

4.2 Comparison with Other Methods

Since the disease gene prioritization algorithms rank genes based on their predicted association against a particular query phenotype, to make a reasonable comparison with CIPHER [5], PRINCE [8] and RWRH [9], the three algorithms were applied to predict the disease genes for each phenotype and the predictions are compared with the results of BiRW phenotype-wise. In the experiment, a disease phenotype was used as a query by an algorithm to rank the genes by their association scores against the query phenotype. For PRINCE and BiRW, the phenotype similarity network was transformed by a logistic function [8]. For all the methods, a 100-fold cross-validation on the OMIM May-2007 Version was performed for parameter tuning, and then the methods were applied to predict the associations in an independent set of associations added into OMIM between May-2007 and May-2010.

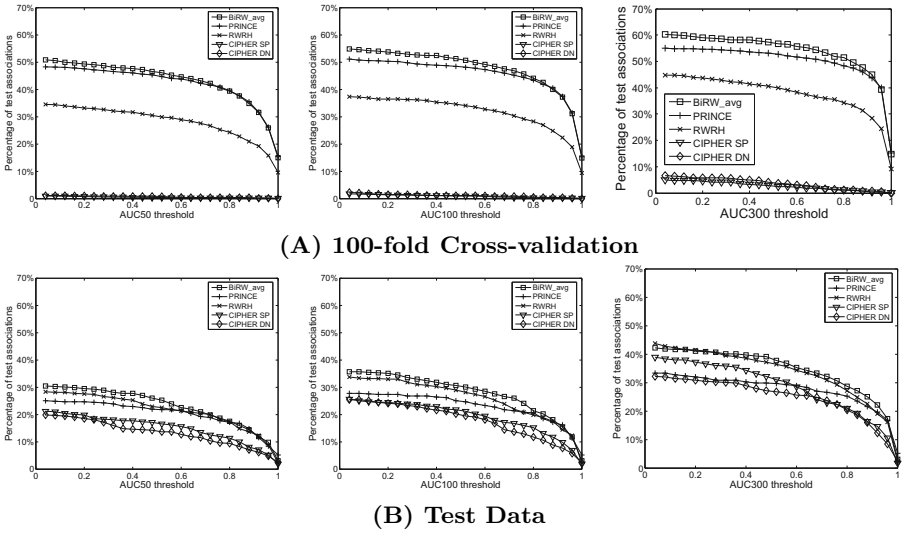

Fig. 3. Performance of predicting OMIM associations. The plots show the percentage of phenotypes, for which a given method achieved a ROC score exceeding a threshold in cross-validation and testing.

There are 1126 disease phenotypes with at least one known causal gene in OMIM version May-2007. In the 100-fold cross-validation, the 1126 disease phenotypes were randomly divided into 100 subsets. In each cross-validation trial, the OMIM associations of the 1% disease phenotypes in a subset were removed, and then used as queries to rank the candidate genes. The hyper-parameters α for both PRINCE and BiRW were chosen from $\{0.1, 0.2, ..., 0.9\}$, and l and r were taken to be between 1 step to 5 steps. The three hyper-parameters of RWRH are set to be the optimal parameters $(0.5, 0.7, 0.5)$ suggested by the experiments in [9]. The test set contains new associations of 518 phenotypes in OMIM May-2010 Version. ROC score (Area Under the Curve of Receiver Operating Characteristic) was used as the global performance measure. The higher the target genes of a query phenotype in the ranking, the better the performance. Specifically, for each phenotype query, the target genes were labeled as positives and the other genes were labeled as negatives. AUCs were computed by the positions of the positives in the ranking list. We reported the AUC with up to 50, 100 and 300 false positives since the top part of AUC is more important.

