Gene 518 (2013) 42–51

Contents lists available at SciVerse ScienceDirect

Gene

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

Spotlight: Assembly of protein complexes by integrating graph clustering methods

Chia-Hao Chin [a,1], Shu-Hwa Chen [a,b,1], Chun-Yu Chen [c], Chao A. Hsiung [c], Chin-Wen Ho [d], Ming-Tat Ko [a,*], Chung-Yen Lin [a,c,e,f,**]

[a] Institute of Information Science, Academia Sinica, No. 128 Yan-Chiu-Yuan Rd., Sec. 2, Taipei 115, Taiwan
[b] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan
[c] Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes. No. 35 Keyan Rd. Zhunan, Miaoli County 350, Taiwan
[d] Department of Computer Science and Information Engineering, National Central University, No.300, Jung-da Rd, Chung-li, Tao-yuan 320, Taiwan
[e] Institute of Fisheries Science, College of Life Science, National Taiwan University, No. 1, Roosevelt Rd. Sec 4, Taipei, Taiwan
[f] Research Center of Information Technology Innovation, Academia Sinica, No. 128 Yan-Chiu-Yuan Rd., Sec. 2, Taipei 115, Taiwan

ARTICLE INFO

Available online 26 December 2012

Keywords:
Network biology
Topology
Protein complex
Algorithm

ABSTRACT

As is generally assumed, clusters in protein–protein interaction (PPI) networks perform specific, crucial functions in biological systems. Various network community detection methods have been developed to exploit PPI networks in order to identify protein complexes and functional modules. Due to the potential role of various regulatory modes in biological networks, a single method may just apply a single graph property and neglect communities highlighted by other network properties.

This work presents a novel integration method to capture protein modules/protein complexes by multiple network features detected by different algorithms. The integration method is further implemented in a web-based platform with a highly effective interactive network analyzer. Conventionally adopted methods with different perspectives on network community detection (e.g., CPM, FastGreedy, HUNTER, MCL, LE, SpinGlass, and WalkTrap) are also executed simultaneously.

Analytical results indicate that the proposed method performs better than the conventional ones. The proposed approach can capture the transcription and RNA splicing machineries from the yeast protein network. Meanwhile, proteins that are highly associated with each other, yet not described in both machineries are also identified. In sum, a protein that is closely connected to components of a known module or a complex in the network view implies the functional association among them. Importantly, our method can detect these unique network features, thus facilitating efforts to discover unknown components of functional modules/protein complexes.

Availability: *Spotlight* is freely accessible at http://hub.iis.sinica.edu.tw/spotlight. Video clips for a quick view of usage are available in the website online help page.

1. Introduction

Elucidating protein complexes and functional modules is essential for understanding genome functions. A protein complex comprises a small set of proteins that are closely associated with each other, and also present in the same scenario. Meanwhile, as a group of proteins, a functional module participates in a specific process, while each binding event may occur in the same or different time and place (Spirin and Mirny, 2003). Several protein–protein interaction databases have emerged with the advent of high-throughput technologies such as yeast two-hybrid assays and affinity purification along with tandem mass spectrometry. Previous studies analyzed the graph topology of protein–protein interaction (PPI) networks, in which the proteins are denoted as nodes and pairwise interactions are denoted as linking edges. According to their results, protein complexes and functional modules tend to be densely connected to each other while having fewer connections to the other proteins in a network (Barabási and Oltvai, 2004; Rives and Galitski, 2003; Spirin and Mirny, 2003). Above observations also imply the rationale to identify protein complexes and functional modules by detecting communities/clusters from a high coverage PPI network.

