Physica A 496 (2018) 53–61

Contents lists available at ScienceDirect

Physica A

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

Discovering disease-associated genes in weighted protein–protein interaction networks

Ying Cui [a,b,d,*], Meng Cai [c,d], H. Eugene Stanley [d]

[a] School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, China
[b] Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an 710071, China
[c] School of Economics and Management, Xidian University, Xi'an 710071, China
[d] Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA

HIGHLIGHTS

- Weight of links is taken into consideration in the construction of a PPI network.
- Disease genes show distinct topological properties from non-disease genes.
- An improved forest-based model was applied as classifier.
- Weighted networks perform better than unweighted networks.

ARTICLE INFO

Article history:
Received 8 October 2017
Received in revised form 20 November 2017
Available online 26 December 2017

Keywords:
Disease gene discovering
Topological properties
Weighted PPI network
Machine learning

ABSTRACT

Although there have been many network-based attempts to discover disease-associated genes, most of them have not taken edge weight – which quantifies their relative strength – into consideration. We use connection weights in a protein–protein interaction (PPI) network to locate disease-related genes. We analyze the topological properties of both weighted and unweighted PPI networks and design an improved random forest classifier to distinguish disease genes from non-disease genes. We use a cross-validation test to confirm that weighted networks are better able to discover disease-associated genes than unweighted networks, which indicates that including link weight in the analysis of network properties provides a better model of complex genotype–phenotype associations.

1. Introduction

Networks provide a ubiquitous and efficient tool for the analysis of biological systems [1–3]. Researchers found that a disease phenotype rarely results from an aberration in a single gene, but is a consequence of various pathological processes that interact in a complex network [4]. The requirement of discovering disease genes from molecular networks leads to the development of "Network medicine" [4], which recapitulates the molecular complexity of human disease and offers network-based computational methods to unravel how the molecular complexity manipulates human disease.

Researches dedicating to systematically capture the properties of disease-associated genes in molecular networks have demonstrated that genes related to the same or similar diseases, tend to cluster and interact with each other in these networks [5–7]. These findings promote the development of network-based approaches for identifying and prioritizing

* Corresponding author at: School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, China.
E-mail addresses: ycui@xidian.edu.cn (Y. Cui), mcai@xidian.edu.cn (M. Cai).

https://doi.org/10.1016/j.physa.2017.12.080
0378-4371/

candidate disease genes by using biological network data, such as protein–protein interaction (PPI) networks [8–10], disease phenotype networks [11–13], regulatory networks [14–16] and co-expression networks [17–19], etc.

Although has contributed a lot to disease diagnose and therapy [4], discovering disease-associated genes by using biological and biomedical networks, is still a challenging task in human genetics. Current network-based approaches for identifying disease-related genes have important limitations [20]. Plenty of network-based methods depend on sophisticated integrated data source [11–19,21,22], which lead to time consumption and increasing computing complexity. Our response is to propose a novel network-based approach to discover disease-related genes by using only PPI network data. A PPI network consists of physical interactions between proteins and is widely used in discovering disease genes [23]. The connections between proteins and human diseases confirm that proteins that physically interact with each other share a common function [5,24]. Thus, an aberration in one protein tends to replicate similar disease phenotypes. Accordingly, PPI network is a powerful data source for discovering disease-related genes.

Furthermore, in most exiting network-based methods, all the connections between nodes are binary with values being either 1 or 0, which means the networks they used to identify disease genes are unweighted. Unweighted networks can only reflect whether there are any interactions between vertices, but fail to display different interaction weights between nodes. However, as has long been appreciated, many molecular networks, such as PPI networks, are intrinsically weighted, their edges are not merely binary entities, but have different weights that record their strengths relative to one another. The weight of edges in molecular networks plays an important role in deciphering the topological properties of PPI networks. In order to take into account the existing heterogeneity in the capacity and intensity of connections, we employ a weighted PPI network to analyze the distinct topological properties between disease and non-disease genes.

In this paper, we propose a hybrid network-based method for the discovery of disease-related genes. We consider the heterogeneity of interactions between genes and construct a weighted PPI network for the purpose to better capture the network topological properties. We then analyzed topological properties of both weighted and unweighted PPI networks. The analysis results show that disease genes have discriminatory network properties which enable their distinction from non-disease genes in both weighted and unweighted PPI network. We constructed four different classification models based on KNN, SVM, Random Forest and CForest, respectively. The topological properties are combined in tandem to use as inputs of the classifier. We use grid-search and 10-fold cross validation to find the optimal parameters for every classifier and CForest is chosen as the best classification model according the prediction performance. The computational simulation results reported that the weighted and unweighted networks achieve 88.56% and 83.57% classification accuracy, respectively. It demonstrated that, by considering the weight of edges, we successfully improved the discovery of disease genes and contributed a deeper understanding into complex genotype–phenotype relationships.

2. Materials and methods

2.1. Data sources

We downloaded high-quality protein–protein interactions from HIPPIE v2.0 (the Human Integrated protein–protein Interaction rEference) [25]. HIPPE database collects human PPIs with experimental annotations from seven major expert-curated databases [26–32]. For each PPI, HIPPIE assigned a stringent confidence score to reflecting its reliability and authenticity. This score is computed by integrating diverse experimental evidence and applying basic network node importance evaluating algorithms. HIPPIE map all source database entries to gene names, Entrez gene ids and UniProt ids or accessions. In this work, we downloaded 71 823 "high confidence" PPIs (confidence score is equal or greater than 0.73) involving 11 813 proteins.

We obtained the list of disease-associated genes and non-disease genes supplied by the Online Mendelian Inheritance in Man (OMIM) [33]. The genemap file of OMIM has 16 161 records with gene symbols, MIM number and related disease phenotypes. We use phenotype mapping key, which appears in parentheses after a disorder, to select gene–disease relations with key (3). This group of gene–disorder associations has well-known molecular basis and a mutation to support. We got 10 980 genes involving 3700 disorder phenotypes, indicating that most diseases are polygenic.

Gene Symbols are used to match the disease and non-disease gene lists from OMIM to protein–protein interactions from HIPPIE. We got 1608 genes that have at least one related disease phenotype and one PPI as disease gene samples, and 3645 genes with interactions in HIPPIE but no disease association record in OMIM as non-disease gene samples. These 1608 disease genes are considered to be positive samples, and 1608 non-disease genes are randomly selected to be negative samples.

2.2. Weighted PPI network construction

We constructed a weighted PPI network by the 71 823 high-quality PPIs we got from HIPPIE. This weighted network is modeled as an undirected graph $G_w = (V, E, w_e)$, where $V = \{v_1, v_2, \ldots, v_M\}$ is the set of nodes, $E = \{e_1, e_2, \ldots, e_N\}$ is the set of edges and w_e is the set of edge weights. Two nodes, referring to proteins in PPI network, are connected by a weighted edge if there is an interaction between them. Based on the confidence score, each edge is assigned a weight

$$w(e_i) = \frac{s_i}{\sqrt{\sum_{i-1}^{N}(s_i - \bar{s})^2/(N-1)}}, \tag{1}$$