Biochimica et Biophysica Acta 1844 (2014) 2214–2221

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

Identifying and prioritizing disease-related genes based on the network topological features

Zhan-Chao Li [a,*], Yan-Hua Lai [b], Li-Li Chen [b], Yun Xie [a], Zong Dai [b], Xiao-Yong Zou [b,**]

[a] School of Chemistry and Chemical Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, People's Republic of China
[b] School of Chemistry and Chemical Engineering, Sun Yat-Sen University, Guangzhou 510275, People's Republic of China

ARTICLE INFO

Article history:
Received 21 May 2014
Received in revised form 22 July 2014
Accepted 14 August 2014
Available online 23 August 2014

Keywords:
Gene ontology
Graph theory
Parkinson's disease
Support vector machine
Topology

ABSTRACT

Identifying and prioritizing disease-related genes are the most important steps for understanding the pathogenesis and discovering the therapeutic targets. The experimental examination of these genes is very expensive and laborious, and usually has a higher false positive rate. Therefore, it is highly desirable to develop computational methods for the identification and prioritization of disease-related genes. In this study, we develop a powerful method to identify and prioritize candidate disease genes. The novel network topological features with local and global information are proposed and adopted to characterize genes. The performance of these novel features is verified based on the 10-fold cross-validation test and leave-one-out cross-validation test. The proposed features are compared with the published features, and fused strategy is investigated by combining the current features with the published features. And, these combination features are also utilized to identify and prioritize Parkinson's disease-related genes. The results indicate that identified genes are highly related to some molecular process and biological function, which provides new clues for researching pathogenesis of Parkinson's disease. The source code of Matlab is freely available on request from the authors.

1. Introduction

Many complex diseases, such as cancer, diabetes, Parkinson's disease (PD) and cardiovascular disorder, have a serious impact on human health. Despite extensive research in uncovering the pathogenesis of disease, there are still many diseases without a known molecular basis. Identifying and prioritizing disease-related genes are the most important for understanding the pathogenesis and discovering the therapeutic targets, because these complex diseases are the results from a complex interaction of multiple genes. Generally, two experimental approaches can be used to identify disease-related genes: linkage analysis and association studies [1]. However, the two methods have some drawbacks. For example, family-based linkage analysis usually associates diseases with chromosomal regions containing tens or even hundreds of genes [1,2]. The experimental examination of these genes is very expensive and laborious [3]. Population-based association studies usually require carefully select candidate genes that are biologically related to the disease of interest. However, the selection is very difficult and often limited by the scope of experts. In addition, association studies may also identify a number of false positives when the investigative diseases have complex inheritance pattern [4,5]. Therefore, it is highly desirable for developing computational methods to narrow down the genomic regions, improve the rates of true positives and further accelerate the process of identifying disease-related genes.

Currently, it is generally considered that the simple genotype–phenotype relationship cannot be applied to the majority of human diseases. The perturbations in the biomolecular networks such as protein–protein interaction (PPI) network are the real cause for some diseases. Recent researches have revealed that genes related to the same or similar diseases are not randomly located in the PPI network. They usually tend to exhibit high connectivity, locate closely to each other and form physical and/or functional modules [6–10]. Many computational approaches based on the PPI network have been dedicated to identify and prioritize candidate diseases genes. These methods can be broadly classified into two categories: unsupervised [3,11–23] and supervised [24–30]. The unsupervised methods prioritize candidate disease genes based on the proximity of them to known disease genes using different scoring strategies. Unfortunately, most of the methods usually have some following limitations: (1) need to set some auxiliary parameters. It is a very difficult task because a priori information about these parameters is usually very rare. (2) Need to use the information of gene expression, gene ontology, protein domains, phenotype or biological pathway, etc. Incompleteness and reasonable integration of the information are the main problems to be solved. (3) Need to improve efficiency. Their running time becomes extremely long when they take into account the global information in the PPI network, which usually requires extensive computation and a lengthy process of algorithm

* Corresponding author. Tel./fax: +86 760 88207939.
** Corresponding author. Tel.: +86 20 84114919; fax: +86 20 84112245.
E-mail addresses: zhanchao8052@gmail.com (Z.-C. Li), ceszxy@mail.sysu.edu.cn (X.-Y. Zou).

http://dx.doi.org/10.1016/j.bbapap.2014.08.009
1570-9639/

convergence. (4) Need to provide a rank list of gene. In order to determine whether a specific gene is disease related or not, a threshold needs to be set according to experience. To conquer these drawbacks, supervised methods have been proposed and used to identify disease-related genes based on the difference of topology between disease genes and non-disease genes in PPI network using sophisticated machine learning techniques. However, the supervised methods have the following disadvantages: (1) usually consider only the local topological information, which is ideally suited for detecting neighbor genes of known disease genes. Obviously, the methods ignore indirect function associations between candidate genes and known disease genes. A gene may be a disease gene when the gene and disease genes are included in the same disease-related biological pathway or functional module, even though the neighbor relationship does not exist between them. (2) Only supply a list of disease-related gene, which cannot quantitatively describe the degree of correlation between disease genes and disease.

