IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 13, 2018, accepted June 14, 2018, date of publication July 2, 2018, date of current version July 25, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2852275

Viewing the Meso-Scale Structures in Protein-Protein Interaction Networks Using 2-Clubs

SONGWEI JIA[1], LIN GAO[2], YONG GAO[3], JAMES NASTOS[4], XIAO WEN[1], XIAOTAI HUANG[1], AND HAIYANG WANG[5]

[1]School of Software, Xidian University, Xi'an 710071, China
[2]School of Computer Science and Technology, Xidian University, Xi'an 710071, China
[3]Department of Computer Science, The University of British Columbia, Okanagan Campus, Kelowna, BC V1V 1V5, Canada
[4]Department of Computer Science, Okanagan College, Kelowna, BC V1Y 4X8, Canada
[5]School of Computer Science and Technology, Xi'an University of Technology, Xi'an 710048, China

Corresponding author: Lin Gao (lgao@mail.xidian.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61702397, Grant 61532014, Grant 61432010, Grant 61702396, Grant 61472299, and Grant 61602354, in part by the China Postdoctoral Science Foundation under Grant 2016M602777 and Grant 2016M600769, and in part by the Fundamental Research Funds for the Central Universities under Grant JBX171008.

ABSTRACT Many current analyses on protein-protein interaction (PPI) networks concentrate on developing the algorithms for identifying the functional modules within the PPI networks. However, understanding the internal structure in the functional modules is needed to define their roles within the entire network. We propose the use of a two-club meso-scale structure that possesses refined inner topological structural properties, useful for deciphering the aforementioned obscure structural supporting evidences. In this paper, we: 1) illustrate the feasibility and advantages of modeling functional modules as two-clubs in PPI networks by taking statistics on the diameter distribution of benchmark functional modules within several golden standard sets; 2) categorize the two-clubs into six subcategories through the use of well-defined internal graph-theoretic characterizations; and 3) analyze these six subcategories based on factors, such as their structure, and various metrics, such as topological centralities, the numbers of involved transcription factors and essential genes, and the numbers of matched protein complexes or GO terms. Our structure-driven analysis allows us to predict the roles of identified functional modules from their network structure alone. Our subcategories of coteries and social circles serve as a classification scheme for determining which functional modules are central, regional, and satellite and which are coordinating. Experimental results show that in order to precisely control and examine the PPI networks for further research, a clear understanding of how the internal topology of modules affects their function is essential.

INDEX TERMS Graph clustering, 2-club, functional modules, protein-protein interaction networks.

I. INTRODUCTION

Protein is the most basic unit of life support, and as a result, protein-protein interaction (PPI) networks are among the most fundamental biological networks. PPI networks are widely used as skeleton networks of which various domain-specific data categories are integrated to comprehensively characterize the properties of life activities. PPI networks have recently found applications in various hot fields of computational biology [1]–[3] such as prioritizing causal genes for diseases [4]–[7], identifying drug targets [8]–[11] and other applications [12]–[15], thus a thorough understanding of PPI networks is paramount for making further progress in systems biology.

The current mainstream analyses of PPI networks mainly focus on how to develop various efficient and effective algorithms for detecting functional modules within entire networks [16]–[18]. Functional modules usually consist of some interacting proteins which are densely connected within themselves but sparsely connected with the remaining PPI networks. As a result, they tend to reveal clear biological functions. Algorithms that detect functional modules within PPI networks can help understand PPI networks from global

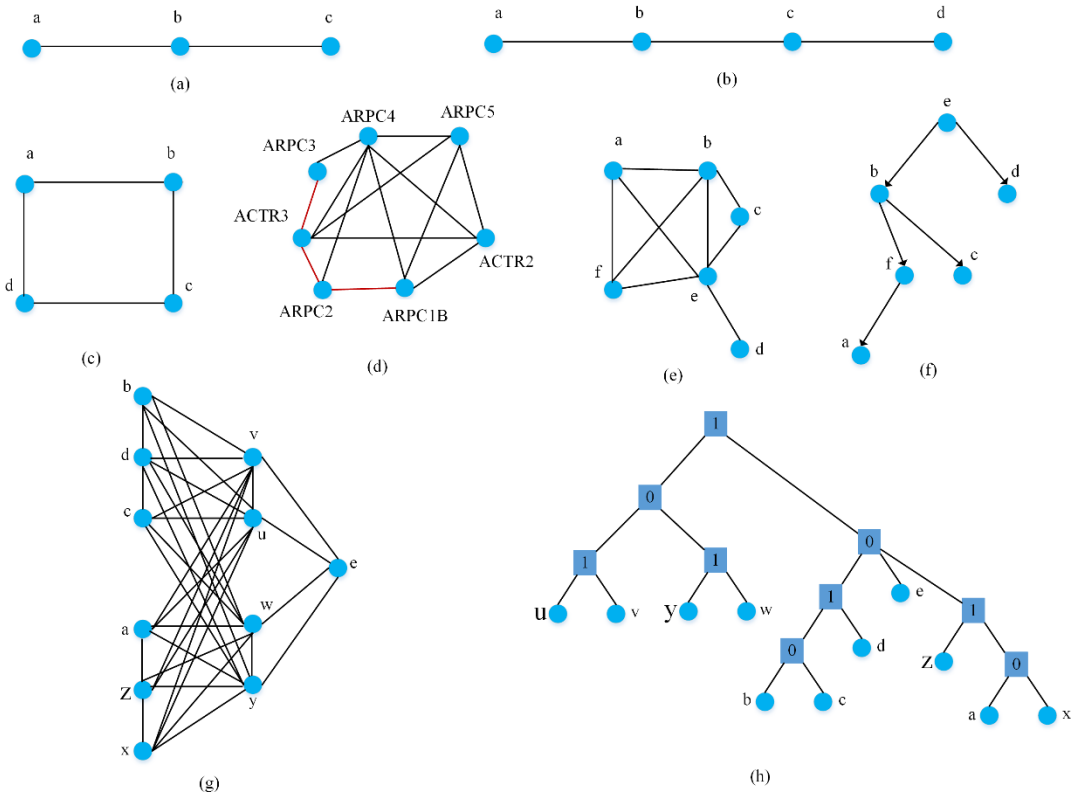


FIGURE 1. Illustrative examples of terminologies. (a) A P_3; (b) a P_4; (c) a C_4; (d) Arp2/3 complex consisting of 7 proteins and 14 interactions; (e) a (P_4, C_4)-free graph; (f) the corresponding comparability tree of (e); (g) a cograph; (h) the corresponding cotree of (g).

levels to meso-scale levels successfully. However, these do not provide valuable information on how to study PPI networks from meso-scale levels to local levels such as the levels of individual proteins and their own interactions. In order to analyze PPI networks from meso-scale levels to local levels, we need to give attention to the refined inner topological structural properties of functional modules [19]. Many of the developed algorithms for detecting dense sub graphs from entire networks are based on edge density [12], [16], [17], which inevitably neglects the interior structural properties. The well-known clique structures with the property of connecting every pair of vertices have been used to model functional modules [20], [21] in the early stages of this research field. Subsequent related researches [22]–[24] show clearly that it is not accurate to use stringent cliques to model functional modules (also called clusters or communities in other networks) mainly for the following reasons: (i) in general the cliques are all smaller than the expected functional modules; (ii) the numbers of cliques are larger than expected functional modules; (iii) cliques require every pair members to directly connect with each other, a requirement of which is too restrictive for matching expected functional modules in PPI networks.

