Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 13-023

Reconstructing Disease Phenome-genome Association by Bi-Random Walk

MaoQiang Xie, TaeHyun Hwang, and Rui Kuang

August 16, 2013

Reconstructing Disease Phenome-genome Association by Bi-Random Walk

MaoQiang Xie[1], TaeHyun Hwang[2], Rui Kuang3[3,*]

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

Abstract

Promising results were recently reported in utilizing network information in phenotype-similarity network and gene-interaction network with graph-based learning to derive new disease phenotype-gene associations. However, a more fundamental understanding of how the network information is relevant to disease phenotype-gene associations is lacking. In this paper, we analyze the circular bigraphs (CBGs) in OMIM phenotype-gene association networks, and introduce a bi-random walk (BiRW) algorithm to capture the CBG patterns in the networks for unveiling the associations between the complete collection of disease phenotypes (phenome) and genes. BiRW performs separate random walk simultaneously on gene interaction network and phenotype similarity network to explore gene paths and phenotype paths in CBGs of different sizes.

In the analysis of OMIM associations, we discovered that 81% of the associations are covered by CBG patterns of path-length up to 3 with variability by 21 disease classes, and there is a clear correlation between the CBG coverage and the predictability of the phenotype-gene associations. Some prominent examples are cancers, nutritional diseases, dermatological diseases, bone diseases, cardiovascular diseases and respiratory diseases. Experiments on recovering known associations in cross-validation and predicting new associations in a test set validated that BiRW effectively improved prediction performance over existing methods by ranking more known associations in the top 100 out of more than 12,000 candidate genes. The investigation of the global disease phenome-genome association map also revealed interesting new predictions and phenotype-gene modules by disease classes.

Availability: http://compbio.cs.umn.edu/BiRW

1 Introduction

Since Gregor Mendel discovered that phenotypes are inheritable from ancestors in nineteenth century, it is now well accepted that phenotypes are determined by genetic material under environmental influences. In this post-genomic era, numerous genomic studies on large patient cohorts such as genome-wide association studies [1, 2] have been conducted to determine the molecular mechanisms of genetic diseases. The objective of such studies is to perform a high-throughput scanning for a list of genes that are involved with the disease under study. In parallel to this high-throughput scanning approach, another strategy is to predict a list of candidate genes based on the knowledge of already determined disease phenotype-gene associations such as those in Online Mendelian Inheritance in Man (OMIM), a database of human genes and genetic disorders. This knowledge-based strategy takes the advantage of the availability of large phenotypic and molecular networks. The human disease phenotype network [3] provides information on phenotype similarities computed by text mining of the full text and clinical synopsis of the disease phenotypes in OMIM [4]. Large molecular networks such as the human protein-protein interaction network [5,6] or functional linkage network [7] provide functional relations among proteins (genes). The problem of predicting phenotype-gene associations in a real phenotype-gene association subnetwork is illustrated in

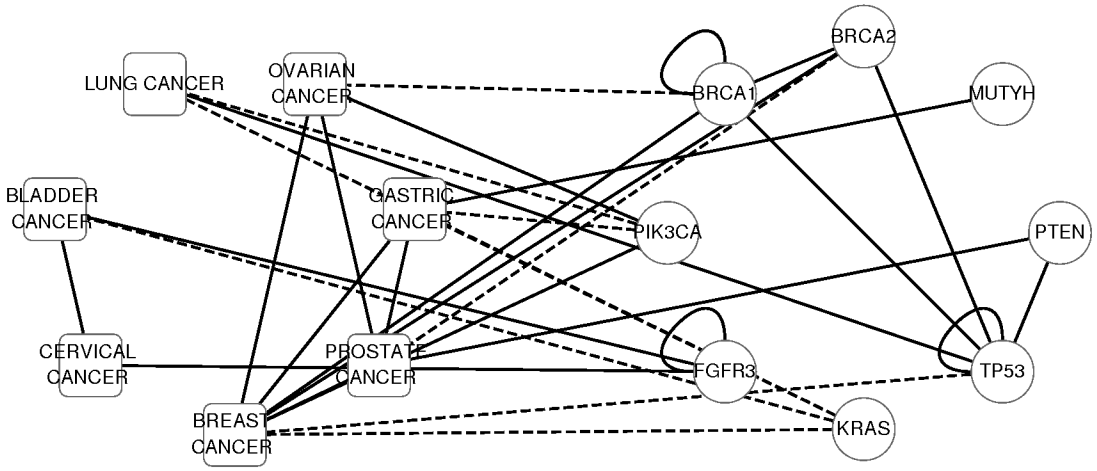


Figure 1. Predicting missing associations in disease phenotype-gene association network. This real subnetwork includes seven cancer phenotypes and eight disease genes. Similar phenotypes and interacting genes are connected. The solid lines represent the OMIM association known before May, 2007. The dash lines represent the new OMIM associations added after May, 2007.

Fig. 1. 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, many network-based approaches have been proposed to utilize the disease modules and gene modules in the networks to prioritize disease genes for a disease phenotype [7–15], to find related disease phenotypes for a gene set [16] or to detect disease-gene modules [17]. Despite of the impressive results in the studies, few attempts have been made to explain the network-based prediction approaches by graph patterns.