Among the various community structure detection (or graph clustering) methods applied to the PPI network to detect protein complexes and functional modules include random walk based

Abbreviations: CPM, Clique percolation method; *CS*, community score function; *CSS*, Consensus method; DAVID, Database for Annotation, Visualization and Integrated Discovery (DAVID); *E*, edge set; FN, False negative; FP, False positives; *G*, giving graph; GO, gene ontology; GUI, Graphical User Interface; *IC*, integrated clusters; *KC*, Known complex; LE, Leading eigenvector; Lsm, Like Sm; MCL, Markov cluster; MIPS, Munich Information Center for Protein Sequences; *PC*, Predicted complex; PPI, Protein–protein Interaction; PSI-MI, Proteomics Standards Initiative, Molecular Interaction; *S*, cluster; SGD, Saccharomyces Genome Database; Sm, a family of RNA-binding proteins; snRPNs, Small nuclear ribonucleoproteins; *TC*, total clusters; TP, true positives; *V*, vertex; *W*, edge weight.

* Corresponding author.
** Correspondence to: C.Y. Lin, Institute of Information Science, Academia Sinica, No. 128 Yan-Chiu-Yuan Rd., Sec. 2, Taipei 115, Taiwan.
E-mail addresses: mtko@iis.sinica.edu.tw (M.-T. Ko), cylin@iis.sinica.edu.tw (C.-Y. Lin).
[1] These authors contributed equally to this work.

0378-1119/$ – see front matter.
http://dx.doi.org/10.1016/j.gene.2012.11.087

methods (Enright et al., 2002; Pons and Latapy, 2005; van Dongen, 2000), edge betweenness-based methods (Dunn et al., 2005; Girvan and Newman, 2002; Luo et al., 2007), clique percolation methods (Adamcsek et al., 2006; Zhang et al., 2006), and core-attachment based methods (Chin et al., 2010; Leung et al., 2009; Liu et al., 2009; Wu et al., 2009). While relying on widely divergent approaches, these methods have their own unique strengths and limitations. Additionally, while various regulatory modes are presented in biological networks, a single method may just encompass a single graph property and disregards communities that may be highlighted by other network properties. Bench researchers have difficulty in justifying the applicability of various algorithms on their interesting targets. Despite the development of some consensus clustering methods to solve this problem (Asur et al., 2007; Lancichinetti and Fortunato, 2012; Zhang et al., 2009), such approaches failed to include overlapping community detection methods while attempting to integrate the partition methods that divide the entire graph into smaller subgraphs and assign each node to one cluster. Therefore, this work presents a novel integration method, capable of grabbing network community structures from the input protein network. The proposed clustering approach, in which graphs are integrated, performs superior to other conventionally adopted methods in terms of protein complex harvesting and gene ontology (GO) term enrichment. Moreover, the proposed integration method allows for the successful retrieval of the transcription machineries from the yeast protein network, as well as those proteins that are closely related to the transcription process yet are not included in the complex.

The proposed integrated graph clustering method is implemented into a web-based protein complex detection scheme with an interactive network analyzer called Spotlight. With an intuitive, zoomable graphical interface, Spotlight displays the PPI network clustering results with rich and updated annotations of proteins and their linking edges (i.e. the interactions) if the input PPIs are described by standard UniProt ID or yeast SGD IDs. For user convenience, the proposed approach includes other conventionally adopted network clustering methods (e.g., CPM (Palla et al., 2005), FastGreedy (Clauset et al., 2004), HUNTER (Chin et al., 2010), LE (Newman, 2006), MCL (van Dongen, 2000), SpinGlass (Reichardt and Bornholdt, 2006), and WalkTrap (Pons and Latapy, 2005)) in Spotlight platform to easily perform network analysis, as well as view/export, and link results to further functional analysis processes. The Spotlight-based integrating graph clustering method outperforms other network clustering methods by exploiting unique network properties that imply the functional association among proteins. Importantly, the proposed method facilitates research efforts to discover unknown functional modules/protein complex structures as well as novel complex components/regulators components from a PPI network.

