Laplacian Normalization and Random Walk on Heterogeneous Networks for Disease-gene Prioritization

Zhi-qin Zhao[1], Guo-sheng Han[1,§], Zu-guo Yu[1,2,*], Jinyan Li[3,*]

[1] *Hunan Key Laboratory for Computation and Simulation in Science and Engineering and Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, Xiangtan, Hunan 411105, China.*
[2] *School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Q4001, Australia.*
[3] *Advanced Analytics Institute & Centre for Health Technologies, University of Technology Sydney, Broadway, NSW 2007, Australia.*

[§] *Joint first author.*

Abstract

Random walk on heterogeneous networks is a recently emerging approach to effective disease gene prioritization. Laplacian normalization is a technique capable of normalizing the weight of edges in a network. We use this technique to normalize the gene matrix and the phenotype matrix before the construction of the heterogeneous network, and also use this idea to define the transition matrices of the heterogeneous network. Our method has remarkably better performance than the existing methods for recovering known gene-phenotype relationships. The Shannon information entropy of the distribution of the transition probabilities in our networks is found to be smaller than the networks constructed by the existing methods, implying that a higher number of top-ranked genes can be verified as disease genes. In fact, the most probable gene-phenotype relationships ranked within top 3 or top 5 in our gene lists can be confirmed by the OMIM database for many cases. Our algorithms have shown remarkably superior performance over the state-of-the-art algorithms for recovering gene-phenotype relationships. All Matlab codes can be available upon email

[*] Corresponding author, Email: yuzg1970@yahoo.com and jinyan.li@uts.edu.au

request.

Keywords: Random walk with restart, Laplacian normalization, Disease genes and phenotypes, Heterogeneous network, Leave-one-out cross-validation, Shannon information entropy

1. Introduction

Early statistical methods [1, 2] have been developed for effective identification of genetic variants, but it is still a challenge for these methods to find the genes that are truly associated with the concerned diseases [2]. The real challenge part is how to narrow down the candidate disease-causing genes at a genomic locus to a small pool. To deal with this problem, many approaches have been proposed to rank candidate genes based on a wide variety of information, including Gene Ontology (GO) annotations [3, 4], protein domain databases [5, 6], protein sequence-based features [7, 8, 9], gene expression profiles [9, 10, 11], functional annotations [3, 4], the published literature descriptions [4, 9, 10, 12, 13], and protein-protein interactions (PPIs) [9, 11, 14, 15].

Recently, there has been intensive interest in developing algorithms to identify gene-phenotype relationships or gene-disease relationships. Lage *et al.* (2007) proposed a Bayesian model to integrate PPIs and phenotype similarities. Wu *et al.* (2008) developed a regression model to explore phenotype similarities using gene proximities. Vanunu *et al.* (2010) proposed a network propagation method, called *PRIoritizatioN and Complex Elucidation* (PRINCE), to characterize the mutual disease status of the genes. These methods open a new window of identifying genes that are responsible for diseases even when their genetic bases are unknown. Most of these algorithms are based on a single data source, or merge separate lists of candidate disease genes derived from a single data source [9, 19, 20]. The bias and noise of a single data source and the data incompleteness can easily cause an inflated uncertainty in the result by these methods.

To improve the performance, integrating multiple data sources is introduced as a new approach to the elimination of the bias and noise. For example, Li and co-workers proposed the *random walk with restart on heterogeneous network* (RWRH) algorithm [21] and the *random walk with restart on multigraph gene networks* (RWRM) working on *Complex Heterogeneous Network* (CHN) algorithms [22] to integrate the different data sources. Random walk with restart [15] simulates a random walker, either starting on a seed gene node or on a set of seed gene nodes, who moves to its immediate neighbors or returns to the seed gene nodes randomly at each step. Every gene node in the graph is ranked according to the probability of the random walker reaching to this node. If the score of the gene node is high, it is likely a disease gene. Chen *et al.* (2012) also proposed a random walk algorithm, called *Network-based Random Walk with Restart on the Heterogeneous network* (NRWRH), for the prediction of new drug-target interactions. Other approaches to the noise and bias elimination include a *maximum flow model* [24], a method *Based on Regression to Identify Disease GEnes* (BRIDGE) [25], and a diffusion-based method [26].

Our work is inspired by the Laplacian normalization idea, an effective idea which is also explored recently by PRINCE [18] and NRWRH [23]. Laplacian network normalization is a technique to normalize the weight of edges in a network based on the degrees of their end-points. It is a good method because the degree information is appropriate to calculate the probability of observing an edge between the same end-points in a random network with the same node degrees in the given network [18]. We take Laplacian network normalization to redefine the key steps of random walk-based methods to propose a new algorithm for disease-gene prioritization.

Our algorithm is named LapRWRH (Laplacian normalization based Random Walk with Restart on Heterogeneous network). It has two options: LapRWRH1 and LapRWRH2. Option LapRWRH1 constructs a gene-gene interaction matrix and a phenotype-phenotype similarity matrix from a known gene-phenotype relationship matrix, and uses this Laplacian idea to normalize these matrices before merging them into a big weight matrix for a heterogeneous network. This is different step from the weighted linear combination idea used by [23]. Our method then calculates traditional transition probabilities for this heterogeneous network and operates random walk with restart [15] on this network for disease gene prioritization. For the second option LapRWRH2, the Laplacian idea is not only used to normalize the gene-gene interaction matrix and the phenotype-phenotype similarity matrix, it is also used to redefine the transition matrix of some subnetworks, including the gene network and the phenotype network, and the transition matrix of the gene-phenotype network and that of the phenotype-gene network. This is an entirely new method.