We postulate that the relation among phenotype-gene associations can be characterized by circular bigraph patterns (CBGs). Based on the observation of high frequency of CBGs in OMIM associations, we propose a bi-random walk algorithm (BiRW) to capture the CBG patterns in the networks for unveiling 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 many overlaying circular bigraphs, i.e. each phenotype-gene association is likely to be paired with some other phenotype-gene association(s) with their phenotypes and genes closely related in the phenotype network and the gene network, respectively. The assumption is supported by the phenotype-gene modules reported in [17] as well as the observation of frequent CBGs. Thus, the reconstruction of the complete phenome-genome association can be achieved by maximizing the number of circular bigraphs balanced with the known associations. BiRW iteratively adds new associations into the network by bi-random walk to evaluate the number of recovered circular bigraphs with a decay factor penalizing longer paths in the CBG patterns. Note that different from the algorithms for disease gene prioritization [7, 10–13], which rank genes for a particular query phenotype, BiRW is a global approach to reconstruct the missing associations for all the phenotypes simultaneously.

2 Methods

The disease phenotype-gene association network is a heterogeneous network composed of a phenotype network, a gene network and the phenotype-gene associations modeled by a bipartite graph (Fig. 1). Let

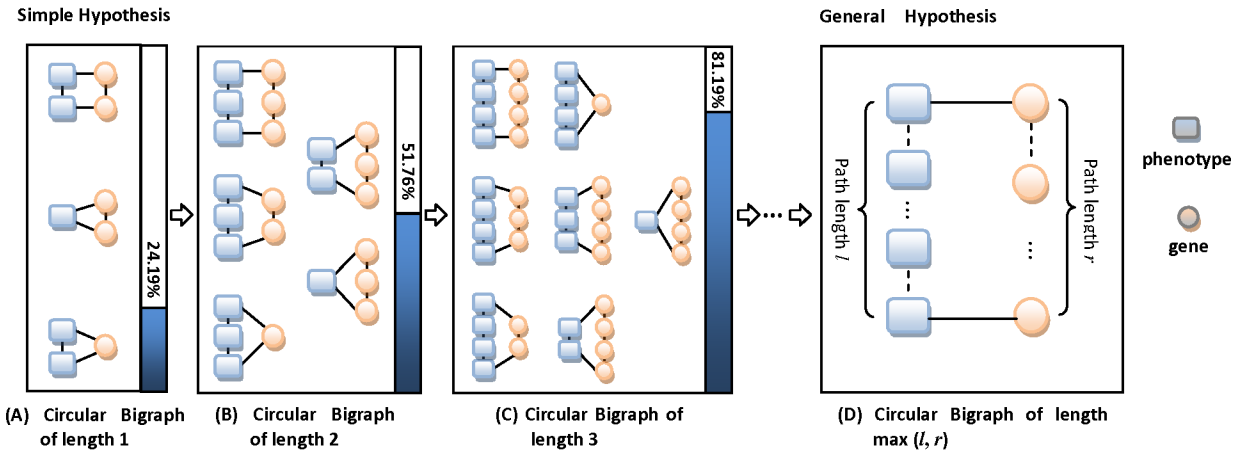


Figure 2. Circular Bigraphs (CBG) in the disease phenome-genome association network. A circular bigraph is composed of a phenotype path and a gene path with two ends connected by phenotype-gene associations. (A), (B) and (C) CBGs with phenotype or gene paths of length 1, 2 and 3. CBGs of length 1 represent the simple hypothesis. (D) The general hypothesis that CBG with phenotype or gene paths of arbitrary length. The gray bars on the right of the figures report the percentage of OMIM disease phenotype-gene associations that are covered by CBGs up to a certain path length.

$P_{(m\times m)}$, $G_{(n\times n)}$ and $A_{(m\times n)}$ be the affinity 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 heterogeneous 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 concept of circular bigraphs (CBG) and then the bi-random walk algorithm (BiRW) for learning $R_{(m\times n)}$.

2.1 Circular Bigraphs

A CBG is defined as a subgraph composed of a phenotype path $\{p_1, p_2, ..., p_m\}$ and a gene path $\{g_1, g_2, ..., g_n\}$ with two ends connected by two phenotype-gene associations (p_1, g_1) and (p_m, g_n) (Fig. 2). A CBG indicates a vicinity relation between two associations (p_1, g_1) and (p_m, g_n) by generalizing the relations between p_1 and p_m by their distance in the phenotype network and the relation between g_1 and g_n in the gene network. The smallest CBGs directly represent the simple hypothesis by the existing methods as illustrated in Fig. 2A. The triangle with two phenotype nodes and one gene node follows the assumption "similar phenotypes may share the same causal gene". The triangle with one phenotype node and two gene nodes follows the assumption "causal genes of the same disease phenotype tend to interact". The rectangle with two phenotype nodes and two gene nodes follows the assumption "genes associated with similar phenotypes tend to interact". In Fig. 2B-D, we generalize the hypothesis to CBGs of length (l, r) by exploring the affinity relations captured by the phenotype path and the gene path of longer lengths. We hypothesize that in the unknown complete disease phenome-genome association network, most of the associations tend to be captured by many very small circular bigraphs. Simple calculations of the CBGs in the OMIM phenome-genome association network confirm that, among 1393 OMIM disease phenotype-gene associations, only 24.19% associations are covered by simple hypothesis (Fig. 2A), while more than 81% of the known OMIM associations are covered by at least one CBG of path length up to 3

