Physica A xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Physica A

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

Q1 Topology association analysis in weighted protein interaction network for gene prioritization

Q2 Shunyao Wu, Fengjing Shao *, Qi Zhang, Jun Ji, Shaojie Xu, Rencheng Sun, Gengxin Sun, Xiangjun Du, Yi Sui

Qingdao University, Qingdao 266071, China

HIGHLIGHTS

- Topology associations between disease and essential genes are analyzed in weighted protein interaction network.
- Network propagation with dual flow is proposed for gene prioritization.
- Weak tie effect exists in protein interaction network.

ARTICLE INFO

Article history:
Received 21 February 2016
Received in revised form 2 May 2016
Available online xxxx

Keywords:
Weighted protein interaction network
Disease genes
Essential genes
Topology association
Gene prioritization

ABSTRACT

Although lots of algorithms for disease gene prediction have been proposed, the weights of edges are rarely taken into account. In this paper, the strengths of topology associations between disease and essential genes are analyzed in weighted protein interaction network. Empirical analysis demonstrates that compared to other genes, disease genes are weakly connected with essential genes in protein interaction network. Based on this finding, a novel global distance measurement for gene prioritization with weighted protein interaction network is proposed in this paper. Positive and negative flow is allocated to disease and essential genes, respectively. Additionally network propagation model is extended for weighted network. Experimental results on 110 diseases verify the effectiveness and potential of the proposed measurement. Moreover, weak links play more important role than strong links for gene prioritization, which is meaningful to deeply understand protein interaction network.

1. Introduction

The flourish of disease gene prediction is generally recognized as the breakthrough that discovers pathogenesis and promotes the diagnosis quality. However, it still depletes inevitably time and resources with biomedical experiments and clinical evidence. Due to the fulfill of human genome project, the explosions of omics data, such as metabolic data, proteomic data and protein interactions, offer the brand-new opportunity to design intelligent models and reveal the disease mechanism [1,2].

In the beginning, gene classification was proposed to predict disease genes based on machine learning theory [3–9]. Generally, gene classification trained models via sequence-based features and topological features, and then automatically identified candidate disease genes. The main defect lied in negative samples selected from unknown genes, which would debase the performance of classifiers [5,9].

* Corresponding author.
E-mail address: sfj@qdu.edu.cn (F. Shao).

http://dx.doi.org/10.1016/j.physa.2016.05.043
0378-4371/

Please cite this article in press as: S. Wu, et al., Topology association analysis in weighted protein interaction network for gene prioritization, Physica A (2016), http://dx.doi.org/10.1016/j.physa.2016.05.043

Table 1
Statistics of interactors in the protein interaction network.

	Number of interactors
D	2551
E	1904
$E \bigcap D$	295
D^-	2256
E^-	1609
O	9858

To overcome above challenge, gene prioritization became another potential strategy for disease gene prediction [10–15]. Given one hereditary disease and its known disease genes, the potential disease genes could be searched out according to the topological similarities between unknown genes and known disease genes. Molecule networks, such as protein interaction networks, were utilized to uncover the mechanics of diseases, for molecule interactions usually reflected function linkages among molecules [16]. Perturbations of molecule networks probably led to human diseases [17,18], and neighbors of disease genes were likely to cause the same or similar diseases [19]. Overall, network-based methods for gene prioritization have been well studied by complex network theory [11,20,21], while some problems are still in suspense. For instance, existing methods are confined to detect potential disease genes within neighborhoods of known disease genes [21], or even mistakes non-disease hub proteins as potential disease genes [15].

Majority studies on gene prioritization are concerned with un-weighted protein interaction networks. Besides that, empirical analyses for the topological properties of disease proteins also mainly focus on them. More precisely, these networks are constructed with strong protein interactions, which are specified by a given protein interaction threshold. Obviously the analysis complexity is considerably minimized at expense of weak protein interactions. Currently, weighted molecule networks have stepped into the public view by degrees [22–24]. Weak links among the neural network count little for link prediction [22], whereas they are essential to stabilize the conjunctions between functional modules in molecule networks [25]. Therefore, the strengths of molecular interactions remain to be explored for deeply understanding protein interaction networks.

In this paper, topology associations between disease and essential genes are analyzed in the weighted protein interaction network. Since empirical analysis exhibits disease proteins are weakly connected to essential proteins, network propagation is extended for weighted protein interaction network. This extended model superiorly outperforms the previous methods supported by experimental results. In addition, this paper states that weak links impact disease gene prediction. Compared to the previous methods, network propagation with dual flow could make better use of weak links.

2. Materials and methods

2.1. Human disease gene list and housekeeping gene list

The disease gene list is downloaded from the Online Mendelian Inheritance in Man database (OMIM) [26]. 2931 disease genes verified by the presence of a mutation with tag '3' are selected from 6285 entries.

Housekeeping genes are obtained from the research of Chang et al. [27]. They are universally expressed in normal tissues or cells and vital to maintaining fundamental life activities. Thus, housekeeping genes can be deemed as essential genes [28].

2.2. Gene–disease associations

110 hereditary diseases and corresponding disease genes are obtained from Kohler et al.'s work[1] [11]. They have collected the associations between genetic diseases and disease genes from OMIM, domain knowledge and medicinal literatures. The collection contains 794 gene–disease associations, but only involves 681 unique genes (one gene may cause a few disease).

2.3. Protein interactions

The human protein interactions are downloaded from STRING database (version 9.05) [29], which provides a score to evaluate the reliability between any two interactors. The STRING database collects protein interactions by utilizing various different kinds of active prediction methods including Neighborhood, Gene Fusion, Text Mining, Co-occurrence, Co-expression, Experiments and Databases [29]. This paper extracts direct human binding interactions from STRING action database. All the binding interactions are physical interactions and gained by Experiments. The protein interaction network is constructed with 14 018 interactors and 164,087 interactions inferred from experimental data.

[1] http://www.cell.com/cms/attachment/2024884754/2044549085/mmc1.zip.

Please cite this article in press as: S. Wu, et al., Topology association analysis in weighted protein interaction network for gene prioritization, Physica A (2016), http://dx.doi.org/10.1016/j.physa.2016.05.043