SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •
Special Focus on Biomolecular Network Analysis and Application

July 2016, Vol. 59 070101:1–070101:10
doi: 10.1007/s11432-016-5584-y

Integrating phenotypic features and tissue-specific information to prioritize disease genes

Yue DENG[1,2], Lin GAO[1*], Xingli GUO[1] & Bingbo WANG[1]

[1]*School of Computer Science and Technology, Xidian University, Xi'an* 710071*, China;*
[2]*School of Software, Xidian University, Xi'an* 710071*, China*

Received March 31, 2016; accepted April 18, 2016; published online June 6, 2016

Abstract Prioritization of candidate disease genes is crucial for improving medical care, and is one of the fundamental challenges in the post-genomic era. In recent years, different network-based methods for gene prioritization are proposed. Previous studies on gene prioritization show that tissue-specific protein-protein interaction (PPI) networks built by integrating PPIs with tissue-specific gene expression profiles can perform better than tissue-naïve global PPI network. Based on the observations that diseases with similar phenotypes are likely to have common related genes, and genes associated with the same phenotype tend to interact with each other, we propose a method to prioritize disease genes based on a heterogeneous network built by integrating phenotypic features and tissue-specific information. In this heterogeneous network, the PPI network is built by integrating phenotypic features with a tissue-specific PPI network, and the disease network consists of the diseases that are associated with the same phenotype and tissue as the query disease. To determine the impacts of these two factors on gene prioritization, we test three typical network-based prioritization methods on heterogeneous networks consisting of combinations of different PPIs and disease networks built with or without phenotypic features and tissue-specific information. We also compare the proposed method with other tissue-specific networks. The results of case studies reveals that integrating phenotypic features with a tissue-specific PPI network improves the prioritization results. Moreover, the disease networks generated using our method not only show comparable performance with the widely used disease similarity dataset of 5080 human diseases, but are also effective for diseases that are not in the dataset.

Keywords gene prioritization, tissue-specific network, phenotype, PPI network, disease network

Citation Deng Y, Gao L, Guo X L, et al. Integrating phenotypic features and tissue-specific information to prioritize disease genes. Sci China Inf Sci, 2016, 59(7): 070101, doi: 10.1007/s11432-016-5584-y

1 Introduction

To study the causes of human genetic disease, researchers often have to deal with a large number of candidate genes identified by positional genetic studies. Since biological experimental verification of candidate genes is expensive and time-consuming, computational methods for gene prioritization attract a great deal of attention, with the goal to identify the most promising genes within the list of candidate genes.

Various computational disease gene prioritization methods are proposed in recent years [1–3]. Most methods rank candidate genes based on their similarities to known disease genes using the principle of guilt by association. Because the completeness of human protein-protein interaction (PPI) network is

* Corresponding author (email: lgao@mail.xidian.edu.cn)

rapidly approaching completion owing to advances in high-throughput techniques, many network-based methods are introduced recently [4,5]. Some methods, such as CIPHER [6], PRIoritizatioN and Complex Elucidation (PRINCE) [7] and Random Walk with Restart on Heterogeneous Network (RWRH) [8], mainly use a PPI network and known disease-gene associations to construct the network. Other methods integrate different types of biological data to improve the results of prioritization. These data include gene expression profiles [9, 10], functional annotations [11–13], and data from model organisms [14, 15].

However, these methods ignore the fact that the majority of hereditary diseases tend to affect only a single or a few tissues [16, 17]. For example, analyzing a brain disease using the same tissue-naïve global PPI network as that for a skin disease is not appropriate because the genes expressed in the brain and skin are quite different. Thus, the concept of the tissue-specific network, which consists of genes that are expressed in a certain tissue, is proposed. Using tissue-specific networks, researchers can study the disease-gene associations that are specific to a single gene in a single tissue. Researchers build tissue-specific networks by integrating PPI networks and tissue-specific gene expression profiles. Magger et al. [18] build tissue-specific PPI networks for 60 tissues. The gene expression profiles for different tissues are downloaded from the Gene Expression Omnibus. The global PPI network is constructed by integrating interactions in the HPRD database [19] and high throughput experiments. Experiments show that using tissue-specific PPI networks can achieve prioritization results that are better than those using global PPI network. Barshir et al. [20] construct 16 tissue-specific PPI networks. Gene expression data are collected from three major resources: Su et al. [21], the Human Protein Atlas (HPA) [22], and the Illumina Body Map 2.0 RNA-seq data [23]. The global PPI network is assembled from four public PPI databases: BioGRID [24], DIP [25], IntAct [26], and MINT [27]. Subsequent comparative analysis [28] reveals that each tissue-specific network contains a core sub-network that is common to all tissues, with only a small fraction being tissue-specific. By integrating thousands of gene expression datasets using a Bayesian model, Greene et al. [29] build functional interaction networks for 144 human tissues and cell types. Li et al. [30] use DNA methylation data to weight the constructed tissue-specific PPI networks. The experimental results show that combining a tissue-specific PPI network with DNA methylation can impove the accuracy of the prioritization of disease genes. Ganegoda et al. [31] propose a novel method to construct tissue-specific gene networks, in which phenotype details are integrated to predict disease genes. Jacquemin et al. [32] construct a three-layer heterogeneous network composed of a disease network layer, a tissue-specific PPI network layer, and a protein complex layer. A random walk algorithm is run on this network to identify disease-related protein complexes.

