TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 08/10 pp500-512
Volume 20, Number 5, October 2015

Computational Approaches for Prioritizing Candidate Disease Genes Based on PPI Networks

Wei Lan, Jianxin Wang*, Min Li*, Wei Peng, and Fangxiang Wu

Abstract: With the continuing development and improvement of genome-wide techniques, a great number of candidate genes are discovered. How to identify the most likely disease genes among a large number of candidates becomes a fundamental challenge in human health. A common view is that genes related to a specific or similar disease tend to reside in the same neighbourhood of biomolecular networks. Recently, based on such observations, many methods have been developed to tackle this challenge. In this review, we firstly introduce the concept of disease genes, their properties, and available data for identifying them. Then we review the recent computational approaches for prioritizing candidate disease genes based on Protein-Protein Interaction (PPI) networks and investigate their advantages and disadvantages. Furthermore, some pieces of existing software and network resources are summarized. Finally, we discuss key issues in prioritizing candidate disease genes and point out some future research directions.

Key words: candidate disease-gene prioritization; protein-protein interaction network; human disease; computational tools

1 Introduction

Hereditary diseases are usually caused by some mutations of single or multiple genes in human genome. According to the number of mutated genes which are related to diseases, genetic disorders can be classified to single gene disorders (Mendelian disorder) and polygenic disorders (complex disorder)[1]. More than 4000 human diseases, such as duchenne muscular dystrophy, polycystic kidney disease, and sickle cell anemia, are caused by single gene mutation. However, most of these diseases are uncommon, for example, the incidence of Huntington's disease is approximate 1/15 000[2, 3]. In contrast, polygenic disorders such as cancers which are influenced by multiple genes and environmental factors are more common in public health. Discovering genes causing these diseases would assist biomedical researchers to understand the underlying mechanisms of actions, further conduce to disease diagnosis and treatment.

With the development of high-throughput techniques, a great deal of biological data has been and will continue to be generated. At present, there are two main genetic mapping approaches to generate candidate disease genes. The first is linkage analysis that aims at finding out the rough location of the gene relative to another DNA sequence called a genetic marker, which has its position already known. The second is Genomic-Wide Association Studies (GWAS) which are successful in detecting associations between variants and genetic disorders[4]. Both of these two types of approaches can find thousands of candidate disease

• Wei Lan, Jianxin Wang, Min Li, and Wei Peng are with the School of Information Science and Engineering, Central South University, Changsha 410083, China. E-mail: jxwang@mail.csu.edu.cn; limin@mail.csu.edu.cn.
• Fangxiang Wu is with the Department of Mechanical Engineering and Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada.
∗ To whom correspondence should be addressed.
Manuscript received: 2015-07-06; accepted: 2015-08-06

genes, but how to identify the most likely disease genes from these candidates is a great challenge for molecular biologists and medical geneticists. As limitations of experimental approaches such as time and labour, it is appealing to develop efficient computational methods to tackle this obstacle.

Recently, some computational approaches have been proposed to prioritize candidate disease genes from Protein-Protein Interaction (PPI) network[5–7]. The PPI network is one of the most important biological networks which has been widely used to predict protein functions[8–10], detect protein complexes[11, 12], identify essential proteins or genes[13, 14], and discover network motifs[15]. A PPI network can be presented as an undirected graph $G(V, E)$ where a set of nodes (V) denote proteins together while a set of edges (E) denote interactions between proteins. The emergence of disease is usually viewed as a consequence of perturbation of a PPI network[16, 17], as shown in Fig. 1. Mutation of nodes, removal of edges or anomaly in modules all can cause different diseases. In this review, our aim is to summarize these approaches which are used to prioritize candidate disease genes based on PPI networks and try to assist readers to keep up with recent and important developments in this filed. The paper is organized as follows: In the second section, we introduce some available data resources for candidate gene prioritization. In the third section, we present the recent computational approaches. In the fourth section, existing and available tools based on PPI data are summarized. Then the paper concludes with highlighting the key issues and the future discussions of this filed.

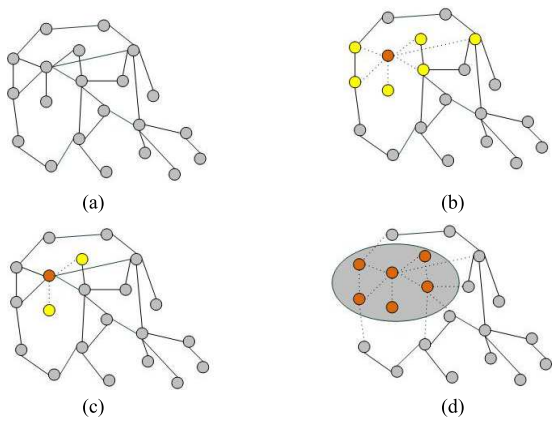


Fig. 1 The association between disease and PPI networks due to the perturbation of PPI networks. Abnormal of nodes, edges, and modules may all lead to disease. (a) The normal PPI network; (b) The node variation in the PPI network; (c) The interaction removal in the PPI network; and (d) The module abnormality of the PPI network.

2 Biological Data Resources

Recently, with the rapid increase of biological data, some specific databases have been built to store and manage these data. These available resources have greatly improved disease gene prioritization by providing various biological data to researchers for constructing computational models. In the rest of this section, we describe the public biological data in terms of different categories.

2.1 Disease gene data

Genomic-wide association studies provide an effective way to explore the genetic basis of complex traits. The disease data have been dramatically increased with the development of genotyping technology. Many disease repositories have been built to collect and store these data. Here, we describe three important disease databases, as shown in Table 1. Online Mendelian Inheritance in Man (OMIM)[18] is a comprehensive database of human genes and genetic disorders, which is maintained by McKusick-Nathans Institute of Genetic Medicine, School of Medicine, Johns Hopkins University. Up to July 2015, there are 23 034 entries: 14 972 for gene description, 86 for the combination of genes and phenotypes, 4499 for molecular basis known phenotype description, 1654 for molecular basis unknown phenotype description or locus, and 1823 for phenotypes with suspected Mendelian basis. The phenotypes of OMIM database primarily describe single gene Mendelian disorders, and also the complex traits for which mutation in a single gene result in a significant contribution to the phenotype. The Genetic Association Database (GAD)[19, 20] is a web-based knowledgebase of human complex diseases and disorders. The summary data are extracted from published papers. There are 5526 diseases classified into 19 categories such as cancer, aging, immune, and so on. The GAD contains 167 130 records, out of which 82 285 records are given a description of whether they are reported to be associated or not with the disease phenotype for that specific record or not. The DisGeNET database[21] integrates associations from several sources that cover different biomedical aspects of diseases and thus provides comprehensive gene-disease association information. The DisGeNET contains 100 729 associations between 9313 genes

Table 1 An overview of disease gene database.

