

RESEARCH

Open Access



# Predicting disease-related genes using integrated biomedical networks

Jiajie Peng<sup>1</sup>, Kun Bai<sup>2,7</sup>, Xuequn Shang<sup>1</sup>, Guohua Wang<sup>2</sup>, Hansheng Xue<sup>2</sup>, Shuilin Jin<sup>3</sup>, Liang Cheng<sup>4</sup>, Yadong Wang<sup>2\*</sup> and Jin Chen<sup>5,6\*</sup>

From The 27th International Conference on Genome Informatics  
Shanghai, China. 3-5 October 2016

## Abstract

**Background:** Identifying the genes associated to human diseases is crucial for disease diagnosis and drug design. Computational approaches, esp. the network-based approaches, have been recently developed to identify disease-related genes effectively from the existing biomedical networks. Meanwhile, the advance in biotechnology enables researchers to produce multi-omics data, enriching our understanding on human diseases, and revealing the complex relationships between genes and diseases. However, none of the existing computational approaches is able to integrate the huge amount of omics data into a weighted integrated network and utilize it to enhance disease related gene discovery.

**Results:** We propose a new network-based disease gene prediction method called SLN-SRW (Simplified Laplacian Normalization-Supervised Random Walk) to generate and model the edge weights of a new biomedical network that integrates biomedical data from heterogeneous sources, thus far enhancing the disease related gene discovery.

**Conclusions:** The experiment results show that SLN-SRW significantly improves the performance of disease gene prediction on both the real and the synthetic data sets.

**Keywords:** Disease gene prediction, Laplacian normalization, Supervised random walk, Integrated network

## Background

One crucial step toward understanding the molecular basis of diseases, such as cancer, diabetes, and cardiovascular disorders, is to identify the predisposing or virulence genes of these diseases, which will lead to early disease diagnosis and effective drug design [1]. With the availability of the big biomedical data, researchers tend to get insights into human diseases by identifying genes that might cause or relate to them. Given the fact that experimentally identifying of the complete list of disease-related genes is generally impractical due to the high cost, computational methods have been proposed in the last decades to predict the relationships between genes

and human diseases [2–10]. However, these tools, including filtering methods based on a set of criteria [11], text mining of biomedical literature [12], integration of genomic data [13–15], semantic similarity [16–21] based disease gene prioritization [22] and network analysis based and highly robust approach [8, 23–26], remain pre-eminent [27].

A human cell consists of several functionally interdependent molecular components. A human disease rarely results from an abnormality in a single gene but reflects the perturbations of the complex molecular network induced by different kinds of factors, such as genetic variations, pathogens and epigenetic changes [28]. The molecular network links molecular states to physiological states associated with diseases in a whole system view [29]. Therefore, network-based approaches may offer better targets for drug development, and may lead to multiple potential biological and clinical applications including disease gene discovery [28].

\*Correspondence: ydwang@hit.edu.cn; jinchen@msu.edu

<sup>2</sup>School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

<sup>5</sup>Institute of Biomedical Informatics, College of Medicine, University of Kentucky, Lexington, KY 40536, USA

Full list of author information is available at the end of the article

The network-based approaches for disease gene identification can be loosely grouped into three categories. The simplest approach, named direct neighbor counting, is to check whether two genes are connected directly in a molecular network. The idea is that if a gene is connected to one of the known disease genes, it may be associated with the same disease [30]. The experimental result demonstrates that using molecular networks can effectively increase the likelihood of identifying candidate disease genes. The direct neighbor counting method, however, does not consider the situation that two genes are not connected directly but still have certain biological associations. To address this problem, Kruthammer et al. employed the shortest path length approach to measure the closeness between a disease gene and a candidate gene. This method has been successfully applied to predict the genes associated Alzheimer's disease, and the prediction results agree with the manually curated candidates [31]. Since both the direct neighbor counting method and the shortest path method are local distance measurements, they largely ignore the global structure of the whole molecular network and cannot fully capture the complex relationships between network nodes [32]. Subsequently, methods have been proposed to predict the gene-disease relation using the global network structure, such as random walk with restart (RWR) [33], propagation flow [34], Markov clustering [35] and graph partitioning [36]. The performance evaluation on HPRD [37], OPHID [38] and OMIM [39] dataset shows that RWR is the best among the then-existing methods [5].

Rapidly evolving bio-technologies promote collecting multiple types of biological data, including diverse genome-scale data, clinical phenotype data, environment data, and data of daily activities [40], making it feasible and attractive to build integrated biomedical networks from multiple sources, rather than focus on one single data set. The integrated network that includes multiple, heterogeneous types of resources, greatly extends the scope and ability for disease gene prediction [41]. For example, BioGraph [42] uses data from 21 publicly available curated databases to identify relations between heterogeneous biomedical entities. The work by Ganegoda et al. runs RWR on a integrated network, and has successfully identified disease-related genes with significant improved performance [43].

Using integrated networks for gene-disease relationship discovery is still a difficult task due to the existence of multiple biomedical entities in the integrated networks. In a network built using a single type of biomedical data, there is only one type of nodes and one type of edges. For example, in a protein-protein interaction network, nodes and edges represent proteins and protein interactions respectively. The integrated network, on the contrary, contains multiple types of nodes and edges representing different

biomedical entities (such as genes, diseases, and ontology terms) and relationships (such as DNA-protein binding and gene ontology annotation). In order to differentiate these edge types, edge weights in the integrated biomedical network should be appropriately assigned [44].

In this article, we present a new algorithm called SLN-SRW (Simplified Laplacian Normalization-Supervised Random Walk) to define edge weights in an integrated network and use the weighted network to predict gene-disease relationships. Comparing with the existing approaches, SLN-SRW has the following advantages:

