<table>
<thead>
<tr>
<th>Title</th>
<th>Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Li, Yongjin; Patra, Jagdish Chandra</td>
</tr>
<tr>
<td>Date</td>
<td>2010</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10220/7252">http://hdl.handle.net/10220/7252</a></td>
</tr>
<tr>
<td>Rights</td>
<td>© 2010 The Author.</td>
</tr>
</tbody>
</table>
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 encompasses 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., 2005). Using 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. The algorithm prioritizes the genes and 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 over 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

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 (Giannourakis 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 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 encompasses 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., 2005). Using 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. The algorithm prioritizes the genes and 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 over 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

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 (Giannourakis 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 encompasses 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., 2005). Using 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
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 obsolete 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 was 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 $\mathcal{A}_{\text{gene}}$, $\mathcal{A}_{\text{phenotype}}$ and $\mathcal{R}_{\text{bipartite}}$ 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 $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{\text{gene}} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{A}_{\text{phenotype}} \end{bmatrix}$, where $\mathbf{B}^T$ represents the transpose of $\mathbf{B}$.

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 $p_i$ be the initial probability vector and $\mathbf{p}$ 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 $s$ can be given by

$$p_{s+1} = (1 - \gamma)\mathbf{M}^s \mathbf{p} + \gamma \mathbf{p}_0,$$

where $\mathbf{M}$ is the transition matrix of the graph. $\mathbf{M}_i$ 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 $\gamma$. After some steps, the probability will reach a steady state. This is obtained by performing the iteration until the difference between $p_i$ and $p_{s+1}$ (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 $p_{i\infty}(i) > p_{s+1}(i)$, then node $i$ is more proximate to seed nodes than node $j$.

Let $\mathbf{M} = \begin{bmatrix} \mathbf{M}_c & \mathbf{M}_{c\rightarrow p} \\ \mathbf{M}_{p\rightarrow c} & \mathbf{M}_p \end{bmatrix}$ be the transition matrix of the heterogeneous network, where $\mathbf{M}_c$ and $\mathbf{M}_p$ are intra-subnetwork transition matrix and $\mathbf{M}_{c\rightarrow p}$, $\mathbf{M}_{p\rightarrow c}$ are inter-subnetwork transition matrix. Let $\lambda$ be the jumping probability, that is the probability of the random walker jumping from gene network to phenotype network or vice 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 $\lambda$, 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 $g_j$ can be described as

$$\begin{cases} \lambda p_{c\rightarrow p}(ji) = p_{c\rightarrow p}(gg) = \frac{1}{\sum\lambda B_{ij}} B_{ij}, & \text{if } \sum\lambda B_{ij} \neq 0, \\
0, & \text{otherwise}. \end{cases}$$

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

$$\begin{cases} \lambda p_{p\rightarrow c}(ij) = p_{p\rightarrow c}(gg) = \frac{1}{\sum\lambda B_{ij}} B_{ij}, & \text{if } \sum\lambda B_{ij} \neq 0, \\
0, & \text{otherwise}. \end{cases}$$

The element of $\mathbf{M}_{c\rightarrow p}$ at $i$-th row and $j$-th column is $p_{c\rightarrow p}(ij)$, the probability of the random walker transition from $g_i$ to $g_j$. It is defined as

$$\begin{cases} \lambda A_{c\rightarrow p}(ij) = \sum\lambda A_{c\rightarrow p}(ii) \frac{1}{\lambda}, & \text{if } \sum\lambda A_{c\rightarrow p}(ii) \neq 0, \\
0, & \text{otherwise}. \end{cases}$$
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 8,919 genes. The phenotypic similarity matrix between 5,080 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 from this phenotype to disease genes and use this phenotype as LOO1 in Table 1.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>LOO1</th>
<th>LOO2</th>
<th>ab initio</th>
</tr>
</thead>
<tbody>
<tr>
<td>RWRH</td>
<td>814</td>
<td>245</td>
<td>201</td>
</tr>
<tr>
<td>CIPHER-SP</td>
<td>709</td>
<td>153</td>
<td>140</td>
</tr>
<tr>
<td>CIPHER-DN</td>
<td>765</td>
<td>165</td>
<td>157</td>
</tr>
</tbody>
</table>