We apply leave-one-out cross-validation to examine the performance of our new algorithms on recovering gene-phenotype relationships whose susceptible chromosomal loci are known, on a genome-wide scan of the susceptible chromosomal locus of a known phenotype, and on an *ab initio* prediction to identify causative genes for those phenotypes whose genetic mechanism is unknown. The performance of our methods is remarkably better than the state-of-the-art methods RWRH [21] and CIPHER [17]. **We analyzed the receiver operating characteristic (ROC) curve [9] and achieved higher AUC (area under the curve ROC) value.** We also compared the area value under the ROC curve with RWRH, and We also compare the Shannon information entropies [27] of the three transition matrices used by the three algorithms to draw the observation that the smaller the entropy is, a higher number of disease genes at a top list can be predicted. Furthermore, we report that some of the most probable gene-phenotype relationships as top-3 or top-5 genes predicted by our method exist in the OMIM database indeed.

Methods

Data Sets

The gene-phenotype relationship matrix (GP) is a 8919×5080 matrix. It is a year 2010 version of the OMIM database [28] (http://www.ncbi.nlm.nih.gov/omim/) downloaded via BioMart [29] (http://www.biomart.org/biomart/martview) for a fair performance comparison with the state-of-the-art methods. Element $GP(i, j)$ at row i and column j of GP denotes the relationship between disease gene i and phenotype j. In this matrix, there are only 1428 known gene-phenotype relationships between 937 genes (of the 8919 genes) and 1126 phenotype entries (of the 5080 phenotypes). The relationships between other genes and other phenotypes were unknown in 2010. For each of the 1428 known gene-phenotype relationships, we define a candidate gene set as the union of this disease gene and its 99 nearest genes in the chromosome [21].

Random Walk Algorithms

Our algorithms LapRWRH1 and LapRWRH2 share the same framework which consists of the following five steps:

- Step 1: Construct and normalize the gene-gene interaction matrix and the phenotype-phenotype similarity matrix where the genes and phenotypes are exactly from the gene-phenotype relationship matrix GP;
- Step 2: Construct a heterogeneous network by merging the gene interaction matrix, the phenotype similarity matrix, and the gene-phenotype relationship matrix;
- Step 3: Calculate the transition matrices related to the heterogeneous network;
- Step 4: Set initial probabilities and perform random walk with restart;
- Step 5: Obtain stable probabilities to rank the candidate genes.

The difference between LapRWRH1 and LapRWRH2 is at Step 3. LapRWRH1 uses traditional methods to calculate the transition matrices. However, LapRWRH2 applies the Laplacian normalization idea to determine the transition matrices. Next, we present details for each of the five steps.

Construction and normalization of the gene-gene interaction matrix and the phenotype-phenotype similarity matrix: The genes and the phenotypes are exactly from the gene-phenotype relationship matrix GP. The gene-gene interaction matrix G is a 8919×8919 matrix. Element $G(i,j)$ at row i and column j of G is the number of phenotypes commonly shared by genes i and j. This idea follows the hypothesis that if two genes have a higher number of common phenotypes, they should have a higher number of interactions [23, 30]. The phenotype similarity matrix P is a 5080×5080 matrix, which is similarly constructed as constructing G. That is, element $P(i,j)$ at row i and column j of P is the number of genes commonly shared by phenotypes i and j. Laplacian normalization (i.e., Eq.(1)) is then used to normalize G and P. The normalized G and P are denoted by $\hat{G}$ and $\hat{P}$ respectively.

Laplacian normalization: Suppose that $A = [A(i,j)], i,j = 1,2,\cdots,N$, is a symmetric matrix, D is a diagonal matrix defined as: $D(i,i)$ is the sum of row i of A and $D(i,j) = 0$ for $i \neq j$. A is normalized by $\hat{A} = D^{-1/2}AD^{-1/2}$ which also yields a symmetric matrix. The elements of $\hat{A}$ are defined by

$$\hat{A}(i,j) = \frac{A(i,j)}{\sqrt{D(i,i)D(j,j)}}. \tag{1}$$

This process is called Laplacian normalization of A. It is often used for the normalization of a weight matrix of a network [18, 23]. In fact, this technique normalizes the weight of an edge based on the degrees of its end-points. These degrees of end-points in a network are closely related to the probability of observing an edge between the same end-points in a random network with the same node degrees in the given network [18]. This normalization process is thus good for the transition matrices needed by random walk algorithms.

Construction of the heterogeneous network: We first use the three matrices $\hat{G}$, $\hat{P}$ and GP to construct three networks, namely a gene network, a phenotype network, and a gene-phenotype network. In the gene network, there exists a link (edge) between genes i and j, if the corresponding proteins i and j have interaction. The phenotype network is a graph presentation of the phenotype similarity matrix. The gene-phenotype relationships can be represented as a bipartite graph. Edges in this bipartite graph connect the phenotype entries with their relevant genes. We construct a heterogeneous network (HN) by connecting the gene network and the phenotype network using a bipartite graph, following the traditional method. As mentioned, there are 1428 gene-phenotype links between the 937 genes and the 1126 phenotype entries. A simple example of a heterogeneous network is illustrated in Figure 1 [21]. The weight matrix of this heterogeneous network is represented as $A_1 = \begin{bmatrix} \hat{G} & GP \\ PG & \hat{P} \end{bmatrix}$, where PG represents the transposition of GP. We divide the nodes of the heterogeneous network into two types. Those nodes for connecting the gene network and the phenotype network are called bridging nodes, and the other nodes are named internal nodes.