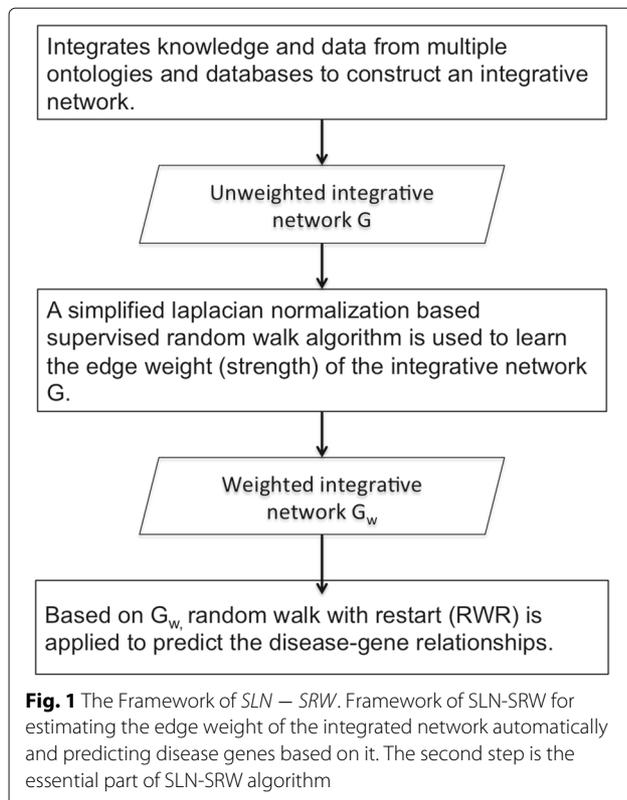
- \* SLN-SRW is the first approach, to the best of our knowledge, to predict gene-disease relationships based on a weighted integrated network with its edge weight being computed to precisely describe the importance of different edge types.
- \* The performance of random walk may be strongly affected by the super hub nodes in an integrated network. SLN-SRW adopts a Laplacian normalization based method to avoid such bias.
- \* To prepare inputs for SLN-SRW, we constructed a new heterogeneous integrated network based on three widely used biomedical ontologies, i.e. Human Phenotype Ontology [45], Disease Ontology [46], and Gene Ontology [47, 48], and biological databases such as STRING [49]. This integrated network combines biomedical knowledge from ontologies that are manually curated and big biomedical data that have been deposited in databases. Based on these two distinctively different types of information, this network forms a foundation for disease gene discovery.

## Methods

We propose SLN-SRW to compute and model the edge weight of an integrated network and then predict disease genes. To achieve the goal, SLN-SRW consists of three steps. First, it integrates knowledge and data from multiple ontologies and databases to construct an integrated network  $G(V, E)$ , where  $V$  is a set of nodes and  $E$  is a set of edges that connect the nodes in  $V$ . Second, it uses a Laplacian normalization based supervised random walk algorithm to learn the edge weight of network  $G$ , resulting in a weighted integrated network  $G_w$ . Third, it employs the RWR method on  $G_w$  to predict disease-gene relationships. The diagram of the whole process of SLN-SRW is shown in Fig. 1. We will introduce the key steps of SLN-SRW in the rest of this section.

### Step 1. Integrating heterogeneous knowledge and data sources for integrated network construction

In the first step of SLN-SRW, an integrated network is constructed based on eleven heterogeneous data sources



in four distinct forms, i.e. ontologies, networks, unified vocabularies, and relational databases. The data sources are listed in Table 1, and they are mainly used for relation extraction, name mapping, and unified vocabulary. The data sources can be grouped into two categories: 1) Curated data that were collected from literature and other high quality data sources, such as Search Tool for the Retrieval of Interacting Gene/Proteins (STRING) and Online Mendelian Inheritance in Man (OMIM), and 2) Curated ontologies that constructed manually by the domain expert, such as Gene Ontology (GO) and Disease Ontology (DO).

The workflow for constructing the integrated network out of the heterogeneous data sources is shown in Fig. 2. Specifically, the network construction process has the following four steps:

1. **Extracting information from heterogeneous data sources.** Ontology parser and database parser have been developed for ontology and database data extraction respectively. The ontology parser processes the OBO file and the ontology annotation file, since HPO, DO and GO are all in Open Biomedical Ontologies (OBO) format. The database parser processes files in Tab Separated Values (TSV), Comma Separated Values (CSV), and Extensible

Markup Language (XML) format. The outputs of the two parsers are pair-wise relations and their properties between two biomedical entities.

2. **Unifying biomedical entity IDs.** The same pair-wise relation may be extracted from multiple data sources with different identifiers. To avoid confusion, we provide a distinct ID number for each biomedical entity by mapping all types of identifiers to the ones provided in the Unified Medical Language System (UMLS). The challenge is that some types of identifiers cannot be direct mapped to UMLS. For example, only a small part (61%) of HPO and DO term can be mapped to UMLS. Therefore, we adopted ClinVar [50] to map all the HPO terms to UMLS, and utilized SIDD [51] to map all the disease names in DO to MeSH ID, provided that there are direct mappings between MeSH ID and UMLS. Please see Additional file 1 for more details. After unifying the entity IDs from multiple data sources, each entity only has one identifier in the database. We removed the identifiers that cannot be mapped to UMLS.
3. **Constructing the integrated network.** The binary relations extracted from multiple data sources form an integrated network  $G$ , in which nodes are biomedical entities (i.e. ontology terms and genes), and edges are the relationships between the entities, which have seven different types: GO term - GO term, GO term - gene, DO term - DO term, DO term - gene, HPO term - HPO term, HPO term - gene, and gene - gene.
4. **Edge initial weight assignment.** We assign the initial edge weight  $t(u, v)$  to every edge  $\langle u, v \rangle$  according to its edge type and the evidence code associated to the edge, where both  $u$  and  $v$  are nodes in  $G$ . Specifically, for the edge types that have edge confidence scores in the source databases, we use the confidence scores directly. For the edge types that do not have confidence scores but are associated with evidence codes, we manually assign initial edge weights based on their edge evidence codes (see Additional file 2 for the manually assigned weights). The edge initial weights are between 0 and 1, and the experimentally verified edges have higher initial weights than computational predictions. For example, an edge between a GO term and a gene with evidence code "EXP" has a high weight (1.0), whereas an edge with "IEA" code has a low weight (0.4), since "EXP" indicates the GO-gene relationship has been experimentally verified while "IEA" means computational prediction. Note that for the edges that have two or more evidences in  $E$ , the initial weights are calculated as the maximal weight of all the valid evidence codes.

