Computational Biology and Chemistry 52 (2014) 34–42

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

A new method for predicting essential proteins based on dynamic network topology and complex information

Jiawei Luo *, Ling Kuang

School of Information Science and Engineering, Hunan University, Changsha 410082, China

ARTICLE INFO

Article history:
Received 24 April 2014
Received in revised form 19 August 2014
Accepted 20 August 2014
Available online 23 August 2014

Keywords:
Centrality measures
Essential proteins
Dynamic network topology
Protein complex

ABSTRACT

Predicting essential proteins is highly significant because organisms can not survive or develop even if only one of these proteins is missing. Improvements in high-throughput technologies have resulted in a large number of available protein–protein interactions. By taking advantage of these interaction data, researchers have proposed many computational methods to identify essential proteins at the network level. Most of these approaches focus on the topology of a static protein interaction network. However, the protein interaction network changes with time and condition. This important inherent dynamics of the protein interaction network is overlooked by previous methods. In this paper, we introduce a new method named CDLC to predict essential proteins by integrating dynamic local average connectivity and in-degree of proteins in complexes. CDLC is applied to the protein interaction network of *Saccharomyces cerevisiae*. The results show that CDLC outperforms five other methods (Degree Centrality (DC), Local Average Connectivity-based method (LAC), Sum of ECC (SoECC), PeC and Co-Expression Weighted by Clustering coefficient (CoEWC)). In particular, CDLC could improve the prediction precision by more than 45% compared with DC methods. CDLC is also compared with the latest algorithm CEPPK, and a higher precision is achieved by CDLC. CDLC is available as Supplementary materials. The default settings of active threshold and alpha-parameter are 0.8 and 0.1, respectively.

1. Introduction

Proteins undoubtedly play important roles in every organism's cell. In particular, essential proteins are indispensable because deleting even one of them is lethal to an organism or renders it infertile (Winzeler et al., 1999; Kamath et al., 2003). Essential proteins can also be regarded as drug targets for new antibiotics because of their indispensability (Clatworthy et al., 2007).

To find essential proteins, researchers have used several experimental approaches, such as single gene knockouts (Giaever et al., 2002), RNA interference (Cullen and Arndt, 2005) and conditional knockouts (Roemer et al., 2003), to discover essential proteins. Although many essential proteins have been identified this way, the process is time consuming and expensive. Therefore, with the fast accumulation of available protein–protein interaction (PPI) data, studies on computational methods for predicting essential proteins are attracting increased attention.

In 2001, Jeong et al. pointed out that proteins highly connected with other proteins in PPI network have higher potential to be essential than randomly selected proteins (Jeong et al., 2001). This phenomenon is called the centrality–lethality rule (Jeong et al., 2001). In recent years, most investigators have proven the positive correlation between the essentiality of proteins and the topological centrality (Hahn and Kern, 2005; Batada et al., 2006; Vallabhajosyula et al., 2009; Estrada, 2006). Based on the latter finding, a variety of centrality measures for identifying essential proteins have been proposed. Among these centrality measures, Degree Centrality (DC) (Jeong et al., 2001), Betweenness Centrality (BC) (Freeman, 1977), Closeness Centrality (CC) (Wuchty and Stadler, 2003), Subgraph Centrality (SC) (Estrada and Rodriguez-Velazquez, 2005), Eigenvector Centrality (EC) (Bonacich, 1987) and Information Centrality (IC) (Stephenson and Zelen, 1989) are the most classic ones. Other useful centrality measures, such as Bottle Neck (BN) (Przulj et al., 2004; Yu et al., 2007), Local Average Connectivity (LAC) (Li et al., 2011) and Sum of ECC (SoECC) (Wang et al., 2012), are also designed to detect essential proteins. The performances of topology-based approaches are known to closely depend on the quality of PPI networks. However, many false positive and false negative interactions exist in PPI networks.

* Corresponding author. Tel.: +86 731 88821971.
E-mail address: luojiawei@hnu.edu.cn (J. Luo).

http://dx.doi.org/10.1016/j.compbiolchem.2014.08.022
1476-9271/

To solve this problem, Li et al. (2010) constructed a relatively reliable weighted PPI network by taking advantage of gene annotations. In this weighted network, the performance of the topology-based methods could be optimized to a great extent. This result confirms that topological centrality is sensitive to the noise in PPI networks. To further improve the prediction precision, researchers attempt to combine biology information with network topology (He and Zhang, 2006; Zotenko et al., 2008; Chua et al., 2008). Consequently, several new methods have been proposed. For example, PeC considers both the edge clustering coefficient and gene expression data (Li et al., 2012), Co-Expression Weighted by Clustering coefficient (CoEWC) (Zhang et al., 2013) integrates clustering coefficient and gene expression data, and another integrated approach proposed by Luo and Ma (2013) combines edge clustering coefficient with complex centrality.