Random walk with restart: We take two strategies to obtain the transition matrix: One is to use the traditional method, the other is to calculate transition matrix via the Laplacian normalization idea. Random walk with restart (RWR) is a ranking algorithm [15], which has been used for candidate gene prioritization in the previous work [15, 20]. RWR simulates a random walker, either starting on a seed node or on a set of seed nodes, and moving to its immediate neighbors or returning to the seed nodes randomly at each step. All the nodes in the graph can be ranked according to the probability of the random walker reaching to the corresponding node.

Let A be the weight matrix of a network. Based on the topology of the network, the traditional transition matrix M with element $\mathbf{M(i,j)}$ is defined as

$$M(i,j) = \begin{cases} \frac{A(i,j)}{\sum_j A(i,j)}, & \text{if } A(i,j) \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Using the Laplacian normalization idea, element $M(i,j)$ is calculated through two steps:

$$M^{'}(i,j) = \begin{cases} \frac{A(i,j)}{\sqrt{\sum_i A(i,j) \sum_j A(i,j)}}, & \text{if } A(i,j) \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

$$M(i,j) = \begin{cases} \frac{M^{'}(i,j)}{\sum_j M^{'}(i,j)}, & \text{if } A(i,j) \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Let $\mathbf{p}_0$ be the initial probability vector, in which equal probabilities are assigned to all the source nodes with the sum of the probabilities equal to 1. The probability vector at step $s+1$ is updated by

$$\mathbf{p}_{s+1} = (1-\gamma)M^T\mathbf{p}_s + \gamma\mathbf{p}_0. \tag{5}$$

Here the parameter $\gamma \in (0,1)$ is the restart probability, and M^T is the transpose matrix of M. According to the Perron-Frobenius theorem, the eigenvalues of stochastic matrix M are in the range of $[-1,1]$. These probabilities can reach to a steady state after a sufficiently large number of iteration. We conduct the iteration until the difference between $\mathbf{p}_s$ and $\mathbf{p}_{s+1}$ (measured by the L_1 norm) falls below 10^{-10}. The steady probability $\mathbf{p}_\infty$ gives a measurement of the proximity to the seed nodes. If $\mathbf{p}_\infty(i) > \mathbf{p}_\infty(j)$, then node i is more proximate than the node j to the seed nodes [21, 22, 23].

Random walk with restart on heterogeneous network: Let $M = \begin{bmatrix} M_{\hat{G}} & M_{GP} \\ M_{PG} & M_{\hat{P}} \end{bmatrix}$ be the transition matrix of the heterogeneous network, where $M_{\hat{G}}$ is the transition matrix of the gene network G, $M_{\hat{P}}$ is the transition matrix of the phenotype network P, M_{GP} is the transition matrix from G to P, and M_{PG} is the transition matrix from P to G. Let parameter λ be the jumping probability. When the random walker moves to a bridging node, it may jump to the other subnetwork with the probability λ or move back to the other nodes in its home subnetwork with the probability $1-\lambda$.

Traditionally, the transition probability from gene g_i to phenotype p_j is defined as

$$M_{GP}(i,j) = \begin{cases} \frac{\lambda GP(i,j)}{\sum_j GP(i,j)}, & \text{if } \sum_j GP(i,j) \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Similarly, the transition probability from p_i to g_j is defined as

$$M_{PG}(i,j) = \begin{cases} \frac{\lambda GP(j,i)}{\sum_j GP(j,i)}, & \text{if } \sum_j GP(j,i) \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

For $M_{\hat{G}}$, element $M_{\hat{G}}(i,j)$ at the ith row and jth column is the transition probability of the random walker moving from node g_i to g_j. It is defined as

$$M_{\hat{G}}(i,j) = \begin{cases} \frac{\hat{G}(i,j)}{\sum_j \hat{G}(i,j)}, & \text{if } \sum_j GP(i,j) = 0 \\ \frac{(1-\lambda)\,\hat{G}(i,j)}{\sum_j \hat{G}(i,j)}, & \text{otherwise.} \end{cases} \tag{8}$$

For $M_{\hat{P}}$, element $M_{\hat{P}}(i,j)$ at the ith row and jth column is the transition probability of the random walker moving from p_i to p_j, defined as

$$M_{\hat{P}}(i,j) = \begin{cases} \frac{\hat{P}(i,j)}{\sum_j \hat{P}(i,j)}, & \text{if } \sum_j GP(j,i) = 0 \\ \frac{(1-\lambda)\,\hat{P}(i,j)}{\sum_j \hat{P}(i,j)}, & \text{otherwise.} \end{cases} \tag{9}$$

As introduced, we also propose to use Laplacian normalization to calculate transition matrices of heterogeneous networks. With this normalization idea, the transition probability $M_{GP}(i,j)$ from gene g_i to phenotype p_j is calculated via two steps, defined as

$$M^{'}_{GP}(i,j) = \\ \begin{cases} \frac{GP(i,j)}{\sqrt{\sum_i GP(i,j) \sum_j GP(i,j)}}, & \text{if } \sum_i GP(i,j) \sum_j GP(i,j) \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

$$M_{GP}(i,j) = \begin{cases} \frac{\lambda M^{'}_{GP}(i,j)}{\sum_j M^{'}_{GP}(i,j)}, & \text{if } \sum_j M^{'}_{GP}(i,j) \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

