BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 9 2010, pages 1219–1224
doi:10.1093/bioinformatics/btq108

Systems biology

Advance Access publication March 9, 2010

Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network

Yongjin Li* and Jagdish C. Patra

School of Computer Engineering, Nanyang Technological University, Singapore

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Clinical diseases are characterized by distinct phenotypes. To identify disease genes is to elucidate the gene–phenotype relationships. Mutations in functionally related genes may result in similar phenotypes. It is reasonable to predict disease-causing genes by integrating phenotypic data and genomic data. Some genetic diseases are genetically or phenotypically similar. They may share the common pathogenetic mechanisms. Identifying the relationship between diseases will facilitate better understanding of the pathogenetic mechanism of diseases.

Results: In this article, we constructed a heterogeneous network by connecting the gene network and phenotype network using the phenotype–gene relationship information from the OMIM database. We extended the random walk with restart algorithm to the heterogeneous network. The algorithm prioritizes the genes and phenotypes simultaneously. We use leave-one-out cross-validation to evaluate the ability of finding the gene–phenotype relationship. Results showed improved performance than previous works. We also used the algorithm to disclose hidden disease associations that cannot be found by gene network or phenotype network alone. We identified 18 hidden disease associations, most of which were supported by literature evidence.

Availability: The MATLAB code of the program is available at http://www3.ntu.edu.sg/home/aspatra/research/Yongjin_BI2010.zip

Contact: yongjin.li@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 10, 2009; revised on February 7, 2010; accepted on March 4, 2010

1 INTRODUCTION

Elucidating the inherited basis of human disease involves linking genomic variation to clinical phenotype. Establishing this relationship, however, can be challenging for several reasons, the pleiotropy of genes, the genetic heterogeneity of diseases and the limited number of cases (Giallourakis *et al.*, 2005).

Most current efforts at disease–gene identification involving linkage analysis and association studies result in a genomic interval of 0.5–10 cM, containing up to several hundreds of genes (Anne *et al.*, 2002; Botstein and Risch, 2003). These candidate genes need to be further investigated to identify disease-causing genes. A number of methods have been proposed to prioritize candidate genes based on different kinds of genomic data, such as sequence-based features (Adie *et al.*, 2006; López-Bigas and Ouzounis, 2004; Turner *et al.*, 2003), functional annotation data (Freudenberg and Propping, 2002; Perez-Iratxeta *et al.*, 2002) and protein interaction data (Köhler *et al.*, 2008; Xu and Li, 2006).

These algorithms typically prioritize candidate genes based on their similarity to known disease genes. Though these methods perform well, they still have some limitations. The first limitation comes from the incompleteness and noise of genomic data sources. Some integration algorithms have been proposed to solve this problem (Aerts *et al.*, 2006; Linghu *et al.*, 2009; Li and Patra, 2010). The other problem is the ambiguous boundary between different diseases. Clinical disease often encompass a variety of phenotypes and biological mechanisms, making it difficult to define the boundary between diseases. Traditionally, diseases have been categorized on the basis of pathophysiology or on etiology, but often these characterizations break down and more *ad hoc* approaches aroused, resulting in the celebrated debate between splitters and lumpers (McKusick, 1969). The ambiguous boundary between different diseases prevents the direct inference of gene–disease association. For example, the Leber's congenital amaurosis (LCA) turns out to be highly heterogeneous on a molecular basis, but these molecular subtypes appear clinically homogeneous (Traboulsi *et al.*, 2006). Using all the LCA genes to prioritize a list of genes responsible to a subtype of LCA may not be correct.

Most recently, two algorithms have been proposed to identify gene–phenotype relationship instead of finding the gene–disease relationship directly (Lage *et al.*, 2007; Wu *et al.*, 2008). Their assumption is that similar phenotypes are caused by functionally related genes (Oti and Brunner, 2007). Lage *et al.* (2007) assign candidate gene to protein complexes and then rank these complexes using phenotypic data. Finally, candidate genes are ranked based on the phenotypes associated with the protein complexes. Wu *et al.* (2008) employ the regression model, named CIPHER, to quantify the concordance between the candidate gene and the target phenotype. Candidate genes are then ranked by the concordance score. CIPHER performed better than Lage *et al.* (2007) on the overlapped benchmark data (Wu *et al.*, 2008).

