Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 416323, 10 pages
http://dx.doi.org/10.1155/2014/416323

Review Article

Approaches for Recognizing Disease Genes Based on Network

Quan Zou,[1] Jinjin Li,[1] Chunyu Wang,[2] and Xiangxiang Zeng[1]

[1] *School of Information Science and Technology, Xiamen University, Xiamen 361005, China*
[2] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China*

Correspondence should be addressed to Xiangxiang Zeng; xzeng@xmu.edu.cn

Received 6 December 2013; Revised 6 January 2014; Accepted 9 January 2014; Published 24 February 2014

Academic Editor: Tao Huang

Diseases are closely related to genes, thus indicating that genetic abnormalities may lead to certain diseases. The recognition of disease genes has long been a goal in biology, which may contribute to the improvement of health care and understanding gene functions, pathways, and interactions. However, few large-scale gene-gene association datasets, disease-disease association datasets, and gene-disease association datasets are available. A number of machine learning methods have been used to recognize disease genes based on networks. This paper states the relationship between disease and gene, summarizes the approaches used to recognize disease genes based on network, analyzes the core problems and challenges of the methods, and outlooks future research direction.

1. Introduction

Although the human genome project has been accomplished and has achieved great success, and new methods that verify gene function with high-throughput have been applied, studying genetic problems that induce diseases is still one of the major challenges facing humanity [1]. The traditional gene mapping method is based on family genetic disease. First, genes inducing diseases are located in a chain interval. Most recent studies at recognizing disease gene that involves linkage analysis or association studies have resulted in a genomic interval of 0.5 cm to 10 cm, which contains 300 genes [2, 3]. Second, using the biological experiment method to identify each gene located in a chain interval requires a large number of human resources and capital support [4]. In addition, recognizing disease gene by checking the genes set in the interval is often not possible [5]. However, the study of candidate association works well when using a set of known functional candidate genes, which have a clear biological relationship to the disease [6]. Selecting known functional candidate genes is not easy and is often limited by a good deal of factors. The selection of functional candidate genes and prioritization candidate genes has been one of the keys in recognizing disease genes because several reorganization approaches are based on the functions of these genes.

In recent years, a number of recognizing disease gene approaches and computer tools have been developed through building mathematic models based on functional annotation, sequence-based features, protein interaction, and disease phenotype [7–16], such as sequence features [15], functional annotation [7, 8, 10, 13], and physical interactions [12, 13, 17]. Based on the above features, an approach to rank candidate disease genes by computing a correlation score that stands for the correlation between genes and diseases has been introduced. However, various factors may affect the association between genes and diseases.

System biology has indicated that diseases with overlapping clinical manifestations are induced by one or more mutations from the same function module [18–21]. Researches in biological experiments of human disease and patterns have found that genes causing similar disease phenotype often interact with each other directly or indirectly [22–24]. These discoveries have shown that positive correlation exists between disease phenotype and disease gene. Many researchers have proposed disease gene prediction methods based on gene interaction and disease phenotype similarity [7, 25–29]. Recently, many approaches making full use of gene interaction and disease similarity have established the gene interaction and disease phenotype similarity network

to predict disease genes. Some typical methods based on networks will be introduced in detail in this paper.

2. Datasets

In the field of biological information, construction dataset is the data foundation of all subsequent work. The validity of datasets directly affects the validity and reliability of the learning algorithm and test. Thus, building a dataset is a basic and important preparatory work.

The recognition of disease gene datasets is mainly obtained from two databases: Online Mendelian Inheritance in Man (OMIM), which is a synthesis database [30–32], and Human Protein Reference Database (HPRD) [33, 34]. Although none of the datasets from OMIM or HPRD are currently complete, they are comprehensive enough [6].

OMIM has the most abundant information, most extensive resources, most comprehensive, authoritative, and timely human genes and genetic disorders based on knowledge composed to support human genetics research and education and the clinical genetics research. OMIM is daily updated and has free access and acquisition at http://www.omim.org/. In OMIM, each item has a short text summary of a generally determined phenotype or gene and a large number of links to other genetic databases [30]. Datasets of disease phenotype and gene-disease phenotype can be obtained from OMIM. However, the data from OMIM need to be disposed to recognize the disease genes [35].

HPRD is a database which curetted proteomic information suited to human proteins. Even though HPRD is updated relatively slow, it is a full-scale resource for studying the relationship between human diseases and genes [36] and is linked to an outline of human signaling paths. HPRD is also available for free at http://www.hprd.org/ [34]. The dataset of gene interaction can be obtained from HPRD [35].

3. Networks

Most research on recognizing disease genes use networks, including the disease phenotype network, protein-protein interaction network, and gene-disease phenotype network, among others. In this study, we introduce only the most commonly used networks. G_{PPI} represents the gene (proteins) interaction network, G_{DP} represents the disease phenotype network, and $G_{P\text{-}DP}$ represents the gene-disease phenotype network [6, 16].

(1) $G_{PPI} = (G, E_G)$ is an undirected graph and denotes the gene-gene interaction. $G = \{g_1, g_2, \ldots, g_n\}$ is the subset of the gene set, and $E_G \subset G \times G$ expresses the interaction of genes with weight. Figure 1(a) shows G_{PPI}. In the gene-gene interaction network, the relationship between genes is obtained from the gene-gene relationship database, which is one of the most important databases in the biological information field.

(2) $G_{DP} = (D, E_D)$ is also an undirected graph and denotes the disease phenotype network. $D = \{d_1, d_2, \ldots, d_n\}$ is the subset of the disease phenotype set, and $E_D \subset D \times D$ represents the similarity of the disease phenotype with weight. Figure 1(b) shows G_{DP}. In the disease phenotype network, the relationship between disease phenotypes is obtained from the phenotype relationship database, which can also be replaced by the disease-disease relationship database.

(3) $G_{P\text{-}DP} = (G, D, E_{P\text{-}DP})$, which is an undirected biograph, is a gene-disease phenotype network. G is the subset of the gene set, and D is the subset of the disease phenotype set. $E_{P\text{-}DP} \subset G \times D$ expresses the link between the known gene and the disease phenotype. Figure 1(c) stands for $G_{P\text{-}DP}$. The association between disease gene and disease phenotype can be obtained from the gene-disease relationship database.