Table 1. In comparison with CIPHER

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 Experiments and Results

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 8,919 genes. The phenotypic similarity matrix between 5,080 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 from this phenotype to disease genes and use this phenotype as LOO1 in Table 1.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>LOO1</th>
<th>LOO2</th>
<th>ab initio</th>
</tr>
</thead>
<tbody>
<tr>
<td>RWRH</td>
<td>814</td>
<td>245</td>
<td>201</td>
</tr>
<tr>
<td>CIPHER-SP</td>
<td>709</td>
<td>153</td>
<td>140</td>
</tr>
<tr>
<td>CIPHER-DN</td>
<td>765</td>
<td>165</td>
<td>157</td>
</tr>
</tbody>
</table>

Table 1. In comparison with CIPHER

3.2 Effect of Parameters

There are three parameters in our algorithm $\gamma$, $\lambda$, and $\eta$. The parameter $\gamma$ 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 $\gamma$ at 0.7.

The parameter $\lambda$ is the jumping probability. It controls the reinforcement between gene network and phenotype network. Large $\lambda$ introduce more mutual dependence of rankings between genes and phenotypes. To investigate the effect of this parameter, we set various values of $\lambda$, 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 $\lambda$ value. When $\lambda$ ranges from 0.5 to 0.9, the performance becomes stable. If the $\lambda$ 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 $\lambda$ 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 $\eta$ controls the impact of two kinds of seed nodes, seed phenotypes and seed genes. If $\eta$ is 0.5, two subnetworks are equally weighted. If $\eta$ is above 0.5, the random walker prefer to return to the phenotypic seed nodes, therefore, the phenotypic gene is given more importance.

We plung the transition matrix $M$ and initial probability $p_0$ into the iterative equation [Equation (1)]. After some steps, the steady probability $p_\infty$ is obtained. Then genes and phenotypes are ranked based on the steady probability $p_{\infty}$ and $v_{\infty}$, respectively.
Table 2. Effect of $\lambda$ value

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>LOO1</th>
<th>LOO2</th>
<th>$ab\ initio$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>789</td>
<td>196</td>
<td>192</td>
</tr>
<tr>
<td>0.3</td>
<td>804</td>
<td>217</td>
<td>196</td>
</tr>
<tr>
<td>0.5</td>
<td>814</td>
<td>245</td>
<td>201</td>
</tr>
<tr>
<td>0.7</td>
<td>815</td>
<td>257</td>
<td>203</td>
</tr>
<tr>
<td>0.9</td>
<td>811</td>
<td>261</td>
<td>203</td>
</tr>
</tbody>
</table>

Table 3. Effect of $\eta$ value

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>LOO1</th>
<th>LOO2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>808</td>
<td>239</td>
</tr>
<tr>
<td>0.3</td>
<td>813</td>
<td>241</td>
</tr>
<tr>
<td>0.5</td>
<td>814</td>
<td>245</td>
</tr>
<tr>
<td>0.7</td>
<td>817</td>
<td>242</td>
</tr>
<tr>
<td>0.9</td>
<td>820</td>
<td>244</td>
</tr>
</tbody>
</table>

network play a more important role in the prioritization of disease genes. To find the effect of $\eta$ value, we run the RWRH algorithm with different $\eta$ values, and calculate the first two measures using leave-one-out cross-validation. As seen from Table 3, the algorithm performs slightly better when $\eta$ 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).

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 611073 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 608907 describes the phenotype of late onset familial Alzheimer’s disease. Wijman 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 (Herr, 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 609636 describes the phenotype of early-onset familial Alzheimer’s disease and the corresponding susceptible 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 163950) 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 $x_{d_i}$ 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 $d_j$ 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 relationships, 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 ophthalmological 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 Levitchenko, 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 phenotype 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 Nijmegen 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


