Computational Biology and Chemistry 58 (2015) 231–240

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research article

CAMWI: Detecting protein complexes using weighted clustering coefficient and weighted density

Amir lakizadeh[a], Saeed Jalili[a,*], Sayed-Amir Marashi[b]

[a] Computer Engineering Department , Tarbiat Modares University, Tehran, Iran
[b] Department of Biotechnology, College of Science, University of Tehran, Tehran, Iran

ARTICLE INFO

Article history:
Received 16 December 2014
Received in revised form 16 June 2015
Accepted 25 July 2015
Available online 30 July 2015

Keywords:
Protein complex
Core-attachment structure
Clustering coefficient
Graph weighted density

ABSTRACT

Detection of protein complexes is very important to understand the principles of cellular organization and function. Recently, large protein–protein interactions (PPIs) networks have become available using high-throughput experimental techniques. These networks make it possible to develop computational methods for protein complex detection. Most of the current methods rely on the assumption that protein complex as a module has dense structure. However complexes have core-attachment structure and proteins in a complex core share a high degree of functional similarity, so it expects that a core has high weighted density. In this paper we present a Core-Attachment based method for protein complex detection from Weighted PPI Interactions using clustering coefficient and weighted density. Experimental results show that the proposed method, CAMWI improves the accuracy of protein complex detection.

1. Introduction

Protein complexes are groups of proteins that interact with each other. They are fundamental building blocks in many biological processes including cell cycle control, molecular transmission, signal transduction and gene expression differentiation, protein folding, transcription, translation, post-translational modification and transportation (Gavin et al., 2002). Therefore protein complexes are key cellular entities in cellular organization and function. Experimental methods such as TAP-MS and co-immunoprecipitation (Bader and Hogue, 2002) for protein complex detection are both expensive and time-consuming (Von Mering et al., 2002). Plus, in TAP-MS, transient low-affinity protein complexes may not be detected (Gavin et al., 2002). On the other hand, in the last decade, the advancement of high-throughput technologies to determine PPI has generated large volumes of PPI experimental data (Von Mering et al., 2002; Orchard et al., 2012). These networks make it possible to develop computational methods for protein complex prediction. So, computational methods can be considered as an alternative to find protein complexes (Li et al., 2010).

Clustering is the main approach to detect protein complexes from PPI networks and it is defined as categorizing data objects into groups (clusters) such that the objects in a cluster are more similar than other clusters. Computational methods for complex detection either use additional biological insights to improve clustering (e.g., core-attachment structure, evolutionary information, functional coherence, mutually exclusive and cooperative interactions), or merely use graph clustering algorithms (Srihari and Leong, 2013).

It has been widely accepted that proteins are organized in a core-attachment structure to form protein complexes (Dezső et al., 2003; Gavin et al., 2006). In a complex, proteins in a core have relatively more interactions among themselves and share a high degree of functional similarity. Attachment proteins, on the other hand, are the surrounding proteins of the core performing related functions. Many computational methods focus on detecting highly-connected subgraphs in PPI networks, but ignore their inherent core-attachment organization (Gavin et al., 2006). However, it is well-known that current PPI networks contain a considerable number of false positive and false negative interactions i.e. noise. In order to overcome the noise of PPI network, a number of methods are developed for assigning “weights” to each pair of proteins in the PPI network (Collins et al., 2007; Yang et al., 2012). This weight can be interpreted as a measure of “reliability” of the interaction between each pair of proteins. By using weighted networks in which each interaction is scored by a “reliability” value, one can reduce the effect of this noise.

* Corresponding author.
E-mail addresses: amir.lakizadeh@modares.ac.ir (A. lakizadeh), sjalili@modares.ac.ir (S. Jalili), marashi@ut.ac.ir (S.-A. Marashi).

http://dx.doi.org/10.1016/j.compbiolchem.2015.07.012
1476-9271/

Several protein complex detection algorithms, such as COACH (Wu, 2009) and CORE (Leung et al., 2009) consider the core-attachment structure but most of them are not able to make use of scored interactions. Recently, a few weighted based methods are introduced, namely CMC (Liu et al., 2009), MCL-CAw (Srihari et al., 2010), HUNTER (Chin et al., 2010) and PEWCC (Zaki et al., 2013). CMC discovered protein complexes from weighted PPI network based on maximal clique concept. MCL-CAw is also a core-attachment based refinement of MCL (Pereira-Leal et al., 2004) method that improves its prediction on weighted yeast PPI networks. HUNTER is designed to detect protein complexes from a PPI network, with the optional possibility to use gene expression data to increase the quality of results. HUNTER does not accept weighted interactions as its input, but rather, it produces the weights directly from Pearson correlation of gene expressions. In PEWCC, also the quality of the interaction data is assessed. Then, protein complexes are detected based on the concept of weighted clustering coefficient. In PEWCC the reliability of protein interactions is computed by a topology-based measure, PE, and weighted interactions are not accepted as input. The method presented by (Ma and Gao, 2012) characterizes a protein complex as a maximal clique in a virtual network which is constructed by graph communicability. There exists several methods that detect protein complexes directly from bipartite TAP data and without using PPI data. For example, (Wu et al., 2012) that consider only non-redundant, reliable bicliques computed from the TAP bipartite graph are regarded as protein-complex cores.