In this work, we propose a RWRH (random walk with restart on heterogeneous network) algorithm to infer the gene–phenotype relationship. We connect the gene network and phenotype network by gene–phenotype relationship and constructed a heterogeneous network. Then, we extend the random walk with restart (RWR) algorithm to the heterogeneous network, using the target phenotype and corresponding disease genes as seed nodes. In the prioritization of candidate genes, we attempt to make better use of the phenotypic data. On benchmark dataset the proposed algorithm

*To whom correspondence should be addressed.

performed better than Wu *et al.* (2008). We also compared with RWR on gene network only (Köhler *et al.*, 2008), and achieved much higher AUC (area under the curve) value.

The RWRH algorithm is inspired by the co-ranking framework (Zhou *et al*., 2007). It ranks phenotypes and genes at the same time. If we set seed nodes as genes and phenotypes associated with one disease, the top ranked phenotype is selected as the most similar phenotype to the query disease. Therefore the disease associate with this phenotype should be the most similar to the query disease. We use this algorithm to disclose the relationship between diseases and found 18 disease associations that cannot be found by gene network or phenotype network alone. Most of these disease associations were supported by various types of evidence.

2 METHODS

In this section, we first introduce various data source used in this work. And then we give detailed description of heterogeneous network construction method and propose the algorithm of RWRH.

2.1 Data source

The protein–protein interaction (PPI) data were derived from Human Protein Reference Database (HPRD; Peri *et al.*, 2003). HPRD contains manually curated scientific information pertaining to the biology of most of the human proteins. Disease-related phenotype can be interpreted as a textual description of a disease's detectable outward manifestations. Same as previous works (van Driel *et al.*, 2006; Wu *et al.*, 2008), a phenotype entry was defined as an MIM record. We excluded the records with the prefix '∗' and '^'. Because the prefix '∗' refers to the record of disease gene, and '^' refers to the obsoleted record. The phenotypic similarity was calculated using MimMiner (van Driel *et al.*, 2006). Gene–phenotype relationship were obtained from the OMIM database (Hamosh *et al.*, 2005), extracted using BioMart (Smedley *et al.*, 2009). Disease category information was taken from a manual classification concerning the physiological system affected (Goh *et al.*, 2007).

2.2 Construction of the heterogeneous network

Three types of data sources are represented by three networks, namely gene network, phenotype network and gene–phenotype network. In the gene network, two genes are connected if the proteins they encode interact with each other according to the HPRD database. The phenotype network is a k nearest neighbor (KNN) graph presentation of the phenotypic similarity matrix, which is calculated using MimMiner (van Driel *et al.*, 2006). Each phenotype entity is connected with its five nearest neighbors, and the edge is weighted by the corresponding similarity score. The gene–phenotype relationship is represented as a bipartite graph. Edges in the bipartite graph connect the phenotype entity with the relevant genes. We construct the heterogeneous network by connecting the gene network and phenotype network using the bipartite graph. A simple example of the heterogeneous network is illustrated in Figure 1.

Suppose $A_{G(n\times n)}$, $A_{P(m\times m)}$ and $B_{(n\times m)}$ are adjacency matrix for gene network, phenotype network and the bipartite graph, respectively, where n and m represent the number of genes and phenotype entities. The adjacency matrix of the heterogeneous network can be represented as $A=\begin{bmatrix} A_G & B \\ B^T & A_P \end{bmatrix}$, where B^T represents the transpose of B.

2.3 RWRH

RWR is a ranking algorithm (Köhler *et al.*, 2008). It simulates a random walker, either starts on a seed node or on a set of seed nodes and moves to its immediate neighbors randomly at each step. Finally, all the nodes in the graph are ranked by the probability of the random walker reaching this node. Let $\mathbf{p}_0$ be the initial probability vector and $\mathbf{p}_s$ be a vector in which the i-th element holds the probability of finding the random walker at node i at step s. The probability vector at step $s+1$ can be given by

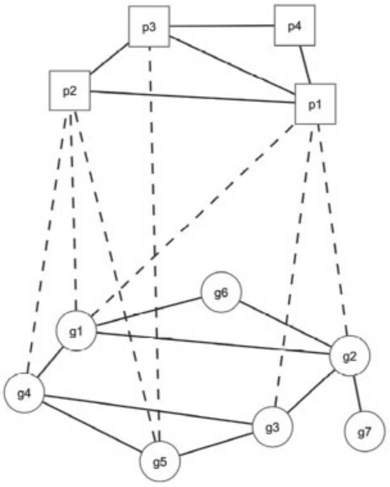

Fig. 1. Illustration of the heterogeneous network. The upper subnetwork is phenotype network, and the lower network is gene network. They are connected by the gene–phenotype relationship. This figure is inspired by Wu *et al.* (2008).

