Methods 83 (2015) 51–62

Contents lists available at ScienceDirect

Methods

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

Essential protein identification based on essential protein–protein interaction prediction by Integrated Edge Weights

Yuexu Jiang [a,b], Yan Wang [a,c,*], Wei Pang [d], Liang Chen [a], Huiyan Sun [a], Yanchun Liang [a,e,*], Enrico Blanzieri [c,*]

[a] Key Laboratory of Symbolic Computation and Knowledge Engineering, College of Computer Science and Technology, Jilin University, Changchun 130012, China
[b] Department of Computer Science, University of Missouri, Columbia, MO, United States
[c] Department of Information Engineering and Computer Science, University of Trento, Povo, Italy
[d] School of Natural and Computing Sciences, University of Aberdeen, Aberdeen, UK
[e] Department of Computer Science and Technology, Zhuhai College of Jilin University, Zhuhai 519041, China

ARTICLE INFO

Article history:
Received 28 January 2015
Received in revised form 9 April 2015
Accepted 10 April 2015
Available online 16 April 2015

Keywords:
Essential protein
Essential protein–protein interaction
Integrated Edge Weights

ABSTRACT

Essential proteins play a crucial role in cellular survival and development process. Experimentally, essential proteins are identified by gene knockouts or RNA interference, which are expensive and often fatal to the target organisms. Regarding this, an alternative yet important approach to essential protein identification is through computational prediction. Existing computational methods predict essential proteins based on their relative densities in a protein–protein interaction (PPI) network. Degree, betweenness, and other appropriate criteria are often used to measure the relative density. However, no matter what criterion is used, a protein is actually ordered by the attributes of this protein *per se*. In this research, we presented a novel computational method, Integrated Edge Weights (IEW), to first rank protein–protein interactions by integrating their edge weights, and then identified sub PPI networks consisting of those highly-ranked edges, and finally regarded the nodes in these sub networks as essential proteins. We evaluated IEW on three model organisms: *Saccharomyces cerevisiae* (*S. cerevisiae*), *Escherichia coli* (*E. coli*), and *Caenorhabditis elegans* (*C. elegans*). The experimental results showed that IEW achieved better performance than the state-of-the-art methods in terms of precision–recall and Jackknife measures. We had also demonstrated that IEW is a robust and effective method, which can retrieve biologically significant modules by its highly-ranked protein–protein interactions for *S. cerevisiae*, *E. coli*, and *C. elegans*. We believe that, with sufficient data provided, IEW can be used to any other organisms' essential protein identification. A website about IEW can be accessed from http://digbio.missouri.edu/IEW/index.html.

1. Introduction

Essential proteins are indispensable for the survival of an organism under certain conditions [1]. Reliable identification of essential proteins is of great significance since it can contribute to a better understanding of the key biological processes of an organism at molecular level, which is useful for guiding drug design, disease diagnosis, and medical treatments. Experimentally, many researchers identify proteins' essentiality by knocking out some particular proteins and checking the viability of the affected organisms [1,2]. However, the cost of such biological wet-lab experiments is normally very high, and more importantly, they are ethically impossible on humans. This makes *in silico* analysis a necessary method of choice to carry out the research. Currently, there is still much work to be done by computational biologists for the effective identification of essential proteins.

Nowadays, due to high-throughput techniques, large-scale protein–protein interaction (PPI) data are available for many organisms, especially for some model organisms such as *Saccharomyces cerevisiae* and *Escherichia coli*. Based on these data, several studies have been conducted, and these studies aim to investigate the relationships between experimentally identified essential proteins and PPI networks. Jeong et al. [3] noted that the essentiality of a protein had high correlation with its centrality in a PPI network, and this observation is formulated as the centrality-lethality rule [3,4]. Guided by this rule, many measures have been proposed for essential protein detection, such as degree centrality [3], betweenness

* Corresponding authors at: Key Laboratory of Symbolic Computation and Knowledge Engineering, College of Computer Science and Technology, Jilin University, Changchun 130012, China.
E-mail addresses: wy6868@hotmail.com (Y. Wang), ycliang@jlu.edu.cn (Y. Liang), blanzier@disi.unitn.it (E. Blanzieri).

http://dx.doi.org/10.1016/j.ymeth.2015.04.013
1046-2023/

centrality [4], closeness centrality [5], subgraph centrality [6], eigenvector centrality [7], and network bottleneck [8]. Basically these methods rank proteins based on their centrality measures in a PPI network to identify their essentiality. In addition to these purely node-centrality based algorithms, a few edge-aided methods have also been developed. For instance, Wang et al. [9] employed the concept of edge clustering coefficient (the NC method) to identify essential proteins in a PPI network. Further improvements of the NC method were achieved by taking gene expression information (PeC) [10] into consideration. Although edge information plays an important role in the prediction processes of these edge-aided methods, the fundamental idea behind these methods is still ranking proteins according to their centrality measure in the PPI network.

