Physica A 398 (2014) 217–228

Contents lists available at ScienceDirect

Physica A

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

Analysis of human genes with protein–protein interaction network for detecting disease genes

Shun-yao Wu [a,b], Feng-jing Shao [a,b,*], Ren-cheng Sun [b], Yi Sui [b], Ying Wang [b], Jin-long Wang [c,d]

[a] College of Automation Engineering, Qingdao University, Qingdao 266071, China
[b] College of Information Engineering, Qingdao University, Qingdao 266071, China
[c] School of Computer Engineering, Qingdao Technological University, Qingdao 266033, China
[d] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

HIGHLIGHTS

- We firstly take essential genes as reference to analyze human genes.
- Nonessential disease genes are topologically more important than other genes.
- Disease genes are not in the periphery but closer to the center than other genes.
- The influence of disease genes on essential genes is weaker than other genes.
- The new topological features are beneficial for disease genes prediction.

ARTICLE INFO

Article history:
Received 31 August 2013
Received in revised form 20 December 2013
Available online 28 December 2013

Keywords:
Human genes
Protein–protein interaction network
Essential genes
Disease genes prediction

ABSTRACT

The topological features of disease genes and non-disease genes were widely utilized in disease genes prediction. However, previous studies neglected to exploit essential genes to distinguish disease genes and non-disease genes. Therefore, this paper firstly takes essential genes as reference to analyze the topological properties of human genes with protein–protein interaction network. Empirical results demonstrate that nonessential disease genes are topologically more important and closer to the center of the network than other genes (unknown genes, which are deemed as non-disease genes in disease genes prediction). Although disease genes are closer to essential genes, we find that the influence of disease genes on essential genes is similar with other genes, or even weaker. Further, we generate new topological features according to our findings and validate the effectiveness of combining the additional features for detecting disease genes. In addition, we find that the *k*-shell index (*ks*) of protein–protein network follows a power law distribution, and the function of the proteins with the largest *ks* may deserve further research.

1. Introduction

Detecting disease genes from human genome is one of the most significant tasks in bioinformatics, which is of great importance to understand disease pathogenesis and improve clinical practice [1]. Recently, in order to speed up the discovery of disease genes, developing timely and relevant algorithms based on machine learning and complex networks has received increasing attention [2–7].

* Corresponding author at: College of Information Engineering, Qingdao University, Qingdao 266071, China. Tel.: +86 18653210169.
E-mail address: sfj@qdu.edu.cn (F.-j. Shao).

0378-4371/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.physa.2013.12.046

One popular and meaningful strategy is gene classification, which could automatically detect whether a gene is disease or not. Several kinds of features, such as gene/protein characteristics [2] or protein topological properties in the protein–protein interaction network [3], were integrated for modeling gene classifiers. Subsequent study demonstrated the performance of topological features was better than other kinds of features [4]. Thus, it is valuable to explore gene classification with topological features.

Besides disease genes and non-disease genes, essential genes are another important group utilized in gene classification. Although essential genes were utilized to select negative samples (non-disease genes),[1] they were neglected to generate features for gene classification. On one hand, previous studies usually identified useful features through comparing topological properties of candidate genes with disease genes, such as the average distance to disease genes [3]. On the other hand, many interesting findings about essential genes were observed, such as they usually encode proteins with high degree and are localized in the center of the protein–protein interaction network [8–11]. However, to the best of our knowledge, few studies researched on comparing topological properties of candidate genes with essential genes, which can make us further understand the characteristic of disease genes and probably improve the performance of disease genes prediction.

Unlike previous studies, we take essential genes as reference to analyze the differences between disease genes and other genes. From empirical results, we draw a conclusion that compared with other genes, the majority of disease genes are topologically more important and not localized in periphery. It means that assuming essential genes as the center of the network, the disease genes are closer to the center than other genes. And we demonstrate although disease genes encode more central proteins, the influence on essential genes is similar with other genes according to their topological properties, or even weaker. Further, on the basis of our findings, we propose three new topological features, *k*-shell, the average distance to the center and the influence on essential genes, and integrate them into disease genes prediction. Experimental results demonstrate the effectiveness and potential of our approach.

The rest of the paper is organized as follows. In Section 2, we introduce the methodology for our work. We propose three questions for analysis of human genes, and describe the procedure of disease genes prediction. In Section 3, we analyze the human genes with the protein–protein interaction network to find the answers of three questions. In Section 4, we define three new topological features based on our findings, and integrate them into disease genes prediction. Finally, we conclude the paper and discuss some future work in Section 5.

2. Methodology

2.1. Topological features for analysis of human genes

In order to explore the differences between disease genes and other genes, we employed degree, *k*-shell index, the average distance, $1N_{\mathcal{X}}$ and $2N_{\mathcal{X}}$ to analyze the topological properties by taking essential genes as reference. Through making a brief introduction of topological features as below, we propose three questions for analysis of human genes: (1) *do the majority of disease genes encode high-degree proteins*? (2) *are disease genes localized in the center or periphery*? (3) *if disease genes are topologically important, can disease genes have a serious influence on essential genes*?

2.1.1. Evaluating topological importance by degree

Degree is the number of links connecting to a node. It is a typical topological property, and is widely used to analyze the importance of proteins in the protein–protein network.

Proteins with high degrees usually have important functions, and are mainly encoded by essential genes or disease genes. Many studies demonstrated essential genes were topologically more important than disease genes [3,8,9,11]. However, there is an argument about the topological importance of disease genes. Goh et al. [9] found that the reason why disease genes was topologically important was that a small fraction of disease genes were also essential genes, whereas the majority of disease genes (nonessential disease genes) were topologically neutral. However, Goh et al. utilized mouse lethal orthologs of human genes as human essential genes, but neglected the fact that 60% of disease genes have not reported a knockout for their mouse orthologs [10].

Besides mouse lethal genes, human housekeeping genes are another main candidate for essential genes [8,12–14]. Most important of all, housekeeping genes are identified from all the human genes by microarray meta-analysis, which ensures the completeness of the data. Therefore, we selected housekeeping genes as essential genes to analyze the topological importance of disease genes, so as to find out whether degree can be utilized for detecting disease genes [3–5].

2.1.2. Analyze network hierarchy structure by k-shell decomposition and average distance

According to the analysis of the correlation between high-degree proteins and disease genes, Goh et al. drew a conclusion that the majority of disease genes were localized in the periphery. However, degree cannot serve as an appropriate indicator

[1] As there are obvious differences between essential genes, disease genes and other genes (the rest of human genes, which are usually deemed as non-disease in gene classification), essential genes should be excluded when selecting negative samples [8]. In this paper, the rest of genes are called other genes in analysis of human genes, while deemed as non-disease genes in disease genes prediction.