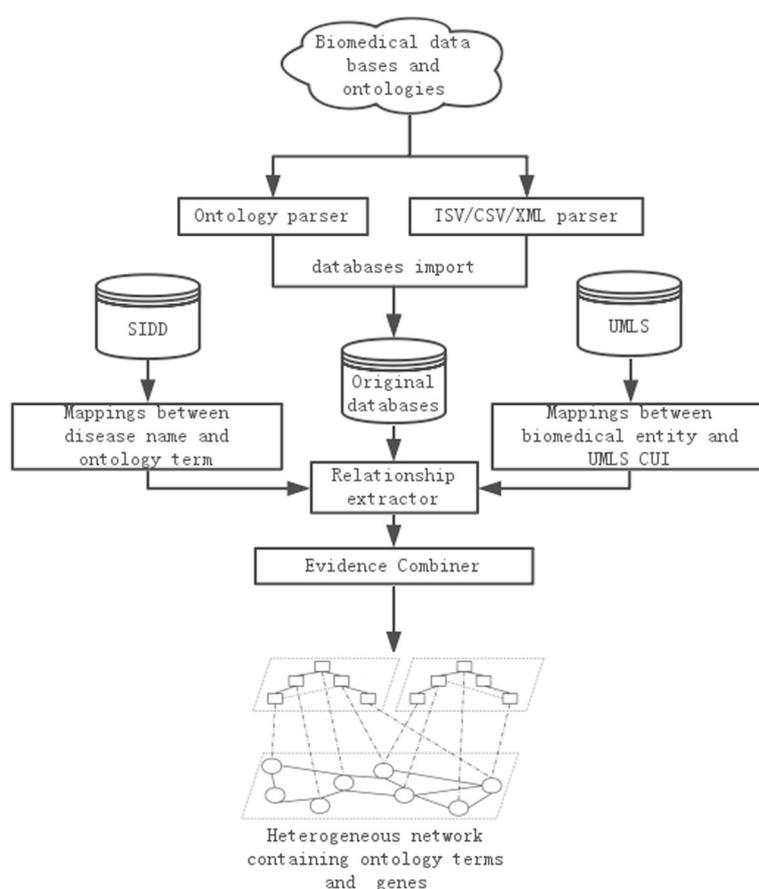
**Table 1** Integrated databases and ontologies. The first column, second column, and third column represent the abbreviation of the data source, simplify the description of the data source and the relationship extracted from the data source respectively. Eleven data sources are used to construct the integrated network. Specific types of nodes and edges are extracted from various data sources and integrated into a network

Abbreviation	Data sources	Relationship
STRING	Search Tool for the Retrieval of Interacting Gene/Proteins	gene-gene
CTD-DG	The Comparative Toxicogenomics Database - Curated Disease-Gene Interactions	disease-gene
OMIM	Online Mendelian Inheritance in Man Disease Subtypes	disease-gene
ClinVar	Clinical Variants and phenotypes	Disease/Phenotype-gene
HGNC	HUGO gene Nomenclature Committee Database	gene name mapping
MeSH	Medical Subject Headings	Unified vocabulary
UMLS	Unified Medical Language System	Unified vocabulary
SIDD	Semantically Integrated Disease-associated Database	disease name mapping
DO	Human Disease Ontology	DO term-gene/ DO term-DO term
HPO	Human Phenotype Ontology	HPO term-gene/ HPO term-HPO term
GO	Gene Ontology	GO term-gene/GO term-GO term

### Step 2. Weighing the importance of different types of edges in integrated network

Given an integrated network  $G$  with manually assigned initial edge weights, the aim of this step is to automatically re-assign all the edge weights, such that the weighted

network  $G_w$  can be used for more precise disease gene prediction. To achieve this goal, we develop a new edge weight optimization algorithm based on supervised random walks (SRW) [52]. SRW combines the information from network structure with the node and edge level



**Fig. 2** The workflow of constructing the integrated network. Work flow of constructing the integrated network based on multiple data sources

attributes, which can guide the random walk on the network. By running SRW, we expect to re-assign weights to all the edges, such that the random walker from a disease node is more likely to visit the nodes representing the associated genes. However, the training process of supervised random walks (i.e. RWR) can be significantly affected by the hub nodes in the network. To solve this problem, we propose a Laplacian normalization method to weigh the importance of different types of edges in an integrated network described as follows.

Given an integrated network  $G(V, E)$ , let node  $v_d \in V$  represent a kind of disease and let  $V_g \subset V$  be the set of the candidate genes of  $v_d$ , then the disease gene prediction problem can be converted to a problem to predict all the new edges between  $v_d$  and a subset of nodes in  $V_g$ , where a critical step is to weigh the edges such that a random walker from  $v_d$  has higher probabilities to reach the known disease genes in  $V_g$  than to reach the other genes. To provide the training set for learning the edge weight, we generate a positive set  $V_p$  and a negative set  $V_n$  for every disease node  $v_d$ , where  $V_p$  includes known disease genes associated with  $v_d$  and  $V_n$  includes genes not associated with  $v_d$ .

The approach to weigh the importance of different edge types consists of the following three steps:

1. **Laplacian normalization on edge weights.** To avoid the biases caused by the hub nodes in the integrated network, we adopt the Laplacian normalization method [53] to normalize all the edge weights. Given an edge  $(u, v) \in E$ , the edge weight of edge  $(u, v)$  is normalized by all the edges connecting to node  $u$  or node  $v$ . Mathematically, the laplacian normalized edge weight  $a(u, v)$  is defined as:

$$a(u, v) = \frac{f(u, v)}{\sqrt{\sum_{i \in N(u)} f(u, i) \sum_{j \in N(v)} f(v, j)}} \quad (1)$$

where  $N(x)$  is the set of neighbors of node  $x$ ;  $f(x, y) = 1 / (1 + e^{-w \cdot t(x,y)})$ ;  $w$  is the edge type importance vector for graph  $G$  that we will learn in the next step using an optimization process, and its length is equal to the number of possible edge types (in our case, seven);  $t(x, y)$  is the vector of the initial weight of edge  $\langle x, y \rangle$ , which has the same length as  $w$ .  $t(x, y)$  is all zero except one cell because each edge can have one and only one kind of edge type. Note that the edge type is decided by the type of nodes connected by it. For example, gene - gene and HPO term - gene are two different types of edges in the integrated network.  $a(u, v)$  integrates and normalizes both the edge type importance  $w$  and the initial edge weight  $t$ ; it can be used to model the random walk transition probability.