The transition probability $M_{PG}(i,j)$ from p_i to g_j is also calculated via two steps, defined as

$$M^{'}_{PG}(i,j) = \\ \begin{cases} \frac{GP(j,i)}{\sqrt{\sum_i GP(j,i) \sum_j GP(j,i)}}, & \text{if } \sum_i GP(j,i) \sum_j GP(j,i) \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

$$M_{PG}(i,j) = \begin{cases} \frac{\lambda M^{'}_{PG}(i,j)}{\sum_j M^{'}_{PG}(i,j)}, & \text{if } \sum_j M^{'}_{PG}(i,j) \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The element of $M_{\hat{G}}$ at the ith row and jth column is the transition probability of the random walker moving from node g_i to g_j. The two steps to compute this probability are:

$$M^{'}_{\hat{G}}(i,j) = \frac{\hat{G}(i,j)}{\sqrt{\sum_i \hat{G}(i,j) \sum_j \hat{G}(i,j)}}, \tag{14}$$

if $\sum_i \hat{G}(i,j) \sum_j \hat{G}(i,j) \neq 0$; and

$$M_{\hat{G}}(i,j) = \begin{cases} \frac{M^{'}_{\hat{G}}(i,j)}{\sqrt{\sum_j M^{'}_{\hat{G}}(i,j)}}, & \text{if } \sum_j GP(i,j) = 0 \\ \frac{(1-\lambda)\, M^{'}_{\hat{G}}(i,j)}{\sqrt{\sum_j M^{'}_{\hat{G}}(i,j)}}, & \text{otherwise.} \end{cases} \tag{15}$$

Similarly, the element of $M_{\hat{P}}$ at the ith row and jth column is the transition probability of the random walker from node p_i to p_j. Its calculation is via two steps:

$$M^{'}_{\hat{P}}(i,j) = \frac{\hat{P}(i,j)}{\sqrt{\sum_i \hat{P}(i,j) \sum_j \hat{P}(i,j)}}, \tag{16}$$

if $\sum_i \hat{P}(i,j) \sum_j \hat{P}(i,j) \neq 0$; and

$$M_{\hat{P}}(i,j) = \begin{cases} \frac{M^{'}_{\hat{P}}(i,j)}{\sqrt{\sum_j M^{'}_{\hat{P}}(i,j)}}, & \text{if } \sum_j GP(j,i) = 0 \\ \frac{(1-\lambda)\, M^{'}_{\hat{P}}(i,j)}{\sqrt{\sum_j M^{'}_{\hat{P}}(i,j)}}, & \text{otherwise.} \end{cases} \tag{17}$$

We use $\mathbf{u}_0$ and $\mathbf{v}_0$ to denote the initial probability of the gene network and the initial probability of the phenotype network, respectively. $\mathbf{u}_0$ and $\mathbf{v}_0$ are constructed in a way to ensure that equal probabilities are assigned to all the seed nodes in the network and that the sum of the probabilities equals to 1. The initial probability vector of the heterogeneous network is denoted as $\mathbf{p}_0 = \begin{bmatrix} (1-\eta)\mathbf{u}_0 \\ \eta\mathbf{v}_0 \end{bmatrix}$, ($\eta \in (0,1)$). The parameter η is used to weight the importance of each subnetwork. Two subnetworks are equally weighted when η is set as 0.5. If the phenotype network is more important than the gene network, then η is set above 0.5. In this case, it implies that the random walker prefers to jump to the phenotype seed nodes. On the other hand, if η is set as thresholds less than 0.5, then the random walker tends to jump to the gene seed nodes.

By repeatedly substituting the transition matrix M and initial probability $\mathbf{p}_0$ into the iterative Eq.(5), we can obtain a steady probability $\mathbf{p}_\infty = \begin{bmatrix} (1-\eta)\mathbf{u}_\infty \\ \eta\mathbf{v}_\infty \end{bmatrix}$. Then, all the genes and phenotypes can be ranked according to the steady probabilities $\mathbf{u}_\infty$ and $\mathbf{v}_\infty$, respectively.

Shannon entropy

Information theory [27] has strong connections to probabilistic modelling. An entropy is a measurement of the average uncertainty of an outcome. The Shannon entropy s associated with a probability distribution p_m is defined as

$$s = -\sum_m p_m \ln(p_m)$$

where the sum extends over all possible outcomes m. $p_m \ln(p_m)$ is defined as zero if $p_m = 0$. ln is the natural logarithm. In this case, the unit of entropy is a 'nats'.

Given a $N \times N$ transition probability matrix M, we calculate the Shannon entropy of matrix M by

$$S = -\sum_{i=1}^{N}\sum_{j=1}^{N} M^{'}(i,j)\ln(M^{'}(i,j)), \quad M^{'}(i,j) \neq 0 \tag{18}$$

where, $M^{'}(i,j) = \frac{M(i,j)}{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N} M(i,j)}$, hence, $\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N} M^{'}(i,j) = 1$.

The Shannon entropy of transition matrices can be used to evaluate the algorithms for the prediction of potential gene-phenotype relationships. We believe that a smaller Shannon entropy of the transition matrix can imply a better performance of the algorithm.

Results and Discussion

Our experimental results are grouped into four parts. First, we report our comparing results with two state-of-the-art methods RWRH [21] and CI-

PHER [17] to show the tremendous improvement on the disease gene prioritization. Second, we describe the effects of parameters on our algorithms. Third, we present the Shannon entropy of the three transition probability matrices, and the fourth part reports the results on predicting potentially new gene-phenotype relationships.