In 2005, Pereira-Leal et al. [11] pointed out that essential proteins tended to be more frequently connected to other essential proteins rather than to non-essential proteins in *S. cerevisiae* PPI networks. They found that after removing all the non-essential proteins from a PPI network, approximately 97% of the essential proteins were still connected, and this suggested a close interaction relationship among essential proteins. He et al. [12] tried to explain the reason why highly connected nodes tend to be essential and proposed the concept of essential protein–protein interactions. They argued that the essentiality of proteins came from the essentiality of protein–protein interactions rather than the proteins *per se*, changing substantially the perspective of the problem. Regarding this, some researchers have taken this direction by scoring the relatedness of proteins connected by edges in a PPI network [9,10,13]. Some of these measures are based on the topology of a PPI network, such as the number of triangles an edge belongs to [9], while other measures are obtained by integrating other biological information, such as Gene Ontology similarity and gene co-expression degree (the EW method) [10,13].

In this paper, we presented a novel essential-protein prediction strategy. Unlike other state-of-the-art methods which directly rank proteins, our method (IEW) predicted essential proteins based on Integrated Edge Weights. By integrating several widely used PPI topological information and biological data, IEW can overcome the possible failure of one or more attributes. We took a comprehensive evaluation on the *S. cerevisiae*, *E. coli*, and *Caenorhabditis elegans* datasets, and proved that IEW was a more accurate and robust method than its competitors. Furthermore, the predicted high-ranked edges tend to be highly biologically significant in *S. cerevisiae*, *E. coli,* and *C. elegans* PPI networks.

2. Methods

After collecting data from various sources, the whole workflow of our method is divided into four parts, as shown in Fig. 1. First, we assess a particular interaction by using five different measurements. Second, we integrate these five measurements into a final weight so that we can rank all the links of a PPI network and obtain a list of the essential interactions. Third, we predict essential proteins based on the obtained essential interaction list. Finally, we use three evaluation methods to test our prediction strategy.

2.1. Protein–protein relationship evaluation

The IEW model aims to evaluate the relationship between two proteins from various perspectives. To achieve this purpose, we integrated into our final model five measures ranging on topology information, gene expression information, physical interaction, gene annotation, and degree of conservation. Among them, topology information, gene expression information, and gene annotation information have been widely used, while to the best of our knowledge physical interaction information and degree of conservation are used for the specific purpose of this research for the first time.

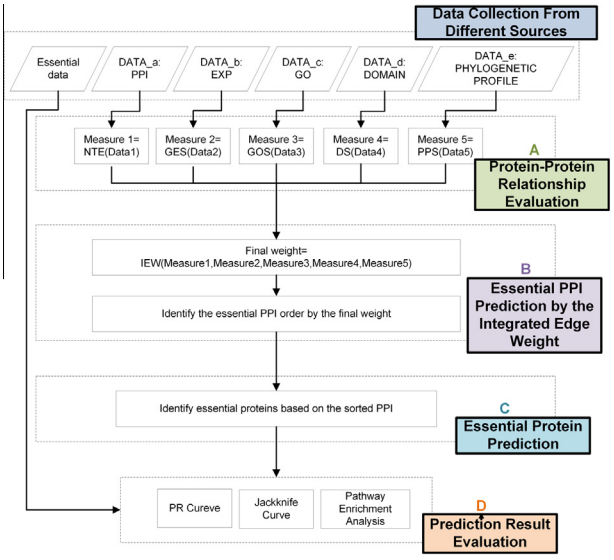

Fig. 1. The overall workflow of the proposed method. After data collection, our method can be divided into four steps: (A) protein–protein relationship evaluation; (B) essential PPI prediction by the Integrated Edge Weight; (C) essential protein prediction; (D) prediction result evaluation.

2.1.1. Measure 1: number of triangles

Topological characteristics of PPI networks encode important information related to the lethality of the absence of a protein. According to the centrality-lethality rule, we considered that essential protein–protein links should tend to be more cliquish. Estrada [14] reported that the proteins selected by any of the spectral measures of centrality tended to form clusters of highly interconnected nodes, and these clusters contained a large number of triangles as measured by the clustering coefficient. Therefore, in this research we used the number of triangles (NTE) as one of the measures to determine the significance and centrality of an edge.

In an undirected graph $G = (N, E)$, where N is the set of the proteins (nodes) in the network, and E is the set of the interactions (edges), the NTE of an edge (u, v) is defined as:

$$\mathrm{NTE}(u, v) = |C_u \cap C_v| + 1, \tag{1}$$

where C_u (or C_v) denotes the set of neighbours of node u (or v) in a PPI network; $|C_u \cap C_v|$ is the number of neighbours shared by nodes u and v, which coincides with the number of triangles that the edge (u, v) belongs to. We add the value "1" at the end of the equation to make the result always bigger than zero. This is to prevent that NTEs between every two proteins are equal to zero, which will cause problems in the normalisation process.

2.1.2. Measure 2: gene expression similarity

Gene expression data are perhaps the most easily obtained and widely used biological data. Studying co-expression patterns [15] can provide useful insights to analysing the underlying cellular processes. Because the co-expressed genes have a high probability to encode interacting proteins [16], we chose gene expression similarity as one of the five measurements. In our method, we used Pearson Correlation Coefficient as the gene expression similarity testing method. The gene expression similarity (GES) of proteins u and v are calculated as follows: