Methods 110 (2016) 54–63

Contents lists available at ScienceDirect

Methods

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

Essential protein discovery based on a combination of modularity and conservatism

Bihai Zhao [b], Jianxin Wang [a,*], Xueyong Li [a,b], Fang-Xiang Wu [c]

[a] School of Information Science and Engineering, Central South University, Changsha 410083, China
[b] Department of Mathematics and Computer Science, Changsha University, Changsha 410003, China
[c] Department of Mechanical Engineering and Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada

ARTICLE INFO

Article history:
Received 15 April 2016
Received in revised form 5 June 2016
Accepted 8 July 2016
Available online 9 July 2016

Keywords:
Essential protein
Multiple network
Modularity
Conservatism

ABSTRACT

Essential proteins are indispensable for the survival of a living organism and play important roles in the emerging field of synthetic biology. Many computational methods have been proposed to identify essential proteins by using the topological features of interactome networks. However, most of these methods ignored intrinsic biological meaning of proteins. Researches show that essentiality is tied not only to the protein or gene itself, but also to the molecular modules to which that protein belongs. The results of this study reveal the modularity of essential proteins. On the other hand, essential proteins are more evolutionarily conserved than nonessential proteins and frequently bind each other. That is to say, conservatism is another important feature of essential proteins. Multiple networks are constructed by integrating protein-protein interaction (PPI) networks, time course gene expression data and protein domain information. Based on these networks, a new essential protein identification method is proposed based on a combination of modularity and conservatism of proteins. Experimental results show that the proposed method outperforms other essential protein identification methods in terms of a number essential protein out of top ranked candidates.

1. Introduction

Essential proteins are those proteins to result in lethality or infertility of a cell if one of them has been deleted [1]. Identification of essential proteins is very important for understanding the basic requirements to sustain a living organism [2]. Furthermore, comprehensive analyses of essential proteins facilitate understanding relationships deeper among mutations in genes and human diseases in order to reveal general principles of human diseases [3,4]. As traditional experimental methods, such as single gene knockouts [5], RNA interference [6] and conditional knockouts [7] are time-consuming and costly, computational approaches for identifying essential proteins offer a new alternative and have become an active research topic in bioinformatics and system biology. Recent developments in experiments have resulted in the publication of many high-quality, large-scale PPI data sets, which provide fundamental and abundant data for computational approaches to the prediction of essential proteins. Computational methods for essential protein identification are typically based on two types of measures: topological centrality measures and heterogeneous source fusion measures.

Jeong et al. [8] have discovered that highly connected proteins in PPI network tend to be essential, which suggests the close correlation between the connectivity and indispensability of a given protein. Since then, many centrality methods have been proposed to predict essential proteins based on their topological properties, such as Eigenvector Centrality (EC) [9], Information Centrality (IC) [10], Closeness Centrality (CC) [11], Betweenness Centrality (BC) [12], Subgraph Centrality (SC) [13], Sum of Edge Clustering Coefficient Centrality (NC) [14], Local Average Connectivity (LAC) [15] and so on. Tang et al. [16] put forward a weighted degree centrality for essential protein prediction. By defining and computing the value of each protein's topology potential, Li et al. [17] obtained a more precise ranking which reflects the importance of proteins from PPI networks.

These centrality methods highly depend on the accuracy of PPI networks. However, a significant proportion of PPI networks obtained from high-throughput biological experiments have been found to contain false positives [18–21]. To overcome these limitations, some researchers have proposed methods to predict essential proteins by integrating their topological properties with their biological properties. Hsing et al. [22] developed a method

[*] Corresponding author at: Computer Building 303, School of Information Science and Engineering, Central South University, Changsha 410083, China.
E-mail address: jxwang@mail.csu.edu.cn (J. Wang).

http://dx.doi.org/10.1016/j.ymeth.2016.07.005
1046-2023/© 2016 Elsevier Inc. All rights reserved.