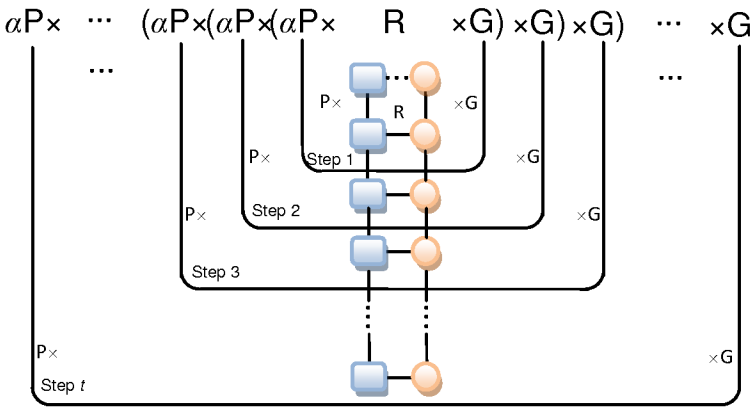


Figure 3. Illustration of bi-random walk. P and G are the affinity matrices of the phenotype network and the gene network, respectively. A is the bipartite graph of the known phenotype-gene association from OMIM. By iteratively extending the phenotype path and the gene path (achieved by multiplying P on the left or G on the right in each step), the algorithm explores the CBGs 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 CBGs formed by introducing this additional connection. In other words, a potential association is evaluated by its distance to the known associations in the phenotype network and the gene network.

(Fig. 2C). CBGs of small path lengths are prevalent in the human disease phenome-genome association network. Some real CBG examples of different path lengths are given in Fig. S1.

2.2 Balanced Bi-Random Walk

Based on the observation of high CBG frequencies in OMIM associations, BiRW is specifically designed to capture the CBG patterns in the networks for unveiling phenome-genome association. BiRW iteratively performs random walk in the phenotype network and the gene network. Steps of random walk on the phenotype (gene) network essentially explores the closeness of the phenotypes (genes) in the network by counting the number of paths of a certain length going from one phenotype (gene) to another. With the results from the previous step of bi-random walk, each bi-random walk evaluates the number of circular bigraphs recovered by connecting phenotypes and genes with a decay factor introduced to penalize longer paths. BiRW aims to explore the CBGs by bi-random walk on both phenotype network and gene network to evaluate potential candidate associations (Fig. 3). 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 explores the CBGs weighted by a decay factor $\alpha \in (0, 1)$. The decay factor down-weights the importance of a CBG as the path length is getting longer. 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 $PAG_{(i,j)}$ represents the number of CBGs 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 $P \dots (P(P \cdot A \cdot G)G) \dots G = P^t A G^t$, CBG patterns with path length t can be evaluated. To achieve the best solution $R_{(m\times n)}$, we formulated the problem as $R = PRG$, assuming P is column-normalized, G is row-normalized, and the elements in R add to 1. PRG can be rewritten in a vector form $P \otimes G\hat{R}$ where $\otimes$ is the Kronecker product and $\hat{R}$ is the vector obtained by concatenating the columns in R. Each step of bi-random walk is the same as a random walk on the Markov matrix $P \otimes G$. Thus, the step of evaluating the CBGs 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 [18], 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 to balance the BiRW and the priori knowledge. The complete form of the model is as follows,

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

The decay factor α also plays the role to balance the objective of capturing CBGs 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 1 with the current R. The process also converges to a unique solution [19]. Candidate associations can then be selected by the magnitude of the scores in R.

2.3 Unbalanced Bi-Random Walk

As illustrated in Fig. 3, the steps to walk on the phenotype network and the gene network explicitly summarize the CBGs 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 CBGs of small path lengths are informative for predicting associations, excessively counting CBGs of long path lengths could introduce false positives. For example, [20] suggested that genes within two-steps in a PPI network are more functional cohesion. 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 parameters (l, r) as the numbers of maximal iterations in the following left/right random walk on the networks,

$$\begin{aligned} \text{Left Walk:} \quad & R_t = \alpha P \cdot R_{t-1} + (1 - \alpha)A \\ \text{Right Walk:} \quad & R_t = \alpha R_{t-1} \cdot G + (1 - \alpha)A \end{aligned} \tag{2}$$

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 (1) to a closed-form but they carry the CBG 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.

2.4 BiRW Algorithm

The BiRW algorithm takes P, G, A, the decay factor α, left/right walk steps l, r as the inputs and outputs the predicted associations R. The BiRW algorithm is outlined as follows,

BiRW(P, G, A, α, l, r)

1 $\bar{P} = D_P^{-\frac{1}{2}} P D_P^{-\frac{1}{2}}$

2 $\bar{G} = D_G^{-\frac{1}{2}} G D_G^{-\frac{1}{2}}$

3 $R_0 = A = \frac{A}{sum(A)}$

4 **for** $t = 1$ **to** $max(l, r)$

5 **if** $t <= l$

6 $R_{t\_left} = \alpha \bar{P} \cdot R_{t-1} + (1-\alpha)A$

7 **if** $t <= r$

8 $R_{t\_right} = \alpha R_{t-1} \cdot \bar{G} + (1-\alpha)A$

9 $R_t = (\delta_{t \le l} \cdot R_{t\_left} + \delta_{t \le r} \cdot R_{t\_right})/(\delta_{t \le l} + \delta_{t \le r})$

10 **return** (R)

Note that P is normalized as $\bar{P} = D_P^{-\frac{1}{2}} P D_P^{-\frac{1}{2}}$ 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. A is normalized with elements adding up to 1. Line 6 and line 8 are the left and right random walks. In Line 9, the propagation results are combined, where $\delta_{t \le x}$ is 1 if $t \le x$ and 0 otherwise. The algorithm will terminate as the number of iterations exceeds max(l, r).