4. Methods

In previous research, various methods, such as CIPHER, RWRH, Prince, Meta-path, Katz, Catapult, Diffusion Kernel [5], and ProDiGe, were used to recognize disease genes. In the current paper, we introduce several types of typical recognition disease gene methods.

4.1. CIPHER. CIPHER [6] is a tool for predicting and prioritizing disease genes. Furthermore, CIPHER is applied to general genetic phenotypes, which do better in the genome-wide scan of disease genes; furthermore, they are extendable for exploring gene cooperatives in complex diseases. CIPHER is based on an assumption that if two genes have the closest connection in the gene interaction, then the two genes can lead to more similar phenotypes. A regression model can be formulated according to this assumption. A score assessing how likely a gene is associated with a specific phenotype is obtained from the regression model. To construct the regression model, the similarity between phenotypes, interaction between proteins and genes, and list of associations between known disease gene and phenotype must be prepared. The next paragraph expresses the procedures of prioritizing disease genes.

For a given query phenotype and candidate genes, CIPHER first combines the gene interaction network, disease phenotype network, and gene-disease phenotype network into a single network. The similarity scores of the query phenotype with all known phenotypes in the disease phenotype network are derived directly from the phenotype network and the topological distances between the candidate genes. All known disease genes in the gene interaction network are counted and grouped on the basis of their phenotypes. The correlation between phenotypes and disease genes is obtained and acts as the concordance score for each candidate gene by using the regression model. Finally, all candidate genes for the query phenotype are ranked in line with the concordance scores. On account of different neighborhood systems, two ways are available to define the topological distance: direct neighbor and shortest path. Thus, there are two versions of CIPHER which are CIPHER-SP and CIPHER-DN [6].

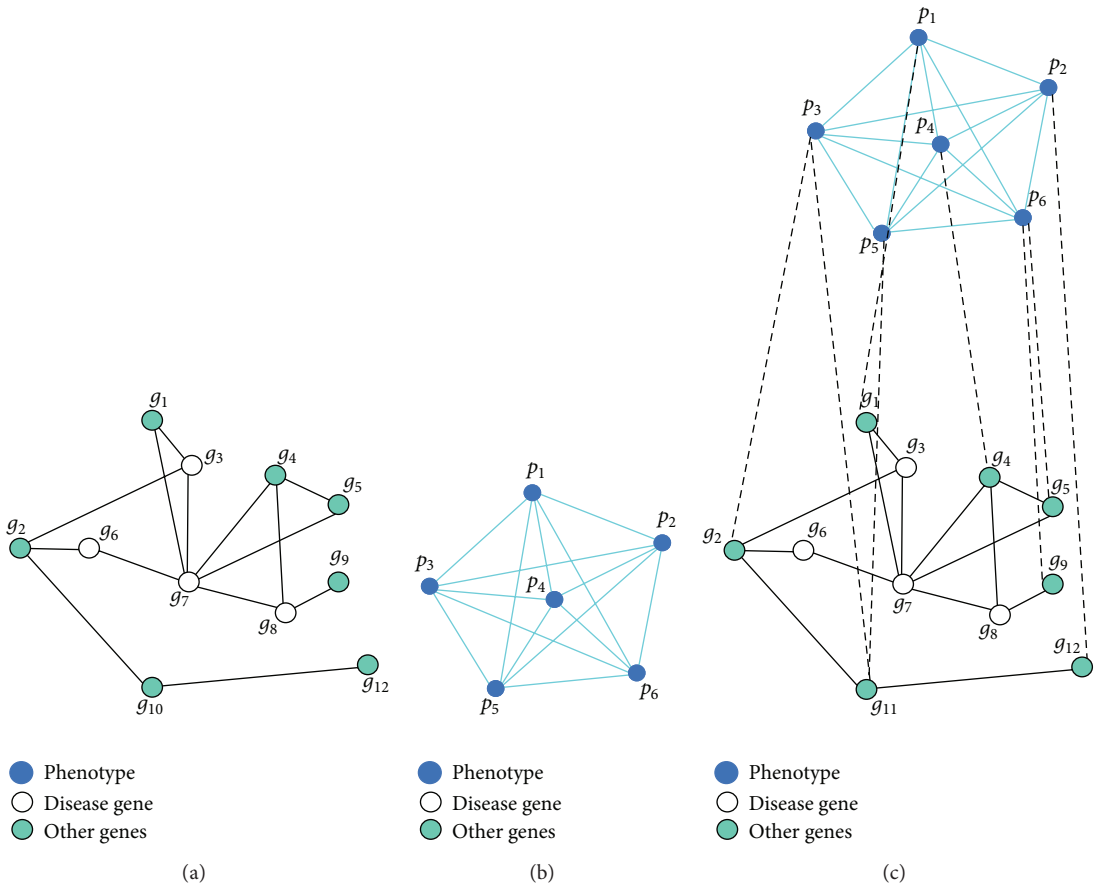

Figure 1: Illustration of the network by using a specific example: (a) is a gene-gene network, (b) is a disease phenotype network, and (c) is a gene-disease phenotype network [6].

The similarity scores of the query phenotype and all phenotypes in the disease phenotype network are calculated by the following formulation:

$$S_{pp'} = C_p + \sum_{g \in G(p)} \sum_{g' \in G(p')} \beta_{pg} e^{-L_{gg'}^2}. \quad (1)$$

In the above formulation, $S_{pp'}$ is the similarity score of the query phenotype and another phenotype in the disease phenotype network. $L_{gg'}$ is the topological distance between the candidate genes and g' in the gene interaction network. There exit two ways to define the topological distance, $L_{gg'}$, on the basis of how to consider indirect the interaction. One way to define the topological distance is shortest path; $L_{gg'}$ is the graph theory shortest path length between genes g and g' in the gene interaction network. The other way to define the topological distance is direct neighbor; $L_{gg'}$ is infinity when g and g' are indirect neighbors. $G(p)$ indicates all disease genes belonging to phenotype p. C_p is a constant and can act as the basal similarity between p and other phenotypes whose causative genes are not connected to those of p in the gene interaction network. β_{pg} is the coefficient of the regression model and stands for the level of gene g contributing to the similarity of the phenotype p to any other phenotype p'. To denote the association between a phenotype and a gene, the following formulation (2) is defined:

$$\Phi_{gp'} = \sum_{g' \in G(p')} e^{-L_{gg'}^2}. \quad (2)$$

The following vector is used to denote the similarities between the query phenotype and all phenotypes in the disease phenotype network: $S_p = (S_{pp_1}, S_{pp_2}, S_{pp_3}, \ldots, S_{pp_n})$.