Database	Disease	Genes	Records	URL	Reference
OMIM	Over 7000	14 972	23 034	http://www.omim.org/	[18]
GAD	5526	about 8000	167 130	http://geneticassociationdb.nih.gov	[19, 20]
DisGeNET	6029	9313	100 729	http://ibi.imim.es/DisGeNET/web/v02/home	[21]

and 6029 diseases including Mendelian, complex, and environmental diseases.

2.2 PPI data

With the development of high-throughput technologies such as yeast two-hybrid map, immunoprecipitation, and mass spectrometry technology, a mass of PPI information has been obtained[22–24]. The alternative way to gain PPI data is based on co-occurrence in Ref. [25] and some databases have been built to store literature-based PPI information such as the Biomolecular Interaction Network Database (BIND)[26], the Database of Interacting Proteins (DIP)[27], the Molecular INTeraction (MINT) database[28], and the protein InterAction (IntAct) database[29]. Table 2 presents an overview of PPI database.

More recent PPI curation efforts have attempted larger scale curation of data, such as the Biological General Repository for Interaction Datasets (BioGRID)[30] and the Human Protein Reference Database (HPRD)[31]. The BioGRID (http://thebiogrid.org/)[30] is a database of physical and genetic interactions, which contains 36 species and 471 829 interactions in the recent version. The Homo sapiens has 17 624 proteins and 130 503 interactions. The HPRD (http://hprd.org/)[31] is a repository of human protein information, which provides PPI data, Post-Translational Modifications (PTMs), and tissue expression data. The latest version of HPRD contains 30 047 proteins and 41 327 interactions. In addition, Search Tool for the Retrieval of Interacting Genes (STRING)[32] and Human Integrated Protein-Protein Interaction rEference (HIPPIE) databases[33] try to construct comprehensive PPI dataset by integrating different organisms or data resources. The STRING database (http://string-db.org) contains protein interaction information in approximate 1100 organisms where protein-protein associations are weighted in term of scoring scheme. The text-mining information has been updated into new version[32]. The HIPPIE database (http://cbdm.mdc-berlin.de/tools/hippie) provides comprehensive PPI dataset by integrating multiple experimental PPI resources with normalized scoring scheme[33].

3 Prioritizing Disease Genes Based on PPI Networks

In the following, we review computational approaches of disease gene prioritization from PPI networks. The core concept of disease gene prioritization from a PPI network is that genes associated with phenotypically close disorders are likely to locate closely to each other in the PPI network[34–36].

3.1 Distance-based methods

The assumption of disease gene prioritization is mostly based on the principle of "guilt-by-association"[34]. Hence, the distance between candidate disease genes and known disease genes seems a good way for disease gene prioritization. If a candidate disease gene has a long distance to the known disease

Table 2 An overview of protein-protein interaction database.

Database	Species	Number of entries	URL	Reference
BIND	About 1500 species	200 000 interactions	http://bind.ca	[26]
DIP	10 species	25 612 proteins and 75 400 interactions	http://dip.doe-mbi.ucla.edu	[27]
MINT	6 species	35 511 proteins and 241 458 interactions	http://mint.bio.uniroma2.it/mint	[28]
IntAct	Over 15 species	65 200 proteins and 312 217 interactions	http://www.ebi.ac.uk/intact	[29]
BioGRID	51 species	471 829 interactions	http://thebiogrid.org	[30]
HPRD	Homo Sapiens	30 047 proteins and 41 327 interactions	http://hprd.org	[31]
STRING	Over 1100 species	5 million proteins and 200 million interactions	http://string-db.org	[32]
HIPPIE	Homo Sapiens	72 916 interactions	http://cbdm.mdc-berlin.de/tools/hippie	[33]

genes, then it has a low chance to be a disease gene. The simplest method to measure the distance between two given genes is to detect whether their corresponding proteins are connected directly in the PPI network. Oti et al.[37] proposed a direct neighbour-based method to predict disease candidates with known disease loci. Hsu et al.[38] introduced a nearest-neighbour-based method to prioritize disease genes. They used the interconnectedness (ICN) to evaluate the closeness between the candidate disease genes and the known disease genes.

Considering that disease genes are generally involved in same pathways instead of physically interacting, the direct neighbour method fails to capture this kind information. Some researchers use the shortest path method to measure the closeness between two proteins. Zhu et al.[39] proposed a Vertex Similarity (VS) method based on shortest paths to rank orphan disease genes. In consideration of the fact that gene with similar function shared with similar disease, Li et al.[40] proposed two shortest path methods, called SPranker and SPGOranker, to prioritize disease-causing genes in protein interaction networks.

Apart from local distance measures such as the shortest path method, the global network distance is an alternative method to measure the distance between candidate disease genes and known distance genes. By using Random Walk with Restart (RWR) and diffusion kernel methods, Köhler et al.[41] proposed two methods to assess the similarity of two genes in the PPI network for disease gene prioritization. Similar work has been done by Erten et al.[42] They developed the VAVIEM method for disease gene identification. The RWR and Pearson Correlation Coefficient (PCC) are combined to measure similarity of two proteins. Le and Kwon[43] further proposed a neighbour-favouring weight reinforcement to improve the performance of RWR in disease gene prioritization. They showed that when only the interactions between the nearest neighbours and known disease gene are reinforced, the performance is optimal. More recently, Zhang et al.[44] proposed a method, named ESFSC, based on RWR to rank candidate disease genes. The innovation of ESFSC is enlarging seed nodes with known disease genes and their k-nearest neighbour nodes.

3.2 Disease-specific or tissue-specific based methods

It is well known that the PPI data contain various noises[45–47]. Using PPI data as a single resource for disease gene prioritization will lead to the potential bias of results. An effective method to overcome this shortage is integrating different data resources, such as gene expression, Gene Onthology, etc[48].

The reconstruction of disease-specific PPI networks is based on the fact that biological networks are highly modular. Comparing to generic PPI networks, disease-specific PPI networks can further reveal the underlying mechanisms and features of diseases and contribute to identifying potential disease genes in a more accurate way. Wang et al.[49] reconstructed a Disease-Aging Network (DAN) by integrating disease-gene association, aging-gene association, and human PPI data. The topological properties of DAN are analysed thoroughly and diseases can be classified into two catalogues: disease genes in catalog I are significantly close to aging genes, but in catalog II are not. Lee et al.[50] constructed a PPI network of abnormally expressed genes of three kinds of diseases (schizophrenia, bipolar disorder, and major depression) by incorporating microarray and PPI data. The abnormally expressed genes are selected by analysing microarray data with t-test. Some significant disease genes are identified by analysing topological features of disease-specific PPI network. Zhao et al.[51] constructed an axial spondyloarthropathy specific PPI network by combining OMIM database, proteomics and microarray experiment data. The topological and pathway features of the PPI network are analysed and some new insights of pathogenesis are found. He et al.[52] built context-specific PPI networks by selecting genes existing in the PPI networks and expressed at the same time. The information flow method is used to identify dysfunctional modules and the candidate disease genes are prioritized via integrating semantic similarity and module analysis. Based on the hypothesis that integrating PPI networks with mRNA expression profiles may contribute to delineate dysregulated molecular sub-networks which contain disease-causing genes, Lee et al.[53] proposed an approach to identify acute myeloid leukemia disease genes. This method overlays expression values of each gene on its corresponding protein in PPI network and identifies significant sub-networks by calculating Perturbation Score (PS) of each sub-network.