Due to shortcomings of using cliques to model functional modules, relevant clique relaxation technologies [25], [26] have been explored and proposed, some of which mainly include density relaxation, distance relaxation and degree relaxation. In this paper we consider clique relaxation technologies focusing on restricted sub graph relaxation [27]–[34] as it can lead to natural detection of functional modules with valuable structural properties. We recall the concepts of P_3, P_4 and C_4 sub graphs in Section 2 and provide illustrative examples for these in Fig. 1. A network composed only of cliques can be characterized as a network with no induced P_3. Nastos and Gao [27] relax cliques into (P_4, C_4)-free graphs inspired by Freeman's functional module definition [24] which was mainly proposed to capture the essence of communal structure in social networks based only on the information of social ties between individual pairs. (P_4, C_4)-free graphs, which do not contain any P_4s or C_4s as induced sub graphs, are also called quasi-threshold graphs [27], [35]. There are several characterizations of (P_4, C_4)-free graphs [36]. Nastos and Gao [27] mainly focus on the property that every quasi-threshold graph has the underlying structure of its associated comparability tree, which is the rooted tree representation of a quasi-threshold graph with forest-like structure. They further utilize the forest-like structure to analyze the member status within functional modules and the in-depth relationships among members. Although the proposed (P_4, C_4)-free clique-relaxation structure has valuable structural properties, the strict requirement that it should not contain any induced

P_4 or C_4 may still be too stringent for our needs as there exist some practical functional modules containing C_4 structures. To overcome such drawbacks of (P_4, C_4)-free graphs, Jia *et al.* [28] proposed a new functional module definition: cograph community, which is equivalent to a P_4-free graph. P_4-free graphs have an unique tree structure representation called a cotree. Based on the valuable structural properties of cotrees, one can further analyze structure-equivalent subgroups within functional modules and infer phylogenomics [37], [38] in computational biology. Although P_4-free graphs have a unique cotree structure, there are still many clear practical functional modules which may contain some P_4s as induced sub graphs. For example the Arp2/3 complex contains P_4s and specifically, the three red edges denoted in Fig. 1(d) compose a P_4. Since the strategy of fitting a network to a close P_4-free graph will suppress the identification of functional modules which contain P_4s as induced sub graphs, Jia *et al.* [29] make a further relaxation by proposing the 2-club structure to model functional modules. Unlike (P_4, C_4)-free graphs having forest-like structural properties and P_4-free graphs having unique cotree structure, 2-clubs do not strictly require a specific internal structure. However, one can still study their refined structural properties according to the different diameters of their corresponding smallest spanning trees, as we exhibit in this paper. Thus we can try to take advantage of the refined structural properties of 2-clubs to analyze functional modules in PPI networks. More details about these mentioned terminologies, are given in Section 2.

In order to decipher how topological structural properties of functional modules support their roles in an entire network, we firstly illustrate the feasibility and advantages of modeling functional modules as 2-clubs in PPI networks by taking statistics on the diameter distribution of benchmark functional modules within several golden standard sets. Secondly, we categorize the 2-clubs into six subcategories: coterie 1 (CT1), coterie 2 (CT2), coterie 3 (CT3), coterie 4 (CT4), social circle 1 (SC1) and social circle 2 (SC2) based on what internal graph-theoretic properties these modules exhibit. We analyze these six subcategories from the perspective of their structural properties, various percentages and topological centralities, the numbers of involved transcription factors (TFs) and essential genes, as well as the numbers of matched protein complexes or GO terms. Our structure-driven analysis reveals that the subcategories CT4 and SC1 tend to play the roles of central functional modules, SC2 tend to play the roles of holding central functional modules together, CT2 tend to play the roles of regional central functional modules, and CT1 tend to play the roles of satellite functional modules around the central modules and regional central ones. Understanding the relationship between topological structural properties and the roles played by functional modules is a sufficient condition for designing a precise strategy in examining PPI networks thereby enhancing further research on a given biological system.

This paper is organized as follows: section 2 introduces the relevant graph-theoretic terminology and related datasets. Section 3 illustrates the feasibility and advantages of using 2-clubs to analyze functional modules in PPI networks. Section 4 introduces the 2-club detecting algorithm DIVANC. The roles of functional modules are inferred in section 4. The conclusions are presented in section 5.

II. TERMINOLOGIES AND DATASETS

In this section, we introduce terminology and datasets used in this paper.

A. TERMINOLOGIES

Generally, a network will be equivalently referred to as a *graph* and we represent a network as $G = (V, E)$ where V is the set of vertices and E is the set of edges. We describe the terminologies used in this paper as follows.

1) INDUCED SUBGRAPH

For a graph $G = (V, E)$, a sub graph $H = (V', E')$ is an *induced sub graph* of G if $V' \subseteq V$, $E' \subseteq E$, where for every pair u and v of V, $uv$is in E' if and only if uv is in E.

2) P_3 AND CLIQUE

A P_3 is a set of three vertices $\{a, b, c\}$ with edges ab and bc, inducing a path (that is, ac is not an edge), and a corresponding example is described in Fig. 1(a). A *clique* in a graph is a set of pairwise connected vertices, which is complete with all possible edges, also called a complete sub graph. Cliques are in fact P_3-free graphs as there is no sub graph equivalent to a P_3. Cliques have long been considered as the standard graph-theoretic ideal module model.

3) P_4 AND C_4

A P_4 is a set of four vertices $\{a, b, c, d\}$ with edges ab, bc and cd, also inducing a path and a corresponding example is given in Fig. 1(b). A C_4 is a set of four vertices $\{a, b, c, d\}$ with edges ab, bc, cd and da, inducing a cycle which is sometimes called a square or a 4-cycle. Corresponding examples are given in Fig. 1(c). Graphs restricting these sub graphs have been exhaustively studied as P_4 and C_4 are typical 4-sets of vertices in a network. The structure of every 4-set of vertices is shown to completely characterize many important global structural properties of a graph in algorithmic graph theory [27].

4) (P_4, C_4)-FREE GRAPH AND COMPARABILITY TREE

A graph is called a (P_4, C_4)-*free graph* if it does not contain any P_4 or C_4 as an induced sub graph, and sometimes is referred to as quasi-threshold graph. In fact, (P_4, C_4)-free graphs serve as a relaxation to P_3-free graphs (cliques) since both P_4 and C_4 are one-vertex extensions of a P_3. The most interesting relevant feature of (P_4, C_4)-free graphs, is its *associated rooted tree* representation for every (P_4, C_4)-free graph. We call the rooted tree its associated *comparability tree*. Within the forest-like structure of associated comparability trees, every vertex connects to every descendant in the tree with an edge. To illustrate, we provide a simple

(P_4, C_4)-free graph and its associated comparability tree in Figs. 1(e) and 1(f).

5) COGRAPH AND COTREE

A graph is called a *cograph* (also known as a P_4-free graph), if it does not contain a P_4 as an induced sub graph. Obviously a cograph can be obtained by further relaxing a (P_4, C_4)-free graph while just only allowing C_4s to be contained within it. Every cograph has a unique rooted tree representation which we refer to as its *cotree* in normalized form [39]. The leaves of a cotree are all vertices of its corresponding cograph, and each internal tree vertex represents either the disjoint union or a complete join operation of its children. We label each internal vertex of a cotree as follows: the root is labeled 1 (join), the children of a vertex with label 1 are labeled 0 (union) and the children of a vertex labeled 0 are labeled 1. For illustration we provide a simple cograph and its associated cotree in Figs. 1(g) and 1(h).