In the same way, the following vector is used to denote the closeness between the genes and the phenotypes in the disease phenotype network: $\Phi_g = (\Phi_{gp_1}, \Phi_{gp_2}, \Phi_{gp_3}, \ldots, \Phi_{gp_n})$. Synthesizing Formulas (1) and (2) and two vectors extends Formula (1) to the following form:

$$S_p = C_p + \sum_{g \in G(p)} \beta_{pg} \Phi_g. \quad (3)$$

In this regression model, the concordance score is defined by Formula (4):

$$CS_{pg} = \frac{\operatorname{cov}(S_p, \Phi_g)}{\sigma(S_p)\sigma(\Phi_g)}, \quad (4)$$

where cov and σ are the covariance and standard deviation, respectively. The candidate genes for the query phenotype are

ranked according to the values obtained from Formula (4). If a gene that does not connect to any disease genes exists, then Formula (4) cannot be used and the gene will rank at the tail.

4.2. PRINCE. PRINCE is another approach based on networks for ranking candidate disease genes for a given disease and inferring the complex associations between genes. PRINCE is on account of formulating constraints on the ranking function that involved usage of prior information and its smoothness over the network.

Before using PRINCE, the gene disease composed of the phenotype network (set of gene-disease associations), gene-gene interaction network (set of gene-gene association), and at least a query disease phenotype is prepared. $G = (V, E, w)$ denotes the gene-gene interaction network, where V is the set of genes, E is the set of interactions, and w is a weight function denoting the reliability of each interaction. Given a query disease phenotype, PRINCE ranks all the genes in V.

Suppose a gene $v \in V$, the direct neighborhood of gene v is denoted by $N(v)$. The prioritization candidate disease gene function is denoted by $F : V \rightarrow \mathfrak{R}$, and $F(v) = q$ reflects the relevance of v to q. Another function is defined as prior knowledge function denoted by $Y : V \rightarrow \{0, 1\}$. In the prior knowledge function, gene v is related to $q, V(v) = 1$; otherwise, $V(v) = 0$. PRINCE computes function F that is smooth over the network. Thus, function F is a combination of two conditions:

$$F(v) = \alpha \left[\sum_{u \in N(v)} F(u)\, w'(v, u) \right] + (1 - \alpha)\, Y(v), \quad (5)$$

where the parameter $\alpha \in (0, 1)$ weighs the relative importance of gene v to gene u. w' is a normalized form of w. Formally, a diagonal matrix D is defined, and $D(i, i)$ is the sum of row i of W. W is normalized by $W' = D^{-1/2} W D^{-1/2}$, which obtains a symmetric matrix. Here, $W'_{ij} = W_{ij} / \sqrt{D(i, i) D(j, j)}$. Formula (5) can be expressed in linear form as follows:

$$F = \alpha W' F + (1 - \alpha)\, Y \Longleftrightarrow F = \left(I - \alpha W'\right)^{-1} (1 - \alpha)\, Y, \quad (6)$$

where F and V are viewed as vectors of size $|V|$. W' is a matrix whose values are given by w'. Given that the eigenvalues of W' are set in $[-1, 1]$, $\alpha \in (0, 1)$, and the eigenvalues of $(I - \alpha W')$ are positive. In addition, $(I - \alpha W')^{-1}$ exists.

The above linear system can be solved accurately because an iterative propagation-based algorithm works fast and is guaranteed to converge to the system solution for larger networks. Formula (6) is transferred to an iterative algorithm and is denoted as follows:

$$F^t = \alpha W' F^{t-1} + (1 - \alpha)\, Y, \quad (7)$$

where $F^1 = Y$. Every node propagates the information received in the previous iteration to its neighbors. Finally, the values obtained from Formula (7) rank all the candidate disease genes for a query disease phenotype.

4.3. RWRH. Random walk with restart on heterogeneous network (RWRH) is extended from the random walk with restart algorithm to the heterogeneous network. The heterogeneous network is constructed by connecting the gene-gene interaction network and disease phenotype network by using the gene-disease phenotype relationship information. In brief, the gene-disease phenotype network is the heterogeneous network. RWRH prioritizes the genes and the phenotypes simultaneously, which is inspired by the coranking framework [37]. Given a query disease, seed nodes as genes and phenotypes are associated with the disease, and the top ranked phenotype is the most similar to the query disease.

Random walk is defined as an iterative walker transition from its current node to a randomly selected neighbor, starting at a given source node v in the network. However, RWRH allows the restart of the walk in every time step at node v with probability r. P_0 is the probability vector at step 0, indicating that it is the initial probability vector with the sum of probabilities equal to 1. Similarly, P_s is the probability vector at step s, in which the ith element holds the probability of finding the random walker at node i at step s. The probability vector at step $s + 1$ is denoted as follows:

$$P_{s+1} = (1 - \gamma)\, M^T P_s + \gamma P_0, \quad (8)$$

where M is the transition matrix of the heterogeneous network; M_{ij} is the transition probability from node i to node j; $\gamma \in (0, 1)$ is the restart probability in every time step. After several iterations, P_∞ reaches a steady-state that is obtained by performing the iteration until the change between P_s and P_{s+1} falls below 10^{-10}. P_∞ is the measure of closing to seed nodes. In vector P_∞, when $P_\infty(i) > P_\infty(j)$, node i is more likely to be the seed node than node j.