Comparison with RWRH and CIPHER

For a fair comparison with the RWRH [21] and CIPHER [17] algorithms, the same data sets, parameters and evaluation measures were used. We had three performance comparisons. First, we employed leave-one-out cross-validation to examine the performance of the three algorithms for recovering the gene-phenotype relationships. At each round of validation, we removed a gene-phenotype relationship (link or edge). The phenotype and the remaining disease genes (if any) related to this phenotype were used as seed nodes. The candidate gene set consisted of the held-out disease gene and the 99 nearest genes in the chromosome. This means that the susceptible chromosomal locus of the held-out disease gene is known. We note that if a gene is related to multiple phenotypes, this gene is tested multiple times corresponding to one of the multiple gene-phenotypes edges each time. Under such a case, the candidate gene lists are all the same, but the seed nodes of the algorithm are different.

A list of candidate genes were ranked after the random walk step Eqn.(5) became stable, forming a candidate gene rank list. Because not all candidate genes were contained in the set of 8919 genes, we only ranked the candidate genes which could be found in the set of 8919 genes. If the held-out disease gene is ranked as top one in the list, we considered it as a successful prediction. Since there were 1428 phenotype-gene relationships, we obtained 1428 candidate gene *rank lists* in total.

When parameters were set as $\gamma = 0.7$, $\lambda = \eta = 0.5$, our algorithms LapRWRH1 and LapRWRH2 successfully ranked 981 and 985 disease genes as top one, respectively. This result was tremendously better than the performance by RWRH or CIPHER, as there were only 814, 709 and 765 successful predictions by R-

WRH, CIPHER-SP and CIPHER-DN, respectively. This evaluation process is denoted by LOO1; the result is summarized at the first column of Table 1.

Table 1: Comparison with RWRH and CIPHER. The parameters λ, γ, and η, are set as 0.5, 0.7, and 0.5, respectively.

Algotithms	LOO1	LOO2	*ab initio*
LapRWRH1	**981**	**914**	**572**
LapRWRH2	**985**	**967**	**623**
RWRH	814	245	201
CIPHER-SP	709	153	140
CIPHER-DN	765	165	157

Our second assessment is on the performance when the susceptible chromosomal locus is assumed to be unknown. Such an experiment is useful because some newly found phenotypes are only linked to some experimentally validated disease genes, but with unknown susceptible chromosomal locus. In this case, there is no good candidate gene set for our algorithms. We decided to use a genome-wide scan to find genes which are likely to be involved in the newly found phenotypes. In the same way as LOO1, we removed one known gene-phenotype link (edge) each time. The phenotype and the remaining disease genes related to this phenotype were then used as seed nodes. Particularly for this experiment, all the genes in the gene network (genome-wide scan), excluding the seed genes, were used as our candidate genes. When parameters were set as $\gamma = 0.7$, $\lambda = \eta = 0.5$, LapRWRH1 and LapRWRH2 could rank 914 and 967 disease genes as top one in their candidate gene lists. However, only 245, 153 and 165 disease genes could be ranked as top one in their candidate gene lists by RWRH, CIPHER-SP and CIPHER-DN, respectively. This is a truly superior performance of our algorithms. This evaluation process is denoted by LOO2; the result is summarized at the second column of Table 1.

Third, we conducted an experiment of *ab initio* prediction [17] to identify causative genes for those phenotypes whose genetic mechanism is totally unknown. For each of the 1126 phenotype entries in the 1428 gene-phenotype relationships, we removed all the links (edges) from this phenotype to its disease genes, and used this phenotype entry as seed node to run the random walk step Eq.(5). If one of the disease genes associated to this phenotype is ranked as top one among all the 8919 genes in the gene network, we considered it as a successful prediction. For this experiment, there were 572 and 623 successful predictions by our algorithms LapRWRH1 and LapRWRH2, respectively, while only 201, 140 and 157 were successfully predicted by RWRH, CIPHER-SP and CIPHER-DN. See the third column of Table 1.

These experiments demonstrate that our algorithms are remarkably superior to the two state-of-the-art algorithms RWRH and CIPHER for (i) recovering gene-phenotype relationships whose susceptible chromosomal loci are known, (ii) genome wide scan of the susceptible chromosomal locus of a known phenotype, and (iii) *ab initio* prediction to identify causative genes for those phenotypes whose genetic mechanism is totally unknown. This huge performance improvement is brought by the Laplacian normalization idea to normalize the gene matrix, the phenotype matrix, and the transition probability matrices of the heterogeneous matrix.

Effect of parameters

Three parameters λ, γ and η can effect the performance of our two algorithms. As parameter γ, the restart probability, has only slight effect on the results [15], it is fixed as the typical value 0.7 for this study.

Parameter λ is the jumping probability. To understand its effect, we set various values of λ changing from 0.1 to 0.9 with step 0.2. The corresponding performances of LapRWRH1, LapRWRH2 and RWRH [21] are compared under Table 2. It can be seen that when the λ value was around 0.5, the performances of our two algorithms had little change. When λ ranged from 0.1 to 0.9, the performance of our two algorithms was always much better than RWRH. These results imply that our LapRWRH1 and LapRWRH2 can capture the mutually reinforcing relationship between the gene network and the phenotype network. We note that for the extreme case of $\lambda = 1$, the random walker cannot reach to any of the nodes outside the bipartite graph (instead only the nodes in the gene network or the phenotype network). We did not test this situation.

Table 2: Comparison with RWRH. Parameter γ was fixed as 0.7 as set by [15].