6) DIAMETER AND 2-CLUB

The *diameter* of a graph is the maximum distance among the shortest paths between every pair of vertices in the graph. A *2-club* is a special case of *k-club* with k = 2, and a k-club is a vertex set that induces a sub graph of diameter at most k. A *k-clique* of a graph is a sub graph in which every pair of vertices in this sub graph are joined by a path of length *k* or less, and these paths might venture outside of this *k*-clique. Similar to *k-clique*, k-clubs can also be viewed as a relaxed clique as a clique is precisely a 1-club. Note that k-club is much tighter than k-clique due to the fact that k-clubs require all the vertices in the shortest paths connecting any given pair of vertices within a k-club have to belong to this k-club but this requirement is not a requirement for k-cliques. For a thorough treatment of these graph types, their characterizations and their computational properties, the reader is referred to [40].

7) COTERIE AND SOCIAL CIRCLE

Before introducing the terminologies *coterie* and *social circle*, we describe some preliminary related terminologies. A tree is a connected graph without cycles. A spanning tree is a tree that includes all of the vertices of G. A smallest spanning tree of G is a spanning tree of G with smallest diameter [41].

According to the different diameters of corresponding smallest spanning tree, a 2-club can be classified into the sub-classes [41]: coterie and social circle. A coterie is a 2-club that has a smallest spanning tree of diameter 2; a social circle is a 2-club that has a smallest spanning tree of diameter 3.

8) NEIGHBORHOOD AFFINITY SCORE

Neighborhood affinity score is defined as a measure for evaluating the coverage level between two functional modules. For example, a detected functional module M_A with $|V_{M_A}|$ proteins or genes is thought to be matched with a benchmark functional module M_B with $|V_{M_B}|$ proteins or genes if the score of neighborhood affinity $NA(M_A, M_B) = |V_{M_A} \cap V_{M_B}|^2 \Big/ (|V_{M_A}| \times |V_{M_B}|) \geq \omega$, where the threshold is usually set as 0.2 or 0.25 in [42].

B. RELATED DATASETS

1) THE DETAILS OF PPI NETWORKS

There are many databases containing detailed information of various species' protein-protein interactions. In this paper we mainly focus on the *H.sapiens* PPI and the *S.cerevisiae* PPI networks since they are both representative species for further research. *H.sapiens* PPI networks are basic skeleton networks commonly used for integrating various domain-specific information such as disease genes and drug targets. On the other hand, *S.cerevisiae* is one of simple model species, thus researches on *S.cerevisiae* PPI networks have profound meaning. We obtained the *H.sapiens* PPI network data from the databases HPRD [43] and DIP [44], while the *S.cerevisiae* PPI network data was obtained from the database DIP [44] and we also use the widely used *S.cerevisiae* Krogan PPI network [45]. We selected these four networks for two reasons: that, firstly, they are both well-known for a fairly low false positive rate and widely used in practice, and secondly, we conduct our studies on refined inner structural properties of functional modules in PPI networks and the related structural studies are very sensitive to false interactions. There are 9269 proteins with 36917 interactions from HPRD for *H.sapiens* PPI networks (*Hsa*HPRD), 4564 proteins with 6864 interactions from DIP for *H.sapiens* PPI networks (*Hsa*DIP), while 4980 proteins with 22076 interactions from DIP for *S.cerevisiae* PPI networks (*Sce*DIP) and 2708 proteins with 7123 interactions within Krogan *S.cerevisiae* PPI networks (*Sce*Krogan).

2) GOLDEN STANDARD SETS ABOUT FUNCTIONAL MODULES

In the later part of this paper, we assess the relative importance of the detected 2-clubs from PPI networks by comparing their own numbers of the 2-clubs which can match at least a protein complex or a GO term. Here we use six widely used golden standard sets of benchmark functional modules including complexes and GO terms [46]. For *H.sapiens* PPI networks, there are 1294 benchmark complexes collected in the databases of Comprehensive Resource of Mammalian Protein Complexes (CORUM) [47] and 1204 benchmark complexes stored in Human Protein Complex Database with a Complex Quality Index (PCDq) [48], while there are also 4457 annotated GO terms purified from the GO database [49]. For *S.cerevisiae* PPI networks, there are 203 benchmark complexes distributed in the databases of Munich Information Center for Protein Sequences (MIPS) [50] and 305 benchmark complexes collected in the Saccharomyces Genome Database (SGD) [51], while there are also 1050 referenced Go terms purified from GO database [49].

3) THE SETS OF TFS AND ESSENTIAL GENES

A TF is a protein that controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence [52]. Their function is to regulate-turn on and off-genes in order to make sure that they are expressed in the right cell at the right time and in the right amount throughout the life cycle of the cell and the organism. In Section 5, we will analyze the numbers of TFs distributed among detected functional modules and we take TFs from the database ITFP: an Integrated TF Platform of mammalian transcription factors, where there are 2312 TFs for *H.sapiens*.

An essential gene is a gene which varies from being 'absolutely required for survival' to those 'strongly contributing to fitness' and robust competitive growth [53]. We also analyze the numbers of essential genes distributed among detected functional modules and we take essential genes from the database DEG 10: an update of the Database of Essential Genes that includes both protein-coding genes and noncoding genomic elements [54], where there are 2570 essential genes for *H.sapiens*.

III. MODELLING FUNCTIONAL MODULES AS 2-CLUBS

Using 2-clubs as a model for a cluster is feasible when analyzing social networks and biological networks ([30], [55], [56]). However, so far to our knowledge 2-clubs have been used for analyzing biological networks only in few papers ([29], [30], [56]). From the viewpoint of qualitative analysis, PPI networks are very suitable to analyze using 2-clubs due to the following reasons: (i) the protein interaction networks themselves are sparsely connected, incomplete and may exhibit high false-positives (ii) the 2-club is relaxed based on nothing else but the index of distance. Hence 2-clubs are fit to be used for modeling the functional modules with certain peculiar features of protein interaction networks. The peculiar features are just that the vertices do not strongly connect with each other but are at smaller distances from each other in PPI networks.

From the viewpoint of quantitative analysis, in order to verify the feasibility of modeling functional modules as 2-clubs we make analysis on the diameter size distributions among the benchmark functional modules of aforementioned six golden standard sets. Before calculating the diameter size distributions, we first have a brief look at the configurations of six golden standard sets respectively. For each golden standard set, we divide them into six groups in order to distinguish subtle difference in the distributions of diameters. The first ones are the benchmark functional modules containing some proteins which do not belong to input PPI networks and after removing those unidentified proteins the remaining proteins are connected with each other, denoted as FM1 for simplicity. The second ones are those function modules where after removing unidentified proteins the remaining proteins turn into more than one connected component, denoted as FM2; the third ones are those function modules where after removing unidentified proteins there is nothing but all isolated proteins, denoted as FM3. The fourth ones are those function modules where all of the composed proteins belong to input networks, then all the proteins are composed of only one connected component, denoted as FM4. The fifth ones are those function modules where all of the composed proteins belong to input networks but their proteins are composed of more than one connected components, denoted as FM5. The sixth ones are those function modules where all of the composed proteins belong to input networks and among those composed proteins there is no any connected components but all isolated proteins, denoted as FM6. For the sake of intuition, we show six relative percentages of their own corresponding aforementioned groups in each golden standard set. In *Hsa*HPRD, we respectively show the percentages of FM1, FM2, FM3, FM4, FM5 and FM6 from CORUM in Fig. 2(a), from PCDq in Fig. 2(b), where we only show the percentages of FM4, FM5 and FM6 from *H.sapiens* GO terms in Fig. 2(c) since there is no any functional modules belonging to FM1, FM2 and FM3. Similarly in *Hsa*DIP, we respectively show the related percentages for the FM1, FM2, FM3, FM4, FM5 and FM6 from CORUM, PCDq and *H.sapiens* GO terms in Figs. 2(d-f). In Figs. 3(a-c), we respectively show the related percentages for those different benchmark functional modules in *Sce*DIP from MIPS in Fig. 3(a), from SGD in Fig. 3(b) and from *S.cerevisiae* GO terms in Fig. 3(c). In *Sce*Krogan, we show the related percentages in Figs. 3(d-f) similar to those shown in Figs. 3(a-c).