In view of the above facts, in order to expedite the identification of disease-related genes, the current study is devoted to developing a powerful method to identify and prioritize candidate disease genes. In the method, we propose the novel network topological features with local and global information based on the graph theory. Support vector machine (SVM) is employed to construct model, and further used to identify and score disease-related genes. We compare the current method with the state-of-the-art technique published in [24] by using the constructed benchmark dataset, and validate the effectiveness through 10-fold cross-validation test and leave-one-out cross-validation. The developed method is utilized to identify and prioritize PD-related genes. Through enrichment analysis, we find that the identified genes are highly related to some molecular processes and biological functions, which provide new clues for researching pathogenesis of PD.

2. Materials and methods

2.1. Construction of human PPI network and collection of disease genes

We download the human PPI data from the database of HIPPIE (Human Integrated Protein–Protein Interaction rEference) [31]. In the database, each PPI is assigned a confidence score representing the reliability of the PPI by integrating multiple experimental PPI datasets. After removing all self-connecting interactions, repeated interactions, interactions with score 0 and interactions contained in the non-largest connected network component, we finally obtain a largest connected network component containing 14,086 unique proteins and 153,749 PPI between the proteins (can be obtained from Supplementary File 1). Please note that the current study only focuses on genes that belong to the largest connected network component, because the topological features cannot be calculated for genes that do not locate in the largest component.

In order to build a high-quality genome-scale dataset of disease-related genes, we collect the data for disease genes from the database of OMIM (Online Mendelian Inheritance in Man) [32], which is a comprehensive and publicly accessible dataset of genotype–phenotype relationship in humans, according to the following order: (1) obtain the information of human disease-gene associations from the annotations of morbidmap in OMIM. (2) Acquire Gene IDs of human disease-related genes by the mapping between Gene ID and Mim ID on the basis of the file of mim2gene.txt (can be obtained from the database of OMIM). (3) Retrieve accession numbers (ACs) of proteins (i.e. disease-related gene products) according to the mapping between UniProt [33] AC and Gene ID with the file of HUMAN_9606_idmapping_selected.tab (can be obtained from the database of UniProt). (4) Map disease-related gene products to our largest connected network component, and the corresponding products are known as "disease-gene set". It is very difficult or even impossible to compile a list of disease-unrelated genes, because there is no database dedicated to the collection of these genes. In accordance with previous studies [24], we exclude disease-related genes contained in the largest connected network component and the remaining genes are called "control-gene set". Therefore, the final "disease-gene set" and "control-gene set" contain 2701 disease-related genes (Supplementary file 2) and 11,385 disease-unrelated genes, respectively.

2.2. Characterization of gene

It is the most important to grasp the difference of topological features between disease-related genes and disease-unrelated genes for identifying and prioritizing genes associated with disease. We use graph theory to capture the differences deeply buried in the human PPI network. In mathematics, graph theory has been widely utilized to investigate graphs consisted of nodes and edges. For the current study, the largest connected network component is modeled as a graph, in which nodes represent proteins, edges denote PPI between two proteins and edge weighted is interaction confidence score. We present six types of topological features with local and global information to characterize gene, and defined as follows:

1. The path weight proportion of disease genes to all genes (*PWPDG*) whose distances to a given gene i is equal to L ($L = 1, 2, \ldots, 10$). The type of topological features can be calculated by Eqs. (1)–(2):

$$PWPDG(i)_1 = \frac{\sum_{d \in N} w_{id}^{dis}}{\sum_{j \in N} w_{ij}} \tag{1}$$

$$PWPDG(i)_2 = \frac{\sum_{k,d \in N} \left(w_{ik} + w_{kd}^{dis} \right)}{\sum_{k.j \in N} \left(w_{ik} + w_{kj} \right)} \tag{2}$$

where, w_{ij} and *wdis id* are the confidence scores between gene i and gene j, as well as between gene i and disease gene d, respectively. N indicates gene set, and the distance between gene contained in the N and the given gene i is 1, 2, …, 10, respectively. Please note that a maximum value of L is 10, because the maximal length between any two genes contained in the largest connected component is 10. In the type of feature, 10 topological features can be obtained. The type of topological features mainly considers the proportion of disease genes in all neighbor genes. The higher values of topological features mean that the gene interacts with more disease genes and has a higher likelihood of disease-related.

2. The average path weight of disease genes (*APWDG*) with distances L to a given gene i. These topological features can be computed by Eqs. (3)–(4):

$$APWDG(i)_1 = \frac{\sum_{d \in N} w_{id}^{dis}}{|N^{dis}|} \tag{3}$$

$$APWDG(i)_2 = \frac{\sum_{k,d \in N} \left(w_{ik} + w_{kd}^{dis} \right)}{|N^{dis}|} \tag{4}$$

where, N^{dis} indicates the set of disease gene and the distance between these genes and the given gene i is L. $|N^{dis}|$ is the number of disease gene contained in the set. According to Eqs. (3) and (4), we can calculate 10