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

where M is the transition matrix of the graph. M_{ij} is the transition probability from node i to node j. The calculation of M is described later. The parameter $\gamma\in(0,1)$ is the restart probability. At each step, the random walker can return to seed nodes with probability γ.

After some steps, the probability will reach a steady state. This is obtained by performing the iteration until the difference between $\mathbf{p_s}$ and $\mathbf{p_{s+1}}$ (measured by the L_1 norm) fall below 10^{-10}. The steady-state probability $\mathbf{p}_\infty$ gives a measure of proximity to seed nodes. If $\mathbf{p}_\infty(i)>\mathbf{p}_\infty(j)$, then node i is more proximate to seed nodes than node j.

Let $M=\begin{bmatrix} M_G & M_{GP} \\ M_{PG} & M_P \end{bmatrix}$ be the transition matrix of the heterogeneous network, where M_G and M_P are intra-subnetwork transition matrix and M_{GP}, M_{PG} are inter-subnetwork transition matrix. Let λ be the jumping probability, that is the probability of the random walker jumping from gene network to phenotype network or vise versa. It regulates the reinforcement between two subnetworks. If $\lambda=0$, the genes and phenotypes are ranked independently. As seen from Figure 1, not all the genes are connected to phenotypes. When the random walker is in the gene network, he can jump to the phenotype network or stay in the gene network. If he is on the node connecting to phenotypes, he will jump to the phenotype network with probability λ, or move to other nodes in gene network with probability $1-\lambda$. Otherwise, he cannot jump to the phenotype network and will only move to other nodes in the gene network. The transition probability from g_i to p_j can be described as

$$(M_{GP})_{i,j}=p(p_j|g_i)=\begin{cases} \lambda B_{ij}/\sum_j B_{ij}, & \text{if } \sum_j B_{ij}\neq 0 \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

Similarly, the transition probability from p_i to g_j can be described as

$$(M_{PG})_{i,j}=p(g_j|p_i)=\begin{cases} \lambda B_{ji}/\sum_j B_{ji}, & \text{if } \sum_j B_{ji}\neq 0 \\ 0, & \text{otherwise.} \end{cases} \qquad (3)$$

The element of M_G at i-th row and j-th column is $p(g_j|g_i)$, the probability of the random walker transition from g_i to g_j. It is defined as

$$(M_G)_{i,j}=\begin{cases} (A_G)_{i,j}/\sum_j (A_G)_{i,j}, & \text{if } \sum_j B_{ij}=0 \\ (1-\lambda)(A_G)_{i,j}/\sum_j (A_G)_{i,j}, & \text{otherwise.} \end{cases} \qquad (4)$$

The element of M_P at i-th row and j-th column is the probability of the random walker transition from p_i to p_j. It is defined as

$$(M_P)_{i,j} = \begin{cases} (A_P)_{i,j}/\sum_j (A_P)_{i,j}, & \text{if } \sum_j B_{ji} = 0 \\ (1-\lambda)(A_P)_{i,j}/\sum_j (A_P)_{i,j}, & \text{otherwise.} \end{cases} \tag{5}$$

Let $\mathbf{u}_0$ and $\mathbf{v}_0$ represent the initial probability of gene network and phenotype network, respectively. The initial probability of gene network $\mathbf{u}_0$ is constructed such that equal probabilities are assigned to all the seed nodes in the gene network, with the sum of the probabilities equal to 1. This is equivalent to letting the random walker begin from each of the seed nodes with equal probability. Similarly, the initial probability of phenotype network $\mathbf{v}_0$ is given. The initial probability vector for heterogeneous network is represented as $\mathbf{p}_0 = \begin{bmatrix} (1-\eta)\mathbf{u}_0 \\ \eta \mathbf{v}_0 \end{bmatrix}$. The parameter $\eta \in (0,1)$ is used to weight the importance of each subnetwork. If η is 0.5, two subnetworks are equally weighted. If η is above 0.5, the random walker prefer to return to the phenotypic seed nodes; therefore, the phenotype is given more importance.