We can expediently calculate their diameters of benchmark functional modules after knowing the detailed configuration percentages in each golden standard set. Here we only pay attention to the diameters of FM1, FM2, FM4 and FM5 in PPI networks since FM3 and FM6 are all isolated proteins. In particular, for FM2 and FM5 we cannot obtain their diameters directly since they have more than one connected components. Thus we approximately calculate the diameters of the largest connected component among all the existing connected components of FM2 and FM5. Corresponding to the configuration percentages as shown in Figs. 2(a-c), we further respectively show their diameter distributions for FM1, FM2, FM4 and FM5 from CORUM, PCDq, and *H.sapiens* GO terms within *Hsa*HPRD in Figs. 2(g-i), while show those related diameter distributions within *Hsa*DIP in Figs. 2(j-l). Similar to Figs. 2(g-l), we show the related diameter distributions for those different benchmark functional modules from MIPS, SGD and *S.cerevisiae* GO terms within *Sce*DIP and *Sce*Krogan in Figs. 3(g-l).

Specifically, we just show their percentages among total FM4 and FM5 according to their own diameter sizes in Fig. 2(i). For example, as shown in Fig. 2(a) there are 1294 benchmark functional modules from CORUM in all, among them FM1, FM2, FM3, FM4, FM5 and FM6 account for 4.17%, 8.04%, 5.49%, 51.08%, 22.26% and 8.96%. Further as shown in Fig. 2(g), among the total FM1, the benchmark functional modules with diameter 1 account for 46.3%, 33.33% with diameter 2 and 20.37% with diameter 3; among the total FM2, 34.62% with diameter 1, 30.77% with diameter 2, 29.81% with diameter 3, 3.85% with diameter 4 and 0.96%

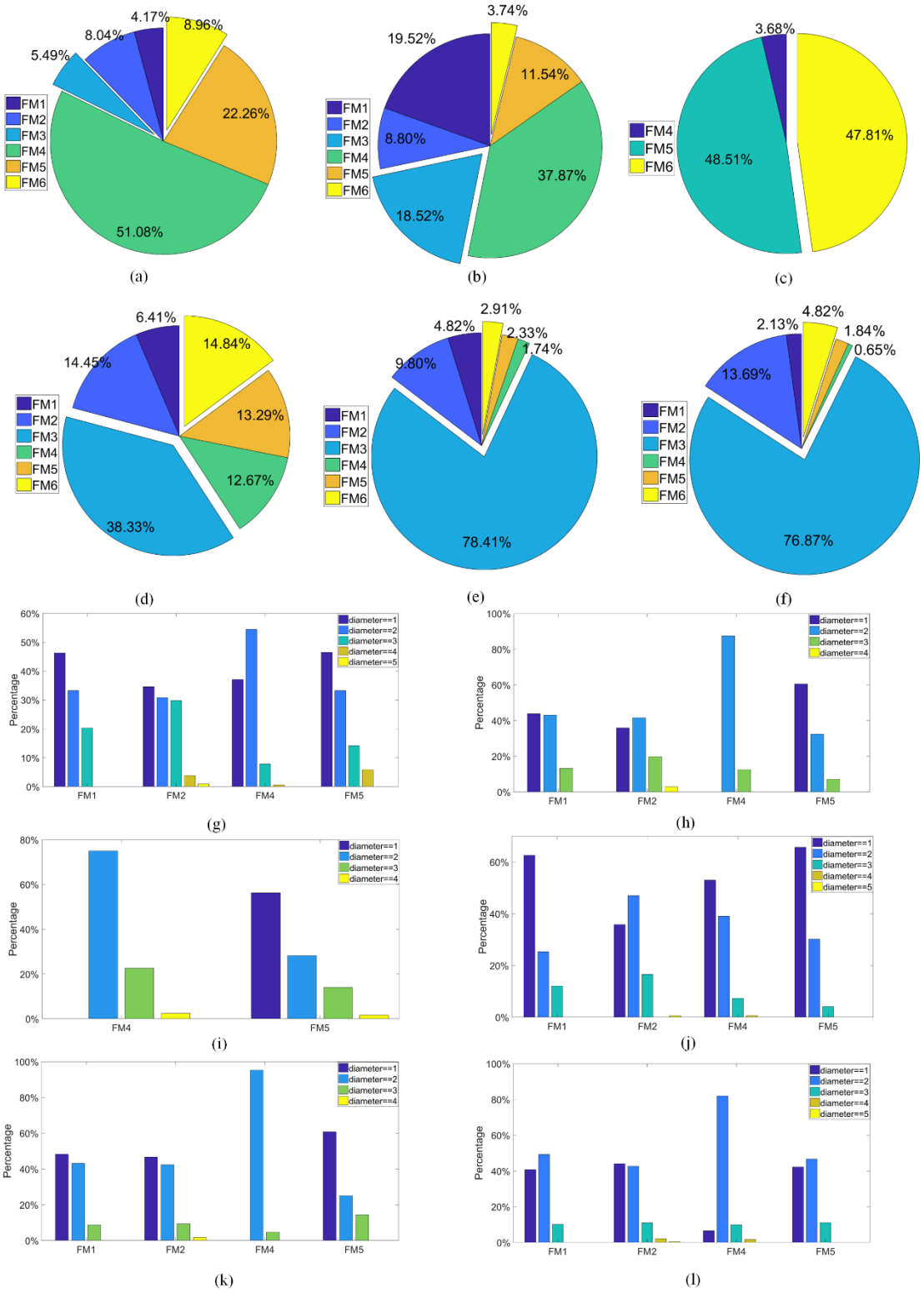


FIGURE 2. The percentages of FM1, FM2, FM3, FM4, FM5 and FM6 within CORUM, PCDq, *H.sapiens* Go terms and their related diameter distributions. (a-c) the related percentages within the three golden standard sets in *HsaHPRD* respectively; (d-f) the related percentages in *HsaDIP* similar to those shown in Figs. 2(a-c); (g-i) the diameter distributions of the benchmark functional modules within the three golden standard sets in *HsaHPRD* respectively; (j-l) the related diameter distributions in *HsaDIP* similar to those shown in Figs. 2(g-i). Specially, we just show the diameter distribution of FM4 and FM5 in Fig. 2(i) since there is no FM1, FM2 and FM3 in *H.sapiens* Go terms.

