Journal of Theoretical Biology 412 (2017) 107–112

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Combining sequence and Gene Ontology for protein module detection in the Weighted Network

Yang Yu[a,c,*], Jie Liu[a], Nuan Feng[b,**], Bo Song[a], Zeyu Zheng[c]

[a] Software College, Shenyang Normal University, Shenyang 110034, PR China
[b] College of Information Technology, Shenyang Institute of Technology, Fushun 113122, PR China
[c] Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, PR China

ARTICLE INFO

Keywords:
Protein complex
Protein interaction
Gene Ontology
The Weighted Network

ABSTRACT

Studies of protein modules in a Protein-Protein Interaction (PPI) network contribute greatly to the understanding of biological mechanisms. With the development of computing science, computational approaches have played an important role in locating protein modules. In this paper, a new approach combining Gene Ontology and amino acid background frequency is introduced to detect the protein modules in the weighted PPI networks. The proposed approach mainly consists of three parts: the feature extraction, the weighted graph construction and the protein complex detection. Firstly, the topology-sequence information is utilized to present the feature of protein complex. Secondly, six types of the weighed graph are constructed by combining PPI network and Gene Ontology information. Lastly, protein complex algorithm is applied to the weighted graph, which locates the clusters based on three conditions, including density, network diameter and the included angle cosine. Experiments have been conducted on two protein complex benchmark sets for yeast and the results show that the approach is more effective compared to five typical algorithms with the performance of f-measure and precision. The combination of protein interaction network with sequence and gene ontology data is helpful to improve the performance and provide a optional method for protein module detection.

1. Introduction

Modularity is a ubiquitous phenomenon in various network systems (Lorenz et al., 2011). A biological network manifests a modular organization and consists of different functional modules. Much of a cell's activity is organized as a network composed of lots of the interacting modules (Chen et al., 2014; Segal et al., 2003), and altering the connections between the different modules may affect changes in cellular properties and functions. A protein module, composed of interdependencies of proteins, is a group of proteins giving rise to the target-specific function whose function is separable from those of other modules (Sharma et al., 2015). Since biologists have found that cellular functions and biochemical events are coordinately carried out by each other in protein modules, and the modular structure of a complex network is critical to functions, identifying such functional modules (or complexes) in PPI networks is of utmost importance as it assists in understanding the structural and functional properties of a biological network and also aids in describing the evolutionary orthology signal. Moreover, it is proposed that a disease is a result of the breakdown of a particular functional module (Barabási et al., 2011), and it has been demonstrated that the modular structure is of great significance in aiding the diagnosis, prevention, and therapy of deadly diseases, especially in cancer research (Segal et al., 2004; Thiagalingam, 2006). Recently a novel concept of modular pharmacology (MP) has emerged (Wang et al., 2012) in pharmacological research. Therefore, based on the above reasons, it is extremely important and necessary to identify functional modules in networks.

In the recent past, a variety of classic clustering approaches, such as density-based clustering (Adamcsek et al., 2006; Altaf-Ul-Amin et al., 2006; Bader and Hogue, 2003), hierarchical clustering (Arnau et al., 2005; Holme et al., 2003; 2010), partition-based clustering, (King et al., 2004; Pfeiffer and Pfeiffer, 2007) and flow simulation-based clustering (Cho et al., 2007; Enright et al., 2002; Pereira-Leal et al., 2002), have been introduced to identify protein complexes from protein interaction data. In recent years, a number of new approaches (Hwang et al., 2008; Inoue et al., 2010; Lecca and Re, 2015; Nepusz et al., 2012; Wu et al., 2009; Yu et al., 2015), utilizing some novel computational models to identify protein modules in a PPI network, has been emerging. Especially, the sources of other biological information have been recently employed to the detection of protein modules

* Corresponding author at: Software College, Shenyang Normal University, Shenyang 110034, PR China.
** Corresponding author.
E-mail addresses: yuyangsd1204@126.com (Y. Yu), 279901222@qq.com (N. Feng).

http://dx.doi.org/10.1016/j.jtbi.2016.10.010
Received 6 January 2016; Received in revised form 25 October 2016; Accepted 26 October 2016
Available online 29 October 2016

in PPI networks (Andreopoulos et al., 2009; Feng et al., 2010; Kouhsar et al., 2016; Lakizadeh et al., 2015; Li et al., 2015; Maraziotis et al., 2007). Though using computational approaches to detect protein functional modules in PPI networks has received considerable attention and researchers have proposed many detection ideas and schemes over the past few years, how to efficiently identify protein modules by means of multiple sources of biological information is still a vital and challenging scientific problem in computational biology.

Based on author's knowledge, there are few methods based on the primary sequence information in the feature selection in the weighted PPI network constructed from the gene ontology information. Thus, in this paper inspired by this observation, we present a novel algorithm called CSeq-GO (Combining Sequence and Gene Ontology for Protein Module Detection) to discover protein complexes from the weighted PPI network. The proposed algorithm consists of mainly three parts: weighted graph construction, feature selection and protein module detection. Moreover, the included angle cosine as the similarity measure is introduced to locate protein complex based on the sequence biological information. The topological properties are based on the fact that proteins are relatively connected densely in the complex (Bader and Hogue, 2003) and protein amino acid background frequency is virtually the axiomatic fact that “sequence specifies structure,” which gives rise to an assumption that knowledge of the amino acid sequence might be sufficient to estimate the interacting property between two proteins for a specific biological function. Therefore, the topological and biological features are both of considerable importance for a complex. This algorithm is helpful to capture more biological clusters and experiments conducted on the two public datasets show that the proposed algorithm outperforms five state-of-the-art clustering algorithms in terms of f-measure and precision.