2. **Edge weight optimization - problem formation.** In order to learn the optimal  $w$  for all the seven edge types in an integrated network, we minimize an optimal function defined in Eq. 2, such that the random walker in the network is more likely to reach the genes in  $V_p$  than the genes in  $V_n$ .

$$w = \arg \min_w O(w) \\ = \arg \min_w \left( \frac{1}{2} \|w\|^2 + \lambda \sum_{v_d \in D} \sum_{v_p \in V_p, v_n \in V_n} h(S_{v_n} - S_{v_p}) \right) \quad (2)$$

where  $\|w\|$  is the euclidean norm; and  $D$  is a set of starting nodes representing the diseases in the training set. For each disease node  $v_d \in D$ ,  $V_p$  and  $V_n$  representing the positive training set and the negative training set respectively.  $S_{v_p}$  ( $S_{v_n}$ ) is the association value between  $v_d$  and  $v_p \in V_p$  ( $v_d$  and  $v_n \in V_n$ ), which can be calculated by running RWR on  $G$  [54].  $\lambda$  is the weight penalty score deciding to what extent the constraints can be violated. Given the value of  $S_{v_n} - S_{v_p}$ ,  $h()$  is a loss function that returns a non-negative value:

$$h(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{1 + e^{-\frac{x}{b}}}, & x \geq 0 \end{cases} \quad (3)$$

where  $b$  is a constant positive parameter,  $x = S_{v_n} - S_{v_p}$ . The smaller the  $b$  is, the more sensitive the loss function is (see Additional file 3). If  $S_{v_n} - S_{v_p} < 0$ , the association between a disease and a gene in the positive training set is stronger than the association between the same disease and a gene in the negative training set, so  $h() = 0$ . Otherwise, the constraint is violated, so  $h() > 0$ .

3. **Edge weight optimization - our solution.** To optimize edge type importance parameter  $w$  to minimize Eq. 2, we adopt a widely used meta-heuristics method called the gradient based optimization method [20], which has been successfully adopted to solve the link prediction problem in social networks and collaboration networks [52]. To make the story complete, we briefly describe the gradient-based optimization method in the following text. First, we construct a stochastic transition matrix  $Q'_{uv}$  of RWR using Eq. 1.

$$Q'_{uv} = \begin{cases} \frac{a(u, v)}{\sum_w a(u, v)}, & \text{if } (u, v) \in E \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

And then, based on the transition matrix  $Q'_{uv}$ , RWR can be described as:

$$Q_{uv} = (1 - \alpha)Q'_{uv} + \alpha \mathbf{1}(v = s) \quad (5)$$

where  $u$  and  $v$  represent two arbitrary nodes in  $G$ ;  $\alpha$  is the restart probability, which is a user given threshold (in this case, we find the best value based on the training data set); and node  $s$  is a disease node, which is the starting node of random walk. Let  $p_i^{(k)}$  be the probability to reach node  $i$  from  $s$  after  $k$  iterations. The probability vector at the  $k$ th iteration can be represented as  $P^{(k)} = (p_1^{(k)}, p_2^{(k)}, \dots, p_{|V|}^{(k)})^T$ . The stationary probability vector  $P$ , which can be obtained after certain iterations, is the solution of the following equation:

$$P^T = P^T Q \quad (6)$$

The next step is to apply a gradient based method to identify  $w$  to minimize  $O(w)$  in Eq. 2. The derivative of  $O(w)$  can be calculated as follows.

$$\begin{aligned} \frac{\partial O(w)}{\partial w} &= 2w + \sum_{v_n, v_p} \frac{\partial h(S_{v_n} - S_{v_p})}{\partial w} \\ &= 2w + \sum_{v_n, v_p} \frac{\partial h(S_{v_n} - S_{v_p})}{\partial (S_{v_n} - S_{v_p})} \left( \frac{\partial S_{v_n}}{\partial w} - \frac{\partial S_{v_p}}{\partial w} \right) \end{aligned} \quad (7)$$

$\frac{\partial S_{v_x}}{\partial w}$  can be calculated as follows:

$$\frac{\partial S_{v_x}}{\partial w} = \sum_{v_i} Q_{v_i v_x} \frac{\partial S_{v_i}}{\partial w} + S_{v_i} \frac{\partial Q_{v_i v_x}}{\partial w} \quad (8)$$

This derivative can be repeatedly computed based on the estimate obtained in the previous iteration. The iteration stops when  $\frac{\partial S_{v_i}}{\partial w}$  and  $S_{v_i}$  do not change. The initial value of  $\frac{\partial S_{v_i}}{\partial w}$  is 0. The  $S_{v_i}$  is initialized as  $\frac{1}{|V|}$ . The initialization process is the same as the traditional SRW method.  $\frac{\partial Q_{v_i v_x}}{\partial w}$  can be calculated as follows. Particularly,  $\frac{\partial Q_{v_i v_x}}{\partial w} = 0$ , if edge  $(v_i, v_x)$  does not exist in the network.

$$\frac{\partial Q_{v_i v_x}}{\partial w} = (1 - \alpha) \frac{\frac{\partial a(v_i, v_x)}{\partial w} \left( \sum_{v_j} a(v_i, v_j) \right) - a(v_i, v_x) \sum_{v_j} \frac{\partial a(v_i, v_j)}{\partial w}}{\left( \sum_k a(v_i, v_j) \right)^2} \quad (9)$$

$$\frac{\partial a(v_i, v_x)}{\partial w} = \frac{\frac{\partial f(v_i, v_x)}{\partial w} \pi(f(v_i, v_x)) - f(v_i, v_x) \frac{\partial \pi(f(v_i, v_x))}{\partial w}}{\pi(f(v_i, v_x))^2} \quad (10)$$

where  $\pi(f(v_i, v_x))$  and  $\frac{\partial \pi(f(v_i, v_x))}{\partial w}$  are:

$$\pi(f(v_i, v_x)) = \sqrt{\sum_{v_j \in N(v_i)} f(v_i, v_j) \sum_{v_y \in N(v_x)} f(v_x, v_y)} \quad (11)$$

$$\frac{\partial \pi(f(v_i, v_x))}{\partial w} = \frac{\sum_{v_j \in N(v_i)} \sum_{v_y \in N(v_x)} \left( \frac{\partial f(v_j, v_i)}{\partial w} f(v_j, v_x) + \frac{\partial f(v_y, v_x)}{\partial w} f(v_j, v_i) \right)}{2 \sqrt{\sum_{v_j \in N(v_i)} f(v_j, v_i) \sum_{v_y \in N(v_x)} f(v_y, v_x)}} \quad (12)$$

where  $N(v)$  is the set of neighbors of node  $v$ . After we get the solution of Eq. 7, we can apply a gradient descent based method and minimize  $O(w)$ .

Practically, the process of obtaining  $w$  has four steps (Fig. 3). Firstly, we initial the  $O(w)$  based on the initial parameters. Secondly, the derivative  $\frac{\partial O(w)}{\partial w}$  is calculated in step 2. The power iteration is applied to calculate  $\frac{\partial S_{v_i}}{\partial w}$  and  $\frac{\partial Q_{v_i v_x}}{\partial w}$  respectively. Thirdly, based on the derivative, we can update the gradient to obtain an updated parameter  $w$ . Fourthly, taking the updated  $w$  as input, step 4 calculates the stationary probability of the RWR. In the process, the iteration for derivative calculation (step 2 in Fig. 3) and the RWR algorithm (step 4 in Fig. 3) are the two key steps. After estimating the edge weight of the integrated network, we can directly apply RWR on the weighted integrated network to predict the relation between diseases and genes.