2.5 Interpretation of BiRW

BiRW explores CBG patterns on the phenotype network and the gene network. The algorithm assumes that the probability of a phenotype being associated with a gene is proportional to the the number of weighted CBGs that are explored by bi-random walk and connecting the phenotype and the gene. An equivalent statement is that a potential association between a phenotype and a gene is evaluated by its distance to the other associations in the phenotype network and the gene network. Actually, the closer the potential association to the other associations, the more highly weighted CBGs will be created by this association. Thus, BiRW is a global strategy to complete the association map, instead of prioritizing genes for a specific disease phenotype. Accordingly, it provides a global optimality in the predicted associations.

A more mathematical interpretation of BiRW is that, by connecting a phenotype and a gene, how many weighted new CBG patterns are curated. This interpretation is closely related to global network alignment algorithms [18, 19, 21–23], applied to aligning protein-protein interaction networks. In our context, the objective is to maximize the number of CBGs (conserved interactions) between two networks. Given phenotype network P and gene network G, $P \otimes G$ 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. BiRW learns an association matrix R to minimize the following regularization framework,

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

In this regularization framework, the first term enforces a smoothness on R where phenotypes (i, j) and gene (u, v) should form a CBG 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 regularization. The matrix form of the above objective function is

$$\min_{R} \quad \alpha \hat{R}^T (D - (P \otimes G))\hat{R} + (1-\alpha)||\hat{R} - \hat{A}||^2, \tag{3}$$

where D is the diagonal matrix with the row sum of $P \otimes G$ as the diagonal entries. This objective function is minimized by the algorithm in equation (1) [24]. IsoRank algorithm adopts the same idea for computing a global network alignment between PPI networks [19,21]. Despite the mathematical similarity with IsoRank, BiRW algorithm explicitly interprets iterative random walks on the two networks as steps to extend the CBG paths, and thus, introduces the flexibility to perform unbalanced left/right walk to capture CBG patterns more precisely. Actually, in our experiments, BiRW with left/right walks performed much better than direct application of IsoRank (see results in Table S1A).

2.6 Related Work

BiRW is most related to the network-based algorithms for disease gene prioritization [7–15]. CIPHER scores the association between a gene g and the phenotype p by computing the correlation coefficient between the gene-phenotype profile of g and the phenotype similarity profile of p [10]. The gene-phenotype profile of g is computed by a logistic function based on the direct neighbors of g or the path length between g and causal genes of each phenotype. PRINCE performs label propagation on the PPI network to prioritize disease genes [12]. 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 [13] runs the same label propagation algorithm on the combined heterogeneous network of all the three networks to rank genes for a query phenotype. MINProp [11] is based on a principled way to integrate three networks in an optimization framework and performs iterative label propagation on each individual subnetwork. MAXIF [15] maximizes the information flow in the phenome-genome association network to identify the sink genes for a source phenotype.

These disease gene prioritization algorithms rank genes based on their predicted association against a particular query phenotype p 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. The mathematical difference between BiRW and the other algorithms lies in the formulation of using the known associations in A. CIPHER and PRINCE use the known associations to decide an initial set of genes that are associated with a query phenotype. RWRH, MINProp and MAXIF directly use A as part of the large network for random walk or maximum flow computation. BiRW treats the associations 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.

3 Results

In the experiments, we first analyzed OMIM disease gene associations by reporting the CBG patterns in the network. We then performed cross-validation and test of an independent test set of OMIM associations to evaluate the performance of BiRW. Finally, we analyzed the predicted disease phenome-genome association by examining association modules by each disease class.

Table 1. CBGs in OMIM phenotype-gene associations. OMIM associations in 21 disease classes that are covered by CBGs up to a certain length (1-3). The coverage by percentage and the AUC in cross-validation in Sect. 3.3 are reported.

Disease Classes	CBG Len=1	CBG Len=2	CBG Len=3	Assoc #	Coverage	Pheno #	AUC
Bone	14	25	30	35	85.7%	28	0.922
Cancer	21	67	87	89	97.7%	55	0.933
Cardiovascular	15	32	53	60	88.3%	50	0.899
Connective Tissue	7	16	22	27	81.5%	20	0.798
Dermatological	34	59	77	79	97.5%	65	0.901
Developmental	9	18	23	27	85.2%	24	0.853
Ear, Nose, Throat	4	5	12	17	70.6%	15	0.771
Endocrine	9	26	46	53	86.8%	40	0.845
Gastrointestinal	6	13	23	28	82.1%	15	0.785
Hematological	18	36	57	62	91.9%	53	0.859
Immunological	15	30	45	49	91.8%	38	0.769
Metabolic	17	34	76	128	59.4%	118	0.495
Muscular	19	35	52	61	85.3%	48	0.826
Neurological	42	93	142	163	87.1%	135	0.832
Nutritional	1	1	6	6	100%	2	0.781
Ophthamological	15	30	51	77	66.2%	62	0.806
Psychiatric	0	2	8	11	72.7%	10	0.516
Renal	12	19	30	39	76.9%	34	0.811
Respiratory	3	6	10	11	90.9%	7	0.889
Skeletal	8	25	43	46	93.5%	45	0.854
Multiple	37	78	123	140	87.9%	103	0.906
Total	306	650	1016	1208	84.1%	967	0.826