FIGURE 3. The percentages of FM1, FM2, FM3, FM4, FM5 and FM6 within MIPS, SGD and *S.cerevisiae* GO terms and their related diameter distributions. (a-c) the related percentages within the three golden standard sets in Sce DIP respectively; (d-f) the related percentages in *Sce* Krogan similar to those shown in Figs. 3(a-c); (g-i) the diameter distributions of the benchmark functional modules within the three golden standard sets in Sce DIP respectively; (j-l) the related diameter distributions in *Sce* Krogan similar to those shown in Figs. 3(g-i).

with diameter 5. For each one of the aforementioned six golden standard sets, we use a pair of panels in Fig. 2 to show their configuration percentages and their diameter size distributions clearly.

As shown in Figs. 2(g-l) and Figs. 3(g-l), the prominent mazarine and wathet blue bars represent the percentages of benchmark functional modules with diameter 1 and diameter 2. And meanwhile, integrating the related

percentages shown in Figs. 2(a-f) and Figs. 3(a-f), it is not difficult to find that there are overwhelming majority benchmark functional modules with diameter 1 or diameter 2. Our model 2-club with diameter 2 at most has the sufficient potential ability to match most of benchmark functional modules. Thus it is not only feasible but also preponderant to model functional modules as 2-clubs in PPI networks since other than appropriate diameter size, they also have available refined inner topological structural properties for further analysis.

IV. METHODS FOR OBTAINING FUNCTIONAL MODULES

In this section we describe how to detect model functional modules as 2-clubs from PPI networks efficiently. We choose appropriate algorithm DIVANC to detect 2-clubs and categorize them into six subcategories for further deciphering the structural supporting evidences of functional modules in PPI networks.

A. ALGORITHM DIVANC FOR DETECTING 2-CLUBS

Among several competitive algorithms [30], [57]–[59] for finding maximum 2-clubs, the algorithm DIVANC [29] possesses superior advantages for internal and external reasons. As for the internal reasons described in [29], DIVANC is not only inspired by the intrinsic triad-rich property within PPI networks but also verified to be efficient and effective by comparing with several other mainstream functional module detection algorithms in PPI networks. Additionally, it has the ability to detect overlapping 2-clubs and this just corresponds to the fact that real functional modules in PPI networks are overlapping. Regarding to external reasons, most of the other maximum 2-club finding algorithms are ultimately turned into solving corresponding integer programming problems and just for this reason they either cannot handle large-scale PPI networks efficiently or need previously fixing several parameters such as the minimum and maximum sizes of 2-clubs. Besides, those other 2-club finding algorithms can just detect non-overlapping functional modules, but this does not conform to the real instances about functional modules in PPI networks.

DIVANC is an edge division algorithm based on the edge niche centrality [29]. The edge niche centrality measure is defined for an edge ij to be: EdgeNiche(ij) = P_4 (ij) + min(deg(i)-1,deg(j)-1) / (C_{ij}+1), where P_4 (ij) is the P_4-centrality [28] of edge ij (which counts the number of induced paths on four vertices the edge ij participates in), deg(x) is the degree of vertex x, and C_{ij} is the *edge embeddedness* of edge ij (that is, the number of triangles in a network the edge ij participates in). By its very definition, the edge niche centrality which is established on the most basic paths and circles is more comprehensive than those other centralities which are only just based on paths or circles solely. A complete treatment of this centrality measure is given in [29], but the reader may continue without such details. The relevant understanding of the DIVANC algorithm needed here is that the algorithm performs (as many divisive community-finding algorithms do) by deleting edges according to some centrality measure, until some stopping condition is reached. In the case of DIVANC, the algorithm terminates when the connected components remaining have diameter 2, which are referred to as maximum 2-clubs. A Java implementation of this algorithm can be freely obtained from https://github.com/william0701/DIVANC.git. We provide the steps of the DIVANC algorithms below, and refer the reader to **[29]** for further details.

Algorithm 1 DIVANC

Input: PPI network;
Output: connected 2-clubs;
1:Calculate the edge niche centrality score for each edge of the input PPI network;
2:***While*** there are connected component of diameter greater than 2 do
3:Remove the edge with the highest niche centrality score among all the edges;
4: Re-calculate the scores of those edges affected by the removal of the edge;
5:***End while***
7:***For*** each of the current connected components ***do***
8:Apply the 2-hop overlapping strategy;
9:***End for***

As reported in [29], for a input network (G, V) the worst-case time complexity of DIVANC is $O\left(\bar{k}^2 |E| + \bar{k}^4 T + \bar{k} |E|\right)$ where $\bar{k}$ represents average degree of the vertices, $|E|$ is the number of total edges, T represents the total number of edges iteratively removed. Within the formula of complexity, the first term is the time to compute edge niche centrality for all edges, the second term is the time to remove T edges iteratively and the third term is the time to search and add overlapping vertices. Especially T is not a real parameter but just the number of total removed edges which does not need to be previously fixed. Since practical networks do not only have a small average degree $\bar{k}$ but also T is at most the sum of $|E| - (|V| - 1)$ and the number of detected 2-clubs, the time complexity of DIVANC $O\left(\bar{k}^2 |E| + \bar{k}^4 T + \bar{k} |E|\right)$ is very low. Except for the very low time complexity, the space complexity $O(|E|)$ is also very low. Whether the running time or the space complexity of DIVANC is both very low, thus it is able to work effectively and efficiently.

B. CATEGORIZING 2-CLUB INTO SIX SUBCATEGORIES

Refined inner topological structural properties provide infrastructures for functional modules to play corresponding potential roles. In order to infer the roles of functional modules accurately, it is necessary to explore the subtle structural differences within functional modules. Thus we need further to divide coteries and social circles into more elaborate subcategories. Inspired by Golumbic [35] dividing 2-clubs into coteries and social circles according to the different diameters of their corresponding smallest spanning trees, we further

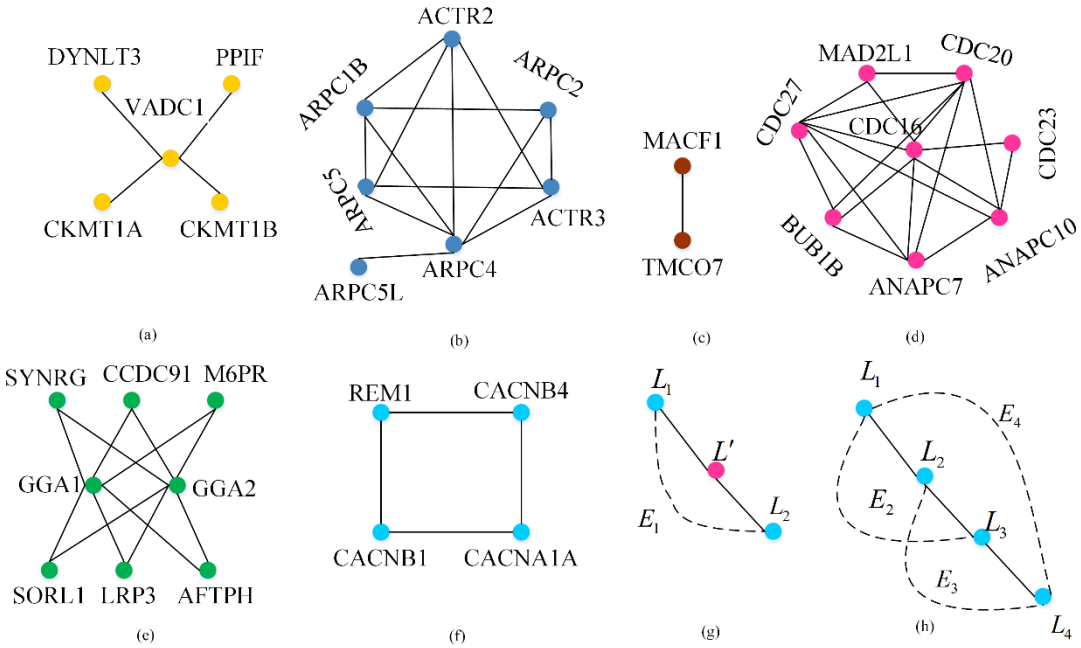


FIGURE 4. The examples of six subcategories and the diagrammatic drawing of smallest spanning trees for coteries and social circles. (a-f) the examples of CT1, CT2, CT3, CT4, SC1 and SC2; (g-h) the diagrammatic drawing of smallest spanning trees for coteries and social circles.

divide coteries into four subcategories and social circles into two subcategories.