Existing methods for identifying essential proteins all regard the PPI network as a static graph. Given that interactions in static PPI networks are obtained at different time points and under different conditions, a static graph can not completely reflect the real network (Chen and Yuan, 2006; Przytycka et al., 2010). In fact, PPI networks change over time, in different environments, and at different stages of cell cycle, which means that dynamism exists in PPI networks (Przytycka et al., 2010; Han et al., 2004; De Lichtenberg et al., 2005). Based on this fact, Tang et al. (2011) constructed a time-course network with gene expression data and the result indicated that the performance of identifying functional modules based on the time-course PPI network is better than that based on a static PPI network. Wang et al. (2013) built a new dynamic PPI network using a method similar to that adopted by Tang et al. (2011) and concluded that the dynamic PPI network can contribute to the discovery of protein complexes. The main difference between the two approaches is the process of selecting the active threshold of proteins.

The discovery of functional modules and protein complexes is more effective when applying algorithms to a dynamic PPI network. Thus, we suppose that using a dynamic PPI network topology can enable better prediction of essential proteins than using a static PPI network. In other words, the effects of topology-based methods for discovering essential proteins can be improved by implementing these methods in dynamic PPI network. We elaborate the theoretical basis of our hypothesis as follows. At a certain time point, the dynamic PPI network can be represented as a temporal network (Tang et al., 2011; Wang et al., 2013). In these temporal networks, interactions form in the same situation and the included proteins are highly expressed (active). Therefore, dynamic PPI network is of higher quality. Given this high quality, topology centralities in dynamic PPI network are more reliable, which can contribute to a better identification of essential proteins. The proof of our hypothesis is shown in Section 3.6. Then to further enhance the effect of predicting essential proteins, we integrate dynamic network topology and biology information. The integration of dynamic local average connectivity and complex information is chosen to predict essential proteins and this new method is named Combine Dynamic LAC with Complex centrality (CDLC). To evaluate the effect of our CDLC method, we compare CDLC with five centrality methods (DC, LAC, SoECC, PeC and CoEWC). Considering that both LAC and SoECC can outperform all six classic methods (DC, BC, CC, SC, EC, and IC) (Li et al., 2011; Wang et al., 2012), we select only DC for comparison because it is the most accessible among the six methods. The comparison results suggest that CDLC performs better than the other five methods (DC, LAC, SoECC, PeC, and CoEWC). Particularly, CDLC can achieve more than 45% improvement in prediction precision compared with DC. We also compare CDLC with the latest essential protein predicting approach named CEPPK (Li et al., 2014), and results show that CDLC outperforms CEPPK.

2. Methods

The static PPI network is generally considered as an undirected graph G (V, E), where V is the set of nodes and E is the set of edges. The nodes in graph G denote the proteins and the edges represent the interactions between proteins.

2.1. Construction of dynamic PPI network

Tang et al. (2011) and Wang et al. (2013) analyzed the feasibility of integrating the time-course gene expression data and the PPI data to build the dynamic PPI network. Both of their methods focus on the active time points of proteins and introduce an active threshold to judge whether a protein is in its active form at each time point. The difference between these two methods is the key to selecting the active threshold. Wang et al. (2013) also pointed out that it is reasonable to regard the time points with the highest expression value of a protein as the active time points, and the time points with expression values near to the highest one can also be considered as active time points since small differences are allowable and noise exists in data. Basing on that, we construct the dynamic PPI network in the following way, and the effects of different thresholds on several characteristics of the dynamic network will be introduced in Section 3.2.

Gene expression data are always in the form of a matrix. Thus, we use a matrix EXP with the size of $m \times n$ to denote it, where m is the number of probes, which are used to obtain the expression value of each gene, and n is the total number of time points contained in the expression experiment. Let c be the number of cycles included in the expression experiment and n_c be the number of time points in each cycle yields $n = c \times n_c$. Considering the existence of unavoidable noise in the expression array (Wang et al., 2013), for each gene involved in the gene expression profiles, we regard the mean of its c expression values obtained at the same time points in each of the c cycle as the final expression value of this gene. In this way, we can obtain a new expression matrix EXP_new with the size of $m \times n_c$. Based on this new expression matrix, the procedure of constructing dynamic PPI network is described as follows:

Step 1. The expression value in each row of EXP_new is normalized by dividing each of the n_c expression values in row j ($j = 1, 2, \ldots, m$) by the maximal value in the same row. Consequently, every expression value ranges from 0 to 1.

Step 2. A proper threshold to determine whether a gene is active at each time point is obtained. If the expression value of a gene at time point i ($i = 1, 2, \ldots, n_c$) is greater than the active threshold, we say that the gene is active at time point i.

Step 3. n_c temporal networks corresponding to n_c different time points are constructed according to the activity of gene (proteins) and the static PPI network. For each time point, we determine each interaction in the static PPI network. If both genes (proteins) corresponding to this interaction are active at time point i ($i = 1, 2, \ldots, n_c$), we add this interaction and this pair of proteins to the temporal network related to time point i. After considering all the interactions in the static PPI network, the construction of the n_c temporal networks is finished.

The purpose of defining the fixed active threshold in Step 2 is to filter out the genes whose expression level is not sufficiently high at some of the n_c time points. The set of the n_c temporal networks obtained in Step 3 refers to the abovementioned dynamic PPI network.

2.2. Dynamic local average connectivity

According to Hart et al. (2007) and Dezső et al. (2003), in many cases, essentiality of proteins is not a product of an individual