We plunge the transition matrix M and initial probability $\mathbf{p}_0$ into the iterative equation [Equation (1)]. After some steps, the steady probability $\mathbf{p}_\infty = \begin{bmatrix} (1-\eta)\mathbf{u}_\infty \\ \eta \mathbf{v}_\infty \end{bmatrix}$ is obtained. Then genes and phenotypes are ranked based on the steady probability $\mathbf{u}_\infty$ and $\mathbf{v}_\infty$, respectively.

3 EXPERIMENTS AND RESULTS

In this section, we first compare the proposed RWRH algorithm with CIPHER (Wu *et al.*, 2008). Then, we investigate the effect of parameters. After that, we compared the algorithm with RWR on gene network only (Köhler *et al.*, 2008). Finally, we identified some hidden disease associations.

3.1 Comparison with CIPHER

To compare with CIPHER (Wu *et al.*, 2008), we used the same data and the same evaluation measures as CIPHER. The gene network contains 34 364 interactions between 8919 genes. The phenotypic similarity matrix between 5080 phenotype entities are calculated using MimMiner (van Driel *et al.*, 2006). There are 1428 gene–phenotype links between 937 genes and 1216 phenotype entities.

We use leave-one-out cross-validation to examine how well the algorithm recovers the gene–phenotype relationship. In each round of validation, we remove a gene–phenotype link. The phenotype and the rest of disease genes related to this phenotype are used as seed nodes. We defined the candidate gene set as the held out disease gene and the 99 nearest genes in the chromosome. We use the random walk algorithm [Equation (1)] to rank the candidate genes. If the held-out disease gene is ranked as top 1, we consider it a successful prediction. We use the number of successful predictions as a measure to compare different algorithms. We set $\gamma = 0.7$, $\lambda = \eta = 0.5$, and successfully ranked 814 known disease genes as top 1. It is much better than CIPHER. There were 709 and 765 success predictions for CIPHER-SP and CIPHER-DN, respectively. The result is shown as LOO1 in Table 1.

Some phenotypes already have experimental validated disease genes, but no susceptible chromosomal locus has been newly found. Therefore no candidate gene is available. In this case, genome-wide scan is needed to find genes likely to be involved in the phenotype. Similar to the above experiment, each time we remove a known gene–phenotype link and use the phenotype and the rest of disease genes associated with this phenotype as seed nodes. In this experiment, all the genes in the gene network except seed genes are used as candidate genes. Finally, 245 disease genes are ranked top 1. In contrast, only 153 disease genes have been ranked at the top by CIPHER-SP and 165 disease genes by CIPHER-DN (LOO2 in Table 1).

Table 1. In comparison with CIPHER

Algorithms	LOO1	LOO2	*ab initio*
RWRH	814	245	201
CIPHER-SP	709	153	140
CIPHER-DN	765	165	157

LOO1, locus and several related genes are known; LOO2, locus unknown, but no related genes are known; *ab initio*, locus unknown, no related genes are known, but phenotype is known.

The genetic mechanism of some phenotypes is totally unknown. No known disease genes or suspectable chromosomal locus have been found related to this kind of disease phenotype. Identifying causative genes for this kind of phenotype from the whole-genome is called *ab initio* prediction (Wu *et al.*, 2008). In the gene–phenotype bipartite graph, 1216 phenotype entities are connected to 973 disease genes. For each of these 1216 phenotype entities, we remove all the links from this phenotype to disease genes and use this phenotype entity as seed node to run the random walk algorithm [Equation (1)]. If one of the disease genes associated to the phenotype is ranked top 1 among all 8919 genes in the gene network, we consider it a successful prediction. As seen from Table 1, there are 201 successful predictions by our algorithm, while CIPHER-SP and CIPHER-DN successfully predicted 140 and 157 cases, respectively.

3.2 Effect of parameters

There are three parameters in our algorithm γ, λ and η. The parameter γ is the restart probability. It has been shown that this parameter only has slight effect on the results (Köhler *et al.*, 2008). In this work, we fix γ at 0.7.