Specifically, we categorize coteries into coterie 1 (CT1), coterie 2 (CT2) and coterie 3 (CT3), coterie 4 (CT4). Additionally, we also divide social circles into social circle 1 (SC1) and social circle 2 (SC2). Coteries have a smallest spanning tree of diameter 2, which is equivalent to the fact that they must have a star-like vertex which directly connects to each of the other remaining vertices [35]. The mere existence of a star-like vertex is not sufficient enough to capture complete internal structural properties of a module. For examples, as shown in Figs. 4(a-d), each of those four 2-clubs has a star-like vertex and they are all coteries, but there are still significant differences among them. Thus we can further categorize coteries into four subcategories: CT1, CT2, CT3 and CT4. CT1 are characterized as those structures which have the property that if we remove the star-like vertex, all of remaining vertices will break up into all isolated vertices; CT2 have the property that if we remove the star-like vertex, the remaining vertices neither compose an entire connected component nor turn into all an overall isolated vertices. CT3 are the trivial case in which the module consists of exactly two adjacent vertices. CT4 characterize those modules with the property that if we remove the star-like vertex, the remaining vertices still form a connected component. We list an ordinary examples of CT1, CT2, CT3, and CT4 detected by DIVANC in *Hsa*HPRD in Figs. 4(a-d).

Similarly, the property of possessing smallest spanning tree of diameter 3 does not sufficiently distinguish all social circles. As proved in [41], the possible cycles in a social circle are 3-cycles or 4-cycles, which are cycles consisting of three or four edges. The structural difference between the social circles with 3-cycles and 4-cycles mainly affect their own stability and ruling ability of functional modules around their local regions in PPI networks. Thus we can categorize social circles into SC1 implying those containing 3-cycles, and SC2 implying those whose smallest cycle are 4-cycles. As shown in Figs. 4(e-f), we illustrate the examples of SC1 and SC2.

V. INFERRING THE ROLES OF FUNCTIONAL MODULES

The refined inner topological structural properties of functional modules do not only decide their own biological functions but also can influence the mutual relationships with other functional modules in entire PPI networks. In this section we infer their roles of functional modules based on their own refined inner topological structural properties and simultaneously integrating the analyses from the viewpoints of related percentages of subcategory among total detected functional modules and total matched functional modules, the subcategory difference in average vertex betweenness, clustering coefficient, closeness centrality, degree and eccentricity, the average numbers of matched protein complexes and GO terms, the average numbers of containing IFs and essential genes. We here specifically call a functional module as a matched functional module if it can match at least one of the benchmark protein complexes or GO terms according to their neighborhood affinity score, for more details about neighborhood affinity please see Section 2.1.

A. CT4 AND SC1 PLAYING THE ROLES OF CENTRAL FUNCTIONAL MODULES

The evidences supporting CT4 and SC1 playing the roles of central functional modules are illustrated as follows.

1) ANALYSIS ABOUT REFINED STRUCTURAL PROPERTIES

CT4 are the 2-clubs with the property that each 2-club has a star-like vertex and a smallest spanning tree of diameter 2, furthermore if the star-like vertex is removed, the remaining vertices still can form a connected component. SC1 are the 2-clubs with the property that they have smallest spanning trees of diameter 3 and containing 3-cycle smallest cycles. In order to clearly exhibit the refined structural properties of coterie and social circle in the form of smallest spanning tree, we display them as shown in Figs. 4(g) and 4(h). The smallest spanning tree of coterie consists of three-layer vertices, while that of social circle consists of at most four-layer vertices. As shown in Fig. 4(g), the layer L' is a very special layer in which there is only one vertex referred to as the star-like vertex in original coteries. The essential difference among those coterie subcategories lie in the different numbers of edges from E_1 which connect those vertices within the layers L_1 and L_2. As shown in Fig. 4(h), the smallest spanning tree of social circle consists of four-layer, for example L_1, L_2, L_3 and L_4. E_2, E_3 and E_4 respectively contains the edges which connect the vertices belonging to the layers of L_1 and L_3, L_2 and L_4, L_1 and L_4. The difference between SC1 and SC2 mainly lies in that for SC1 there is at least one edge among those edges from E_2 or E_3, while for SC2 there must be at least one edge from E_4 with no edges from E_2 or E_3. For CT4, they possess smallest spanning trees of diameter 2, which means their constituent vertices are organized in a fairly compact manner. Furthermore, if we remove CT4's star-like vertices they can still be revealed as connected components, indicating that CT4 have significantly strong robustness. CT4 could be put up in the absence of star-like vertices and still remain connected, letting alone the absence of other ordinary vertices. For SC1, although their smallest spanning trees of diameter 3 are larger than those smallest spanning trees of diameter 2 for CT4, the 3-cycle smallest cycles just can make up for the negative influence in robustness by the slightly larger diameters of their own smallest spanning trees to a certain extent. The contained 3-cycle smallest cycles in SC1 reveal coherent and condense relationships among those constituent vertices. The unique interesting graph-theoretic characteristics of CT4 and SC1 imply that both of their topological structures are very robust and can still maintain their complete functions even if suffering from serious attacks in entire networks. Thus their own graph-theoretic characteristics provide the most basic potential structure for CT4 and SC1 acting as central functional modules from the aspect of basic topological structure.