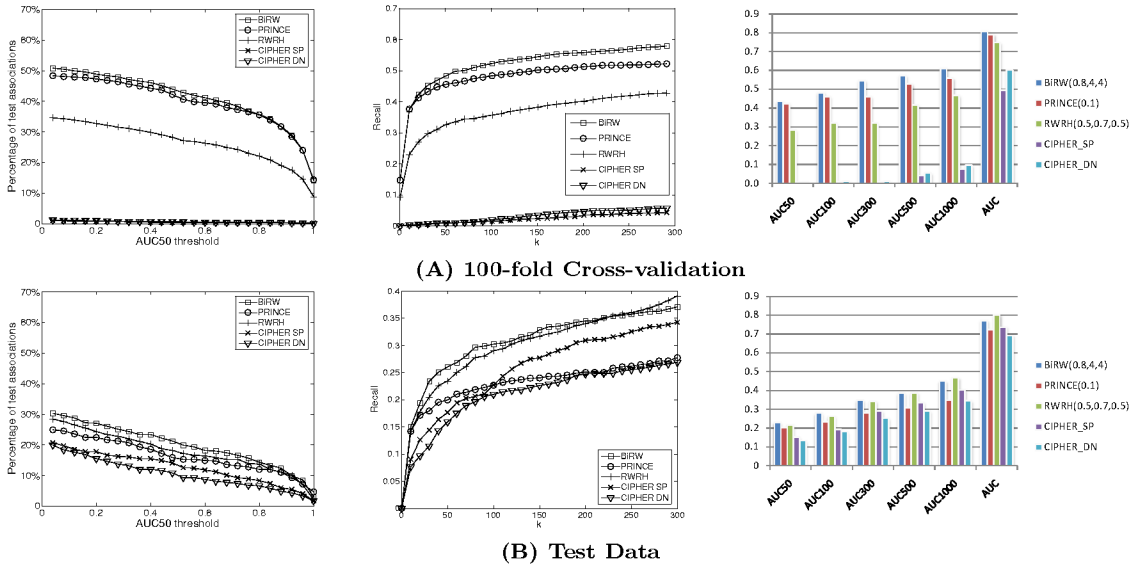
3.1 Data Preparation

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 [3]. 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 [5]. The PPI network contains 34,364 curated binary interactions between 8919 genes.

3.2 Analysis of CBGs in OMIM Associations

To analyze the CBGs in the OMIM disease phenome-genome association network, we calculated the frequencies of the CBG patterns in OMIM (May-2007) phenotype-gene association network. In the phenotype similarity network, the five nearest nodes of each node were selected as neighbors [13]. We first report the percentage of the 1393 OMIM associations covered in CBGs categorized by the path length in Fig. 2. Only 24.19% of the associations are covered by CBG patterns of path length 1 (Fig. 2A). When larger CBGs are considered, the percentage is significantly higher. Specifically, 51.76% of the associations are covered by CBG patterns of path length up to 2 and 81.19% of the associations are covered by CBG of path length up to 3. When we randomly created 1393 links between 5080 disease phenotypes and 8919 genes, there is no occurrence of CBG of path length 1 in 1000 runs. The significance is expected because the diameter of the phenotype network is 33, and there are many components in the PPI network with the largest diameter 14.

To further analyze the relation between CBGs and disease phenotypes, we calculated the CBG statistics for the 21 disease phenotype classes that are manually curated by [25] The statistics are reported in Table 1. Some of the classes such as nutritional diseases, cancers, dermatological diseases, skeletal diseases, hematological diseases and immunological diseases have a high CBG coverage on their OMIM



(A) 100-fold Cross-validation

(B) Test Data

Figure 4. Performance of predicting OMIM associations. The plots compare the performance of all the methods on predicting the target genes of query phenotypes in cross-validation (A) and testing on the test set (B). The plots on the left show the total number of phenotypes, for which a given method achieved a AUC_{50} exceeding a threshold. The plots in the middle show the recalls at different top-k cutoff. The plots on the right compares the average AUCs across all the test phenotypes by the compared methods.

associations while some other classes such as metabolic diseases, ophthamological diseases, ear-nose-throat diseases and psychiatric diseases have lower coverage. For example, above 97% associations in cancers, nutritional diseases and dermatological diseases, are covered by CBGs with path length up to 3, while only 66% of those in metabolic disease are covered. A more detailed break-down of the CBGs in each disease class by lengths is also given in a table in Fig. S2.

3.3 Comparison with Other Methods by Phenotype-gene Association Prediction

BiRW was compared to CIPHER [10], PRINCE [12] and RWRH [13], three of the best performing algorithms for disease gene prioritization. Since these algorithms rank genes based on their predicted association against a particular query phenotype, to make a reasonable comparison, 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 [12]. 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.

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.