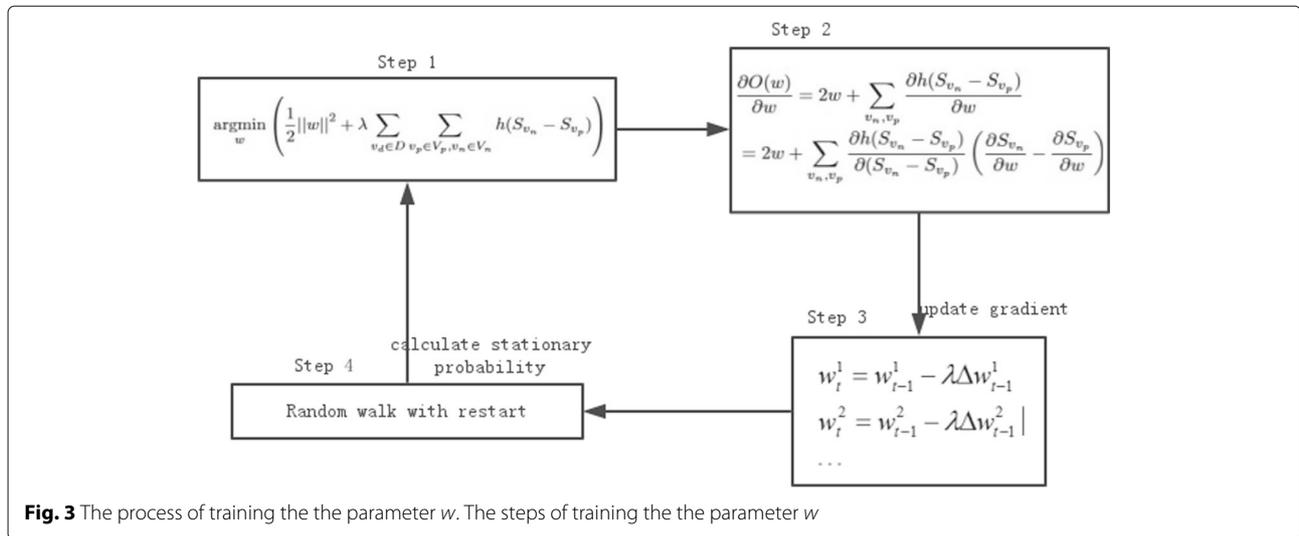
## Results and discussion

We compare SLN-SRW with SRW and RWR, where the latter has been widely used in network-based disease gene prediction, on a real and a synthetic data sets. SLN-SRW was implemented with Java 7 in Linux.

### Data preparation

As shown in Table 1, eleven data sources, i.e. STRING [49], CTD [55], OMIM [56], ClinVar [50], HGNC [57], MeSH [58], UMLS [59], SIDD [51], DO [60], HPO [61] and GO [62], are used to construct the integrated network  $G$ , which has 78,786 nodes and 504,517 edges.

To test the performance of SLN-SRW, we select 430 disease-gene edges from the integrated network as the positive set. The rules for data selection are similar to the rules used in [42]. In the positive set, there are 16 diseases, each of which has at least five known disease-associated genes in the integrated network. More detail about the positive set is listed in Additional file 4. The disease-gene pairs included in the negative set are generated in two steps. First, we select a disease  $d$  from the positive set.



**Fig. 3** The process of training the the parameter  $w$ . The steps of training the the parameter  $w$

Second, we repeatedly and randomly select genes that do not connect to  $d$  in the integrated network  $G$ . The number of the randomly selected genes is the same as the number of genes that connect to  $d$  in the positive set. We repeat the process until all disease nodes in the positive set are elected. Note that the positive set is removed from the integrated network in the testing process. Both positive and negative sets are evenly divided into two parts randomly, one for training and the other for testing.

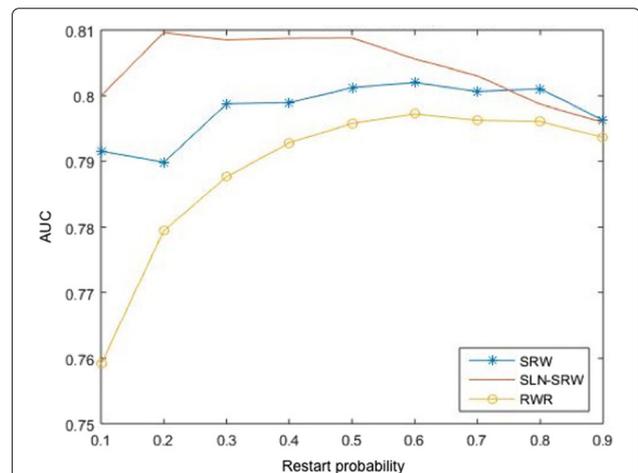
A synthetic data set is generated following the rules in [52]. Specifically, we generated a scale-free network with 1,000 nodes using the Copying model [63]. The generation process starts with three connected nodes. We connect a new node  $u$  to any of the existing nodes, which are selected at random with probability 0.8 or with probability proportional to the node degree. Parameter  $b$  is equal to 0.03 in all the experiments. For each edge in the network, we set  $w = \{1, -1\}$  as the gold standard labeled as  $w'$ . Then, we randomly choose one of the original three nodes as the start point  $v$ . Based on the edge strength determined by  $w'$ , we run RWR starting from  $v$  and ranked the other nodes via the stationary probability. We select the top 20 nodes that directly connect with  $v$  as the positive training set, and select the nodes that do not connect with  $v$  are the negative set. Note that both the positive set and the negative set are removed from the integrated network in the testing process. In the subsection "Performance evaluation on synthetic data set", we test whether  $w'$  can be estimated precisely.

**Disease gene prediction**

The parameters in SLN-SRW and SRW method are estimated based on the training set. The RWR method does not need the training set for edge weight assignment. Alternatively, the training set is used to estimate the best

restart probability in RWR. Finally, the performance of all the three methods is tested based on the testing set.