The phenotypic features of genes and diseases play a significant role in studies on human diseases [33]. It is proven that genes with similar phenotypes tend to form biological modules, and genes with related functions lead to the same or similar phenotypes when mutated [34–36]. Thus, previous studies show that phenotype data constitute a powerful resources for disease gene prioritization. Phenolyzer [37] first maps input phenotypes to related diseases, then a machine learning model prioritizes candidate genes based on the knowledge of known disease genes. Phen-Gen [38] combines disease symptoms and sequencing data to identify the genes that cause rare disorders. Chen et al. [39] construct a phenome-interactome network by integrating a PPI network with the phenotype similarities between diseases. Potential disease genes are then predicted by maximizing the information flow in this network. Xie et al. [40] build a phenotype-gene association network using the Online Mendelian Inheritance in Man (OMIM) [41] database, and run a bi-random walk algorithm on it to prioritize the disease genes.

In this study, we propose a method to prioritize disease genes based on two-layer heterogeneous network built by integrating phenotypic features and tissue-specific information. The PPI network layer of the heterogeneous network is built by integrating phenotypic features with a tissue-specific PPI network, and the disease network layer consists of the diseases that are associated with the same phenotype and same tissue of the query disease.

To determine the impact of phenotypic features and tissue-specific information, we tested three typical network-based prioritization methods on multiple heterogeneous networks consisting of the combinations of different PPIs and disease networks built with or without phenotypic features and tissue-specific information. We also compare the proposed method with another tissue-specific network construction

method called the tissue-specified genes (TSG) method [31] to determine its effectiveness.

Case studies show that integrating phenotypic features to tissue-specific PPI networks improves the prioritization results. Not only can the constructed disease network achieve results that are comparable with those from the widely used disease similarity dataset of 5080 human diseases, but it is also effective for diseases that are not in the dataset.

2 Methods

A two-layer heterogeneous network model is used in our study, which comprises a PPI network and a disease network. We integrate two factors, phenotypic features and tissue-specific information, in the construction of both the PPI and disease networks.

2.1 Construction of the PPI network

Tissue-specific PPI networks built by filtering the global PPI network with tissue-specific expression profiles [18, 28, 29] produce better prioritization results compared with those obtained using a global PPI network. Previous studies also prove that genes with similar phenotypes yield biological modules in terms of diseases, and can therefore be used to predict disease genes [34, 35]. In our study, we integrate phenotypic features with a tissue-specific PPI network to further refine the PPI network.

For a query disease, its associated tissue is obtained as mentioned in [42], in which Lage et al. map approximately 1000 diseases to human tissues based on the co-occurrence of tissue-disease pairs in published literatures in PubMed. The tissue with the highest score among all tissues is considered to be associated with the disease. The tissue-specific PPI network is then constructed for the tissue using the MyProteinNet web server [43]. Using the default setting for MyProteinNet, a global PPI network is generated by combing four public PPI databases: BioGRID [24], IntAct [26], DIP [25], and MINT [27]. Three resources for expression data across 16 tissues are supported by MyProteinNet: the DNA microarray data set of Su et al. [21], the protein immunohistochemistry data in HPA [22], and the Illumina Body Map 2.0 RNA-seq data [23]. All the proteins not expressed in the tissue are removed. We further filter the tissue-specific PPI network by the phenotypic features of the proteins. Proteins that have no common annotation in the Human Phenotype Ontology (HPO) [44] with any known disease proteins of the query disease are removed.

A weight is then assigned to each edge in the resulting network. We use HPO-based phenotypic similarity between two interacting proteins as the weight of the interaction. We use the semantic similarity measure proposed by Resnik et al. [45]. The Resnik measure is based on the information content (IC) of the HPO terms to which the proteins are annotated. The IC of a term t in HPO is defined as follows:

$$\mathrm{IC}\,(t) = -\log\left(p\,(t)\right), \tag{1}$$

where $p(t)$ is the probability of observing t and the descendants of t in all proteins annotated to the phenotypic abnormality (PA) sub-ontology of HPO. The PA sub-ontology is chosen because it contains approximately 99% of all the HPO terms. The Resnik measure defines the similarity between HPO terms t_1 and t_2 as the IC of the most informative common ancestor (MICA) of these two terms:

$$\mathrm{sim}_{\mathrm{Resnik}}(t_1, t_2) = \mathrm{IC}(t_{\mathrm{MICA}}), \tag{2}$$

where t_{MICA} is the common ancestor of t_1 and t_2 that has the highest IC among all the common ancestors. The similarity between two proteins is calculated using the pairwise similarities between each term in the HPO term sets annotating these proteins. The funSimMax method [46] is used to combine multiple term-term similarities into protein-protein similarity. Given protein p_1 annotated by the HPO term set $T_1 = \{t_{11}, t_{12}, \ldots, t_{1m}\}$ and p_2 annotated by $T_2 = \{t_{21}, t_{22}, \ldots, t_{2n}\}$. The similarity matrix $S = (s_{ij})_{m\times n}$ consists of all pairwise similarities of terms in T_1 and T_2. The similarity between p_1 and p_2 is calculated as follows:

$$\mathrm{Sim}(p_1, p_2) = \max\left\{\frac{1}{m}\sum_{i=1}^{m}\max_{1\leqslant j\leqslant n} s_{ij}, \frac{1}{n}\sum_{j=1}^{n}\max_{1\leqslant i\leqslant m} s_{ij}\right\}. \tag{3}$$

2.2 Construction of the disease network

The similarity matrix of 5080 diseases provided by van Driel et al. [35] is widely used as the disease network in previous studies [6–8]. However, it has two drawbacks: first, it cannot be used on diseases that are not included in the dataset; second, even after removing the similarities that are less than 0.3 as suggested by the provider, the disease network using the similarity matrix as its weights is still very dense, and one disease has nearly 2000 similar diseases on average.

Instead of integrating phenotypic features and tissue-specific information into the disease network derived from the similarity matrix, we integrate the phenotype annotations in HPO with the tissue-specific information of diseases to build a new disease network.

The nodes of our proposed disease network for the query diseases are obtained in two steps. First, all the diseases in OMIM that have at least one common annotation in HPO with the query disease are collected. Then, the diseases that do not associate with the same tissue as the query disease are removed.

Any two nodes in the network are linked by an edge. The weight of each edge is the phenotypic similarity between the two diseases. The same method used in weighting the tissue-specific PPI network is used here. Given disease d_1 annotated by the HPO term set $T_1 = \{t_{11}, t_{12}, \ldots, t_{1m}\}$ and d_2 annotated by $T_2 = \{t_{21}, t_{22}, \ldots, t_{2n}\}$. The similarity matrix $S = (s_{ij})_{m\times n}$ consists of all pairwise similarities of terms in T_1 and T_2. The similarity between d_1 and d_2 is calculated as follows:

$$\mathrm{Sim}(d_1, d_2) = \max\left\{\frac{1}{m}\sum_{i=1}^{m}\max_{1\leqslant j\leqslant n} s_{ij}, \frac{1}{n}\sum_{j=1}^{n}\max_{1\leqslant i\leqslant m} s_{ij}\right\}. \tag{4}$$

2.3 Gene prioritization on the heterogeneous network

To construct the heterogeneous network, the PPI and the disease network described in Subsections 2.1 and 2.2 are linked together by adding edges representing known disease-protein associations. Known disease-protein associations are extracted from the OMIM database [41]. It should be noted that the heterogeneous network is constructed for the tissue that the query disease affects, so it is not the same for different diseases.

Three typical network-based prioritization methods are adapted to the constructed heterogeneous network: RWRH [8], PRINCE [7], and Guo's method [47]. All the methods take a query disease as the input, and give a ranked list of candidate genes as the output. The framework of our method is shown in Figure 1.

RWRH [8] performs a random walk on a two-layer heterogeneous network consisting of a disease network and a PPI network. Random walkers start from both the query disease and its associated proteins. In each step, the walker may move randomly to one of its direct neighbors in the same layer or jump to another layer if there is an edge linking two layers. After the random walk achieves a steady distribution, all the nodes in the network are ranked by the probability of the walker reaching this node.

PRINCE [7] uses a PPI network and similarities between diseases as input. PRINCE assigns scores to proteins known to be associated with diseases that are similar to the query disease. The score is then propagated through the PPI network in an iterative process, which simulates a process where proteins pump information to their neighbors in the PPI network.

Guo's method [47] integrates a PPI network and a disease similarity network for prioritizing candidate genes. The association score between a query disease and a candidate gene is defined as the weighted sum of all the association scores between similar diseases and interacting genes.

3 Results and discussion

To assess the contributions made by phenotypic features and tissue-specific information to the prioritization results, three network-based prioritization methods are tested on heterogeneous networks built with the combinations of different PPI networks and disease networks.

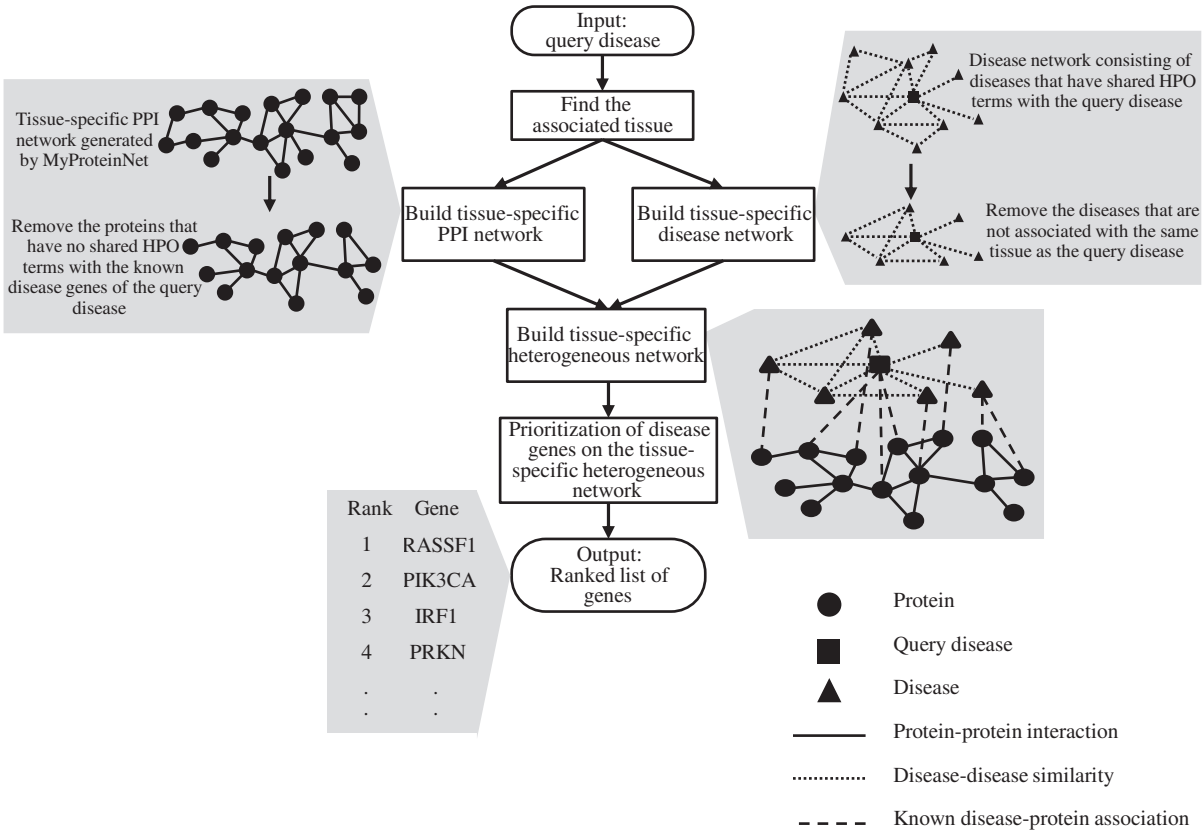

Figure 1 Graphical overview of the prioritization framework.

Three methods of building PPI networks are tested. Our method of generating the PPI network is described in Subsection 2.1. To compare the performance of prioritization methods for PPI networks with or without phenotypic features and tissue-specific information, a tissue-naïve global human PPI network (referred to as GPN) and a tissue-specific PPI network for the tissue affected by the query disease (referred to as TSPN) are generated using the MyProteinNet web server [43].

Three methods of building disease networks are also tested. Besides our method of building the disease network described in Subsection 2.2, two disease networks are constructed based on the disease similarity matrix provided by van Driel et al. [35]. The similarity matrix of 5080 diseases can be obtained from MimMiner [35], in which the similarity between two diseases is calculated using a text-mining approach based on the records of phenotypes contained in the OMIM database. A disease network containing all the 5080 diseases can be constructed using the pairwise similarities as the weights of the edges. After removing the edges whose weights are less than 0.3 as suggested in [35], the resulting disease network is referred to as GDN. All the diseases that do not associate with the tissue affected by the query disease are then removed from GDN, and the resulting network is referred to as TSDN.

