Computational Biology and Chemistry 35 (2011) 298–307

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

A degree-distribution based hierarchical agglomerative clustering algorithm for protein complexes identification

Liang Yu[a,*], Lin Gao[a], Kui Li[a], Yi Zhao[b], David K.Y. Chiu[c]

[a] School of Computer Science and Technology Xidian University, Xi'an 710071, China
[b] Bioinformatics Group, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[c] School of Computer Science, University of Guelph, Guelph, Canada

ARTICLE INFO

Article history:
Received 7 March 2011
Received in revised form 30 May 2011
Accepted 3 July 2011

Keywords:
Protein–protein interaction (PPI) networks
Complexes
Degree distribution
Hierarchical agglomerative algorithm

ABSTRACT

Since cellular functionality is typically envisioned as having a hierarchical structure, we propose a framework to identify modules (or clusters) within protein–protein interaction (PPI) networks in this paper. Based on the within-module and between-module edges of subgraphs and degree distribution, we present a formal module definition in PPI networks. Using the new module definition, an effective quantitative measure is introduced for the evaluation of the partition of PPI networks. Because of the hierarchical nature of functional modules, a hierarchical agglomerative clustering algorithm is developed based on the new measure in order to solve the problem of complexes detection within PPI networks. We use gold standard sets of protein complexes to validate the biological significance of predicted complexes. A comprehensive comparison is performed between our method and other four representative methods. The results show that our algorithm finds more protein complexes with high biological significance and a significant improvement. Furthermore, the predicted complexes by our method, whether dense or sparse, match well with known biological characteristics.

1. Introduction

Protein complexes encompass groups of genes or proteins involved in common elementary biological functions (Barabasi and Oltvai, 2004; Hartwell, 1999). Identifying the component proteins in a complex from biological networks is an important step to understand the organization and interaction of the cellular processes. Furthermore, they are extremely useful to the fundamental understanding of biological organizations, which could be indicated by the network hierarchical structure, modules (internal clusters in network) and topology. As a result, in recent years protein complexes prediction has been an actively pursued research topic. It is based on the protein interaction data generated by high-throughput methods, such as the yeast two-hybrid (Y2H) system (Giot et al., 2004; Ito et al., 2000; Ucar et al., 2006), mass spectrometry (MS) (Gavin et al., 2002; Gingras et al., 2005; Ho et al., 2002; Krogan et al., 2004; Markillie et al., 2005; Tong et al., 2002), and protein microarrays (Ge, 2000; Zhu et al., 2001).

Recently, many network computational methods have been proposed to find protein complexes in protein–protein interaction (PPI) networks. Tong et al. (2002) revealed that protein complexes generally corresponded to highly interconnected subgraphs in protein interaction graphs. Based on that finding, Spirin and Mirny (2003) proposed a method to find functional modules in PPI networks using the fully connected subgraphs (cliques). However, clique has a restrictive nature, which requires complete interconnections between the graph nodes. King et al. (2004) developed the restricted neighborhood search clustering (RNSC) algorithm, which partitions the network's nodes into clusters based on a cost function. However, like many other clustering algorithms (Enright et al., 2002; Spirin and Mirny, 2003), it is stochastic and the results depend on the quality of the initial random seeds. Additionally, there were relatively fewer modules detected by this algorithm than expected. Bader and Hogue (2003) proposed a novel three-stage molecular complex detection (MCODE) algorithm to identify the densely connected regions from PPI networks. Markov clustering method (MCL) (Pereira-Leal et al., 2004) detects the dense regions as clusters. It simulates random walks within the graph structure and partitions PPI networks into many non-overlapping dense clusters. Friedel et al. (2009) proposed an unsupervised algorithm (represented as UBT) for the identification of protein complexes from the purification data.

However, all these methods mainly focus on identifying highly connected subgraphs in PPI networks as complexes. They ignore the sparser modules that may perform useful cellular functions. To address this problem, Gavin et al. (2006) studied the organization of

* Corresponding author.
E-mail addresses: lyu@xidian.edu.cn (L. Yu), lgao@mail.xidian.edu.cn (L. Gao), lkgreatgreen@163.com (K. Li), biozy@ict.ac.cn (Y. Zhao), dchiu@cis.uoguelph.ca (D.K.Y. Chiu).

1476-9271/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2011.07.005

protein complexes and demonstrated that a protein complex generally contains a core and attachment parts. Several recent studies for detecting the core-attachment protein complexes have been performed (Leung et al., 2009; Wu et al., 2009), sufficiently illustrating that not all of the significant protein complexes are dense and many sparser modules indeed correspond to useful functional units. In addition, PPI networks have been identified as modular and hierarchical in nature (Ravasz et al., 2002; Tanay et al., 2004). Cellular functionality is typically treated as having a hierarchical structure. Therefore, the modular and hierarchical network models can be applied reasonably to PPI networks. Extracting these structures from PPI networks may provide valuable information regarding the cellular function.

Inspired by this insight, in the paper we propose a hierarchical agglomerative clustering algorithm to extract protein complexes from PPI networks. Based on the within-module and between-module edges of subgraphs and degree distribution, we present a new formal definition of module in PPI networks. Given the new module definition, an effective method of quantitative measurement is introduced to evaluate the partition results of a PPI network. Because of the hierarchical nature of functional modules, we develop a hierarchical agglomerative clustering (ADHAC) algorithm based on the new measure for complexes detection within PPI networks.