λ	η	LapRWRH1			LapRWRH2			RWRH		
		LOO1	LOO2	*ab initio*	LOO1	LOO2	*ab initio*	LOO1	LOO2	*ab initio*
0.1	0.5	**981**	**919**	**570**	**985**	**958**	**608**	789	196	192
0.3	0.5	**981**	**919**	**572**	**985**	**968**	**620**	804	217	196
0.5	0.1	**981**	**917**	**572**	**985**	**966**	**623**	808	239	201
0.5	0.3	**981**	**916**	**572**	**985**	**968**	**623**	813	241	201
0.5	0.5	**981**	**914**	**572**	**985**	**967**	**623**	814	245	201
0.5	0.7	**981**	**915**	**572**	**985**	**968**	**623**	817	242	201
0.5	0.9	**982**	**911**	**572**	**985**	**963**	**623**	820	244	201
0.7	0.5	**981**	**909**	**573**	**985**	**959**	**621**	815	257	203
0.9	0.5	**979**	**863**	**570**	**983**	**914**	**621**	811	261	203

Parameter η controls the effect of seed genes and seed phenotypes. If $\eta = 0.5$, the two subnetworks are equally weighted. If $\eta > 0.5$, the phenotype network is assumed to be more important, and therefore, the random walker prefers to return to the seed phenotype. On the other hand, if $\eta < 0.5$, the gene network can take a more significant role. To understand the effect of η, we run algorithms LapRWRH1 and LapRWRH2 with η changing from 0.1 to 0.9 with step 0.2. From the middle of the first column of Table 2, we can see that the performance of our two algorithms was stable, and they also always much outperformed over RWRH [21].

ROC curve analysis

Similar to RWRH [21], we also considered those phenotypes associated with at least two disease genes. Here 168 phenotypes, 375 disease genes and 470 phenotype-gene relationships in total were obtained as [21].

By the evaluation process LOO1, we performed leave-one-out cross-validation for each disease gene. The receiver operating characteristic (ROC) curve is used to compare our algorithms with RWRH. The ROC curve plots the true positive rate (Sensitivity) versus the

false positive rate (FPR) (1 - Specificity) subject to the threshold separating the prediction classes [9, 21]. As shown in Figure 2, the ROC curves of algorithms LapRWRH1 and LapRWRH2 are located above that of algorithm RWRH. In order to compare different curves obtained by ROC analysis, we calculated the area under the ROC curve (AUC) for each curve. Both the AUC values of algorithms LapRWRH1 and LapRWRH2 are almost equal to 1, which is higher than that of RWRH (0.96) [21].

Entropy comparison

Our two algorithms LapRWRH1 and LapRWRH2 produced smaller Shannon entropies for the gene and phenotype transition matrices than those by the RWRH algorithm as shown in Table 3. As the Shannon entropy of a heterogeneous network transition matrix is related to parameter λ, we set various values of λ from 0.1 to 0.9 at every 0.2 interval to get the entropies of the heterogeneous network transition matrices. The result is shown in Table 4. We can see that the Shannon entropy of the transition matrix of our heterogeneous network was always smaller than that of the heterogeneous network used by RWRH.

Table 3: The entropies of the gene network and phenotype network transition matrices in the three algorithms

	LapRWRH1	LapRWRH2	RWRH
Gene network	**9.1472**	**9.1471**	10.4101
Phenotype network	**8.6766**	**8.6758**	10.4912

Table 4: Under different parameters, the entropies of the heterogeneous network transition matrix for the three algorithms

λ	LapRWRH1	LapRWRH2	RWRH
0.1	**9.6739**	**9.6735**	11.1195
0.3	**9.7051**	**9.7047**	11.1158
0.5	**9.7064**	**9.7059**	11.0820
0.7	**9.6833**	**9.6828**	11.0240
0.9	**9.6303**	**9.6297**	10.9360

Mapping the disease gene prioritization results in Tables 1 and 2 with the entropy results in Tables 3 and 4, we infer that the smaller the entropy is, the more number of disease genes can be predicted. With this observation, we believe that the Shannon entropy of a network can be used as a criteria for comparing the performance of ranking algorithms in future studies.

Prediction of potentially new gene-phenotype relationships

The gene-phenotype data set GP is a data set obtained from the OMIM database in 2010. There are many new gene-phenotype relationships which have been added to the OMIM database since 2010. We had conducted an experiment to understand whether or not our predicted gene-phenotype relationships from the 2010 data set have already been stored in the OMIM database over the last four years. The procedure is exactly the same as LOO1, except that we define the top-3 genes in a candidate gene list as the most potential genes to form new gene-phenotype relationships. The bench-marking verification of these predicted gene-phenotype relationships against the latest version of the OMIM database is as follows:

- Get the two files mim2gene.txt and genemap.txt from the OMIM database
- Find out the corresponding OMIM entry of the predicted genes at mim2gene.txt.
- Go through genemap.txt to see whether the predicted gene-phenotype relationships are recorded in the latest version of the OMIM database.

We examined the most potential genes for each of the 1428 phenotypes. The prediction results by LapRWRH1 and LapRWRH2 are presented in Tables 5 in comparison with the performance by the RWRH algorithm. We can observe that our LapRWRH1 and LapRWRH2 algorithms have achieved superior performance over RWRH.

As seen from this table, there are 37 of our predicted genes (the most potential genes) by LapRWRH1 (also LapRWRH2) which have actually been annotated in the OMIM database since 2010. However, RWRH could only predict

Table 5: The numbers of predicted gene-phenotype relationships by LOO1 as top-3 or top-5 for the 1428 phenotypes which can be verified by the latest version of OMIM. Parameter γ is fixed as 0.7 as set by [15].