The results produced by the best parameters in the cross-validation of each method is reported in Fig. 3A ($l = 4$, $r = 4$ and $\alpha = 0.8$ for BiRW and $\alpha = 0.1$ for PRINCE). To make a comprehensive comparison, we plot the number of phenotype queries with a AUC higher than a certain threshold in the plots. The BiRW algorithm performed the best. Out of the 1126 phenotypes, BiRW ranked around 55% in top 50 and 63% in top 500. PRINCE also gave decent prediction performance although BiRW consistently outperformed PRINCE in all the measures. RWRH, CIPHER DN (direct neighbor) and SP (shortest path)

Table 2. Statistical significance in performance comparison. A pairwise comparison by paired t-test of the ranking results in 100-fold cross-validation.

(A) p-values for AUC_{50} comparison

	BiRW(0.8,4,4)	PRINCE(0.1)	RWRH(0.5,0.7,0.5)	C-SP	C-DN
BiRW	NaN				
PRINCE	0.046	NaN			
RWRH	4.41e-037	6.08e-030	NaN		
CIPHER SP	6.97e-158	1.87e-150	1.20e-091	NaN	
CIPHER DN	9.81e-158	7.99e-150	6.87e-090	0.836	NaN

(B) p-values for AUC_{100} comparison

	BiRW(0.8,4,4)	PRINCE(0.1)	RWRH(0.5,0.7,0.5)	C-SP	C-DN
BiRW	NaN				
PRINCE	4.73e-004	NaN			
RWRH	7.30e-039	4.27e-027	NaN		
CIPHER SP	1.65e-175	2.09e-160	1.09e-100	NaN	
CIPHER DN	1.65e-176	1.94e-161	2.71e-099	0.79	NaN

produced inferior results in this experiment. The possible reason for the worse results of CIPHER might be because the associations of the test phenotypes were all removed (called *ab initio* experiment) and each cross-validation held out a significant number of known associations. Thus, no direct neighbors were available for the correlation calculation for many phenotype queries by CHPHER. PRINCE, RWRH and BiRW worked much better than CIPHER SP and CIPHER DN because label propagation and bi-random walk both explore more global information of the networks. We also measured the statistical significance of the difference in AUC_{50} and AUC_{100} by paired t-test. The p-values are reported in Table 2. Clearly, BiRW performs significantly better than all other methods at the significance level 0.05.

4.3 Comparison of BiRW Variants

To understand the effect of combining left walk and right walk with different strategies, we compared BiRW_avg, BiRW_seq and BiRW_bl with the same experiments on OMIM data. The results are reported in Table 3. BiRW_avg and BiRW_seq, which perform random walk with a limited number steps, performed significantly better than BiRW_bl, which performs random walk till the convergence to the stationary distribution. The observation partially agrees with the results by [20] [21], which showed that genes within two-steps are more functional cohesion in the PPI network. When the random-walk steps are above 2 in the gene network, results are very close to optimal as long as the number of steps in the phenotype network is properly chosen. Since the results depends on the random-walks in two networks and the decay factor, we found that it is better to treat the steps as parameters as in BiRW_avg and BiRW_seq. It is also interesting that BiRW_avg performed constantly better than BiRW_seq although the difference is only marginal. We suspect that there might be a bias

Table 3. Comparison of the three BiRW Variants. The table reports a comparison of the ranking results by the BiRW variants, BiRW_avg, BiRW_seq and BiRW_bl. The parameters α, m and n of BiRW are chosen by the 100-fold cross-validation. AUCs up to 50, 100, 300, 500, 1000 and all false positives are reported.

(A) 100-fold Cross-validation

	AUC_{50}	AUC_{100}	AUC_{300}	AUC_{500}	AUC_{1000}	AUC
BiRW_avg(0.8,4,4)	0.4349	0.4818	0.5455	0.5721	0.6097	0.8063
BiRW_seq(0.8,4,3)	0.4295	0.4696	0.5323	0.5596	0.5972	0.8019
BiRW_bl(0.8)	0.2730	0.3344	0.4229	0.4608	0.5138	0.7768

(B) Test Data

	AUC_{50}	AUC_{100}	AUC_{300}	AUC_{500}	AUC_{1000}	AUC
BiRW_avg(0.8,4,4)	0.2321	0.2809	0.3498	0.3862	0.4494	0.7708
BiRW_seq(0.8,4,3)	0.2235	0.2651	0.3344	0.3700	0.4344	0.7672
BiRW_bl(0.8)	0.1675	0.2198	0.3167	0.3689	0.4461	0.7754

in choosing the order of left walk and right walk when BiRW_seq performs sequential random walks, and the bias might be data dependent. In BiRW_avg, there is no ambiguity in the order of the bi-random walk and thus, there might be less variation expected in different data.