We apply the RWRH [8], PRINCE [7], and Guo's method [47] to the combinations of different PPI networks and disease networks. Lung cancer (OMIM: 211980), breast cancer (OMIM: 114480), and ovarian cancer (OMIM: 167000) are used as case studies. The associated tissue of breast cancer is prostate associated with only three diseases in the dataset provided by van Driel et al. [35], so the disease network built using our method only has no more than three nodes. We choose breast cancer to test the performance of our proposed disease network when it contains only a few nodes. Ovarian cancer is chosen because it is not included in the dataset provided by van Driel et al. [35], so the effectiveness of the disease network built using our method can be tested.

The proposed method is also compared with the tissue-specified genes (TSG) method introduced by Ganegoda et al. [31]. The tissue-specific gene expression data set is used to construct the tissue-specific gene network. Each interaction in the tissue-specific gene network is then weighed by considering two

Table 1 Topological properties of the PPI networks of lung cancer

Network	#nodes(proteins)	#edges(interactions)
GPN	10372	61846
TSPN	8370	46219
Proposed method	1356	3482

Table 2 Topological properties of the disease networks of lung cancer

Network	#nodes(diseases)	#edges(associations)
GDN	5080	5003221
TSDN	17	29
Proposed method	13	78

factors: the pearson correlation coefficient of the gene expression of two interacting genes, and the relationship between genes and different phenotypes. The disease-tissue associations are obtained from Lage et al. [42]. The disease network is built by selecting the most similar phenotypes for the query disease using the MimMiner approach [35].

3.1 Case study on lung cancer

First, three different PPI networks are generated for lung cancer. The global human PPI network (GPN) and the tissue-specific PPI network (TSPN) for lung are generated using MyProteinNet [43]. In our method, we only keep the nodes in TSPN that have shared HPO annotation with the known diseases genes of lung cancer. All the protein IDs are mapped to Entrez IDs for the calculation of phenotypic similarities later. The topological properties of the PPI networks are shown in Table 1.

Three different disease networks are then generated for lung cancer. The GDN and TSDN are generated from the disease similarities matrix provided by van Driel et al. [35]. Using our method described in Subsection 2.2, all the diseases in OMIM that have at least one common annotation in HPO with lung cancer are collected. Then diseases that are not associated with lung are then removed. The topological properties of the disease networks are shown in Table 2. It can be seen from Table 1 and Table 2 that integrating phenotypic features significantly downsizes the PPI and disease networks, which reduces the amount of calculation.

To evaluate the performance of the prioritization methods on the combination of different PPI networks and disease networks, we used a leave-one-out cross validation procedure. For lung cancer, there are 16 known disease genes listed in OMIM, of which 13 genes are included in the GPN. These 13 genes are also included in the TSPN. In each round of cross-validation, we chose one of the known disease genes and remove the link between this gene and lung cancer from the network. The cross-validation procedure runs for 13 rounds to choose all the 13 genes one at a time.

We test three methods on heterogeneous networks consisting of the combination of three different PPI networks and three disease networks mentioned above. These three methods are also tested on the TSG network [31]. Two criteria are used to evaluate the results. First, we calculate the mean rank (MR) of all the known disease genes using the top 100 genes as the rank list. Second, we calculate the area under the curve (AUC) value below the area of the fraction of disease genes ranking above a particular threshold and the fraction of control genes ranking below this threshold. A comparison is performed between the different heterogeneous networks, and the results are shown in Table 3. The best MR and AUC values for each method are shown in bold.

It is clear that the PPI networks involving tissue information outperform the global network, and integrating phenotypic features can improve the prioritization results. It can also be seen from Table 3 that integrating tissue-specific information and phenotypic features with the PPI network is more helpful than with the disease network. It is also clear that the proposed method produces better results than the TSG method.

Table 3 Comparison of the performance of three methods on different heterogeneous networks of lung cancer

PPI network	Disease network	MR (RWRH)	AUC (RWRH)	MR (PRINCE)	AUC (PRINCE)	MR (Guo)	AUC (Guo)
GPN	GDN	18.15	0.75	18.69	0.738	16.85	0.758
GPN	TSDN	17.69	0.767	18	0.758	16.31	0.764
GPN	Proposed method	17.23	0.768	18.08	0.763	15.85	0.788
TSPN	GDN	13.77	0.797	13.38	0.794	12.31	0.821
TSPN	TSDN	13.54	0.802	13	0.798	12.08	0.835
TSPN	Proposed	13.69	0.808	13.08	0.806	12.23	0.827
Proposed method	GDN	12.23	0.825	12.08	0.822	11.31	0.858
Proposed method	TSDN	12.15	0.83	11.54	0.824	11.15	0.874
Proposed method	Proposed method	**12.08**	**0.84**	**11.46**	**0.834**	**10.92**	**0.881**
TSG	TSG	12.38	0.819	12.15	0.821	11.77	0.848

Note: Smaller MR values and larger AUC values indicate higher performance.