λ	η	LapRWRH1		LapRWRH2		RWRH	
		as top-3	as top-5	as top-3	as top-5	as top-3	as top-5
0.1	0.5	**37**	**74**	**37**	**74**	8	12
0.3	0.5	**37**	**74**	**37**	**74**	8	12
0.5	0.1	**37**	**74**	**37**	**74**	7	12
0.5	0.3	**37**	**74**	**37**	**74**	7	12
0.5	0.5	**37**	**74**	**37**	**74**	7	12
0.5	0.7	**37**	**74**	**37**	**74**	7	12
0.5	0.9	**37**	**74**	**37**	**74**	7	12
0.7	0.5	**37**	**74**	**37**	**74**	8	11
0.9	0.5	**37**	**74**	**37**	**74**	8	11

7 gene-phenotype relationships. Of the 37 confirmed gene-phenotype relationships, 3 were predicted by both our algorithms and RWRH; the remaining 34 relationships were predicted only by LapRWRH1 but not by RWRH. On the other hand, 4 relationships were predicted only by RWRH but not by LapRWRH1 (also LapRWRH2).

One example of the 3 commonly predicted relationships is *220290-10804 (604418)*, where '220290' is a MIM phenotype id for deafness (a disease), '10804' is an NCBI gene entry corresponding to MIM '604418', standing for GJB6 GAP Junction protein, beta 6. The other two correctly predicted phenotype-gene relationships are *254090-1291 (120220)* and *264350-6340 (600761).* In detail, MIM '254090' describes the phenotype of Ullrich congenital muscular dystrophy (a disease), and NCBI gene entry '1291' corresponds to MIM '120220', standing for GOL6A1, collagen, type VI, alpha 1. MIM '264350' describes the phenotype of pseudohypoaldosteronism type I (a disease), and NCBI gene entry '6340' corresponds to MIM '600761', standing for SCNN1G sodium channel, non-voltage-gated 1, gamma subunit.

When condition "top-3" was changed to "top-5", we obtained similar better result. There are 74 predictions by LapRWRH1 (also LapRWRH2) which have already been annotated in the OMIM database since 2010. However, RWRH could find only 12 gene-phenotype relationships in the OMIM database. Here, there are only 4 gene-phenotype relationships which could be identified by both our algorithms and RWRH. But, 70 of the 74 gene-phenotype relationships were predicted only by LapRWRH1 (also LapRWRH2) but not by RWRH. On the other hand, there are only 8 relationships which were predicted only by RWRH but not by LapRWRH1.

We also conducted experiments by changing the procedure from LOO1 to LOO2 (genome-wide scan). It was found that our algorithms loss some performance. So, we suggest to use LOO1 for the prediction of potentially new gene-phenotype relationships. As some of the truly predicted gene-phenotype relationships by RWRH are not covered by the true predictions of our algorithms. It is also an interesting problem to combine our algorithms and RWRH when applied to the latest version of the OMIM database for the prediction of potentially new gene-phenotype relationships.

Conclusion

We have proposed and implemented two options of our algorithm LapRWRH1 and LapRWRH2 to prioritize disease genes and identify potentially new gene-phenotype relationships. These two algorithms are random walk-based methods which work on heterogeneous networks merging gene-gene interaction networks, phenotype-phenotype networks, and gene-phenotype networks. The novel idea of our algorithms is to use Laplacian normalization to normalize the gene interaction matrix and the phenotype similarity matrix before the construction of the heterogenous network, and also use the Laplacian normalization idea to normalize the transition probability matrices. Our algorithms have shown remarkably superior performance over the state-of-the-art algorithms for recovering gene-phenotype relationships whose chromosomal loci are known, for genome-wide scan of disease genes, and for the *ab initio* prediction of causative genes. Our algorithms have also been tested for the identification of potentially new gene-phenotype relationships by predicting the newly-added gene-phenotypes at the OMIM database since 2010. The performance is significant. According to the Shannon information theory, we have drawn an observation that a ranking algorithm has a better prediction performance if the Shannon entropy of the transition matrix is smaller. Disease gene prioritization is a hard research problem. We will conduct a deep study in the future to combine the novel ideas of the existing method to improve more of the prediction performance.

Acknowledgements

This work was mainly supported by the Natural Science Foundation of China (Grants No. 11071282 and 11371016), the Chinese Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) (Grant No. IRT1179), the Lotus Scholars Program of Hunan province of China, Hunan Provincial Innovation Foundation for Postgraduate (CX2012B243), and the Australia Research Council (DP130102124).

References

[1] Lander,E. and Schork,N. Genetic dissection of complex traits. Science, 265(1994) 2037-2048.

[2] Glazier,A.M., Nadeau,J.H. and Aitman,T.J. Finding Genes That Underlie Complex Traits. Science, 298(2002) 2345-2349.

[3] Freudenberg,J and Propping,P. A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics, 18(Suppl. 2)(2002) S110-S115.

[4] Perez-Iratxeta,C., Bork,P. and Andrade,M.A. Association of genes to genetically inherited diseases using data mining. Nat. Genet, 31(2002) 316-319.

[5] Turner,F.S., Clutterbuck,D.R. and Semple.C.A.M. POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol, 4(2003) R75.

[6] Kent,W.J., Hsu,F., Karolchik,D., Kuhn,R.M., Clawson,H., Trumbower,H. and Haussler,D. Exploring relationships and mining data with the UCSC Gene Sorter. Genome Res, 15(2005) 737-741.

[7] Lopez-Bigas,N. and Ouzounis,C.A. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res, 32(2004) 3108-3114.

[8] Adie,E.A., Adams,R.R., Evans,K.L., Porteous,D.J. and Pickard,B.S. Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics, 6(2005) 55.