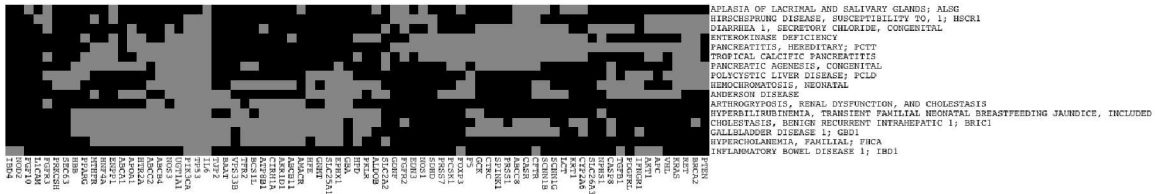


Figure 5. Gastrointestinal disease phenotype-gene association module. The predicted gastrointestinal disease module of associations (grey color) between 64 disease genes including the 22 known disease genes associated with 16 gastrointestinal disease phenotypes.

The three hyper-parameters of RWRH are set to be the optimal parameters $(0.5, 0.7, 0.5)$ suggested by the experiments in [13]. 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. The area under the curve of ROC (AUC) were computed by the positions of the positives in the ranking list. We reported the AUC with up to 50, 100, 300, 500 and 1000 false positives since the top part of AUC is more important. In addition, we also report the recall of the test phenotypes, which measures how many target genes are ranked within top k.

The results produced by the best parameters in the cross-validation of each method is reported in Fig. 4A and Fig. S3A ($l = 4$, $r = 4$ and $\alpha = 0.8$ for BiRW and $\alpha = 0.1$ for PRINCE). The detailed parameter tuning for BiRW, PRINCE and CIPHER are given in Table S1-S3. To make a comprehensive comparison, we plot the number of phenotype queries with a AUC higher than a certain threshold in the left plots of Fig 4. The BiRW algorithm performed the best. Out of the 1126 phenotypes, BiRW ranked around 55% in top 50 and 63% in top 500 (Fig. S3A). PRINCE also gave decent prediction performance although BiRW consistently outperformed PRINCE in all the measures. RWRH, CIPHER DN (direct neighbor) and SP (shortest path) 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 are 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. In addition, the recall are reported in the middle plots in Fig. 4 and the complete results measured by AUCs up to different false positives are reported in the right plots in Fig. 4 and Table S4. 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 S6. Clearly, BiRW performs significantly better than all other methods at the significance level 0.05. In the last column of Table 1, the average AUC of the phenotypes in each disease class by BiRW are also reported. As expected, there is a strong correlation between the CBG coverage and the AUCs. For example, the cancer class has both high coverage and AUC, while psychiatric and metabolic classes have both low coverage and AUC.

With the best parameter from cross-validation, the methods were used to predict the new associations of 518 phenotypes in OMIM May-2010 Version. The results are reported in Fig. 4B and Fig. S3B. BiRW consistently outperformed the other methods although RWRH performed better on the holdout set than the cross-validation. Interestingly, PRINCE performed worse in the test than the cross-validation, which might suggest a possible bias when the neighbors' disease genes are directly initialized as true disease genes. CIPHER DN and SP performed better on the test set since the test cases have other known disease genes and thus, are not as difficult as those in cross-validation. A more detailed comparison of BiRW, RWRH, PRINCE and CIPHER by AUCs is given in Table S5. The p-values by paired t-test are reported in Table S7. Clearly, BiRW and RWRH performed significantly better than all other methods

while BiRW performs better at the the starting part of the AUCs.

3.4 Robustness Analysis with Unbiased PPI Network

It is known that well-studied disease proteins tend to have more interactions in the PPI network and this degree bias could potentially lead to the good performance of the network-based methods. To test whether the methods are robust to the degree bias, we repeated the experiments on an extended PPI network. The extended PPI network with the same degree of interactions for each protein was generated to assess the influence of the degree bias. The extended PPI network was combined from HPRD, OPHID, BIND, and MINT database contain 72,431 undirected binary interactions between 14,433 human proteins [10]. The results are reported in Fig. S4 and Table S8. The BiRW algorithm consistently outperformed the baselines. With the replacement by the unbiased PPI network, all the methods performed actually similarly as in the original experiment. This experiment on the extended PPI network suggests that BiRW is also robust to the degree bias in the networks.

3.5 Analysis of Gastrointestinal Disease-gene Modules

Identification of the essential biological processes of diseases can lead to new insights into possible pathogenic mechanisms, and development of efficient targeted therapeutics. However, the current known associations between human disease traits and genes are too sparse to support such studies. BiRW identifies the essential associations for similar disease phenotypes and tend to generate bi-modules for a collection of phenotypes in the same disease class. Such bi-modules could help discover the core biological mechanisms underlying the human disease classes. To derive such bi-module for a disease class, we collected the known disease genes and the top 3% high-association genes (e.g. top 74 genes) of each phenotype in the predicted phenome-genome association, from which those genes that occurred as a disease gene of at least five phenotypes in the disease class (or all the phenotypes if the disease class contains less than five phenotypes) were included in the module. The genes associated with only a few phenotypes in the disease class were filtered to keep the modules dense. Each disease-gene association module describes the associations between the phenotypes in a disease class and the predicted frequent disease genes of the phenotypes. We focused on the analysis of the gastrointestinal disease-gene module as an example.

The known disease phenotype-gene associations in the gastrointestinal disease class from OMIM May-2010 are very sparse; therefore, no enriched GO biological processes (corrected p-value < 0.05) were found with standard gene set enrichment analysis. Indeed, most gastrointestinal disease phenotypes do not share any disease causative genes in common in OMIM. This observation agrees with the findings from recent studies in GWAS [26, 27], which reported that phenotypically similar diseases that are gastrointestinal-related do not tend to share their disease genes. These previous studies also hypothesized that, due to the unique topological characteristics of the gastrointestinal disease susceptibility genes, the existing network-based methods would also fail to reveal any common disease genes for understanding the underlying biological mechanisms of gastrointestinal diseases.