Table 4 Topological properties of the PPI networks of breast cancer

Network	#nodes(proteins)	#edges(interactions)
GPN	10372	61846
TSPN	7855	42892
Proposed method	1262	3264

Table 5 Topological properties of the disease networks of breast cancer

Network	#nodes(diseases)	#edges(associations)
GDN	5080	5003221
TSDN	3	2
Proposed method	2	1

Table 6 Comparison of the performance of three methods on different heterogeneous networks of breast cancer

PPI network	Disease network	MR (RWRH)	AUC (RWRH)	MR (PRINCE)	AUC (PRINCE)	MR (Guo)	AUC (Guo)
GPN	GDN	14.24	0.742	14.35	0.804	13.82	0.814
GPN	TSDN	15.65	0.714	15.47	0.775	14.65	0.79
GPN	Proposed method	14.29	0.73	14.41	0.78	13.65	0.817
TSPN	GDN	10.24	0.816	10.18	0.829	9.06	0.885
TSPN	TSDN	10.71	0.809	10.24	0.828	9.35	0.864
TSPN	Proposed method	10.76	0.791	11.12	0.818	10.53	0.841
Proposed method	GDN	**9.93**	**0.837**	**9.36**	**0.844**	8.43	0.896
Proposed method	TSDN	10.07	0.828	10.14	0.832	9.14	0.876
Proposed method	Proposed method	10.14	0.816	9.86	0.839	8.57	0.89
TSG	TSG	10	0.819	9.57	0.836	**8.36**	**0.907**

Note: Smaller MR values and larger AUC values indicate higher performance.

3.2 Case study on breast cancer

Three PPI networks and three disease networks are constructed using the same approach used in the case study on lung cancer. The topological properties of the PPI and disease networks are shown in Tables 4 and 5.

For breast cancer, there are 23 known disease genes listed in OMIM, of which 22 genes are included in the GPN. Of those 22 genes, 17 are included in the TSPN and 14 have annotations in HPO. Thus, the cross-validation procedure runs for 17 rounds for GPN/TSPN and 14 rounds for the disease network generated using our method. A comparison is performed between different heterogeneous networks, and the results are shown in Table 6. The best MR and AUC values for each method are shown in bold.

It can be seen from Table 6 that the PPI networks involving tissue information still outperform the

Table 7 Topological properties of the PPI networks of ovarian cancer

Network	#nodes(proteins)	#edges(interactions)
GPN	10372	61846
TSPN	7807	42713
Proposed method	1211	3093

Table 8 Comparison of the performance of three methods on different heterogeneous networks of ovarian cancer

PPI network	Disease network	MR (RWRH)	AUC (RWRH)	MR (PRINCE)	AUC (PRINCE)	MR (Guo)	AUC (Guo)
GPN	Proposed method	13	0.726	13	0.728	11.8	0.758
TSPN	Proposed method	10.6	0.766	10.2	0.771	9.4	0.811
Proposed method	Proposed method	**10**	**0.808**	**9.8**	**0.815**	**8.2**	**0.842**

Note: Smaller MR values and larger AUC values indicate higher performance.

global network. However, the result hardly improves after integrating phenotype information. In the disease-tissue associations provided by Lage et al. [42], breast cancer is associated with the prostate. However, there are only two diseases associated with prostate besides breast cancer: bladder cancer (OMIM:109800) and prostate cancer (OMIM:176807). Thus, the TSDN and disease network generated using our method only have two and three nodes, which cannot provide much additional information from the diseases similar to breast cancer.

Because the disease network used in the TSG method is based on GDN, the TSG method produces better results compared with our proposed disease network in Guo's method.

For diseases related to tissue that only have few associated diseases, we suggest that the disease similarity dataset provided by van Driel et al. [35] should be used as the disease network. Alternatively, other disease-tissue associations should be integrated besides the data provided by Lage et al. [42].

3.3 Case study on ovarian cancer

The disease similarity dataset provided by van Driel et al. [35] is used extensively in existing prioritization methods. However, these methods do not work for diseases that are not included in the dataset. To work around this limitation, we determine the effectiveness of the disease network constructed by integrating phenotypic features and tissue-specific information. Ovarian cancer, which is not in the dataset, is chosen as the third case study.