The parameter λ is the jumping probability. It controls the reinforcement between gene network and phenotype network. Large λ introduce more mutual dependence of rankings between genes and phenotypes. To investigate the effect of this parameter, we set various values of λ ranging from 0.1 to 0.9. The performance of the algorithm is measured using three measures mentioned in the above section. Results are shown in Table 2. The performance is improved with the increase in λ value. When λ ranges from 0.5 to 0.9, the performance becomes stable. If the λ value is too big, the random walker jumps between gene network and phenotype network based on the structure of bipartite graph. But the topological structure of gene network and phenotype network cannot be well utilized. In the extreme case, if $\lambda = 1$, the random walker will not reach any of the nodes outside the bipartite graph (nodes only in gene network or phenotype network). Therefore, we suggest to select the λ value from 0.5 to 0.9. The performance at $\lambda < 0.5$ is comparatively poor, but still much better than CIPHER. Results suggest that the RWRH algorithm successfully captures the mutually reinforcing relationship between gene network and phenotype network.

The parameter η controls the impact of two kinds of seed nodes, seed phenotypes and seed genes. If η is 0.5, two subnetworks are equally weighted. If η is above 0.5, the random walker prefers to return to the seed phenotypes; therefore, the structure of phenotype

Table 2. Effect of λ value

λ	LOO1	LOO2	*ab initio*
0.1	789	196	192
0.3	804	217	196
0.5	814	245	201
0.7	815	257	203
0.9	811	261	203

Table 3. Effect of η value

η	LOO1	LOO2
0.1	808	239
0.3	813	241
0.5	814	245
0.7	817	242
0.9	820	244

network play a more important role in the prioritization of disease genes. To find the effect of η value, we run the RWRH algorithm with different η values, and calculate the first two measures using leave-one-out cross-validation. As seen from Table 3, the algorithm performs slightly better when η is above 0.5. It shows that phenotype network should be given more importance.

3.3 Comparison with RWR on gene network only

To further highlight the importance of phenotype network, we compared the performance of RWRH with RWR on gene network only (Köhler *et al.*, 2008). In RWR algorithm, for one phenotype, at least two genes are required to perform leave-one-out cross-validation. Therefore, in this experiment, only phenotypes associated with at least two disease genes were considered. We obtained 168 phenotypes in total, associated with 470 disease genes.

For each disease gene, we defined the artificial linkage interval to be the set of genes containing the first 99 genes located nearest to the disease gene according to their genomic distance on the same chromosome. We performed leave-one-out cross-validation for each disorder. In each round of cross-validation, we held out one disease gene and remove the link between this gene to the phenotype entry. The rest disease genes and the phenotype entry were used as seed nodes. The held-out gene and all the genes in the artificial linkage are ranked by the RWRH algorithm. We use the receiver operating characteristic (ROC) curve to compare two algorithms, which plots the sensitivity versus 1−specificity subject to the threshold separating the prediction classes (Aerts *et al.*, 2006). Sensitivity refers to the percentage of disease genes that were ranked above a particular threshold. Specificity refers to the percentage of non-disease genes ranked below this threshold. As shown in Figure 2, the curve of RWRH algorithm is above RWR with gene network only. It suggests that the RWRH algorithm obtained both higher sensitivity and higher specificity; therefore, it is better than RWR on gene network only. The AUC value of the RWRH algorithm is 0.96, which is much higher than RWR on gene network only (0.92).

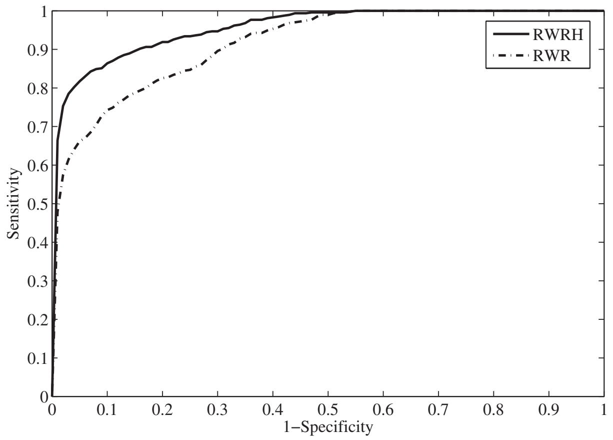

Fig. 2. ROC curve of RWR and RWRH.

3.4 Predict new disease gene of Alzheimer's disease

Alzheimer's disease is the most common form of progressive dementia in the elderly. It is a genetically heterogeneous neurodegenerative disorder. There are 16 disease phenotypes (MIM Record) for Alzheimer's disease, 12 of which with prefix '%'. We use the proposed RWRH algorithm to predict new disease genes for these 12 phenotypes. The target phenotype is used as seed node to run the RWRH algorithm. Top 5 ranked candidate genes have been selected. Results are shown in Supplementary Table S1. Three examples of the novel prediction are given below.