Varying the restart probability  $\alpha$  from 0.1 to 0.9, the AUC (area under receiver operating characteristic curve) scores [64] of all the three methods are shown in Fig. 4. If  $\alpha = 0.2$ , SLN-SRW method reaches the highest AUC score 0.81, whereas SRW and RWR have the highest AUC scores if  $\alpha = 0.6$ , indicating that SLN-SRW can find the disease genes which are far from the disease node. Based on the edge weights learned using the training data, we predicted the disease-gene relationships in the testing set. We compared the performance of all the three methods using the receiver operating characteristic (ROC) curve. In our test, the AUC score of SLN-SRW (0.79) is the



**Fig. 4** The AUC score for each given restart probability for three methods. The AUC score for each given restart probability for three methods. The red, blue and yellow lines are represent SLN-SRW, SRW and RWR method respectively

highest (see Fig. 5). Especially, the true positive rate of SLN-SRW is significantly higher than RWR and SRW while its false positive rate keeps low. This is important for disease gene predict, since researchers usually select candidate disease genes with a stringent threshold, which corresponds to a low false positive rate.

Finally, we ranked the predicted disease genes to check whether the true disease-related genes have higher ranks than the other genes. Figure 6 shows that the prediction result of SLN-SRW contains more known disease-related genes than SRW and RWR at a majority of the top  $k$  levels, indicating that the edge weighing process in SLN-SRW has contributed significantly to the high recall of the results.

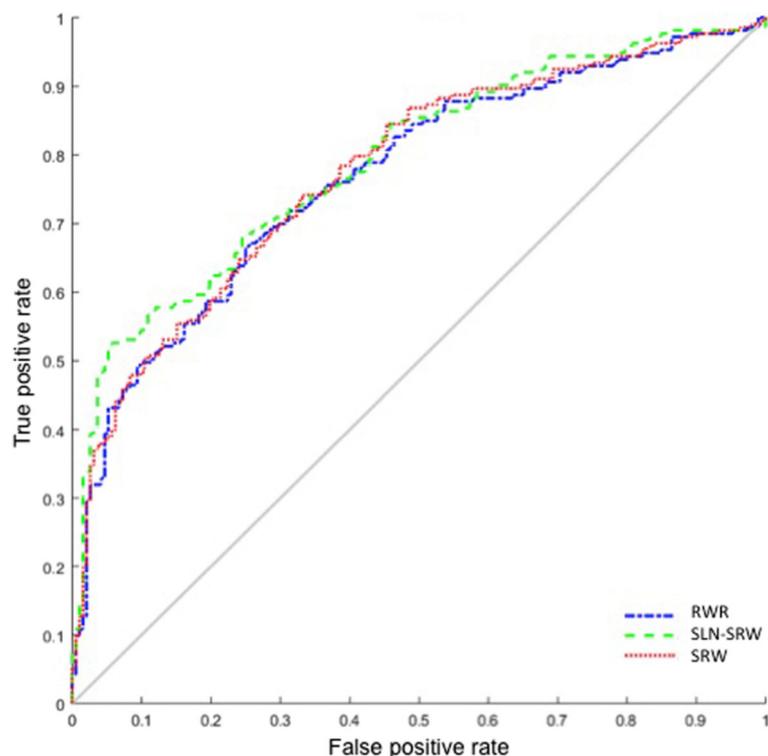
#### Performance evaluation on synthetic data set

To compare SLN-SRW with SRW, we ran both methods on synthetic data, following the method described in below [52]. For SRW and SLN-SRW, we estimated the edge-type parameter based on the synthetic network and the training set described in the “Data preparation” subsection, resulting in  $w^*$ . We measure the performance of SRW and SLN-SRW by comparing the true edge-type parameter  $w'$  with  $w^*$ , using  $error = \sum_i |w'_i - w_i^*|$ . After repeating the experiment 100 times, we find that the

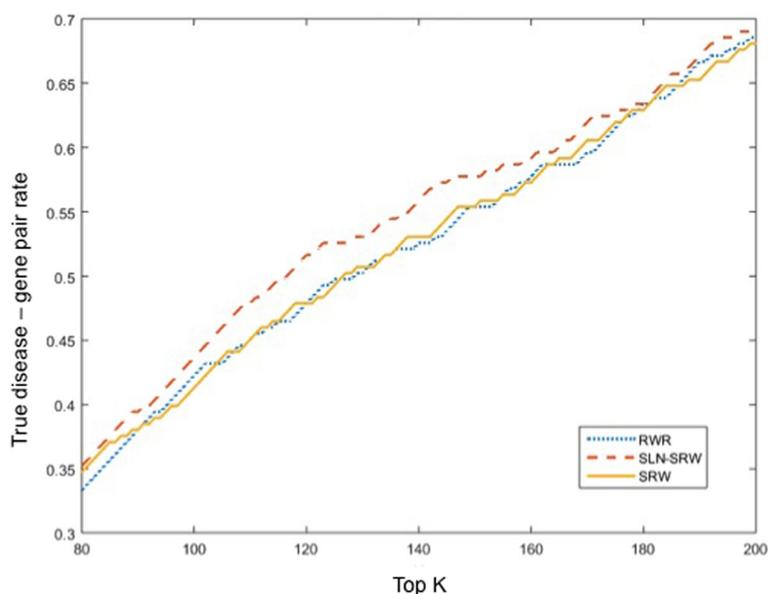
error of SLN-SRW is statistically significantly lower than that of SRW (t-test  $p - value < 0.05$ ) indicating that SLN-SRW performs better than SRW (see Fig. 7). The error of SLN-SRW is also lower in the first and third quartile.

#### Conclusions

Identifying the relationships between diseases and genes is vital for disease diagnosis and drug design. Recently, researchers have started to employ integrated biomedical networks to extend the scope and ability for disease gene prediction. In this article, we proposed a novel network-based method named SLN-SRW to define the weight of edges in an integrated network and then use it to predict the gene-disease relationships. SLN-SRW has the following advantages: 1) it can estimate edge weight by differentiating different edge-types; 2) it adopts a Laplacian normalization based method to avoid the bias caused by the super hub nodes in an integrated network; 3) three widely used biomedical ontologies are used to construct a new heterogeneous integrated network. To demonstrate the advantages of SLN-SRW, we compare it with two existing methods SRW and RWR. The experiment on a real data set shows that SLN-SRW performs best among all the three methods. Furthermore, the experiment on



**Fig. 5** ROC curves for the experimental results on testing set. ROC curves for the experimental results calculated with SLN-SRW (green), SRW (red) and RWR (blue)



**Fig. 6** True disease-gene pair rates. True disease-gene pair rates at different top  $k$  levels

a synthetic data set indicates that the edge weights predicted by SLN-SRW are more precise than SRW. Comparing with the existing methods, SLN-SRW has the unique function to identify disease genes, which are not close to any disease node in the disease-gene networks. This could benefit clinicians on discovering new disease-associated genes that have not been identified by the existing methods. Besides, SLN-SRW provides a novel approach to automatically assign weights to the heterogeneous edge types in the disease-gene networks, whereas the existing methods can only define the edge weights manually.