Experiments using three unweighted and two weighted *S. cerevisiae* PPI networks show that ADHAC identifies statistically significant functional complexes with accuracy. To validate the biological significance, we compare the predicted results with known protein complexes in MIPS (Munich Information Center for Protein Sequence) (Güldener et al., 2005) and GO (Gene Ontology) (Dwight et al., 2002) databases which cover three domains including biological process, molecular function and cellular component. Compared to MCL (Pereira-Leal et al., 2004), MCODE (Bader and Hogue, 2003), CFinder (Chen and Yuan, 2006), and UBT (Friedel et al., 2009), ADHAC identifies more known protein complexes. In addition, ADHAC can discover both dense and sparser biologically significant complexes.

2. Materials and methods

A PPI network can be represented as a weighted graph. In the graph, nodes are proteins and the weight of the edge connecting two nodes is the probability of their interaction. If the graph is unweighted, the weight of the edge is assumed to be 1. In the paper, the graph is synonymous with the network. We consider the graph as an undirected and weighted simple graph (without self-loops).

2.1. Definition of modules

Degree distribution often determines important global characteristics. Hence, the average degree calculated from the degree distribution can be a reliable predictor of the topological properties of PPI networks. For complexes prediction, we consider the groups of nodes within which connections are dense but between which they are sparser. Here we present an effective definition of a module, which is based on the average degree of the network.

The definition for a module is given as follows:

Definition 1. *Given a weighted PPI network G with n nodes and m edges, the adjacency matrix of G is A, which is given by*

$$A_{ij} = \begin{cases} w_{ij} & \text{weight of connection from } i \text{ to } j, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Let S be a subgraph of G, i.e. $S \subset G$. $K_{\text{avg}}(G)$ is the average degree of the graph G, i.e. $K_{\text{avg}}(G) = (1/n)\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}$. If the network structure of G is fixed and known, $K_{\text{avg}}(G)$ is a constant. For the subgraph S, the number of within-module edges of S is equal to $((1/2)\sum_{i,j\in S} A_{ij})$. The number of extension within-module edges of S is defined as M_{id}^{S}, which is given by $M_{\text{id}}^{S} = K_{\text{avg}}(G) \times (1/2)\sum_{i,j\in S} A_{ij}$. The number of between-module edges that connect S to the remaining part of G is defined as $M_{\text{od}}^{S} = \sum_{i\in S, j\in (G-S)} A_{ij}$.

According to Definition 1, the within-module edges are those connecting the internal nodes of a module. The between-module edges are those connecting a module to the remaining part of the network. Generally, the size of a module is much smaller than that of the remaining part of the network. Therefore, for a module S, its within-module edges $((1/2)\sum_{i,j\in S} A_{ij})$ are much more important than its between-module edges $M_{\text{od}}^{S} = \sum_{i\in S, j\in (G-S)} A_{ij}$. A module in a network is usually defined as a densely connected subgraph with more within-module edges than between-module edges. The definition does not reflect the importance of the difference between the within-module and between-module edges of a module. To describe the modules in PPI networks more accurately, we introduce the extension within-module edges $M_{\text{id}}^{S} = K_{\text{avg}}(G) \times (1/2)\sum_{i,j\in S} A_{ij}$. We use the average degree $K_{\text{avg}}(G)$ of the network G to measure the importance of difference between the within-module and between-module edges of a module S. Hence, a new formal definition of modules is presented as follows:

Definition 2. *Given a weighted PPI network as a graph G with n nodes and m edges, a subgraph S ($S \subset G$) is a module if we have $M_{\text{id}}^{S} > M_{\text{od}}^{S}$.*

Note that the new module definition is looser than the strong and the weak definitions of community suggested by Radicchi et al. (2004). Most of the real protein complexes are not dense (Habibi et al., 2010). Therefore, the new definition is useful for detect more significant modules.

2.2. Module filter and measurement criteria

2.2.1. Size of predicted internal module as complex

In this paper, we use a metric based on the module size to filter the partition results. Two factors are considered. One is a higher overlapping proportion between larger complexes (King et al., 2004). The other is a lower p-value for smaller complexes. Let the parameter ω represent the filter size, using the findings from (Chen and Yuan, 2006; King et al., 2004; Zhang et al., 2006). Here, we determine the value of ω experimentally and discard predicted modules with size below this value.

2.2.2. Complex coverage rate and matching rate

To evaluate the reliability of the predicted modules, two parameters are evaluated. The coverage rate, defined as $\text{CR} = N_{\text{m}}/N_{\text{c}}$, evaluates the whole set's property, where (1) N_{m} is the number of complexes that match at least one predicted module using MIPS (Güldener et al., 2005) and GO (Dwight et al., 2002), and (2) N_{c} is the total number of complexes predicted. A higher CR value indicates the algorithm performs effectively.

A predicted complex is compared with each known complex and is given a matching rate or $\text{GM} = N_{\text{pm}}/N_{\text{pc}}$ to show its significance, where N_{pm} is the number of matched proteins, and N_{pc} is the number of proteins contained in the predicted complex.

2.2.3. The p-value measurement

For validation, we compare the predicted modules with known functional classification in MIPS Functional Catalogue Database (FunCatDB) (Ruepp et al., 2004) and GO. We use the hypergeometric distribution for each categorization to model the probability of observing at least k proteins from a predicted cluster of size $|C|$, by chance a known category contains $|F|$ proteins from a group (?)