The first example is MIM 6 11 073 and the corresponding suspectable region is on chromosome 8p12-q22. There are 241 candidate genes in this locus. The second ranked gene is *PRKDC*. It encodes an enzyme, DNA-dependent protein kinase catalytic subunit, also known as *DNA-PKcs*. Deficits in *DNA-PKcs* render neurons vulnerable to adverse conditions of relevance to the pathogenesis of neurodegenerative disorders such as Alzheimer's disease and stroke (Zhang *et al.*, 2007).

MIM 6 08 907 describes the phenotype of late onset familial Alzheimer's disease. Wijsman *et al.* (2004) applied the Bayesian Markov chain Monte Carlo (MCMC) linkage analysis methods to an analysis of late-onset Alzheimer's disease. They identified strong evidence of a late-onset Alzheimer's disease locus on 19p13.2. There are 199 genes in this region. The fourth ranked gene is *LDLR* (low-density lipoprotein receptor). Its ligand *ApoE* is the major genetic modifier of the age of onset of Alzheimer's disease (Herz, 2009). The fifth ranked gene is *PIN1*. It has been identified as the molecular partner of *Tau* and amyloid precursor protein (*APP*), the key factors of Alzheimer's disease (Takahashi *et al.*, 2008).

MIM 6 09 636 describes the phenotype of early-onset familial Alzheimer's disease and the corresponding suspectable region is 7q36. There are 87 genes in this region. The third ranked gene is *CDK5*. It has been proposed that relative resistance to phosphatases might be a common feature of *CDK5* substrates and could contribute to the hyperphosphorylation of *CRMP2* and *Tau* observed in Alzheimer's disease (Cole *et al.*, 2008).

3.5 Disclose hidden disease–disease associations

With the cumulated data in OMIM, people's view of human disease is being changed (Oti *et al.*, 2008). Diseases sharing similar phenotypes may be related to dysfunction of a regulatory network, such as a signaling pathway or a biochemical module, as has been demonstrated for Noonan syndrome (MIM 1 63 950) and related disorders (Gelb and Tartaglia, 2006). Therefore disease association analysis is of great importance for our understanding of the common physiology and pathophysiology of cellular networks shared by diseases. Dysfunction in these common cellular networks or pathways may lead to similar phenotypic consequences (Robinson *et al.*, 2008). Diseases are usually linked into a network for searching of common pathogenetic mechanisms shared by similar diseases. Some groups link diseases together based on their phenotype overlap (van Driel *et al*., 2006; Oti and Brunner, 2007) or clinical diagnosis records (Rzhetsky *et al.*, 2007). This method has two limitations. On the one hand, it may be affected by the standardization and quantification of phenotypic description (Biesecker, 2005). On the other hand, those disease–disease associations that can be easily detected at the molecular level but not at the phenotypic level will be missed. Some others try to find the genetic overlap between diseases. Diseases are linked together if they share disease genes (Goh *et al.*, 2007) or metabolite (Lee *et al.*, 2008), or even biological pathways (Li and Agarwal, 2009). This method is limited by the relative paucity of knowledge of disease-causing genes and the incompleteness and noise of genomic data.

As described in Section 2.3, the RWRH algorithm ranks genes and phenotypes at the same time. In this section, we use the RWRH algorithm to identify disease associations. The disease category and disease ID are obtained from Goh *et al.* (2007). Each disease is represented as a group of disease phenotypes (MIM Record). If we start from the phenotypes and genes associated to a disease, phenotypes of the most relevant disease should be ranked at the top. Therefore the association between diseases is found. Since the RWRH algorithm successfully captures the mutually reinforcing relationship between gene network and phenotype network, it may find some hidden associations that cannot be found by gene network or phenotype network alone. We try to disclose the hidden disease associations using the following procedures. In the first step, for one disease d_i, we set the seed nodes as the disease-associated phenotypes and disease genes. Other phenotypes are ranked based on the ranking score, i.e. the steady probability in $\mathbf{v}_\infty$ described in Section 2.3. In the second step, the top ranked phenotype is selected out. Subsequently, if this phenotype is not linked to any phenotype of d_i in the phenotype network, we find the disease d_j, the top-ranked phenotype it belongs to. Finally, the association between d_i and d_j is found, and there is no overlap phenotype between d_i and d_j.