In the future, SLN-SRW will be applied to networks with different edge densities and qualities to test its robustness.

Furthermore, we will apply SLN-SRW on more recent datasets and examine the results using both biological experiments and literature.

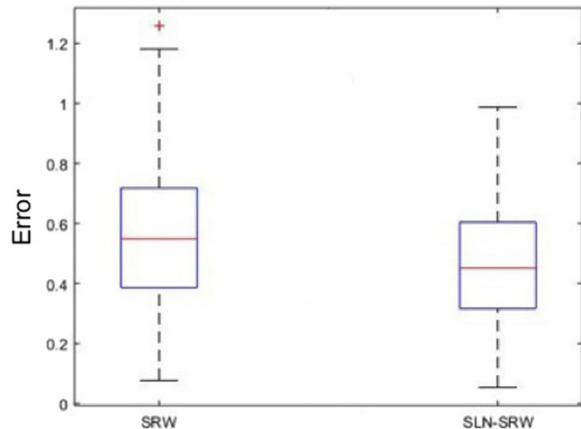
### Additional files

**Additional file 1:** Process of mapping different types of IDs. Additional file 1 is a figure to illustrate how different types of IDs are unified. (PDF 54.5 kb)

**Additional file 2:** Initial weight for difference evidence code. Additional file 2 is a table that lists the weight values for different evidence code. (PDF 42 kb)

**Additional file 3:** Relation between parameter  $b$  and loss value. Additional file 3 is a figure showing the relation between parameter  $b$  and loss value. (PNG 155 kb)

**Additional file 4:** Diseases selected as the evaluation set. Additional file 4 is a table of diseases selected as the evaluation set. (PDF 708 kb)



**Fig. 7** The boxplot of the error score. The boxplot of the error score for SLN-SRW and SRW

### Acknowledgments

We would like to thank Dr. Qingcai Chen, Professor at Harbin Institute of Technology, Shenzhen Graduate School, for invaluable comments and suggestions to the project.

### Declarations

This article has been published as part of *BMC Genomics* Volume 18 Supplement 1, 2016: Proceedings of the 27th International Conference on Genome Informatics: genomics. The full contents of the supplement are available online at <http://bmcgenomics.biomedcentral.com/articles/supplements/volume-18-supplement-1>.

### Funding

This project has been funded by the National Natural Science Foundation of China (Grant No. 61332014, 61272121); the Start Up Funding of the Northwestern Polytechnical University (Grant No. G2016KY0301); the Fundamental Research Funds for the Central Universities (Grant No. 3102016QD003); the National High Technology Research and Development Program of China grant (no. 2015AA020101, 2015AA020108, 2014AA021505). The publication costs for this article were funded by Northwestern Polytechnical University.

**Availability of data and materials**

The datasets during and/or analysed during the current study available from the corresponding author on reasonable request.

**Authors' contributions**

JP, JC and YW conceived the project; JP, KB and JC designed the algorithm and experiments; JC and JP wrote this manuscript; XS, GW, HX, SJ and LC helped to test the algorithm. All authors read and approved the final manuscript.

**Competing interests**

The authors declare that there are no competing interests.

**Consent for publication**

Not applicable.

**Ethics approval and consent to participate**

Not applicable.

**Author details**

<sup>1</sup>School of Computer Science, Northwestern Polytechnical University, Xi'an, China. <sup>2</sup>School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. <sup>3</sup>Department of Mathematics, Harbin Institute of Technology, Harbin, China. <sup>4</sup>College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China. <sup>5</sup>Institute of Biomedical Informatics, College of Medicine, University of Kentucky, Lexington, KY 40536, USA. <sup>6</sup>Department of Energy Plant Research Lab, Michigan State University, East Lansing, MI 48824, USA. <sup>7</sup>Current address: Tencent, Inc., Shenzhen, China.