5 Conclusion

In the paper, we introduced a bi-random walk algorithm (BiRW) for disease gene prioritization. We analyzed the algorithm by comparison with other random walk algorithms for disease gene prioritization with both algorithmic analysis and empirical experiments. We concluded that BiRW is an effective algorithm for disease gene prioritization and the steps of random walks play a crucial role in the performance of the algorithms. In future, we plan to explore other variations of BiRW to more effectively utilize the hidden information in the networks.

Acknowledgement. This work is supported in part by grant III 1117153 from National Science Foundation.

References

1. Consortium The Wellcome Trust Case Control. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678 (2007)
2. Johnson, A., O'Donnell, C.: An open access database of genome-wide association resutls. BMC Med. Gent. 10, 6 (2009)
3. Franke, L., Bakel, H., Fokkens, L., et al.: Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am. J. Hum. Genet. 78, 1011–1025 (2006)

4. Köhler, S., Bauer, S., Horn, D., et al.: Walking the Interactome for Prioritization of Candidate Disease Genes. Am. J. Hum. Genet. 82, 949–958 (2008)
5. Wu, X.B., Jiang, R., Zhang, M.Q., et al.: Network-based global inference of human disease genes. Mol. Syst. Biol. 4 (2008)
6. Linghu, B., Snitkin, E.S., Hu, Z., et al.: Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol. 10, R91 (2009)
7. Hwang, T.H., Kuang, R.: A Heterogeneous Label Propagation Algorithm for Disease Gene Discovery. In: Proc. of SIAM Intl. Conf. on Data Mining, pp. 583–594 (2010)
8. Vanunu, O., Magger, O., Ruppin, E., et al.: Associating genes and protein complexes with disease via network propagation. PLoS Comput. Biol. 6, e1000641 (2010)
9. Li, Y., Patra, J.C.: Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics 26, 1219–1224 (2010)
10. van Driel, M.A., Bruggeman, J., Vriend, G., et al.: A text-mining analysis of the human phenome. Eur. J. Hum. Genet. 14, 535–542 (2006)
11. McKusick, V.A.: Mendelian inheritance in man and its online version, OMIM. Am. J. Hum. Genet. 80, 588–604 (2007)
12. Peri, S., Navarro, J.D., Amanchy, R., et al.: Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 13, 2363–2371 (2003)
13. Chuang, H., Lee, E., Liu, Y., et al.: Network-based classification of breast cancer metastasis. Mol. Syst. Biol. 3, 140 (2007)
14. Singh, R., Xu, J., Berger, B.: Pairwise Global Alignment of Protein Interaction Networks by Matching Neighborhood Topology. Res. in Comp. Mol. Biol. 4453, 16–31 (2007)
15. Li, Z., Zhang, S., Wang, Y., et al.: Alignment of molecular networks by integer quadratic programming. Bioinformatics 23, 1631–1639 (2007)
16. Guo, X., Hartemink, A.J.: Domain-oriented edge-based alignment of protein interaction networks. Bioinformatics 25, i240–i246 (2009)
17. Zaslavskiy, M., Bach, F., Vert, J.P.: Global alignment of protein-protein interaction networks by graph matching methods. Bioinformatics 25, i259–i267 (2009)
18. Singh, R., Xu, J., Berger, B.: Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc. Natl. Acad. Sci. U.S.A. 105, 12763–12768 (2008)
19. Zhou, D., et al.: Learning with Local and Global Consistency. Advanced Neural Information Processing Systems 16, 321–328 (2004)
20. Chua, H., Sung, W., Wong, L.: Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics 22, 1623–1630 (2006)
21. Xu, J., Li, Y.: Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics 22, 2800–2805 (2006)