We found 122 disease associations sharing no phenotype. We further filtered out the disease association pairs sharing disease genes. There are 18 disease associations left, which are shown in Supplementary Table S2. Among these 18 disease associations, 12 disease pairs have been classified in the same disease class. Especially eight of these disease pairs are metabolism diseases. In the human disease network constructed by Goh *et al.* (2007), metabolism diseases were not well connected. We can disclose these relationship, because in the RWRH algorithm phenotype similarity information and gene interaction information are complementarily used. We also found two disease pairs, which are actually subtypes of the same disease, but classified into different disease classes. Diseases 1130 and 72 are two subtypes of oculocutaneous albinism. One is classified as ophthamological disease and the other is classified as dermatological disease (Goh *et al.*, 2007). The other example is Diseases 1325 and 315. Disease 1325 is classified as ‘multiple’, and disease 315 is classified as Connective tissue disorder. In addition, there are interactions between two sets of disease genes from these two diseases. The association between Bartter syndrome and Gitelman syndrome is supported by recent literature. Type III Bartter syndrome is clinically and biochemically overlapping with Gitelman syndrome (Knoers and Levtchenko, 2008).

4 CONCLUSIONS AND DISCUSSIONS

In this article, we integrated gene network and phenotype network to identify gene–phenotype relationships. We constructed a heterogeneous network by connecting the gene network and phenotype network using known gene–phenotype relationships obtained from OMIM (Hamosh *et al.*, 2005). Then we extended the RWRH algorithm. The performance of RWRH algorithm is significantly better than CIPHER (Wu *et al.*, 2008) and RWR method using only gene network (Köhler *et al.*, 2008). It suggests that the RWRH algorithm effectively captures the complementarity between gene network and phenotype network. Another advantage of the RWRH algorithm is robustness to the parameters. Results change slightly with the values of three parameters ranging from 0.5 to 0.9.

We also showed the ability of RWRH algorithm to disclose hidden disease associations. We identified 18 disease associations that cannot be found by gene network or phenotype network alone. Most of them are supported by various types of evidence. Using RWRH algorithm to integrate gene network and phenotype network would be a promising way to identify disease–disease relationship, because both the gene network and phenotypic data are noisy and incomplete and the RWRH algorithm well captures the dependence between two data sources.

Recently, genome wide association studies (GWAS) have been generally used to detect allelic variations that affect susceptibility to complex diseases. A number of bioinformatics algorithms have been proposed to identify disease-related single nucleotide polymorphism (SNP) from GWAS data, including gene-set-based approach (Wang *et al.*, 2007), text-based approach (Raychaudhuri *et al.*, 2009) and pathway-based approach (Eleftherohorinou *et al.*, 2009). The RWRH algorithm can also be used to prioritize candidate genes obtained from GWAS data. We start from the selected candidate SNPs. Candidate genes are seemed as the neighboring genes of selected candidate SNPs. After prioritization, both disease gene and the corresponding SNP can be obtained.

The proposed RWRH algorithm relies on the topology of the heterogeneous network, therefore the low-quality of gene network, phenotype network and gene–phenotype network may limits its performance. The PPI network suffers both high false positive and false negative. Integrating multiple data sources may overcome this limitation. There are some possible integration strategies: (i) to construct a gene functional network by combining multiple genomic data sources (Linghu *et al.*, 2009); (ii) to construct a gene network based on each data source, and then run RWRH algorithm to get a ranking list of candidate genes, finally combine multiple rank lists in to one (Aerts *et al.*, 2006; Li and Patra, 2010); (iii) to construct a heterogeneous network including information of multiple genomic data sources, which means there are possibly more than one links between two genes, and the transition matrix [M in Equation (1)] is determined by multiple data sources.

The phenotype network is also problematic. The similarity between two phenotypes entities are calculated based on the text description in OMIM (Hamosh *et al.*, 2005). But OMIM does not use a controlled vocabulary and is heavily underannotated (Oti *et al.*, 2009). Recently, the ontological description of OMIM phenotypes has been proposed (Robinson *et al.*, 2008). With the availability of well-annotated phenotype data, a higher quality phenotype network may be obtained by using suitable ontological similarity measure.