Published: 25 January 2017

**References**

- Wang X, Gulbahce N, Yu H. Network-based methods for human disease gene prediction. *Brief Funct Genomics*. 2011;10(5):280–93.
- Ala U, Piro RM, Grassi E, Damasco C, Silengo L, Oti M, Provero P, Di Cunto F. Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS Comput Biol*. 2008;4(3):e1.000043.
- Kann MG. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief Bioinformatics*. 2010;11(1):96–110.
- Jiang Q, Wang J, Wu X, Ma R, Zhang T, Jin S, Han Z, Tan R, Peng J, Liu G. LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression. *Nucleic Acids Res*. 2015;43(Database issue):193–6.
- Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics*. 2010;26(8):1057–63.
- Jiang q, Wang G, Zhang T, Wang Y. Predicting human microRNA-disease associations based on support vector machine. *Int J Data Mining Bioinformatics*. 2013;8(3):282–93.
- Browne F, Wang H, Zheng H. A computational framework for the prioritization of disease-gene candidates. *BMC Genomics*. 2015;16(Suppl 9):S2.
- Chen B, Li M, Wang J, Shang X, Wu FX. A fast and high performance multiple data integration algorithm for identifying human disease genes. *BMC Med Genomics*. 2015;8(Suppl 3):S2.
- Chen B, Shang X, Li M, Wang J, Wu FX. Identifying individual-cancer-related genes by re-balancing the training samples. *IEEE Transactions on Nanobioscience*. 2016;15(4):309–315.
- Jiang q, Hao Y, Wang G, Juan L, Zhang T, Teng M, Liu Y, Wang Y. Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst Biol*. 2010;4:1.
- Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. In: *Pacific Symposium on Biocomputing*. Pacific Symposium on Biocomputing. The Big Island of Hawaii: NIH Public Access; 2009. p. 368.
- Yu S, Van Vooren S, Tranchevent LC, De Moor B, Moreau Y. Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. *Bioinformatics*. 2008;24(16):i119–25.
- Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, et al. Gene prioritization through genomic data fusion. *Nat Biotechnol*. 2006;24(5):537–44.
- Hu Y, Zhou W, Ren J, Dong L, Wang Y, Jin S, Cheng L. Annotating the function of the human genome with gene ontology and disease ontology. *BioMed Res Int*. 2016;4130861.
- Zhang T, Hu Y, Wu X, Ma R, Jiang Q, Wang Y. Identifying liver cancer-related enhancer SNPs by integrating GWAS and histone modification ChIP-seq data. *BioMed Res Int*. 2016;6968:2395341.
- Peng J, Uygun S, Kim T, Wang Y, Rhee SY, Chen J. Measuring semantic similarities by combining gene ontology annotations and gene co-function networks. *BMC Bioinformatics*. 2015;16:1.
- Cheng L, Li J, Hu Y, Jiang Y, Liu Y, Chu Y, Wang Z, Wang Y. Using semantic association to extend and infer literature-oriented relativity between terms. *IEEE/ACM Trans Comput Biol Bioinformatics*. 2015;12(6):1219–26.
- Cheng L, Jiang Y, Wang Z, Shi H, Sun J, Yang H, Zhang S, Hu Y, Zhou M. DisSim: an online system for exploring significant similar diseases and exhibiting potential therapeutic drugs. *Sci Rep*. 2016;6:30024.
- Peng J, Wang Y, Chen J. Towards integrative gene functional similarity measurement. *BMC Bioinformatics*. 2014;15(2):1.
- Peng J, Li H, Jiang Q, Wang Y, Chen J. An integrative approach for measuring semantic similarities using gene ontology. *BMC Syst Biol*. 2014;8(Suppl 5):S8.
- Peng J, Li H, Liu Y, Juan L, Jiang q, Wang Y, Chen J. InteGO2: a web tool for measuring and visualizing gene semantic similarities using gene ontology. *BMC Genomics*. 2016;17(5):530.
- Schlicker A, Lengauer T, Albrecht M. Improving disease gene prioritization using the semantic similarity of gene ontology terms. *Bioinformatics*. 2010;26(18):i561–7.
- Peng J, Wang T, Hu J, Wang YW, Chen J. Constructing Networks of Organelle Functional Modules in Arabidopsis. *Curr Genomics*. 2016;5:427–38.
- Cheng L, Shi H, Wang Z, Hu Y, Yang H, Zhou C, Sun J, Zhou M. IntNetLncSim: an integrative network analysis method to infer human lncRNA functional similarity. *Oncotarget*. 2016;7(30):47864–74.
- Hu Y, Zhang Y, Ren J, Wang Y, Wang Z, Zhang J. Statistical approaches for the construction and interpretation of human protein-protein interaction network. *BioMed Res Int*. 2016;5313050.
- Song S, Hao J, Liu Y, Sun J. Improved EGT-Based Robustness Analysis of Negotiation Strategies in Multiagent Systems via Model Checking. *IEEE Trans Human-Mach Syst*. 2015;86(86):1–12.
- Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet*. 2012;13(8):523–36.
- Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet*. 2011;12:56–68.
- Schadt EE. Molecular networks as sensors and drivers of common human diseases. *Nature*. 2009;461(7261):218–23.
- Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. *J Med Genet*. 2006;43(8):691–8.
- Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc Natl Acad Sci U S A*. 2004;101(42):15148–53.
- Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*. 2008;82(4):949–58.
- Li Y, Patra JC. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*. 2010;26(9):1219–24.
- Vanunu O, Magger O, Ruppim E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol*. 2010;6:e1000641.
- Van Dongen S. Graph clustering via a discrete uncoupling process. *SIAM J Matrix Anal Appl*. 2008;30:121–41.
- Navlakha S, White J, Nagarajan N, Pop M, Kingsford C. Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. In: *Research in Computational Molecular Biology*. Springer; 2009. p. 400–17.
- Goel R, Harsha H, Pandey A, Prasad TK. Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. *Mol bioSystems*. 2012;8(2):453–63.

38. Brown KR, Jurisica I. Online predicted human interaction database. *Bioinformatics*. 2005;21(9):2076–82.
39. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015;43(D1):D789–98.
40. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. 2014;11(3):333–7.
41. Wang J, Chen G, Li M, Pan Y. Integration of breast cancer gene signatures based on graph centrality. *BMC Syst Biol*. 2011;5(3):1.
42. Liekens AM, De Knijff J, Daelemans W, Goethals B, De Rijk P, Del-Favero J, et al. BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol*. 2011;12(6):R57.
43. Ganegoda GU, Wang J, Wu FX, Li M. Prediction of disease genes using tissue-specified gene-gene network. *BMC Syst Biol*. 2014;8(Suppl 3):S3.
44. Eronen L, Toivonen H. Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*. 2012;13:1.
45. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, Schriml LM, Kibbe WA, Schofield PN, Beck T, et al. The human phenotype ontology: semantic unification of common and rare disease. *Am J Hum Genet*. 2015;97:111–24.
46. Kibbe WA, Arze C, Felix V, Mitraga E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res*. 2015;43(D1):D1071–8.
47. Consortium GO, et al. Gene ontology consortium: going forward. *Nucleic Acids Res*. 2015;43(D1):D1049–56.
48. Peng J, Wang T, Wang J, Wang Y, Chen J. Extending gene ontology with gene association networks. *Bioinformatics*. 2016;32(8):1185–94.
49. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2015;43(D1):D447–D452.
50. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42(D1):D980–5.
51. Cheng L, Wang G, Li J, Zhang T, Xu P, Wang Y. SIDD: a semantically integrated database towards a global view of human disease. *PLoS ONE*. 2013;8(10):e75504.
52. Backstrom L, Leskovec J. Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the fourth ACM international conference on Web search and data mining. Kowloon: ACM; 2011. p. 635–44.
53. Johnson R, Zhang T. On the Effectiveness of Laplacian Normalization for Graph Semi-supervised Learning. *J Mach Learn Res*. 2007;8(4):1489–1517.
54. Tong H, Faloutsos C, Pan JY. Random walk with restart: fast solutions and applications. *Knowl Inf Syst*. 2008;14(3):327–46.
55. Mattingly C, Rosenstein M, Colby G, Forrest J, Boyer J. The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *J Exp Zool Part A Comparative Exp Biol*. 2006;305(9):689–92.
56. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33(suppl 1):D514–7.
57. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. The HUGO gene nomenclature committee (HGNC). *Hum Genet*. 2001;109(6):678–80.
58. Lipscomb CE. Medical subject headings (MeSH). *Bull Med Libr Assoc*. 2000;88(3):265.
59. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32(suppl 1):D267–70.
60. Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res*. 2012;40(D1):D940–6.
61. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res*. 2014;42(D1):D966–74.
62. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;25:25–9.
63. Kumar R, Raghavan P, Rajagopalan S, Sivakumar D, Tomkins A, Upfal E. Stochastic models for the web graph. In: Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on. Redondo Beach: IEEE; 2000. p. 57–65.
64. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29–36.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