M is the transition matrix of the heterogeneous network. In addition, M consists of four subnetwork transition networks and is denoted as follows:

$$M = \begin{bmatrix} M_G & M_{GP} \\ M_{PG} & M_p \end{bmatrix}, \quad (9)$$

where M_G is the transition matrix of the gene-gene interaction network, which is the intrasubnetwork of the heterogeneous network. M_p is the transition matrix of the disease phenotype network, which is also the intrasubnetwork of the heterogeneous network. M_{PG} and M_{GP} are the intersubnetwork transition matrixes. Supposing the probability of jumping from gene-gene interaction network to the disease phenotype network is λ, the reverse is the same. In the gene-gene interaction network, $\lambda = 0$ if a node is not connected to the phenotype. If a node is directly linked to the disease phenotype network, then the node will jump to the disease phenotype network with probability λ. The node will jump to other nodes in the gene-gene interaction network with probability $1 - \lambda$. Thus, the transition probability from g_i to p_j can be denoted as follows:

$$\left(M_{GP}\right)_{i,j} = P\left(p_j \mid g_i\right) = \begin{cases} \dfrac{\lambda B_{ij}}{\sum_j B_{ij}}, & \text{if } \sum_j B_{ij} \neq 0 \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

In the same way, the transition probability from p_i to g_j can be denoted as follows:

$$(M_{PG})_{i,j} = P(g_j \mid p_i) = \begin{cases} \dfrac{\lambda B_{ji}}{\sum_j B_{ji}}, & \text{if } \sum_j B_{ji} \neq 0 \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

The transition probability from g_i to g_j, which is the element of M_G at the ith row and jth column, can be denoted as follows:

$$(M_G)_{i,j} = \begin{cases} \dfrac{(A_G)_{i,j}}{\sum_j (A_G)_{i,j}}, & \text{if } \sum_j B_{ij} = 0 \\ \dfrac{(1-\lambda)(A_G)_{i,j}}{\sum_j (A_G)_{i,j}}, & \text{otherwise.} \end{cases} \quad (12)$$

The transition probability from p_i to p_j, which is the element of M_P at the ith row and the jth column, can be denoted as follows:

$$(M_P)_{i,j} = \begin{cases} \dfrac{(A_P)_{i,j}}{\sum_j (A_P)_{i,j}}, & \text{if } \sum_j B_{ij} = 0 \\ \dfrac{(1-\lambda)(A_P)_{i,j}}{\sum_j (A_P)_{i,j}}, & \text{otherwise.} \end{cases} \quad (13)$$

In the above four formulations, $A_{G(n\times n)}$, $A_{P(m\times m)}$, and $B_{(n\times m)}$ are the adjacency matrixes for the gene-gene interaction network, disease phenotype network, and gene-disease phenotype network, respectively. The adjacency matrix of the heterogeneous network can be denoted as follows:

$$A = \begin{bmatrix} A_G & B \\ B^T & A_P \end{bmatrix}. \quad (14)$$

The initial probability of the gene-gene interaction network and phenotype network is denoted by u_0 and v_0, respectively. The initial probability of the gene network u_0 makes the equal probabilities to all the seed nodes in the gene network, with the sum of the probabilities equal to 1. The initial probability of the phenotype network v_0 is the same as the gene-gene interaction network. Thus, the initial probability vector of the heterogeneous network is denoted as $P_0 = [(1-\eta)u_\infty \ \ \eta v_\infty]^T$. In the initial probability vector of the heterogeneous network, the parameter $\eta \in (0, 1)$ acts as the judge to weight the importance of each subnetwork. When $\eta = 0.5$, the importance of the gene-gene interaction network and the disease phenotype network are equal. If $\eta > 0.5$, then the importance of the gene-gene interaction network is greater than the disease phenotype network. When $\eta < 0.5$, the gene-gene interaction network is more important than the disease phenotype network is. P_0 and the transition matrix M are substituted into Formula (8). After many iterations, steady-state P_∞ is denoted as $P_\infty = [(1-\eta)u_\infty \ \ \eta v_\infty]^T$. In this way, the steady probabilities u_∞ and v_∞ are used to rank the genes and disease phenotypes. A web server named GeneWanderer, which is a computational method that prioritizes a set of candidate genes according to their probability to become involved in a particular disease or phenotype using HWRH or diffusion kernel, is used.

4.4. Katz. The Katz method is successfully applied to social network link prediction. Predicting the social network link is close to the problem of predicting disease genes. The Katz approach, which is based on a graph, finds the similar nodes for the query nodes in the network [38].

An adjacency matrix A is available in an undirected unweighted graph. The Katz approach counts the number of walks of different lengths that connects i and j. These walks act as the similarity of the two nodes i and j. $(A^l)_{ij}$ is the number of walks of length l that connect i and j. $(A^l)_{ij}$ gives a measure of similarity between i and j. A single similarity measure based on the different walk lengths is necessary. The measure is given below, in which β is a constant that restrains contributions of longer walks:

$$S_{ij} = \sum_{l=1}^{k} \beta_l \left(A^l\right)_{ij}. \quad (15)$$

The above measure is denoted as follows:

$$S = \sum_{l=1}^{k} \beta_l A^l. \quad (16)$$

If $l \to \infty$, $\beta_l \to 0$. In this study, setting $\beta_l = \beta^l$ leads to the well-known Katz method:

$$S^{\text{katz}} = \sum_{l\geq 1} \beta^l A^l = (I - \beta A)^{-1} - I, \quad (17)$$

where β is chosen, such that $\beta < 1/\|A\|^2$. In the case of the Katz method, the connections in the graph are weighed so that A_{ij} is the strength of the connection between nodes i and j. For the choice of k, the sum over infinitely many path lengths is not necessarily considered. According to the experimental results, small values of k ($k = 3$ or $k = 4$) obtain good performance in the task of recommending similar nodes.