2) ANALYSIS ABOUT RELATED PERCENTAGES FOR SUBCATEGORIES

After comprehending the structural properties of CT4 and SC1, we can learn their overall constituent instances in entire PPI networks by analyzing the percentages of related subcategories. As shown in Figs. 5(a-c), we show the percentages of six subcategories among total detected functional modules, the percentage of total matched functional modules among total detected functional modules, the percentages of six subcategories among total matched functional modules respectively in *Hsa*HPRD. We illustrate the three similar percentages of functional modules within *Hsa*DIP in Figs. 5(d-f), within *Sce*DIP in Figs. 5(g-i) and within *Sce*Krogan in Figs. 5(j-l). As shown in Figs. 5(a), the algorithm DIVANC detect 2151 functional modules in *Hsa*HPRD, in 5(d) it detects 1189 functional modules in *Hsa*DIP, while in 5(g) and 5(j), it respectively detects 1128 functional modules in *Sce*DIP and 651 functional modules in *Sce*Krogan. Among those total detected functional modules, the percentages of CT4 and SC1 are very low, for example 5.95% of CT4 and 3.35% of SC1 in *Hsa*HPRD, 7% of CT4 and 1.6% of SC1 in *Sce*DIP. In addition, in order to study the distributed functional modules of entire PPI networks comprehensively, we pay attention to the percentages of total matched functional modules among total detected modules. We can calculate the numbers of total matched functional modules according to the related percentages shown in Figs. 5(b), 5(e), 5(h) and 5(k) easily since the numbers of total detected functional modules are previously known. Furthermore we can also calculate the percentages of six subcategories among those total matched functional modules to contrast the six percentages among total detected functional modules correspondingly. As shown by the green and brown areas in Figs. 5(c) and 5(a), the percentages of CT4 and SC1 among total matched functional modules increase observably comparing to their corresponding percentages among total detected functional modules in *Hsa*HPRD. It can be seen that, the percentages of CT4 increases from 7% to 20.28% as shown in Figs. 5(g) and 5(i). The significant increase about the percentages of CT4 and SC1 among total matched functional modules means more CT4 and SC1 with more opportunities tend to match at least one of the benchmark protein complexes or GO terms with clear and important biological significance. Thus CT4 and SC1 have the unique potential ability to play the roles of central functional modules as the very low percentages of CT4 and SC1 among total detected functional modules along with the corresponding significantly increase the percentages among total matched functional modules. The low percentages of CT4 and SC1 among total detected functional modules agree with the expectation of CT4 and SC1 as the most important central functional modules in entire PPI networks provided that if the corresponding percentages are very high, it means CT4 and SC1 may be very common functional modules and absolutely impossible to be taken as the very important central functional modules.

3) ANALYSIS BASED ON TOPOLOGICAL CENTRALITIES

In order to portray potential roles of six subcategories in entire networks approximately from the viewpoint

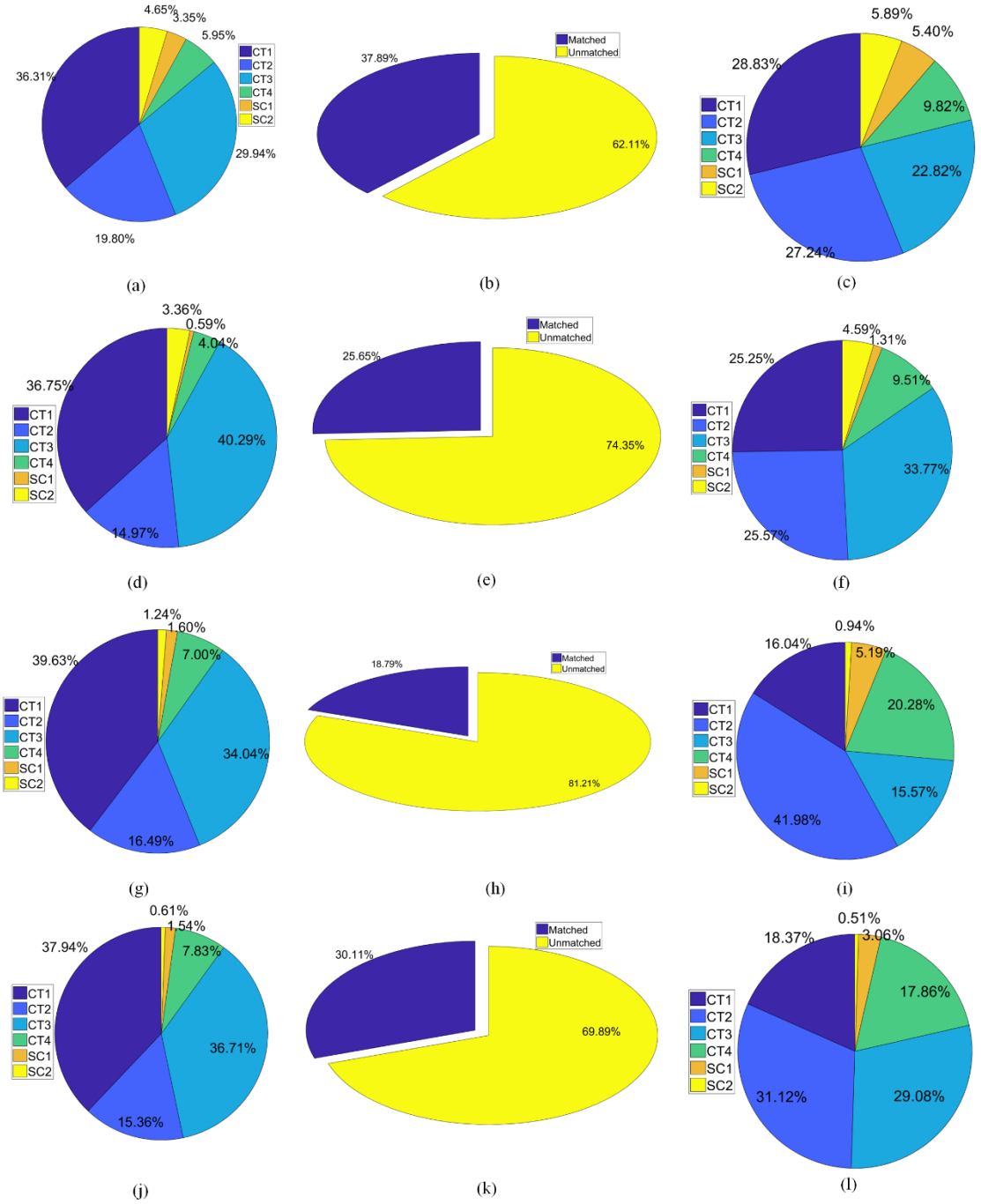


FIGURE 5. Pie charts for the percentages of subcategories. (a-c) the demonstration of percentages for six subcategories among total detected functional modules, the percentage of total matched functional modules among total detected functional modules, the percentages of six subcategories among total matched functional modules respectively in *Hsa*HPRD; (d-f) illustration of the three related percentages of functional modules in *Hsa*DIP as the instances shown in Figs. 5(a-c); (g-i) and (j-l) respectively show the three percentages of functional modules in Sce DIP and *Sce*Krogan.

of topology, we calculate the average vertex betweenness, clustering coefficient, closeness centrality, degree and eccentricity [60] for six subcategories respectively and pin our hopes on drawing the outline of their own roles based on the combination of those mentioned centralities. We use above 5 centralities due to the fact that 5 centralities are all based on the basic patterns of intuition: shortest paths, smallest circles or numbers of neighbor vertices in networks. *Betweenness centrality* is the number of shortest paths linking pairs of vertices and passing through a vertex [61]. *Clustering coefficient* of a vertex is the number of triangles (3-circles) that pass through this vertex, relative to the possible maximum number