As graph clustering is a variant of data clustering, related methods differed mainly in the similarity of the objects handled. Restated, in data clustering, the similarity of any two objects of the input data is well defined; meanwhile, in graph clustering, the similarity of objects is expressed by edges of an input graph. Data clustering can be classified into hard data clustering and soft data clustering. In hard clustering, an object belongs to exactly one cluster; meanwhile, in soft clustering, an object is assigned to multiple clusters with membership weights that are equivalent to one. In contrast to data clustering, graph clustering is classified into overlapping graph and non-overlapping graph. Similar to hard data clustering, an object in a non-overlapping graph clustering outcome belongs to exactly one cluster. Unlike soft data clustering, an object in an overlapping clustering result is assigned to multiple clusters with weighted ones respectively. In contrast to conventionally adopted data clustering methods, consensus clustering (also called ensemble clustering or median partitioning) attempts to integrate multiple data clustering results in order to obtain better results. Consensus clustering is based on the premise of majority rule. Restated, a consensus clustering outcome is, on average, most similar to all of the input clustering results. Therefore, consensus clustering performs poorly when the integrated clustering results significantly differ from each other. To avoid this problem, the proposed method adopts the elitist strategy, in which good clusters are chosen from all clustering results and merged together.

2. Methods

The proposed integrating graph clustering approach attempts to identify good clusters with multiple network features of a PPI network. Therefore, a measure must be designed for qualifying the clustering results concluded from different methods. This section describes the community score function for evaluating the cluster quality. The integration method is introduced as well.

2.1. Community score function

Based on the definition of weak community (Radicchi et al., 2004), Lázár et. al. proposed a measure shown as Formula (1) to judge the quality of a cluster S in a graph G (Lázár et al., 2010).

$$L(S) = \frac{1}{|S|} \times \frac{|E(G[S])|}{\binom{|S|}{2}} \sum_{i \in S} \frac{k_i^{in}(S) - k_i^{out}(S)}{d_i \times s_i}, \quad (1)$$

where $|S|$ denotes the cardinality of S; $|E(G[S])|$ represents the number of edges in the subgraph induced by S; $k_i^{in}(S)$ is the number of neighbors of a vertex i, which are also in S; in contrast, $k_i^{out}(S)$ denotes the number of neighbors of i, which are not in S; d_i represents the number of neighbors of i and s_i is the number of clusters containing i; and the ratio of $|E(G[S])|$ and $\binom{|S|}{2}$ represents the edge density of the cluster. Therefore, the fact that $L(S)$ is higher suggests that the quality of cluster S should be better. Another assumption of this measure is that the number of inward going edges (i.e. $k_i^{in}(S)$) should be greater than that of outward going edges. However, this measure is inappropriate for those clusters containing vertices in high degree. To explain this situation, Fig. 1 describes a simple example. The network shown in Fig. 1 contains four clusters. For cluster A, although the number of outward going edges of vertex v is significantly greater than that of inward going edges of vertex v, cluster A is viewed here as a cluster from this network because the outward going edges are not across different clusters. We believe that the quality of cluster A should be better than that of cluster D because the members of cluster A are more closely associated with each other than that of cluster D. However, according to Formula (1), $L(A)$ 0.6$<L(D)$ 0.73, which contradicts our intuition. To solve this problem, this work proposes the measure (i.e. the community score function) to help us select good clusters from the integrated clustering results.

Our measure is based on two observations of Watts and Strogatz: (1) PPI networks have a small average shortest path between two proteins; (2) the clustering coefficient is significantly higher than would be expected under a random selection (Watts and Strogatz, 1998). Our measure is introduced formally by using graph terms hereinafter to describe it. A vertex and an edge denote a protein and an interaction between two proteins of a PPI network. For an undirected graph G, let $G=(V, E, w)$, where V is a vertex set; E represents an edge set; and w refers to an edge weight function. For a cluster $S \subset V$, a vertex-induced subgraph $G[S]$ is S together with any edge whose endpoints are both in S. Here, the number of closed, three-step walk paths is used to describe neighboring condition of a cluster, along with the average length of shortest paths used to describe the compactness of a cluster. The community score function $CS(S)$ is defined as

$$CS(S) = \frac{\text{the number of closed walks whose step is three in G[S]}}{\text{the average shortest path lenght in G[S]}}.$$