The study of tissue-specific protein interactions is still at the initial stage. Proteins whose genes are translated more efficiently in a specific tissue tend to

have more connections within this tissue as compared to other proteins in the same tissue[54, 55]. Magger et al.[56] pointed out that many present methods use static human PPI data to prioritize disease gene, but the disease impact specific tissue corresponding PPI network may be dramatically different. They construct a tissue-specific PPI network by integrating tissue-specific gene expression data and employ existing prioritization methods to compare experimental result between the tissue-specific PPI network and the generic PPI network. The results demonstrate that tissue-specific PPI network can effectively improve the performance of prioritization. In addition, Li et al.[57] constructed weighted tissue-specific network by combining gene expression and DNA methylation data and page-rank based method used to rank candidate disease genes. Their results demonstrate that the performance of weighted tissue-specific network is better than original protein interaction network.

3.3 Multiple-network methods

It has been demonstrated that phenotypically similar diseases often share a set of functional similar genes[58]. With this observation, disease similarity networks can be constructed by using phenotype similarity data. In addition to phenotype similarity data, other biological data such as Gene Ontology (GO) and pathway information also are used to construct specific network[59].

3.3.1 Integration of phenotypic information

Based on assumption that a group of functionally related genes may be bound up with phenotypically similar diseases[58], some studies have been conducted by combining phenotypically similar profiles and PPI data for disease gene prioritization.

Wu et al.[60] proposed the CIPHER method based on linear regression to predict and prioritize disease genes. The concordance score upon phenotype similarity is employed to evaluate the consistency between the position genes in the PPI network and the variation of phenotype similarity for the phenotype network and used to rank all candidate disease genes. Zhang et al.[61] developed a Bayesian regression method based on the linear relation between disease phenotype and gene proximity to disease gene prioritization. For a query disease and a candidate gene, Bayes factors, which indicate the strength of association between disease similarity and gene proximity, are computed to rank candidate genes. The results showed that Bayess approaches can more effectively enhance performance than CIPHER in identifying disease genes. In addition, Yao et al.[62] developed a HItting-Time-based approach (CIPHER-HIT, as a continuation of CIRHER) to prioritize candidate disease genes. Unlike CIPHER, the CIPHER-HIT captures the global relationships instead of local ones. The Mean-Hitting-Time of the random walk on the heterogeneous network is used to measure the closeness of two nodes for disease gene ranking and the condition of Mean-Hitting-Time is employed to find modularity characteristics for disease subtype inference. Yang et al.[63] also proposed a method called RWPCN (Random Walk on Protein Complex Network) to predict and prioritize disease genes. Different types of computational methods have been proposed for the identification of protein complexes, such as IPCA[64] which has been used successfully in the studies of rheumatoid arthritis[65]. Li and his fellows[66] constructed a multi-graph by merging different data resources (include PPI and three ontologies) and phenotype network by using phenotype data. The extensional random walk with restart, Random Walk with Restart on Multigraphs (RWRM) is employed to prioritize disease gene on multi-graph and phenotype network. The result demonstrates that the performance of RWRM exceeds the state-of-the-art approaches in disease gene identification. In addition, Xie et al.[67] introduced a Bi-Random Walk (BiRW) algorithm to unveil the associations between the complete collection of disease phenotypes (phenome) and genes.

Vanunu et al.[68] presented a network propagation method called PRINCE to identify disease genes and protein complexes by using disease-disease similarity and PPI data. The input is a query disease, and then the resembled disease is selected by computing phenotypic similarity of two diseases. The score function based on network propagation is designed to rank disease genes. In each propagation process, the genes receive the flow from its neighbours which receive the flow from the previous iteration. The final score of each gene is the amount of flows. The performance of PRINCE surpasses random walk and CIHPER. However, the PRINCE computes disease-gene association scores based on the association between the disease similar to the query disease and their involved genes independently. Guo et al.[69] proposed a framework to prioritize candidate disease genes by exploiting modular nature of the genetic diseases and the

consistency between the disease phenotypic overlap and the genetic overlap. The association score between a query disease and a candidate gene is defined as the sum of all association scores between the neighbour of candidate genes and the diseases which are similar to the query disease. The result shows it outperforms the PRINCE. More recently, Ganegoda et al.[70] constructed tissue specific gene network and phenotype-phenotype network and detected the similarity between seed genes and candidate genes.

In addition, information flow based methods and network alignment methods have been developed to prioritize disease genes in phenotype and PPI networks. Chen et al.[71] utilized MAXimizing Information Flow (MAXIF) approach to measure the strength of association between a query disease and a candidate gene. Experimental results have showed that the MAXIF is superior to PRINCE. Wu et al.[72] proposed network alignment based framework, called AlignPI, to identify and predict disease genes from gene network and phenotype network. The network alignment toolkit can be obtained from Ideker lab (http://chianti.ucsd.edu/nct/index.php) and is employed to find pairs of sub-networks called gene module (gene sub-network) and disease module (disease sub-network), by aligning gene and phenotype network. The DAVID tool[73] is used to analyse gene function of module and Fisher exact test is used to calculate the P-value of enriched disease category. The results showed AligPI is better than CIPHER in performance and phenotypic overlap is a general indicator of shared pathogenesis.

3.3.2 Integrating other biological information

The GO is a repository of biological knowledge, which contains three independent sub-ontologies: biological process, cellular component, and molecular function. Li and Patra[74] demonstrated that these sub-ontologies are independent. Therefore, three gene functional similar networks can be constructed by calculating functional similarity. The random walk with restart method is used to rank disease genes in four networks (three sub-ontology networks and one PPI network), and then the rank lists are transformed into discounted rating lists through a Discounted Rating System (DRS). The discounted rating scores of genes in different networks are combined to rank these disease genes. The results showed that the discounted rating system method is comparable with N-Dimensional Order Statistics (NDOS) used in Endeavour in terms of performance and faster than NDOC in terms of the working speed. In addition, Chen et al.[75] constructed gene co-expression network, PPI network, and pathway network by using three specific data resources (PPI, gene co-expression, and pathway networks) for gene prioritization. The importance of two genes similarity in different networks is defined in view of Diffusion Kernel approach (DK) and further Data Integration Rank (DIR) can be calculated in term of similarity. The final DIR score is employed to evaluate the association between candidate gene and known disease genes of specific diseases. It defines a meta score method instead of using the top-K approach to report prioritization result. The informativeness of networks for specific disease is also discussed in that paper.