The adjacency matrix of the heterogeneous network is denoted as follows:

$$A = \begin{bmatrix} A_G & B \\ B^T & A_P \end{bmatrix}. \quad (18)$$

One of the advantages of Katz is A, which can represent the other species if we want to study human disease phenotypes and other species disease phenotypes.

$$B = \begin{bmatrix} P_{HS} & P_S \end{bmatrix}, \qquad A_P = \begin{bmatrix} A_{PHS} & 0 \\ 0 & A_{PS} \end{bmatrix}. \quad (19)$$

Here, A_{PHS} and A_{PS} represent human phenotypes and the other species phenotypes, respectively. P_{HS} and P_S indicate gene-disease phenotype association of human and other species, respectively. When an experiment on human is conducted, set $P_S = 0$ and $A_{PS} = 0$. By synthesizing expressions (18) and (19), we substitute matrix A into Formula (17) and obtain the similarity of genes and phenotypes from the similarity matrix.

Initialize $\theta = 0$, $s(x) = 0$, $\forall x \in U$, and $n(x) = 1$, $\forall x \in U$
For t $= \mathbf{1}, \mathbf{2}, \mathbf{3}, \ldots, \mathbf{T}$**:**
Step 1. Draw a bootstrap sample $U_t \subseteq U$ of size $n+$
Step 2. Train a linear classifier θ_t using the positive training examples A and U_t as negative examples by solving:

$$\min_{\theta' \in R^d} \frac{1}{2}\left\|\theta'\right\|^2 + C_- \sum_{i \in U_t} \xi_i + C_+ \sum_{i \in A} \xi_i$$

Subject to $\xi_i \geq 0, \quad \forall i \in A \cup U_t$

$$\left\langle \Phi(x_i), \theta' \right\rangle \geq 1 - \xi_i, \quad \forall i \in A$$

$$-\left\langle \Phi(x_i), \theta' \right\rangle \geq 1 - \xi_i, \quad \forall i \in U_t$$

Step 3. For any $x \in U \setminus U_t$ update:
(i) $n(x) \leftarrow n(x) + 1$
(ii) $s(x) \leftarrow s(x) + \left\langle \theta_t, \Phi(x) \right\rangle$
Return $s(x) \leftarrow s(x)/n(x), \quad \forall x \in U$.

Algorithm 1: CATAPULT algorithm description.

4.5. CATAPULT. Combining data across species by using positive-unlabeled learning techniques is abbreviated to CATAPULT. And CATAPULT is a supervised machine learning method which uses a biased support vector machine (SVM), where the features are derived from walks in a heterogeneous gene-trait network.

Given a query disease phenotype, a gene is not associated with the query phenotype. Scholars report positive association between genes and phenotypes; however, the negative associations are rarely reported. In the CATAPULT approach, the unlabeled gene-disease phenotype pairs act as negative associations. The characteristics of the dataset are that only the positive associations are known, and the negative associations and a large number of unlabeled gene-disease phenotype pairs as negative associations are unknown. The general idea of CATAPULT is that the examples are not known to be negative. The false positives are not penalized heavily, but the false negatives are penalized heavily.

CATAPULT uses a biased SVM to classify the gene-phenotype pairs of humans with a single training phase. This approach draws a random bootstrap sample of a few unlabeled examples from the set of all unlabeled examples and trains a classifier to classify the bootstrap samples as negatives along with the positive samples. CATAPULT also uses the bagging technique to obtain an aggregate classifier by using positive and unlabeled examples. The algorithm description is shown in Algorithm 1. T denotes the number of bootstraps, A is the set of positive, n_+ denotes the number of examples in A, U denotes the set of unlabeled gene-disease phenotype pairs, C_- is a penalty for false positives, and C_+ is a relatively larger penalty for false negatives. The source code can be downloaded from http://marcottelab.org/index.php/Catapult.

Before applying any supervised machine learning approach, extracting the features for gene-disease phenotypes is essential. The features are derived from the paths in the heterogeneous network. For a given gene-disease phenotype pair, different walks of the same length and walks of different lengths can be used as features for the gene-disease phenotype pair.

4.6. Meta-Path. The meta-path approach mainly uses the technology of multilabel classification. The multilabel classification method is useful for recognizing disease genes. A gene may exhibit many diseases caused by the gene. In the above example, the gene is an instance, and various diseases are different labels. Given an instance, a large space of all possible label sets may exist, which may be exponential to the number of candidate labels. The frequently used approach to solve the above problem is exploiting correlations among different labels. In the network, exploiting the correlations among different labels denoted by nodes is an advantage.

Meta-path is defined as a sequence of relations in the network. The objects in the network are linked through multiple-type associations. Multiple-type associations help exploit the correlations among different labels for multilabel classification. In recognizing the disease genes, the labels of the genes are diseases, and the labels of the diseases are genes. The explanation of the correlations among genes is summed up in this study.

Given a set of meta-paths among the gene nodes acting as labels, $S_l = \{P_1, P_2, \ldots, P_{cl}\}$, the meta-path-based label correlations can be used as follows:

$$\forall i, \quad P\left(Y_i \mid x_i\right) = \prod_{k=1}^{q} P\left(Y_i^k \mid x_i, Y_i^{P_1(k)}, Y_i^{P_2(k)}, \ldots, Y_i^{P_{cl}(k)}\right), \tag{20}$$

where $P_j(k)$ denotes the index set of the genes linked to the kth gene through the meta-path $P_j \in S_l$. x_i denotes the feature vector of node i in the input space. Y_i denotes the association between a gene and a gene set. The set of all candidate genes is denoted as $V_l = \{l_1, l_2, \ldots, l_q\}$, and Y_i is denoted as $Y_i = (Y_i^1, Y_i^2, \ldots, Y_i^q)^T \in \{0, 1\}^q$. In the same way, given a set of meta-paths among disease phenotype nodes acting as instances, $S_I = \{P_1', \ldots, P_{cI}'\}$, the meta-path-based label correlations can be used as follows:

$$P\left(Y \mid X\right) \approx \prod_i P\left(Y_i \mid x_i, Y_{p_1'(i)}, \ldots, Y_{p_{cI}'(i)}\right), \tag{21}$$

(i) $x_l = LabelPathFeature(l_k, Y_i)$
For each meta-path $P_j' \in S_I$:
(1) Get related labels for node l_k through meta-path P_j.
(2) $x_i = Aggregation(\{Y_i \mid i \in C\})$
Return $(\ldots, x_j^T, \ldots)^T$

(ii) $x_I = LabelPathFeature(i, Y)$
For each meta-path $P_j \in S_I$:
(1) Get related labels for node I_i through meta-path P_j'.
(2) $x_i = Aggregation(\{Y_i \mid i \in C\})$
Return $(\ldots, x_j^T, \ldots)^T$

Algorithm 2: The function of construction relational features for meta-path-based label correlations and meta-path-based instance correlations.

where $P_j'(i)$ denotes the index set of disease phenotypes linked to the ith disease phenotype through meta-path $P_j' \in S_I$.