Using the same approach used in the previous case studies, three PPI networks are constructed. The topological properties of the PPI networks are shown in Table 7.

In this case, we combine three PPI networks with a disease network generated using our method to test the performance of PRINCE, RWRH, and Guo's method. The disease network generated using our method consists of 6 nodes and 15 edges. Because the ovarian cancer is not included in the dataset provided by van Driel et al. [35], the TSG method is not used in this case study.

For ovarian cancer, there are six known disease genes listed in OMIM, of which five genes are included in the GPN, TSPN, and PPI network generated using our method. Thus, the cross-validation procedure runs for five rounds. A comparison is performed between different heterogeneous networks, and the results are shown in Table 8. The best MR and AUC values for each method are shown in bold.

It can be seen from Table 8 that the PPI networks that involve tissue information still outperform the global network. The disease network constructed using our method can be used to prioritize the diseases that are not included in the dataset provided by van Driel et al. [35].

4 Conclusion

The purpose of this research is to determine the contributions made by phenotypic features and tissue-specific information to disease gene prioritization. We propose a method to prioritize disease genes based

on a heterogeneous network built by integrating phenotypic features and tissue-specific information about proteins and diseases.

To validate the effectiveness of the proposed method, we build different PPI and disease networks with or without phenotypic features and tissue-specific information. RWRH, PRINCE, and Guo's method are used on the heterogeneous networks built by the combinations of these PPI and disease networks. We also compared the proposed method with another heterogeneous network model that is also based on the tissue-specific network.

The results of the case studies show that integrating phenotypic features with tissue-specific PPI networks improves the prioritization results. The disease network built using our method not only produces results that are comparable with the widely used disease similarity dataset provided by van Driel et al. [35], but is also effective for diseases that are not in the dataset.

Our method can be extended further by applying the following directions: first, more strategies to construct the disease network may be considered, including the human symptoms-disease network [48] and the human disease network [49]; second, other types of network, such as the tissue-specific functional interaction network [29], may be used instead of a PPI network; third, the phenotypic features contained in HPO are accurate but limited in size, so other resources of phenotypic features may be integrated with HPO to further explore the use of phenotypic features in the prioritization of candidate disease genes.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61532014, 91530113, 61432010, 61402349, 61303122, 61303118) and Fundamental Research Funds for the Central Universities (Grant No. BDZ021404).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Ritchie M D, Holzinger E R, Li R, et al. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet, 2015, 16: 85–97
2 Moreau Y, Tranchevent L-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet, 2012, 13: 523–536
3 Piro R M, Di Cunto F. Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J, 2012, 279: 678–696
4 Wang X J, Gulbahce N, Yu H Y. Network-based methods for human disease gene prediction. Brief Funct Genomics, 2011, 10: 280–293
5 Lan W, Wang J X, Li M, et al. Computational approaches for prioritizing candidate disease genes based on PPI networks. Tsinghua Sci Technol, 2015, 20: 500–512
6 Wu X B, Jiang R, Zhang M, et al. Network-based global inference of human disease genes. Mol Syst Biol, 2008, 4: 189
7 Vanunu O, Magger O, Ruppin E, et al. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol, 2010, 6: e1000641
8 Li Y J, Patra J C. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics, 2010, 26: 1219–1224
9 Wang J X, Peng X Q, Peng W, et al. Dynamic protein interaction network construction and applications. Proteomics, 2014, 14: 338–352
10 Gaulton K J, Mohlke K L, Vision T J. A computational system to select candidate genes for complex human traits. Bioinformatics, 2007, 23: 1132–1140
11 Schlicker A, Lengauer T, Albrecht M. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. Bioinformatics, 2010, 26: i561–i567
12 Linghu B, Snitkin E S, Hu Z, et al. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol, 2009, 10: R91
13 Franke L, van Bakel H, Fokkens L, et al. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Amer J Hum Genet, 2006, 78: 1011–1025
14 Robinson P N, Webber C. Phenotype ontologies and cross-species analysis for translational research. PLoS Genet, 2014, 10: e1004268
15 Hwang S, Kim E, Yang S, et al. MORPHIN: a web tool for human disease research by projecting model organism biology onto a human integrated gene network. Nucl Acids Res, 2014, 42: W147–W153
16 Winter E E, Goodstadt L, Ponting C P. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res, 2004, 14: 54–61