2. Material and methods

In this part, the protein complex detection is described in detail. PPI network can be represented as an undirected weighted graph G =(V, E), where V is the nodes set, corresponding with proteins, E is the set of weighted edges, representing interactions between pairs of proteins. In CSeq-GO, the input is the weighted PPI graph and complex is considered as a subgraph in the whole PPI network, which represents a subset of nodes with a specific set of edges connecting among them.

2.1. The weighted graph construction

It is argued that the detection of protein complexes can greatly be improved by taking into account network weights globally (Nepusz et al., 2012). In this paper, gene ontology is employed to construct the weighted graph. Gene Ontology (GO) is a comprehensive resource across species describing gene and gene product biological properties related to biological process, molecular function, and cellular component. It provides us with promising ways to characterize the functional relationship between pairs of proteins and to infer the interaction between them at functional level (Consortium, 2004; Zhang and Tang, 2016).

The reliability of protein interactions is computed by the definition that qualifies the functional correlation of two proteins using Gene Ontology(GO) annotations based on semantic similarity, which has been used in information science to evaluate the similarity between two concepts in a taxonomy (Resnik, 1995). We use semantic similarity to construct the weighted graph based on the gene ontology and protein interaction information.

In this section, the PPI network is transformed into a weighted graph based on gene ontology information, where the weights are computed by the BMA (best-match) strategy of Lin's method (Lin, 1998) by utilizing the tool of FastSemSim. The attribution to each interaction reflects the degree of confidence and represents the confidence level and the related equations are in (1–3).

$$sim_{MAX}(A,\ B) = MAX_{t_1\in GO(A),t_2\in GO(B)}(sim(t_1,\ t_2)) \tag{1}$$

$$sim_{AVG}(A,\ B) = AVG_{t_1\in GO(A),t_2\in GO(B)}(sim(t_1,\ t_2)) \tag{2}$$

$$sim_{BMA}(A,\ B) = \frac{AVG_{t_1}(MAX_{t_2}sim(t_1,\ t_2)) + AVG_{t_2}(MAX_{t_1}sim(t_1,\ t_2))}{2} \tag{3}$$

2.2. Feature selection

The topological features of this paper mainly include the density and the diameter of the subgraph. The density is used based on the theory that proteins of complex in the internal parts links more closely than the external part. The subgraph diameter is selected based on small world characteristics of the network (Chakrabarti, 2005). Based on my previous study (Yu et al., 2013), the biological characteristics of the background frequency of the amino acids is introduced as the biological characteristic.

(1) Density: Node degree is the sum of the edge weight for a node v. Cluster density is defined in (5).

$$dg_w(w) = \sum_{e=(u,v)\in E} w(e) \tag{4}$$

$$den_w(G) = \frac{2*\sum_{e\in E} w(e)}{(|V|*(|V|-1))} \tag{5}$$

$|V|$ is the number of vertexes and $w(e)$ is the weight of the edge e in a cluster.

(2) Network diameter: Network diameter is the number of links in the shortest path between the furthest pair of nodes of a cluster.

(3) Amino acid background frequency: As for biological properties, amino acid background frequency is proposed and calculated in each subgraph and is defined in (6).

$$freq(C_i) = \frac{sum(C_i)}{\sum_{k=1}^{s} len(p_k)} \tag{6}$$

Where C_i is a kind of amino acid among twenty amino acids, $sum(C_i)$ is the count of this amino acid C_i appearing in a subgraph, $\sum_{k=1}^{s} len(p_k)$ is the sum of each protein amino acid sequence length in a subgraph, s is the size of subgraph.

(4) The included angle cosine (Yu et al., 2011): The included angle cosine value $\cos\theta$ measures the intrinsic similarity between two interaction proteins, which is introduced in our method based on the fact that proteins in the same complex have intrinsic similarity.

$$\cos\theta = \frac{\sum_{m=1}^{n} x_{im}x_{jm}}{\sqrt{\sum_{m=1}^{n} x_{im}^2 \sum_{m=1}^{n} x_{jm}^2}} \qquad \cos\theta \in [0\ 1] \tag{7}$$

where n (n=20) is the size of the vector $V = (x_1, x_2, \ldots, x_n)$ for background frequency, x_{im} and x_{jm} are the m^{th} value of the vector $V_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ from protein i and the vector $V_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ from protein j.

2.3. Algorithm description

Our detection part operates mainly in three stages: seed selection, cluster update and a key stage of update judgment. When a cluster is detected in this stage, the cluster is restricted by $\cos\theta$, density and diameter, which is defined in (8). As we know, the larger the value of cosine is, means the more similarity between proteins. If a node v satisfies the following constraint condition at the same time in (8), v will be added to this cluster (subgraph). Usually, $density \geq \lambda$ and λ is typically set as 0.7 in Refs (King et al., 2004) and diameter≤2 are adopted (Li et al., 2008). The algorithm flow and the description are shown in Fig. 1 and Fig. 2.