On the other hand, since the core proteins in a complex have relatively more interactions among themselves and share a high degree of functional similarity, we expect that a core has above weighted density. However, all recent proposed methods do not consider weighted density concept in finding the cores. In this paper, we present a novel method called CAMWI (a Core-Attachment based Method for protein complex detection using Weighted Interactions) which finds protein complexes in two phases. In the first phase, it computes the weighted clustering coefficient value for each protein and based on a threshold, it selects a seed protein for each core. Then CAMWI finds the complex core by expanding the seed. Expanding a core is continued (in a greedy approach) until no increment in the weighted density of core is obtained. In the second phase, each core complex is grown by addition of attachment proteins. Experimental results show that CAMWI significantly improves the accuracy of protein complex detection.

2. Material and method

2.1. Datasets

Presenting an accurate analysis, we evaluate CAMWI in three steps. First, we try to verify useless features of CAMWI. The WI-PHI-best (Kiemer et al., 2007) and DIP (Xenarios et al., 2002) datasets are used in this step. Second, CAMWI is compared with other weighted based methods on "Consolidated" yeast weighted PPI network (Collins et al., 2007). To evaluate the ability of CAMWI on unweighted networks, it is compared with some methods on unweighted yeast PPI network "Gavin+Krogan" (Srihari and Leong, 2013) (by assigning weight 1 to each interaction). This network is obtained by combining two popular previously-reported PPI networks, Gavin (Gavin et al., 2006) and Krogan (Krogan et al., 2006). Finally for a fair comparison, the "PPI-D1" (Liu et al., 2009), "Krogan" (Krogan et al., 2006) and "Yeast-collins-mcc" (Chin et al., 2010) networks are used to compare CAMWI with some methods. The properties of all used networks are shown in Table 1.

In order to evaluate the detected complexes, three protein complex sets are used as benchmark: "CYC2008" (Pu et al., 2009); "MIPS" (Mewes et al., 2004); and "ALOY" (Aloy et al., 2004) dataset.

Table 1
Properties of used PPI networks.

	#Proteins	#Interactions
Consolidated (Collins et al., 2007)	1622	9704
Gavin+Krogan (Srihari and Leong, 2013)	2964	13507
WI-PHI-best (Kiemer et al., 2007)	5955	50000
DIP (Xenarios et al., 2002)	4928	17201
Krogan "core" (Krogan et al., 2006)	2708	7123
PPI-D1 (Liu et al., 2009)	3869	23399
Yeast-collins-mcc (Chin et al., 2010)	3383	26003

Table 2
Properties of used benchmark protein complexes set.

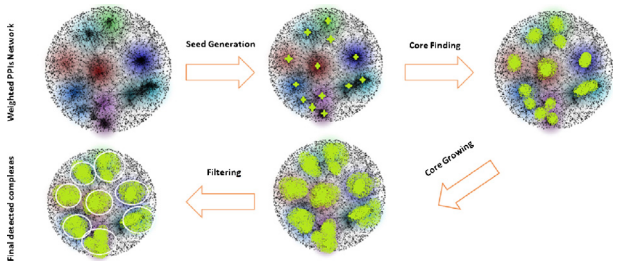
	#Complexes	#Proteins	#Complexes of size			
			<3	3–10	11–25	>25
CYC2008 (Pu et al., 2009)	408	1627	172	204	27	5
MIPS (Mewes et al., 2004)	313	1225	106	138	42	27
Aloy (Aloy et al., 2004)	101	630	23	58	19	1

Note that for an accurate evaluation, we cleaned it such that only the set of derivable benchmark complexes from each of the PPI networks is considered. If a protein be absent in a PPI network, it deletes it from the set of benchmark complexes. By repeated deletes, if the size of a benchmark complex gets below 3, we remove the complex from the benchmark. The properties of original benchmarks are shown in Table 2.

2.2. The CAMWI method

According to Fig. 1, the proposed method consists of four steps. (1) Seeds are found, (2) The concept of weighted density of a subgraph is used to expand the seeds to determine the candidate cores. Note that each candidate core is a high weighted density subgraph around the seed. (3) The protein complexes are found by addition of attachment proteins to the cores. (4) Some resulting complexes from step 3 may be redundant; so, it removes the redundant to generate the final protein complex set. Algorithm 1 shows the main function of the proposed method. Fig. 2 shows a real protein complex of yeast that is predicted correctly in a core-attachment structure by CAMWI.

Algorithm 1. CAMWI algorithm



2.2.1. Seed Generation

"Seed Generation" function (Algorithm 2) reads a weighted PPI network *G(V,E)* and a coefficient α as its inputs and returns "*Seeds*", the set of initial seeds. It assigns wcc_i, the weighted local clustering coefficient (Kalna and Higham, 2007) (Eq. (1)) to each vertice p_i in