17 Chao E C, Lipkin S M. Molecular models for the tissue specificity of DNA mismatch repair-deficient carcinogenesis. Nucl Acids Res, 2006, 34: 840–852
18 Magger O, Waldman Y Y, Ruppin E, et al. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. PLoS Comput Biol, 2012, 8: e1002690
19 Prasad T S K, Goel R, Kandasamy K, et al. Human protein reference database2009 update. Nucl Acids Res, 2009, 37: D767–D772
20 Barshir R, Basha O, Eluk A, et al. The tissuenet database of human tissue protein-protein interactions. Nucl Acids Res, 2013, 41: D841–D844
21 Su A I, Wiltshire T, Batalov S, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Nat Acad Sci USA, 2004, 101: 6062–6067
22 Berglund L, Björling E, Oksvold P, et al. A genecentric human protein atlas for expression profiles based on antibodies. Mol Cell Proteom, 2008, 7: 2019–2027
23 Bradley R K, Merkin J, Lambert N J, et al. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. PLoS Biol, 2012, 10: e1001229
24 Chatr-aryamontri A, Breitkreutz B-J, Oughtred R, et al. The BioGRID interaction database: 2015 update. Nucl Acids Res, 2015, 43: D470–D478
25 Salwinski L, Miller C S, Smith A J, et al. The database of interacting proteins: 2004 update. Nucl Acids Res, 2004, 32: D449–D451
26 Orchard S, Ammari M, Aranda B, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucl Acids Res, 2014, 42: D358–D363
27 Licata L, Briganti L, Peluso D, et al. MINT, the molecular interaction database: 2012 update. Nucl Acids Res, 2012, 40: D857–D861
28 Barshir R, Shwartz O, Smoly I Y, et al. Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. PLoS Comput Biol, 2014, 10: e1003632
29 Greene C S, Krishnan A, Wong A K, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet, 2015, 47: 569–576
30 Li M, Zhang J Y, Liu Q, et al. Prediction of disease-related genes based on weighted tissue-specific networks by using DNA methylation. BMC Med Genomics, 2014, 7: S4
31 Ganegoda G U, Wang J X, Wu F-X, et al. Prediction of disease genes using tissue-specified gene-gene network. BMC Syst Biol, 2014, 8: S3
32 Jacquemin T, Jiang R. Walking on a tissue-specific disease-protein-complex heterogeneous network for the discovery of disease-related protein complexes. BioMed Res Int, 2013, 2013: 455–458
33 Robinson P, Krawitz P, Mundlos S. Strategies for exome and genome sequence data analysis in disease-gene discovery projects. Clin Genet, 2011, 80: 127–132
34 Köhler S, Bauer S, Horn D, et al. Walking the interactome for prioritization of candidate disease genes. Amer J Hum Genet, 2008, 82: 949–958
35 van Driel M A, Bruggeman J, Vriend G, et al. A text-mining analysis of the human phenome. Eur J Hum Genet, 2006, 14: 535–542
36 Brunner H G, van Driel M A. From syndrome families to functional genomics. Nat Rev Genet, 2004, 5: 545–551
37 Yang H, Robinson P N, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods, 2015, 12: 841–843
38 Javed A, Agrawal S, Ng P C. Phen-Gen: combining phenotype and genotype to analyze rare disorders. Nat Methods, 2014, 11: 935–937
39 Chen Y, Jiang T, Jiang R. Uncover disease genes by maximizing information flow in the phenome-interactome network. Bioinformatics, 2011, 27: i167–i176
40 Xie M Q, Hwang T, Kuang R. Prioritizing disease genes by bi-random walk. In: Proceedings of 16th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Kuala Lumpur, 2012. 292–303
41 Hamosh A, Scott A F, Amberger J S, et al. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucl Acids Res, 2005, 33: D514–D517
42 Lage K, Hansen N T, Karlberg E O, et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc Nat Acad Sci, 2008, 105: 20870–20875
43 Basha O, Flom D, Barshir R, et al. MyProteinNet: build up-to-date protein interaction networks for organisms, tissues and user-defined contexts. Nucl Acids Res, 2015, 43: W258–W263
44 Köhler S, Doelken S C, Mungall C J, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. Nucl Acids Res, 2014, 42: D966–D974
45 Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc., 1995. 448–453
46 Schlicker A, Domingues F, Rahnenführer J, et al. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinform, 2006, 7: 302
47 Guo X L, Gao L, Wei C S, et al. A computational method based on the integration of heterogeneous networks for predicting disease-gene associations. PLoS ONE, 2011, 6: e24171
48 Zhou X Z, Menche J, Barabási A-L, et al. Human symptoms-disease network. Nat Commun, 2014, 5: 4212
49 Goh K-I, Cusick M E, Valle D, et al. The human disease network. Proc Nat Acad Sci, 2007, 104: 8685–8690