3.4 Machine learning methods

Machine learning is a useful tool to prioritize candidate disease genes by training classifiers with features of known disease genes and non-disease genes.

Supervised machine learning prioritizes candidate genes based on the differences between disease genes and non-disease genes of biological knowledge. Hindumathi et al.[76] investigated the cervix related cancer by means of combining PPI data and cancer gene data. The topological properties (such as vulnerability, closeness, centroid values, shortest-path betweenness centrality, current flow betweenness centrality, and eigenvector centrality) and gene ontology enrichment analysis are employed to classify the cervix related cancer genes and non-disease genes. Besides topological features, Jia et al.[77] integrated other biological features to predict disease genes of autism spectrum disorders and intellectual disabilities. Eleven features (GO biological processes, GO cellular components, GO molecular functions, transcription factor binding sites, metabolites associated with gene-lists, knockout mouse phenotypes, microRNA targets, structural domains, hub proteins, and gene signatures) are chosen to train three classifiers (two network-based and one attribute-based). The Matthew's Correlation Coefficient (MCC), accuracy, sensitivity, specificity, and Area Under the ROC Curve (AUC) are used to evaluate the classifier's performance. Gene expression data can be viewed as an indicator that a gene is abnormal or not. Nitsch et al.[78] combined different gene expression datasets with PPI data and developed a web-based machine learning method to prioritize disease genes. Four strategies

(three of them are based on random walk while one is based on direct neighbourhood) are designed to train classifiers. Their assumption is that a disease gene is surrounded by highly differentially expressed genes in PPI networks. Chen et al.[79, 80] utilized the biological features of protein complex, expression profile, pathway to identify disease gene based on Bayesian analysis and Markov random field method. In addition, Chen et al.[81] presented a multiple regression model with lasso penalty method to discover genes associated with a disease. The sequence, network, expression, and pathway semantic and phenotypic features are selected to classify disease genes and non-disease genes.

Supervised learning methods based on hypothesis that the separation of disease genes and non-disease genes has been used to prioritize candidate disease genes. However, a limited number of discovered genes are as yet-unidentified genes which should be taken into account for prediction process. Nguyen and Ho[82] proposed a semi-supervised learning approach which takes yet-unidentified genes into consideration to detect disease genes. Several features including topology, keyword, enzyme, sequence length, GO term, protein domain, and biological pathway are chosen to train classifiers. In addition, Mordelet and Vert[83] developed ProDiGe, a multi-task machine learning method to prioritize disease genes. ProDiGe exploits the relative similarity of both known and candidate disease genes to jointly score, instead of scoring independently the different candidate genes. Moreover, this method also gathers the information from known disease genes and important role of genes in similar diseases to rank candidate disease genes. The challenge of machine learning method is how to select useful biological features to train classifier[84]. Therefore, integrating multiple data resource is an effective method to improve performance. However, the redundant or irrelevant biological information may be futile even degrade the performance. How to select useful features from various biological resources would be further research focus. As some topological characters have been used both for the identification of essential proteins and disease genes, we can also borrow some features used in the identification of essential proteins[85–87]. In addition, the different classification algorithms may be suit for different data resources. Hence, utilizing multiple learning algorithms can obtain better predictive performance.

4 Computational Tools for Candidate Disease Gene Prioritization

In the past few years, a large number of computational tools have been developed to assist biologists to prioritize disease genes[88–90]. In this section, we describe some current available candidate disease gene prioritization tools which can be used with PPI data. Table 3 presents brief comparison of these computation tools.

4.1 Gene prioritization web-based tools

The inputs of these web tools can be classified into two categories: candidate disease genes and prior knowledge of disease-related. Some tools, such as ToppNet[91], Gentrepid[93], Endeavour[95], and GUILDify[97], require users to upload a specific candidate gene set for disease gene prioritization. Other tools, including GeneDistiller[96], MetaRanker[94] and PINTA[92], provide whole genomic prioritization sever without any candidates. The user can choose specific

Table 3 A brief comparison of computational tools.

Tools	Input		Output	Reference
	Training data	Candidate genes		
ToppNet	Known genes	List of genes	Ranking and test statistics	[91]
PINTA	Expression dataset	Region, list of genes and genome	Ranking and test statistics	[92]
Gentrepid	Keywords	Region and list of genes	Ranking and selection of candidates	[93]
MetaRanker	Known genes, keywords and expression dataset	Genome	Ranking and test statistics	[94]
ENDEAVOUR	Known genes and keywords	Region, list of genes and genome	Ranking and test statistics	[95]
GeneDistiller	Known genes	Region	Ranking	[96]
GUILDify	Known genes	List of genes	Ranking	[97]
ProphNet	Known genes	List of genes	Ranking	[98]
ProDiGe	Known genes	List of genes	Ranking	[83]

web tools on the basis of their data. In training data, three kinds of data, known disease genes, keywords (disease name), and gene expression, are viewed as priori knowledge. The ToppNet [91] only need provide a set of known genes while others need specify keywords. The outputs of all web tools contain the rank of candidate disease gene. Some tools like ToppNet[91], PINTA[92], and ENDEAVOUR[95] are giving score. The outputs of genomic-based tools contain candidate genes. In addition, some-tools are able to integrate other data resources except PPI data. The PINTA[92] provides comparative analysis with gene expression data and Endeavour[95] can integrate GO, literature, sequence, gene expression, and motif into result.

4.2 Standalone programs

The standalone tools are easy to be overlooked as most of user lack of professional skills. However, it shows advantage in large-scale disease gene prioritization. ProphNet[98] (free MATLAB code can be downloaded from http://genome2.ugr.es/prophnet) is a network-based gene prioritization tool which allows users integrate different biological entities. ProDiGe[83] (free MATLAB code can be downloaded from http://cbio.ensmp.fr/∼jFvert/svn/prodige/html/) is a gene prioritization tool mentioned above.

5 Conclusions and Future Research

The goal of gene prioritization is to identify the most likely disease genes among a large number of candidates to a particular disease. It can assist geneticists and molecular biologists to elucidate the genetic basis of human diseases. Further, it also contributes to the diagnosis, prevention, and treatment of human diseases[99, 100]. With the development of high-throughput technologies, PPI data has increased sharply in the past decade. These giant data is viewed as an important resource for protein complex identification[101, 102], essential protein discovery[103, 104], and disease gene research[34]. Many computational approaches have been proposed to prioritize candidate disease genes based on PPI data. In this paper, we have reviewed the recently advanced computational approaches for candidate disease gene prioritization. Although prioritization methods have achieved a great success in the past few years[105], there are some further researches that are needed to improve these methods.