On contrary, BiRW identified a gastrointestinal disease-gene module that shows many common genes across the phenotypes in the disease class in Fig. 5. Some interesting examples are IL6, TP53, and PIK3CA. IL6 is a previously known causative gene of inflammatory bowel disease but not other gastrointestinal disease phenotypes. However, recent studies discovered that IL6 is involved with other gastrointestinal related disease phenotypes including pancreatitis [28], polycystic liver disease [29], salivary glands [30], congenital diarrhea [31], pancreatic agenesis [32] and gallbladder disease [33]. TP53 and PIK3CA are also not known for association with any gastrointestinal-related disease phenotype but it was recently determined that TP53 and PIK3CA play a role in developing gallbladder [34], pancreatic agenesis [35] and inflammatory bowel disease [36, 37].

We further studied the functional roles of the predicted disease genes in each module with enrichment analysis against Gene Ontology biological processes [38] using DAVID [39]. The enrichment p-values

were adjusted by $Bonferroni$ correction for multiple testing. The GO biological processes enriched by the predicted disease genes in each module with p-value$\leq$0.05 are reported in Fig. S5. Across 19 human disease classes[1], many GO biological process terms known for relevance with the disease classes are significantly enriched, such as "cell migration", "cell proliferation", and "homeostatic process" for gastrointestinal diseases, and "cell proliferation", "programmed cell death", and "apoptosis" for cancers, while these biological process terms cannot be readily identified by very sparse existing disease-gene associations. Another notable example is the enrichment of "behavior", "synaptic transmission", and "transmission of nerve impulse" by the causative genes of psychiatric diseases. Recent studies showed that regulation of synaptic transmission and transmission of nerve impulse are associated with psychiatric disease phenotypes such including autism [40].

4 Discussion

The promise to cure genetic diseases lies in a global understanding of the relation between phenotypes and the biological roles of genes. Laboratory techniques such as PCR and dideoxy sequencing were the earliest approaches to determine the genes in a targeted chromosomal region for understanding a particular phenotype. The recent microarray technologies and the second generation sequencing are more powerful tools for genome-wide studies of phenotype-gene associations. Large scale ontologies to describe the full collection of phenotypes (called phenome) such as PhenomicDB, PhenoGO and Gramene Ontologies are now becoming more stabilized and systematic. The next step is to develop computational techniques for a global inference phenome-genome association based on the previous discoveries and the phenome information.

Towards this goal, we analyzed the patterns of OMIM disease phenotype-gene associations by correlating the associations with phenotype similarity network and gene network. We showed that majority of the known associations are part of short circular bigraphs. This non-random pattern provides the foundation for deriving new associations based on the linkages in the networks. The BiRW algorithm is specifically designed to capture the circular bigraphs, and thus, can more reliably reconstruct the complete disease phenome-genome association. Functional analysis of the reconstructed phenome-genome association by disease classes revealed a global map between GO biological processes and human disease classes. In future, we plan to further investigate the relation between disease classes and the GO biological processes to understand the common molecular mechanisms of human diseases.

The experiments with 100-fold cross-validation are designed to mimic the environment of phenome-genome prediction, where associations with genes are missing for multiple phenotypes. Thus, BiRW is not directly comparable to disease gene prioritization for each individual phenotype by leave-one-out cross-validation. Our purpose is to demonstrate that BiRW will handle the phenome-genome prediction better than algorithms simply prioritizing genes (sometime a small number of genes) for each individual phenotype.

The phenotype network plays a vital role in phenome-genome association analysis since the phenotype network structure implicitly defines (known or unknown) diseases classes. We constructed a larger network of OMIM phenotypes based on Human Phenotype Ontology (HPO) [41]. Our preliminary results suggested that the overlap in HPO terms is not a good measure as phenotype similarities. Thus, constructing an informative phenotype network for phenome-genome association analysis is a more challenging problem than simple application of arbitrary similarity measures.

There are also possible variations of the BiRW algorithm. For example, we tested a variation that sequentially takes the output of left walk as the input for the right walk in each iteration. Our empirical results (not shown) suggest that BiRW performs constantly better or similar than the variation. Potentially, different combinations of left-walk and right walk might lead to difference in the performance.

[1]Nutritional, and Ear,Nose,Throat disease classes were left out since there was no predicted frequent disease genes that passed the selection criteria.

BiRW adopted a simple but effective strategy to combine the results of left walk and right walk.

Acknowledgement

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

References

1. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.

2. Johnson A, O'Donnell C (2009) An open access database of genome-wide association resutls. BMC Med Gent : 6.

3. van DM, Bruggeman J, Vriend G, et al. (2006) A text-mining analysis of the human phenome. Eur J Hum Genet 14: 535-542.

4. McKusick V (2007) Mendelian inheritance in man and its online version, OMIM. Am J Hum Genet 80: 588-604.

5. Peri S, Navarro J, Amanchy R, et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13: 2363-2371.

6. Chuang H, Lee E, Liu Y, et al. (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3: 140.

7. Linghu B, Snitkin E, Hu Z, et al. (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol 10: R91.

8. Franke L, van Bakel H, Fokkens L, et al. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78: 1011-1025.

9. Köhler S, Bauer S, Horn D, et al. (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82: 949-958.

10. Wu X, Jiang R, Zhang M, et al. (2008) Network-based global inference of human disease genes. Mol Syst Biol 4.

11. Hwang T, Kuang R (2010) A heterogeneous label propagation algorithm for disease gene discovery. Proc of SIAM Intl Conf on Data Mining : 583-594.