To perform multilabel collective classification more effectively in heterogeneous information network, both meta-path-based label correlations and meta-path-based instance correlations are performed simultaneously:

$$P(Y \mid X) \approx \prod_i \prod_{k=1}^{q} P\left(Y_i^k \mid x_i, Y_i^{p_j(k)}, \ldots, Y_{P_j'(i)}^k\right), \quad (22)$$

where the gene is set as the label set, and the disease phenotype is set as the instance set. On the contrary, the disease phenotype is set as the label set, and the gene set is as the instance set.

Some research has proposed algorithms based on multilabel collective classification. We briefly introduce the multilabel collective classification algorithm called PIPL. The algorithm roughly includes the following steps.

(1) Meta-path constructions: extracting all nonredundant meta-paths for label correlations and instance correlations.

(2) Training initialization: construction of q extended training sets for all $1 \le k \le q$, $D_k = \{(x_i^k, y_i^k)\}$ by converting each instance x_i to x_i^k by using the functions in Algorithm 2. Training one classifier on each label by using the extended training sets.

(3) Iterative inference: the inference step is an iterative classification algorithm. It updates the testing instance label set predictions and the relational features of label and instance correlations.

5. Evaluation Methods

A comprehensive comparison should be conducted among these methods. In the next several paragraphs, we will introduce some of the key comparisons for recognizing the disease genes reported so far.

Cross-validation is the most frequently used approach in evaluating these methods. However, this method is similar to that used in a previous work, which performs leave-one-out. Each of the known gene-disease phenotype associations is taken as a test case, and a set of genes is assigned as the negative control for each test case. In each round of cross-validation, the disease phenotype is held out, a link between the disease phenotype and one of the associated genes is removed, and the link removed gene is added into the test genes. The rank of the test genes is obtained according to the recognizing methods. Several processing approaches are available for the rank, such as the enrichment score, setting a threshold, precision, recall, and receiver-operating characteristic (ROC).

5.1. Enrichment Score. Suppose the number of test genes is 100. If a recognition disease gene method ranks the actual disease gene as the highest and is sequenced first, then an enrichment of 50-fold exists. The formula of the enrichment score is Enrichment = 50/rank.

5.2. ROC Analysis. ROC analysis denotes the true-positive rate (TPR) versus the false-positive rate (FPR) subject to the threshold dividing the prediction classes. The TPR/FPR is the rate of correctly/incorrectly classified samples of all samples classified to the positive class. To evaluate the scores of disease gene predictions, ROC is explained as a plot of the number of the disease genes above the threshold versus the number of the disease genes below the threshold. The area under the ROC curve for each curve is calculated to compare the different curves obtained by the ROC analysis.

5.3. Setting a Threshold. Concordance score is calculated for each test gene. If the true disease gene ranks first based on the concordance score, then the prediction is successful, and precision is used as the proportion of the successful predictions among all predictions. Another evaluation approach is setting a threshold, in which the highest score of all test genes in this case is not less than the threshold. Thus, *recall* is the fraction of true disease genes predicted among all disease genes.

6. Materials and Results

In the above section, several recognition disease gene methods and evaluation methods have been mentioned. This part introduces the data used and the comparison results.

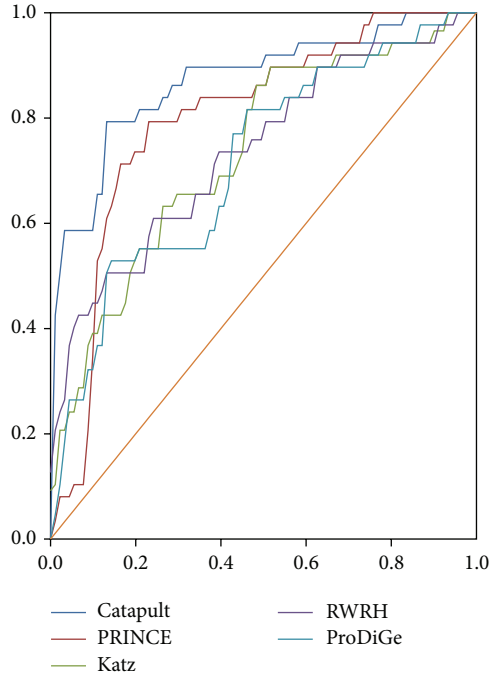

Figure 2: Setting a threshold to compare recognition disease gene methods.

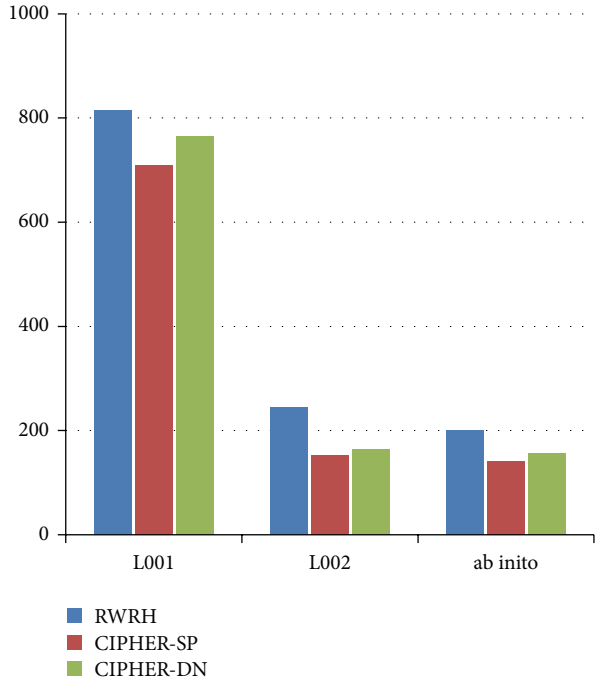

Figure 3: The comparison of different methods.