(1) It cannot be ignored that current available PPI data contain large amounts of false positive and false negative noises. It has been shown that integrating other types of biological data into PPI data can improve performance of disease gene prioritization[106–108]. However, it is still challenging to combine multiple data resources with PPI data appropriately for disease gene prioritization. For example, some approaches score different data resources separately and obtain the final rank of genes by summing these scores. Therefore, methods for effectively integrating various data resources into a PPI network are needed to develop to improve network quality for gene prioritization.

(2) It has been proved that disease progression is dynamic, involving differentially expressed genes and proteins during different periods[109]. Comparing to static PPI network, dynamic PPI network can illustrate how the onset and progression of disease are reflected in the form of differentially expressed genes[110]. It could be a new direction for candidate gene prioritization based on dynamic PPI network in the future.

(3) Different methods make use of different biological data to prioritize disease genes. The criteria of data selection and data integration may be different. These differences give rise to biases among different approaches. Therefore, it is still challenging to design suitable strategy for assessing the performance of different gene prioritization methods.

(4) Some computational tools based on PPI network have been developed to prioritize candidate disease genes. Although these tools have achieved great successes, some improvements are still necessary. For example, some tools only provide the ranking of candidate disease genes in the final report. It would be better to provide more supplemental information such as P values of ranking and topological features of candidate disease genes to enhance confidence of results. The web tools are easy to get result with few simple step. Therefore, if the user has few disease data and without excellent computer skills, it is suggested to use web tools. While the researcher would like to prioritize large-scale disease gene, it should use standalone tools instead. In addition, as different tools have different strengths and weaknesses, it is advisable for user to obtain the result with multiple tools instead of a single one.

References

[1] A. Tenesa and C. S. Haley, The heritability of human disease: Estimation, uses and abuses, *Nature Reviews Genetics*, vol. 14, no. 2, pp. 139–149, 2013.

[2] A. Masoudi-Nejad, A. Meshkin, B. Haji-Eghrari, and G. Bidkhori, Candidate gene prioritization, *Molecular Genetics and Genomics*, vol. 287, no. 9, pp. 679–698, 2012.

[3] F. O. Walker, Huntington's disease, *The Lancet*, vol. 369, no. 9557, pp. 218–228, 2007.

[4] D. Altshuler, M. J. Daly, and E. S. Lander, Genetic mapping in human disease, *Science*, vol. 322, no. 5903, pp. 881–888, 2008.

[5] J. Wang, G. Chen, M. Li, and Y. Pan, Integration of breast cancer gene signatures based on graph centrality, *BMC Systems Biology*, vol. 5, no. Suppl 3, p. S10, 2011.

[6] L. Jiang, S. M. Edwards, B. Thomsen, C. T. Workman, B. Guldbrandtsen, and P. Sørensen, A random set scoring model for prioritization of disease candidate genes using protein complexes and data-mining of generif, omim and pubmed records, *BMC Bioinformatics*, vol. 15, no. 1, p. 315, 2014.

[7] G. Valentini, A. Paccanaro, H. Caniza, A. E. Romero, and M. Re, An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods, *Artificial Intelligence in Medicine*, vol. 61, no. 2, pp. 63–78, 2014.

[8] J. Wang, M. Li, J. Chen, and Y. Pan, A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks, *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 8, no. 3, pp. 607–620, 2011.

[9] W. Peng, J. Wang, J. Cai, L. Chen, M. Li, and F.-X. Wu, Improving protein function prediction using domain and protein complexes in ppi networks, *BMC Systems Biology*, vol. 8, no. 1, p. 35, 2014.

[10] W. Xiong, H. Liu, J. Guan, and S. Zhou, Protein function prediction by collective classification with explicit and implicit edges in protein-protein interaction networks, *BMC Bioinformatics*, vol. 14, no. Suppl 12, p. S4, 2013.

[11] B. Zhao, J. Wang, M. Li, F.-X. Wu, and Y. Pan, Detecting protein complexes based on uncertain graph model, *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 11, no. 3, pp. 486–497, 2014.

[12] Z. H. Yang, Y. Y. Feng, H. F. Lin, and J. Wang, Integrating ppi datasets with the ppi data from biomedical literature for protein complex detection, *BMC Medical Genomics*, vol. 7, no. Suppl 2, p. S3, 2014.

[13] M. Li, R. Zheng, H. Zhang, J. Wang, and Y. Pan, Effective identification of essential proteins based on priori knowledge, network topology and gene expressions, *Methods*, vol. 67, no. 3, pp. 325–333, 2014.

[14] L. Yang, J. Wang, H. Wang, Y. Lv, Y. Zuo, X. Li, and W. Jiang, Analysis and identification of essential genes in humans using topological properties and biological information, *Gene*, vol. 551, no. 2, pp. 138–151, 2014.

[15] J. Wang, Y. Huang, F.-X. Wu, and Y. Pan, Symmetry compression method for discovering network motifs, *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 9, no. 6, pp. 1776–1789, 2012.

[16] R. M. Piro and F. Di Cunto, Computational approaches to disease-gene prediction: Rationale, classification and successes, *FEBS Journal*, vol. 279, no. 5, pp. 678–696, 2012.

[17] C. Zhu, C. Wu, B. J. Aronow, and A. G. Jegga, Computational approaches for human disease gene prediction and ranking, in *Systems Analysis of Human Multigene Disorders*. Springer, 2014, pp. 69–84.

[18] J. Amberger, C. Bocchini, and A. Hamosh, A new face and new challenges for online mendelian inheritance in man (omim®), *Human Mutation*, vol. 32, no. 5, pp. 564–567, 2011.

[19] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, The genetic association database, *Nature Genetics*, vol. 36, no. 5, pp. 431–432, 2004.

[20] Y. Zhang, S. De, J. R. Garner, K. Smith, S. A. Wang, and K. G. Becker, Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information, *BMC Medical Genomics*, vol. 3, no. 1, p. 1, 2010.

[21] A. Bauer-Mehren, M. Bundschus, M. Rautschka, M. A. Mayer, F. Sanz, and L. I. Furlong, Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases, *PloS One*, vol. 6, no. 6, p. e20284, 2011.

[22] R. M. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. D. Robinson, L. O'Connor, M. Li, et al., Large-scale mapping of human protein–protein interactions by mass spectrometry, *Molecular Systems Biology*, vol. 3, no. 1, p. 89, 2007.

[23] M. Dreze, D. Monachello, C. Lurin, M. E. Cusick, D. E. Hill, M. Vidal, and P. Braun, High-quality binary interactome mapping, *Methods in Enzymology*, vol. 470, pp. 281–315, 2010.