12. Vanunu O, Magger O, Ruppin E, et al. (2010) Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 6: e1000641.

13. Li Y, Patra J (2010) Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics 26: 1219-1224.

14. Mehan M, Nunez-Iglesias J, Dai C, et al. (2010) An integrative modular approach to systematically predict gene-phenotype associations. BMC Bioinformatics 11 Suppl 1: S62.

15. Chen Y, Jiang T, Jiang R (2011) Uncover disease genes by maximizing information flow in the phenome-interactome network. Bioinformatics 27: i167-76.

16. Hwang T, Zhang W, Xie M, et al. (2011) Inferring disease and gene set associations with rank coherence in networks. Bioinformatics 27: 2692-2699.

17. Wu X, Liu Q, Jiang R (2009) Align human interactome with phenome to identify causative genes and networks underlying disease families. Bioinformatics 25: 98-104.

18. Zaslavskiy M, Bach F, Vert J (2009) Global alignment of protein-protein interaction networks by graph matching methods. Bioinformatics 25: i259-i267.

19. Singh R, Xu J, Berger B (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc Natl Acad Sci U S A 105: 12763-12768.

20. Xu J, Li Y (2006) Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics 22: 2800-2805.

21. Singh R, Xu J, Berger B (2007) Pairwise global alignment of protein interaction networks by matching neighborhood topology. Res in Comp Mol Biol 4453: 16-31.

22. Li Z, Zhang S, Wang Y, et al. (2007) Alignment of molecular networks by integer quadratic programming. Bioinformatics : 1631-1639.

23. Guo X, Hartemink A (2009) Domain-oriented edge-based alignment of protein interaction networks. Bioinformatics 25: i240-i246.

24. Zhou D, et al. (2004) Learning with local and global consistency. In: Advanced Neural Information Processing Systems. Cambridge, MA, volume 16, pp. 321-328.

25. Goh K, Cusick M, Valle D, et al. (2007) The human disease network. Proc Natl Acad Sci USA 104: 8685-8690.

26. Baranzini S (2009) The genetics of autoimmune diseases: a networked perspective. Current Opinion in Immunology 21: 596–605.

27. Barrenas F, Chavali S, Holme P, et al. (2009) Network properties of complex human disease genes identified through genome-wide association studies. PLoS ONE 4: e8090.

28. Lesina M, Kurkowski MU, Ludes K, Rose-John S, Treiber M, et al. (2011) Stat3/socs3 activation by il-6 transsignaling promotes progression of pancreatic intraepithelial neoplasia and development of pancreatic cancer. Cancer cell 19: 456–469.

29. Strazzabosco M, Somlo S (2011) Polycystic liver diseases: Congenital disorders of cholangiocyte signaling. Gastroenterology .

30. Tucker AS (2007) Salivary gland development. Seminars in cell developmental biology 18: 237–244.

31. Dinan TG, Clarke G, Quigley EMM, Scott LV, Shanahan F, et al. (2008) Enhanced cholinergic-mediated increase in the pro-inflammatory cytokine il-6 in irritable bowel syndrome: Role of muscarinic receptors. Am J Gastroenterol 103: 2570–2576.

32. Kobayashi T, Yamaguchi T, Hamanaka S, Kato-Itoh M, Yamazaki Y, et al. (2010) Generation of Rat Pancreas in Mouse by Interspecific Blastocyst Injection of Pluripotent Stem Cells. Cell 142: 787–799.

33. Menendez A, Arena ET, Guttman JA, Thorson L, Vallance BA, et al. (2009) Salmonella infection of gallbladder epithelial cells drives local inflammation and injury in a model of acute typhoid fever. The Journal of Infectious Diseases 200: 1703–1713.

34. Goldin RD, Roa JC (2009) Gallbladder cancer: a morphological and molecular update. Histopathology 55: 218–229.

35. Rooman I, Real FX (2011) Pancreatic ductal adenocarcinoma and acinar cells: a matter of differentiation and development? Gut 61: 449-58.

36. Laurent C, Svrcek M, Flejou JF, Chenard MP, Duclos B, et al. (2011) Immunohistochemical expression of cdx2, β-catenin, and tp53 in inflammatory bowel disease-associated colorectal cancer. Inflammatory Bowel Diseases 17: 232–240.

37. Krishnan K, Arnone B, Buchman A (2011) Intestinal growth factors: Potential use in the treatment of inflammatory bowel disease and their role in mucosal healing. Inflammatory Bowel Diseases 17: 410–422.

38. Ashburner M, Ball C, Blake J, et al. (2000) Gene ontology: tool for the unification of biology. Nature genetics 25: 25–29.

39. Huang D, Sherman B, Lempicki R (2009) Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc : 44-57.

40. Xu L, Li J, Huang Y, Zhao M, Tang X, et al. (2011) Autismkb: an evidence-based knowledgebase of autism genetics. Nucleic Acids Res .

41. Robinson P, Mundlos S (2010) The human phenotype ontology. Clin Genet 77: 525-534.