[9] Aerts,S., Lambrechts,D., Maity,S., Van LOO,P., Coessens,B., Smet,F., Tranchevent,L.C., Moor,B.D., Marynen,P., Hassan,B., Carmeliet,P. and Moreau,Y. Gene prioritization through genomic data fusion. Nat. Biotechnol., 24(2006) 537-544.

[10] Van Driel,M.A., Cuelenaere,K., Kemmeren,P.P.C.W., Leunissen,J.A.M. and Brunner,H.G. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. Eur. J. Hum. Genet., 11(2003) 57-63.

[11] Franke,L., Van Bakel,H., Fokkens,L., Jong,E.D., Egmont-Petersen,M. and Wijmenga.C. Reconstruction of a Functional Human Gene Network, with an Application for Prioritizing Positional Candidate Genes. Am. J. Hum. Genet., 78(2006) 1011-1025.

[12] Tiffin,N., Kelso,J.F., Powell,A.R., Pan,H., Bajic,V.B. and Hide,W.A. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res., 33(2005) 1544-1552.

[13] Gaulton,K.J., Mohlke,K.L. and Vision,T.J. A computational system to select candidate genes for complex human traits. Bioinformatics, 23(2007) 1132-1140.

[14] Oti,M., Snel,B., Huynen,M.A. and Brunner,H.G. Predicting disease genes using protein-protein interactions. J. Med. Genet., 43(2006) 691-698.

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

[16] Lage,K., Karlberg,E.O., Storling,Z.M., Olason,P.I., Pedersen,A.G., Rigina,O., Hinsby,A.M., Tumer,Z., Pociot,F., Tommerup,N., Moreau,Y. and Brunak,S. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol., 25(2007) 309-316.

[17] Wu,X.B., Jiang,R., Zhang,M.Q. and Li,S. Network-based global inference of human disease genes. Mol. Syst. Biol., 4(2008) Article 189.

[18] Vanunu,O., Magger,O., Ruppin,E., Shlomi,T. and Sharan,R. Associating Genes and Protein Complexes with Disease via Network Propagation. PLoS Comput. Biol., 62010 e1000641.

[19] Stuart,J.M., Segal,E., Koller,D. and Kim,S.K. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science, 302(2003) 249-255.

[20] Li,Y.J and Patra,J.C. Integration of multiple data sources to prioritize candidate genes using discounted rating system. BMC Bioinformatics, 11(Suppl 1)(2010) S20

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

[22] Li,Y.J. and Li,J.Y. Disease gene identificaton by random walk on multigraphs merging heterogeneous genomic and phenotype data. BMC Genomics, 13(Suppl 7)(2012) S27.

[23] Chen,X., Liu,M.X. and Yan,G.Y. Drug-target interaction prediction by random walk on the heterogeneous network. Molecular BioSystems, 8(2012) 1970-1978.

[24] Chen,Y., Jiang,T. and Jiang,R. Uncover disease genes by maximizing information flow in the phenome-interactome network. Bioinformatics, 27(13)(2011) i167-i176.

[25] Chen,Y., Wu,X. and Jiang,R. Integrating human omics data to prioritize candidate genes. BMC Medical Genomics, 6(2013) 57.

[26] Zhu,J., Qin,Y.F., Liu,T.G., Wang,J. and Zheng,X.Q. Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles. BMC Bioinformatics, 14(Suppl 5)(2013) S5.

[27] Cover,T.M. and Thomas,J.A. *Elements of Information Theory.* John Wiley and Sons, Inc. (2012)

[28] Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and Mckusick,V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res., 33(Database Issue)(2005) D514-D517.

[29] Smedley,D., Haider,S, Ballester,B., Holland,R., London,D., Thorisson,G. and Kasprzyk,A. BioMartCbiological queries made easy. BMC Genomics., 10(2009) Article 22.

[30] Xia,Z., Wu,L.Y., Zhou,X.B. and Wong,S.T.C. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst. Biol., 4(Suppl 2)(2010) S6.

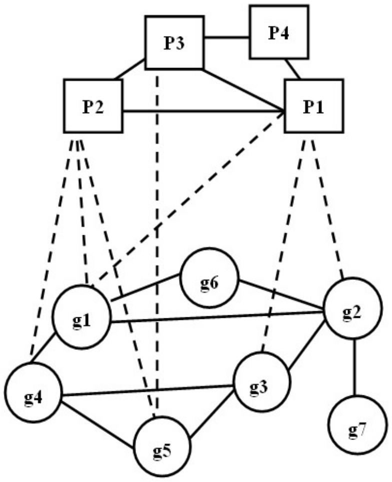

Figure 1: A heterogeneous network. The upper subnetwork is a phenotype network, and the lower subnetwork is a gene network. They are connected by a gene-phenotype relationship [21].

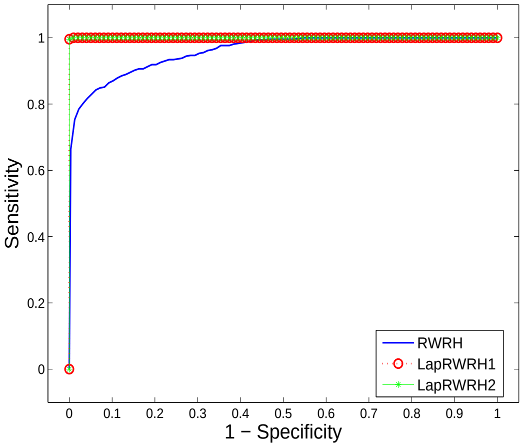

Figure 2: ROC curves of LapRWRH1, LapRWRH2 and RWRH