[24] X. Ding, W. Wang, X. Peng, and J. Wang, Mining protein complexes from ppi networks using the minimum vertex cut, *Tsinghua Science and Technology*, vol. 17, no. 6, pp. 674–681, 2012.

[25] M. E. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A.-R. Carvunis, N. Simonis, J.-F. Rual, H. Borick, P. Braun, M. Dreze, et al., Literature-curated protein interaction datasets, *Nature Methods*, vol. 6, no. 1, pp. 39–46, 2009.

[26] G. D. Bader, D. Betel, and C. W. Hogue, Bind: The biomolecular interaction network database, *Nucleic Acids Research*, vol. 31, no. 1, pp. 248–250, 2003.

[27] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, The database of interacting proteins: 2004 update, *Nucleic Acids Research*, vol. 32, no. suppl 1, pp. D449–D451, 2004.

[28] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, et al., Mint, the molecular interaction database: 2012 update, *Nucleic Acids Research*, vol. 40, no. D1, pp. D857–D861, 2012.

[29] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, et al., The intact molecular interaction database in 2012, *Nucleic Acids Research*, p. gkr1088, 2011.

[30] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, Biogrid: A general repository for interaction datasets, *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[31] T. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, et al., Human protein reference database—2009 update, *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D767–D772, 2009.

[32] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, et al., String v9. 1: Protein-protein interaction networks, with increased coverage and integration, *Nucleic Acids Research*, vol. 41, no. D1, pp. D808–D815, 2013.

[33] M. H. Schaefer, J.-F. Fontaine, A. Vinayagam, P. Porras, E. E. Wanker, and M. A. Andrade-Navarro, Hippie: Integrating protein interaction networks with experiment based quality scores, *PloS One*, vol. 7, no. 2, p. e31826, 2012.

[34] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, Network medicine: A network-based approach to human disease, *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[35] K. Wysocki and L. Ritter, Diseasome an approach to understanding gene–disease interactions, *Annual Review of Nursing Research*, vol. 29, no. 1, pp. 55–72, 2011.

[36] H. Tang, F. Zhong, and H. Xie, A quick guide to biomolecular network studies: Construction, analysis, applications, and resources, *Biochemical and Biophysical Research Communications*, vol. 424, no. 1, pp. 7–11, 2012.

[37] 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.

[38] C.-L. Hsu, Y.-H. Huang, C.-T. Hsu, and U.-C. Yang, Prioritizing disease candidate genes by a gene interconnectedness-based approach, *BMC Genomics*, vol. 12, no. Suppl 3, p. S25, 2011.

[39] C. Zhu, A. Kushwaha, K. Berman, and A. G. Jegga, A vertex similarity-based framework to discover and rank orphan disease-related genes, *BMC Systems Biology*, vol. 6, no. Suppl 3, p. S8, 2012.

[40] M. Li, Q. Li, G. U. Ganegoda, J. Wang, F. Wu, and Y. Pan, Prioritization of orphan disease-causing genes using topological feature and go similarity between proteins in interaction networks, *Science China Life Sciences*, vol. 57, no. 11, pp. 1064–1071, 2014.

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

[42] S. Erten, G. Bebek, and M. Koyutürk, Vavien: An algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks, *Journal of Computational Biology*, vol. 18, no. 11, pp. 1561–1574, 2011.

[43] D.-H. Le and Y.-K. Kwon, Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization, *Computational Biology and Chemistry*, vol. 44, pp. 1–8, 2013.

[44] S.-W. Zhang, D.-D. Shao, S.-Y. Zhang, and Y.-B. Wang, Prioritization of candidate disease genes by enlarging the seed set and fusing information of the network topology and gene expression, *Molecular BioSystems*, vol. 10, no. 6, pp. 1400–1408, 2014.

[45] J. Wang, X. Peng, W. Peng, and F.-X. Wu, Dynamic protein interaction network construction and applications, *Proteomics*, vol. 14, nos. 4&5, pp. 338–352, 2014.

[46] X. Tang, J. Wang, B. Liu, M. Li, G. Chen, and Y. Pan, A comparison of the functional modules identified from time course and static ppi network data, *BMC Bioinformatics*, vol. 12, no. 1, p. 339, 2011.

[47] S. J. Wodak, J. Vlasblom, A. L. Turinsky, and S. Pu, Protein–protein interaction networks: The puzzling riches, *Current Opinion in Structural Biology*, vol. 23, no. 6, pp. 941–953, 2013.

[48] M. Li, X. Wu, J. Wang, and Y. Pan, Towards the identification of protein complexes and functional modules by integrating ppi network and gene expression data, *BMC Bioinformatics*, vol. 13, no. 1, p. 109, 2012.

[49] J. Wang, S. Zhang, Y. Wang, L. Chen, and X.-S. Zhang, Disease-aging network reveals significant roles of aging genes in connecting genetic diseases, *PLoS Comput. Biol.*, vol. 5, no. 9, p. e1000521, 2009.

[50] S.-A. Lee, T. T. Tsao, K.-C. Yang, H. Lin, Y.-L. Kuo, C.-H. Hsu, W.-K. Lee, K.-C. Huang, and C.-Y. Kao, Construction and analysis of the protein-protein interaction networks for schizophrenia, bipolar disorder, and major depression, *BMC Bioinformatics*, vol. 12, no. Suppl 13, p. S20, 2011.

[51] J. Zhao, J. Chen, T.-H. Yang, and P. Holme, Insights into the pathogenesis of axial spondyloarthropathy from network and pathway analysis, *BMC Systems Biology*, vol. 6, no. Suppl 1, p. S4, 2012.

[52] D. He, Z.-P. Liu, and L. Chen, Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach, *BMC Genomics*, vol. 12, no. 1, p. 592, 2011.

[53] E. Lee, H. Jung, P. Radivojac, J.-W. Kim, and D. Lee, Analysis of aml genes in dysregulated molecular networks, *BMC Bioinformatics*, vol. 10, no. Suppl 9, p. S2, 2009.

[54] X. Zhang, R. Zhang, Y. Jiang, P. Sun, G. Tang, X. Wang, H. Lv, and X. Li, The expanded human disease network combining protein-protein interaction information, *European Journal of Human Genetics*, vol. 19, no. 7, pp. 783–788, 2011.

[55] S. Hwang, S.-W. Son, S. C. Kim, Y. J. Kim, H. Jeong, and D. Lee, A protein interaction network associated with asthma, *Journal of Theoretical Biology*, vol. 252, no. 4, pp. 722–731, 2008.

[56] O. Magger, Y. Y. Waldman, E. Ruppin, and R. Sharan, Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks, *PLoS Comput. Biol.*, vol. 8, no. 9, p. e1002690, 2012.

[57] M. Li, J. Zhang, Q. Liu, J. Wang, and F.-X. Wu, Prediction of disease-related genes based on weighted tissue-specific networks by using dna methylation, *BMC Medical Genomics*, vol. 7, no. Suppl 2, p. S4, 2014.

[58] X. Wang, N. Gulbahce, and H. Yu, Network-based methods for human disease gene prediction, *Briefings in Functional Genomics*, vol. 10, no. 5, pp. 280–293, 2011.

[59] J. Wang, X. Peng, M. Li, and Y. Pan, Construction and application of dynamic protein interaction network based on time course gene expression data, *Proteomics*, vol. 13, no. 2, pp. 301–312, 2013.

[60] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, Network-based global inference of human disease genes, *Molecular Systems Biology*, vol. 4, no. 1, p. 189, 2008.

[61] W. Zhang, F. Sun, and R. Jiang, Integrating multiple protein-protein interaction networks to prioritize disease genes: A bayesian regression approach, *BMC Bioinformatics*, vol. 12, no. Suppl 1, p. S11, 2011.

[62] X. Yao, H. Hao, Y. Li, and S. Li, Modularity-based credible prediction of disease genes and detection of disease subtypes on the phenotype-gene heterogeneous network, *BMC Systems Biology*, vol. 5, no. 1, p. 79, 2011.

[63] P. Yang, X. Li, M. Wu, C.-K. Kwoh, and S.-K. Ng, Inferring gene-phenotype associations via global protein complex network propagation, *PloS One*, vol. 6, no. 7, p. e21502, 2011.

[64] M. Li, J. Chen, J. Wang, B. Hu, and G. Chen, Modifying the dpclus algorithm for identifying protein complexes based on new topological structures, *BMC Bioinformatics*, vol. 9, no. 1, p. 398, 2008.

[65] C. Lu, C. Xiao, G. Chen, M. Jiang, Q. Zha, X. Yan, W. Kong, and A. Lu, Cold and heat pattern of rheumatoid arthritis in traditional chinese medicine: Distinct molecular signatures indentified by microarray expression profiles in cd4-positive t cell, *Rheumatology International*, vol. 32, no. 1, pp. 61–68, 2012.

[66] Y. Li and J. Li, Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data, *BMC Genomics*, vol. 13, no. Suppl 7, p. S27, 2012.

[67] M. Xie, T. Hwang, and R. Kuang, Prioritizing disease genes by bi-random walk, in *Advances in Knowledge Discovery and Data Mining*. Springer, 2012, pp. 292–303.

[68] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, Associating genes and protein complexes with disease via network propagation, PhD dissertation, Tel Aviv University, Israel, 2009.

[69] X. Guo, L. Gao, C. Wei, X. Yang, Y. Zhao, and A. Dong, A computational method based on the integration of heterogeneous networks for predicting disease-gene associations, *PloS One*, vol. 6, no. 9, p. e34171, 2011.

[70] G. U. Ganegoda, J. Wang, F.-X. Wu, and M. Li, Prediction of disease genes using tissue-specified gene-gene network, *BMC Systems Biology*, vol. 8, no. Suppl 3, p. S3, 2014.

[71] Y. Chen, T. Jiang, and R. Jiang, Uncover disease genes by maximizing information flow in the phenome–interactome network, *Bioinformatics*, vol. 27, no. 13, pp. i167–i176, 2011.

[72] X. Wu, Q. Liu, and R. Jiang, Align human interactome with phenome to identify causative genes and networks underlying disease families, *Bioinformatics*, vol. 25, no. 1, pp. 98–104, 2009.

[73] D. W. Huang, B. T. Sherman, and R. A. Lempicki, Systematic and integrative analysis of large gene lists using david bioinformatics resources, *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2008.

[74] Y. Li and J. C. Patra, Integration of multiple data sources to prioritize candidate genes using discounted rating system, *BMC Bioinformatics*, vol. 11, no. Suppl 1, p. S20, 2010.

[75] Y. Chen, W. Wang, Y. Zhou, R. Shields, S. K. Chanda, R. C. Elston, and J. Li, In silico gene prioritization by integrating multiple data sources, *PLoS One*, vol. 6, no. 6, p. e21137, 2011.

[76] V. Hindumathi, T. Kranthi, S. Rao, and P. Manimaran, The prediction of candidate genes for cervix related cancer through gene ontology and graph theoretical approach, *Molecular BioSystems*, vol. 10, no. 6, pp. 1450–1460, 2014.

[77] P. Jia, C.-F. Kao, P.-H. Kuo, and Z. Zhao, A comprehensive network and pathway analysis of candidate genes in major depressive disorder, *BMC Systems Biology*, vol. 5, no. Suppl 3, p. S12, 2011.

[78] D. Nitsch, J. P. Gonçalves, F. Ojeda, B. De Moor, and Y. Moreau, Candidate gene prioritization by network analysis of differential expression using machine learning approaches, *BMC Bioinformatics*, vol. 11, no. 1, p. 460, 2010.

[79] B. Chen, J. Wang, M. Li, and F.-X. Wu, Identifying disease genes by integrating multiple data sources, *BMC Medical Genomics*, vol. 7, no. Suppl 2, p. S2, 2014.

[80] B. Chen, M. Li, J. Wang, and F.-X. Wu, Disease gene identification by using graph kernels and markov random fields, *Science China Life Sciences*, vol. 57, no. 11, pp. 1054–1063, 2014.

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

[82] T.-P. Nguyen and T.-B. Ho, Detecting disease genes based on semi-supervised learning and protein–protein interaction networks, *Artificial Intelligence in Medicine*, vol. 54, no. 1, pp. 63–71, 2012.

[83] F. Mordelet and J.-P. Vert, Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples, *BMC Bioinformatics*, vol. 12, no. 1, p. 389, 2011.

[84] J. Wang, W. Peng, and F.-X. Wu, Computational approaches to predicting essential proteins: A survey, *PROTEOMICS-Clinical Applications*, vol. 7, nos. 1&2, pp. 181–192, 2013.

[85] J. Wang, M. Li, H. Wang, and Y. Pan, Identification of essential proteins based on edge clustering coefficient, *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 9, no. 4, pp. 1070–1080, 2012.

[86] X. Tang, J. Wang, J. Zhong, and Y. Pan, Predicting essential proteins based on weighted degree centrality, *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 11, no. 2, pp. 407–418, 2014.

[87] M. Li, J. Wang, X. Chen, H. Wang, and Y. Pan, A local average connectivity-based method for identifying essential proteins from the network level, *Computational Biology and Chemistry*, vol. 35, no. 3, pp. 143–150, 2011.

[88] D. Börnigen, L.-C. Tranchevent, F. Bonachela-Capdevila, K. Devriendt, B. De Moor, P. De Causmaecker, and Y. Moreau, An unbiased evaluation of gene prioritization tools, *Bioinformatics*, vol. 28, no. 23, pp. 3081–3088, 2012.