for predicting highly-connected hub nodes based on the interaction data and Gene Ontology (GO) annotations. Acencio et al. [23] combined network topological properties with genomic features, such as cellular localization and biological process information to identify essential proteins. With the integration of network topology and gene expression, Li et al. [24] proposed a method called PeC to identify essential proteins. Peng et al. [25] proposed an iteration method for predicting essential proteins by integrating the orthology with PPI networks, named by ION. Differently from other methods, ION identifies essential proteins depending on not only the connections between proteins but also their orthologous properties and features of their neighbors. Li et al. [26] proposed a method for evaluating the confidence of each interaction based on the combination of logistic regression-based model and function similarity. Zhang et al. [27] proposed a method, named CoEWC to discovery essential proteins based on the integration of topological properties of PPI networks and the co-expression of interacting proteins. In previous study, we proposed a method to predict essential proteins based on overlapping essential modules, named POEM [28]. In POEM, the original PPI network is partitioned into many overlapping essential modules. The frequencies and weighted degrees of proteins in these modules are employed to score and sort proteins. Li et al. [29] proposed the united complex centrality (UC) to identify essential proteins by integrating the protein complexes with the topological features of PPI networks. Peng et al. [30] proposed an essential protein prediction method, named UDoNC, by combining the domain features of proteins with their topological properties in PPI networks. In UDoNC, the essentiality of proteins is decided by the number and the frequency of their protein domain types, as well as the essentiality of their adjacent edges measured by edge clustering coefficient. Li et al. [31] proposed a priori knowledge-based scheme to discover new essential proteins from PPI networks. Based on this scheme, two essential protein discovery algorithms, CPPK and CEPPK, were developed. CPPK predicts essential proteins based on network topology, while CEPPK predicts essential proteins by integrating network topology and gene expressions. Xiao et al. [32] proposed a framework for identifying essential proteins from active PPI networks constructed with dynamic gene expression. Firstly, they processed the dynamic gene expression profiles by using time-dependent model and time-independent model. Then, they constructed an active PPI network based on co-expressed genes. Finally, they applied six classical centrality measures to the active PPI network. Harmonic centrality (HC) [33] was proposed to predict essential proteins, which is the weighted average of complex centrality and subgraph centrality. It combines PPI network topology and protein complex information. Tang et al. [34] presented CytoNCA, a Cytoscape plugin integrating calculation, evaluation and visualization analysis for multiple centrality measures. Those methods, which integrate network topology and biological information, increase the precision of predicting essential proteins in comparison with those centrality measures only based on network topological features. Even so, computational methods for essential protein identification still cannot get satisfactory results.

Song et al. [35] pointed out that hub proteins in the *Saccharomyces cerevisiae* physical interaction network are more likely to be essential than other proteins. This point of view has been subject to some controversy, while recent work suggests that it arises due to the participation of hub proteins in essential complexes and processes. This conclusion is consistent with the point of Hart et al. [36], which states that essentiality is tied not only to the protein or gene itself, but also to the molecular modules to which that protein belongs. These research results reveal the modularity property of essential proteins.

Another important feature of essential proteins is their conservative property. Researches [37,38] showed that essential proteins evolve much slower than other proteins. That is to say, essential proteins are more evolutionarily conserved than nonessential proteins. The reason is that essential genes are more likely involved in basic cellular processes, thus the negative selection acting on essential genes is more stringent than nonessentials [37]. In order to verify the conservative property of essential proteins, researchers use alignment tool BLAST [39] to search for homologous proteins from different species. If a protein can be found to have homologous proteins in other species, the protein is conservative [40,41]. The term 'phyletic retention' is introduced to describe the homology mapping of a protein in other organisms. Gustafson et al. [42] pointed out the phyletic retention is the most predictive of essentiality. In other words, the conservative property is another important feature of essential proteins.

Inspired by the researches and discoveries mentioned above, we propose a new method for Predicting Essential proteins based on a combination of Modularity and Conservatism, named as PEMC. As currently available PPI datasets contain many false positives, the PEMC integrates network topology with gene expression profile and protein domain information to construct three weighted protein network for essential protein discovery. The PEMC detects overlapping modules from the three networks to generate modular scores for proteins firstly. Then, taking orthologous scores of proteins as the original values, the PEMC method computes conservative scores for proteins by random walks on the PPI networks. Finally, the PEMC method ranks proteins according to a weighted sum of modular scores and conservative scores. Different from current centrality methods, PEMC takes both modularity and conservative property of proteins into account to identify essential proteins. To evaluate the performance of PEMC, we predict essential proteins by using a yeast network from DIP data [43]. Experimental results show that the PEMC method outperforms other previous centrality measures: DC [8], BC [12], CC [11], SC [13], IC [10], NC [14] and four competing methods by integrating network topological features and heterogeneous data sources: PeC [24], CoEWC [27], POEM [28] and ION [25]. We also compare the prediction performance of PEMC with other competing methods based on proteins from Krogan data [44], which is a yeast network compiled from diverse sources of interaction evidence. Results confirm that PEMC gets the best performance on prediction of essential proteins in Krogan data.

2. Methods

The PEMC algorithm consists of three stages, weighted networks construction, essential modules prediction and essential protein discovery. The PEMC method predicts essential modules from multiple individual networks firstly. Modular scores of proteins are generated according to these essential modules. In the stage of essential protein discovery, PEMC gets conservative scores for proteins by random walks on the PPI network combining orthologous information. Final scores of proteins are the linear combinations of modular scores and conservative scores.

2.1. Constructing weighted networks

Considering that PPI data contains a lot of false positives which greatly reduce the essential proteins detection accuracy, we integrate network topology with other heterogeneous data sources, such as gene expression data and protein domain information. On the other hand, different types of interactions or connections have various roles and importance in detecting essential proteins. Therefore, in this study, we separately construct three weighted protein networks (one network based on one data source), such as the co-expression network (gene expression data), the