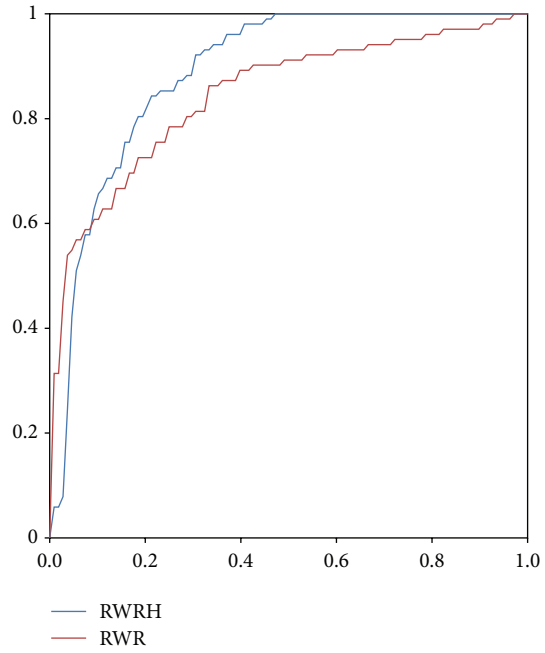

Figure 4: ROC curve of RWR and RWRH.

Figure 2 [17] denotes the comparison of different recognition methods by setting a threshold. In Figure 2, six recognizing methods are shown, including Catapult, Katz, ProDiGe, RWRH, PRINCE, and Degree. HumanNet gene network, which is a part of the OMIM dataset, is the dataset used to compare the different recognition methods. Figure 2 shows that Katz and Catapult do better than the others with the HumanNet gene network and the evaluation method.

RWRH is compared with CIPHER-DN and CIPHER-SP. The evaluation experiment is based on gene network containing 34,364 interactions between 8919 genes, the phenotypic similarity matrix between 5080 phenotype entities calculated by using MimMiner, and 1428 gene-phenotype links between 937 genes and 1216 phenotype entities. The comparison result is denoted by Figure 3 When $\gamma = 0.7$, $\lambda = \eta = 0.5$, RWRH successfully ranks 814 known disease genes as top 1. The result is denoted by L001 in Figure 3. The column of L002 is the result of removing a known gene-phenotype link and using the phenotype and the rest of the disease genes associated with this phenotype as seed nodes. The identification of disease genes for phenotype from the genome is called *ab initio* prediction. The *ab initio* method removes all the links from a phenotype to disease genes and uses the phenotype entity as seed node to run RWRH. If one of the disease genes associated to the phenotype ranks top 1, then the prediction is successful. The result of *ab initio* is shown in Figure 3. From the L001, L002, and *ab initio*, RWRH is better than CIPHER-SP and CIPHER-DN. Figure 4 denotes the result of the comparison between RWR and RWRH. Leave-one-out cross-validation is conducted for each disorder. In each cross-validation, a disease gene is selected, and the links between the phenotype entries and the disease gene are removed. The rest of the disease genes and the phenotype entry are used as seed nodes. The selected disease gene and all disease genes in the artificial linkage are ranked by RWRH and RWR. ROC analysis is used to evaluate the two recognizing approaches.

7. Conclusion

Identifying disease genes is one of the fundaments of medical care and has been a goal in biology. Although traditional linkage analysis and modern high-throughput techniques often provide hundreds of disease gene candidates, identifying disease genes in the candidate genes by using the biological experiment method time-consuming and expensive. To deal with the above issues, the methods based on networks have been proposed. Many methods based on network have been created to recognize disease genes. In this paper, five typical algorithms based on networks, namely, CIPHER, PRINCE, RWRH, Katz, and CATAPULT, are introduced in detail.

Some novel methods have been put forward to recognize and prioritize disease genes. For instance, BRIDGE [39] takes advantage of multiple regression models with penalty to automatically weight different data sources. A researcher employed the ensemble boosting learning technique to combine variant computational approaches for gene prioritization to improve overall performance [40].

Biological relationships are showed by networks, which brings forth new ideas. A network can be used to denote the association between genes and disease to recognize the gene-disease phenotype and to obtain a more complete understanding of the biological system. Networks have been successfully used in biology. However, combining experiments with networks results in the challenge of defining node similarities. Different ways to define node similarity may lead to different effects.

With the development of biology and the emergence of a large number of relevant data, disease gene research based on networks constantly matures. New machine learning methods and technologies will be used to predict disease genes. Research on disease gene recognition will achieve new breakthroughs. The disease gene research will open a new era of medical treatment.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The work was supported by the Natural Science Foundation of China (no. 61370010, no. 61202011, no. 61271346, no. 61172098, and no. 60932008) and the Ph.D. Programs Foundation of Ministry of Education of China (20120121120039).

References

[1] M. D. Adams, J. M. Kelley, J. D. Gocayne et al., "Complementary DNA sequencing: expressed sequence tags and human genome project," *Science*, vol. 252, no. 5013, pp. 1651–1656, 1991.

[2] A. M. Glazier, J. H. Nadeau, and T. J. Aitman, "Genetics: finding genes that underline complex traits," *Science*, vol. 298, no. 5602, pp. 2345–2349, 2002.

[3] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease," *Nature Genetics*, vol. 33, pp. 228–237, 2003.

[4] M. Lin and S. Gottschalk, "Collision detection between geometric models: a survey," in *Proceedings of the IMA Conference on the Mathematics of Surfaces*, 1998.

[5] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[6] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Molecular Systems Biology*, vol. 4, article 189, 2008.

[7] C. Perez-Iratxeta, P. Bork, and M. A. Andrade, "Association of genes to genetically inherited diseases using data mining," *Nature Genetics*, vol. 31, no. 3, pp. 316–319, 2002.

[8] J. Freudenberg and P. Propping, "A similarity-based method for genome-wide prediction of disease-relevant human genes," *Bioinformatics*, vol. 18, supplement 2, pp. S110–S115, 2002.

[9] M. A. van Driel, K. Cuelenaere, P. P. C. W. Kemmeren, J. A. M. Leunissen, and H. G. Brunner, "A new web-based data mining tool for the identification of candidate genes for human genetic disorders," *European Journal of Human Genetics*, vol. 11, no. 1, pp. 57–63, 2003.

[10] F. S. Turner, D. R. Clutterbuck, and C. A. M. Semple, "POCUS: mining genomic sequence annotation to predict disease genes," *Genome Biology*, vol. 4, no. 11, article R75, 2003.

[11] N. Tiffin, J. F. Kelso, A. R. Powell, H. Pan, V. B. Bajic, and W. A. Hide, "Integration of text- and data-mining using ontologies successfully selects disease gene candidates," *Nucleic Acids Research*, vol. 33, no. 5, pp. 1544–1552, 2005.

[12] S. Aerts, D. Lambrechts, S. Maity et al., "Gene prioritization through genomic data fusion," *Nature Biotechnology*, vol. 24, no. 5, pp. 537–544, 2006.

[13] L. Franke, H. Van Bakel, L. Fokkens, E. D. De Jong, M. Egmont-Petersen, and C. Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes," *American Journal of Human Genetics*, vol. 78, no. 6, pp. 1011–1025, 2006.

[14] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard, "SUSPECTS: enabling fast and effective prioritization of positional candidates," *Bioinformatics*, vol. 22, no. 6, pp. 773–774, 2006.

[15] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard, "Speeding disease gene discovery by sequence based candidate prioritization," *BMC Bioinformatics*, vol. 6, no. 1, article 55, 2005.

[16] N. López-Bigas and C. A. Ouzounis, "Genome-wide identification of genes likely to be involved in human genetic disease," *Nucleic Acids Research*, vol. 32, no. 10, pp. 3108–3114, 2004.

[17] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, "Predicting disease genes using protein-protein interactions," *Journal of Medical Genetics*, vol. 43, no. 8, pp. 691–698, 2006.

[18] P. Jiménez, F. Thomas, and C. Torras, "3D collision detection: a survey," *Computers and Graphics*, vol. 25, no. 2, pp. 269–285, 2001.

[19] B. Grisart, W. Coppieters, F. Farnir et al., "Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition," *Genome Research*, vol. 12, no. 2, pp. 222–231, 2002.

[20] G. Thaller, C. Kühn, A. Winter et al., "DGAT1, a new positional and functional candidate gene for intramuscular fat deposition in cattle," *Animal Genetics*, vol. 34, no. 5, pp. 354–357, 2003.

[21] A. Clop, F. Marcq, H. Takeda et al., "A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep," *Nature Genetics*, vol. 38, no. 7, pp. 813–818, 2006.

[22] M. Oti and H. G. Brunner, "The modular nature of genetic diseases," *Clinical Genetics*, vol. 71, no. 1, pp. 1–11, 2007.

[23] L. D. Wood, D. W. Parsons, S. Jones et al., "The genomic landscapes of human breast and colorectal cancers," *Science Signaling*, vol. 318, no. 5853, pp. 1108–1113, 2007.

[24] J. Lim, T. Hao, C. Shaw et al., "A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration," *Cell*, vol. 125, no. 4, pp. 801–814, 2006.

[25] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen, "A text-mining analysis of the human phenome," *European Journal of Human Genetics*, vol. 14, no. 5, pp. 535–542, 2006.

[26] S. Li, L. Wu, and Z. Zhang, "Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach," *Bioinformatics*, vol. 22, no. 17, pp. 2143–2150, 2006.

[27] K. J. Gaulton, K. L. Mohlke, and T. J. Vision, "A computational system to select candidate genes for complex human traits," *Bioinformatics*, vol. 23, no. 9, pp. 1132–1140, 2007.

[28] V. van Heyningen and P. L. Yeyati, "Mechanisms of non-Mendelian inheritance in genetic disease," *Human Molecular Genetics*, vol. 13, supplement 2, pp. R225–R233, 2004.

[29] K. Lage, E. O. Karlberg, Z. M. Størling et al., "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nature Biotechnology*, vol. 25, no. 3, pp. 309–316, 2007.

[30] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 33, supplement 1, pp. D514–D517, 2005.

[31] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Onlined Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 30, no. 1, pp. 52–55, 2002.

[32] A. Hamosh, A. F. Scott, J. Amberger, D. Valle, and V. A. McKusick, "Online Mendelian inheritance in man (OMIM)," *Human Mutation*, vol. 15, no. 1, pp. 57–61, 2000.

[33] L. Baolin and H. Bo, "HPRD: a high performance RDF database," in *Network and Parallel Computing*, vol. 4672 of *Lecture Notes in Computer Science*, pp. 364–374, Springer, Berlin, Germany, 2007.

[34] T. S. Keshava Prasad, R. Goel, K. Kandasamy et al., "Human protein reference database—2009 update," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D767–D772, 2009.

[35] Y. Li and J. C. Patra, "Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network," *Bioinformatics*, vol. 26, no. 9, Article ID btq108, pp. 1219–1224, 2010.

[36] S. Peri, J. D. Navarro, T. Z. Kristiansen et al., "Human protein reference database as a discovery resource for proteomics," *Nucleic Acids Research*, vol. 32, supplement 1, pp. D497–D501, 2004.

[37] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles, "Co-ranking authors and documents in a heterogeneous network," in *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM '07)*, pp. 739–744, Omaha, Neb, USA, October 2007.

[38] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[39] Y. Chen, X. Wu, and R. Jiang, "Integrating human omics data to prioritize candidate genes," *BMC Medical Genomics*, vol. 6, no. 1, article 57, 2013.

[40] P. F. Lee and V. W. Soo, "An ensemble rank learning approach for gene prioritization," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '13)*, pp. 3507–3510, Osaka, Japan, 2013.