[89] L.-C. Tranchevent, F. B. Capdevila, D. Nitsch, B. De Moor, P. De Causmaecker, and Y. Moreau, A guide to web tools to prioritize candidate genes, *Briefings in Bioinformatics*, vol. 12, no. 1, pp. 22–32, 2011.

[90] Y. Tang, M. Li, J. Wang, Y. Pan, and F.-X. Wu, Cytonca: A cytoscape plugin for centrality analysis and evaluation of protein interaction networks, *BioSystems*, vol. 127, pp. 67–72, 2015.

[91] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, Toppgene suite for gene list enrichment analysis and candidate gene prioritization, *Nucleic Acids Research*, vol. 37, no. suppl 2, pp. W305–W311, 2009.

[92] D. Nitsch, L.-C. Tranchevent, J. P. Goncalves, J. K. Vogt, S. C. Madeira, and Y. Moreau, Pinta: A web server for network-based gene prioritization from expression data, *Nucleic Acids Research*, vol. 39, no. suppl 2, pp. W334–W338, 2011.

[93] R. A. George, J. Y. Liu, L. L. Feng, R. J. Bryson-Richardson, D. Fatkin, and M. A. Wouters, Analysis of protein sequence and interaction data for candidate disease gene prediction, *Nucleic Acids Research*, vol. 34, no. 19, pp. e130–e130, 2006.

[94] T. H. Pers, N. T. Hansen, K. Lage, P. Koefoed, P. Dworzynski, M. L. Miller, T. J. Flint, E. Mellerup, H. Dam, O. A. Andreassen, et al., Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes, *Genetic Epidemiology*, vol. 35, no. 5, pp. 318–332, 2011.

[95] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, et al., Gene prioritization through genomic data fusion, *Nature Biotechnology*, vol. 24, no. 5, pp. 537–544, 2006.

[96] D. Seelow, J. M. Schwarz, and M. Schuelke, Genedistiller–distilling candidate genes from linkage intervals, *PLoS One*, vol. 3, no. 12, p. e3874, 2008.

[97] E. Guney, J. Garcia-Garcia, and B. Oliva, Guildify: A web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms, *Bioinformatics*, vol. 30, no. 12, pp. 1789–1790, 2014.

[98] V. Martínez, C. Cano, and A. Blanco, Prophnet: A generic prioritization method through propagation of information, *BMC Bioinformatics*, vol. 15, no. Suppl 1, p. S5, 2014.

[99] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure, Exome sequencing as a tool for mendelian disease gene discovery, *Nature Reviews Genetics*, vol. 12, no. 11, pp. 745–755, 2011.

[100] T. P. O'Connor and R. G. Crystal, Genetic medicines: Treatment strategies for hereditary disorders, *Nature Reviews Genetics*, vol. 7, no. 4, pp. 261–276, 2006.

[101] M. Li, J. Wang, J. Chen, Z. Cai, and G. Chen, Identifying the overlapping complexes in protein interaction networks, *International Journal of Data Mining and Bioinformatics*, vol. 4, no. 1, pp. 91–108, 2010.

[102] N. Zaki, D. Efimov, and J. Berengueres, Protein complex detection using interaction reliability assessment and weighted clustering coefficient, *BMC Bioinformatics*, vol. 14, no. 1, p. 163, 2013.

[103] W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, and Y. Pan, Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks, *BMC Systems Biology*, vol. 6, no. 1, p. 87, 2012.

[104] M. Li, Y. Lu, J. Wang, F.-X. Wu, and Y. Pan, A topology potential-based method for identifying essential proteins from ppi networks, *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 12, no. 2, pp. 372–383, 2015.

[105] Y. Moreau and L.-C. Tranchevent, Computational tools for prioritizing candidate genes: Boosting disease gene discovery, *Nature Reviews Genetics*, vol. 13, no. 8, pp. 523–536, 2012.

[106] J. Wang, M. Li, Y. Deng, and Y. Pan, Recent advances in clustering methods for protein interaction networks, *BMC Genomics*, vol. 11, no. Suppl 3, p. S10, 2010.

[107] K. K.-H. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, et al., Genetic and epigenetic fine mapping of causal autoimmune disease variants, *Nature*, vol. 518, no. 7539, pp. 337–343, 2015.

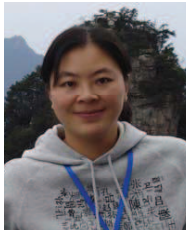
[108] Q. Chen, W. Lan, and J. Wang, Mining featured patterns of mirna interaction based on sequence and structure similarity, *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 10, no. 2, pp. 415–422, 2013.

[109] J. Wang, X. Peng, Q. Xiao, M. Li, and Y. Pan, An effective method for refining predicted protein complexes based on protein activity and the mechanism of protein complex formation, *BMC Systems Biology*, vol. 7, no. 1, p. 28, 2013.

[110] M. Li, W. Chen, J. Wang, F.-X. Wu, and Y. Pan, Identifying dynamic protein complexes based on gene expression profiles and ppi networks, *BioMed Research International*, vol. 2014, p. 375262, 2014.

Wei Lan received his BS and MS degrees from Henan Polytechnical University and Guangxi University, China in 2009 and 2012, respectively. He is currently a PhD candidate in bioinformatics at Central South University. His currently research interest includes bioinformatics and data mining especially in disease gene and noncoding RNA.

Jianxin Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in 2001. He is the vice dean and a professor in School of Information Science and Engineering, Central South University, China. His current research interests include algorithm analysis and optimization, parameterized algorithm, bioinformatics and computer network. He has published more than 150 papers in various international journals and refereed conferences. He is a senior member of the IEEE.



Min Li received the BS degree in communication engineering from Central South University, China, in 2001, MS degree in traffic information and control engineering from Central South University, China, in 2004 and the PhD degree in computer science from Central South University, China, in 2008, respectively. She is an associate professor in School of Information Science and Engineering, Central South University, China. Her current research interests include protein-protein interaction networks, essential proteins discovery, integrative analysis of molecular networks with other biological data, and identifying dynamic network modules.

Wei Peng received the PhD degree in computer science from Central South University, China, in 2013. Currently, she is an associate professor at Kunming University of Science and Technology, China. Her current research interests include molecular systems biology, biological system identification, and data mining.

Fangxiang Wu received the BSc and MSc degrees in applied mathematics from Dalian University of Technology, China, in 1990 and 1993, respectively, the first PhD degree in control theory and its applications from Northwestern Polytechnical University in 1998, and the second PhD degree in biomedical engineering from the University of Saskatchewan, Canada, in 2004. Currently, he is working as an associate professor of bioengineering with the Department of Mechanical Engineering and graduate chair of the Division of Biomedical Engineering at the University of Saskatchewan, Canada. His current research interests include systems biology, genomic and proteomic data analysis, biological system identification and parameter estimation, and applications of control theory to biological system.