ACKNOWLEDGEMENTS

We would like to thank Dr Martin Oti and Prof. Han G. Brunner from Radboud University Niumegen Medical Centre for providing the MimMiner program. We would also like to thank three anonymous reviewers for valuable suggestions.

Conflict of Interest: none declared.

REFERENCES

Adie,E.A. *et al.* (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, **22**, 773–774.

Aerts,S. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.

Anne,M.G. *et al.* (2002) Finding genes that underlie complex traits. *Science*, **298**, 2345–2349.

Biesecker,L. (2005) Mapping phenotypes to language: a proposal to organize and standardize the clinical descriptions of malformations. *Clin. Genet.*, **68**, 320–326.

Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat. Genet.*, **33** (Suppl.), 228–237.

Cole,A. *et al.* (2008) Relative resistance of CDK5-phosphorylated CRMP2 to dephosphorylation. *J. Biol. Chem.*, **283**, 359–375.

Eleftherohorinou,H. *et al.* (2009) Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PLoS One*, **4**, 359–375.

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

Gelb,B.D. and Tartaglia,M. (2006) Noonan syndrome and related disorders: dysregulated RAS-mitogen activated protein kinase signal transduction. *Hum. Mol. Genet.*, **15**, R220–R226.

Giallourakis,C. *et al.* (2005) Disease gene discovery through integrative genomics. *Annu. Rev. Genomics Hum. Genet.*, **6**, 381–406.

Goh,K.-I. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Hamosh,A. *et al.* (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33** (Database Issue), D514–D517.

Herz,J. (2009) Apolipoprotein E receptors in the nervous system. *Curr. Opin. Lipidol.*, **20**, 190–196.

Knoers,N.V. and Levtchenko,E.N. (2008) Gitelman syndrome. *Orphanet J. Rare Dis.*, **3**, Article 22.

Köhler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Lee,D.S. *et al.* (2008) The implications of human metabolic network topology for disease comorbidity. *Proc. Natl Acad. Sci. USA*, **105**, 9880–9885.

Linghu,B. *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.

Li,Y. and Agarwal,P. (2009) A pathway-based view of human diseases and disease relationships. *PLoS ONE*, **4**, e4346.

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

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

McKusick,V.A. (1969) On lumpers and splitters, or the nosology of genetic disease. *Perspect. Biol. Med.*, **12**, 298–312.

Oti,M. and Brunner,H.G. (2007) The modular nature of genetic diseases. *Clin. Genet.*, **71**, 1–11.

Oti,M. *et al.* (2008) Phenome connections. *Trends Genet.*, **24**, 103–106.

Oti,M. *et al.* (2009) The biological coherence of human phenome databases. *Am. J. Hum. Genet.*, **85**, 801–808.

Perez-Iratxeta,C. *et al.* (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316–319.

Peri,S. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.

Raychaudhuri,S. *et al.* (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.*, **5**, 359–375.

Robinson,P. *et al.* (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**, 610–615.

Rzhetsky,A. *et al.* (2007) Probing genetic overlap among complex human phenotypes. *Proc. Natl Acad. Sci. USA*, **104**, 11694–11699.

Smedley,D. *et al.* (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, Article 22.

Takahashi,K. *et al.* (2008) Prolyl isomerase, Pin1: new findings of post-translational modifications and physiological substrates in cancer, asthma and Alzheimer's disease. *Cell. Mol. Life Sci.*, **65**, 359–375.

Traboulsi,E.I. *et al.* (2006) Lumpers or splitters? the role of molecular diagnosis in Leber congenital amaurosis. *Ophthalmic Genet.*, **27**, 113–115.

Turner,F. *et al.* (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, **4**, R75.

van Driel,M.A. *et al.* (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, **14**, 535–542.

Wang,K. *et al.* (2007) Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.*, **81**, 1278–1283.

Wijsman,E.M. *et al.* (2004) Evidence for a novel late-onset alzheimer disease locus on chromosome 19p13.2. *Am. J. Hum. Genet.*, **75**, 398–409.

Wu,X. *et al.* (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol.*, **4**, Article 189.

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

Zhang,P. *et al.* (2007) DNA damage responses in neural cells: Focus on the telomere. *Neuroscience*, **145**, 1439–1448.

Zhou,D. *et al.* (2007) Co-ranking authors and documents in a heterogeneous network. In *IEEE International Conference on Data Mining (ICDM 2007)*. IEEE Computer Society Washington, DC, USA, pp. 739–744.