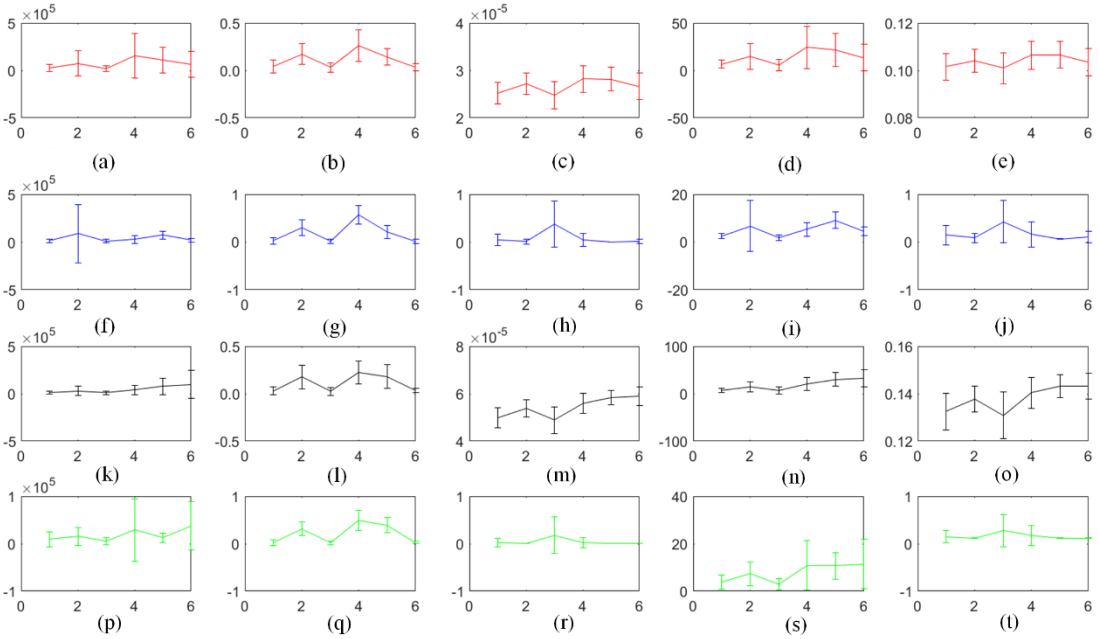


FIGURE 6. Illustrating the average vertex betweenness, clustering coefficient, closeness centrality, degree and eccentricity for subcategories. (a-e) depicting average vertex betweenness, clustering coefficient, closeness centrality, degree and eccentricity for six subcategories with their corresponding standard variances respectively in *Hsa*HPRD; (f-j) illustrating the same five centralities of six subcategories in *Hsa*DIP as the instances in (a-e); (k-o) and (p-t) respectively illustrating the same five centralities of six subcategories in *Sce*DIP and *Sce* Krogan.

of 3-circles that could pass through the vertex [62]. *Closeness centrality* is a measure of how fast information spreads from a given vertex to other reachable vertices in a network [63]. *Degree* is the number of vertices adjacent to a given vertex. *Eccentricity* of a vertex v is calculated by computing the shortest paths between the vertex v and all other vertices in a network, then the longest shortest path is chosen [60].

We can furthermore calculate their average values of corresponding centralities for each subcategory by calculating 5 centralities for each vertex of *Hsa*HPRD, *Hsa*DIP, *Sce*DIP and *Sce*Krogan. As shown in Figs. 6(a-e), we illustrate the average vertex betweenness, clustering coefficient, closeness centrality, degree and eccentricity for six subcategories with their corresponding standard variances respectively in *Hsa*HPRD. And as shown in Figs. 6(f-j), we illustrate the same percentages of six subcategories in *Hsa*DIP, in Figs. 6(k-o) for those subcategories in*Sce*DIP and in Figs. 6(p-t) for those subcategories in*Sce*Krogan. As shown in Figs. 6(a-e) and 6(f-j), for the CT4 and SC1 in *Hsa*HPRD and *Hsa*DIP, they have the largest average vertex betweenness, clustering coefficient, closeness centrality and degree comparing to other subcategories. As shown in Figs. 6(k-o) and 6(p-t), the instances of CT4 and SC1 in *Sce*DIP and *Sce*Krogan are subtle different from the instances in *Hsa*HPRD and *Hsa*DIP. Other than the average eccentricity values of CT4 and SC1 appear remarkable difference, they are largest in *Hsa*HPRD and *Sce*DIP while lowest in *Hsa*DIP and *Sce*Krogan. The specific inconsistent situations in eccentricity for CT4 and SC1 may be caused by the unconnectedness of *Hsa*DIP and *Sce*Krogan themselves. The unconnectedness derives from that we remove the interactions which cannot be verified by experiments to obtain very low false positive rate since the related structural studies are sensitive to false interactions. A higher average betweenness value of a protein in PPI networks can suggest it has a higher relevance as an organizing regulatory molecule. A very high average clustering coefficient suggests that vertex is more likely organizing single and intact functional units or modules. The closeness centrality and eccentricity are used to assess the speed of information spreading from a given vertex to other reachable vertices in a network from different aspects and there are rough inversely correlated relations between them. The largest closeness centrality and lowest eccentricity means CT4 and SC1 are distributed in central regions of entire PPI networks. A vertex with higher degree means there are more neighbor vertices around it with significance of intuition. Of course, only one topological centrality may describe the characteristics of the subcategories unilaterally, and we should consider the above centralities together for describing various subcategories more comprehensively. Thus as shown in Fig. 6, although there are slightly difference in the centralities between CT4 and SC1 among those four PPI networks, however this does not affect CT4 and SC1 possessing topological potential as central functional modules in PPI networks at all.

4) ANALYSIS ABOUT BIOLOGICAL SIGNIFICANCE

As mentioned above, we analyze the relative importance of subcategories from the viewpoints of refined structural properties, related percentages and topological centralities,

while we continue to analyze their importance at biological aspect. We try to assess the relative importance of subcategories in PPI networks based on comparing their own numbers of contained TFs, essential genes, and the corresponding matched benchmark protein complexes or GO terms.

Although the benchmark protein complexes or GO terms with clear biological significance are the most familiar reference sets used for analyzing PPI networks, it is still really needed to consider other biological properties for analyzing the importance of subcategories in PPI networks since there are majority of detected functional modules which cannot match any one of the benchmark protein complexes or GO terms, also called unmatched functional modules, as shown in Figs. 5(b), 5(e), 5(h) and 5(k). If we just only use benchmark protein complexes or GO terms as reference sets, we cannot obtain comprehensive and accurate results since we may not cover the unmatched detected functional modules effectively. Inspired by Will and Helms [18] made research on transcription factor complexes and their roles, we propose a simple intuitive computational strategy for verifying the importance of various subcategories. The specific strategy is that we correlate the relative importance of subcategories in PPI networks with their own numbers of contained TFs and essential genes. The relationship between the relative importance of subcategories and their own numbers of contained TFs and essential genes is roughly positively related.

The relationship that the more TFs and essential genes functional modules contain, the more important roles they may play lies in the following aspects. As we know, TF is a protein which controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence. Their function is to regulate-turn on and off-genes in order to make sure that they are expressed in the right cell at the right time and in the right amount throughout the life of the cell and the organism. In fact, gene expression is controlled through molecular logic circuits that combine regulatory signals of many different factors. In particular, complexion of TFs and other regulatory proteins is a prevailing and highly conserved mechanism of signal integration within critical regulatory pathways [18]. TFs together with other proteins compose the protein complexes which contain TFs referred to as TF complexes. Within TF complexes TFs are to produce decisive output signals which referred to as cooperative factors. Cooperative TFs are found in shorter distance and are more clustered within the protein interaction network than expected by chance. The TFs neither seem to share similar regulatory inputs nor regulate each other, thus they may play the roles as important regulatory drivers. Moreover, cooperative binding events are evolutionary much stronger conserved and show a greater impact on expression compared with individual binding events. In addition, essential gene is some gene which varies from genes absolutely required for survival to those strongly contributing to fitness and robust competitive growth. Thus intuitively we can obtain the rough positively related relationship between the relative importance of subcategories and their own numbers of contained TFs and essential genes as the more TFs and essential genes some functional modules contain, the stronger conservative property and regulatory leadership they may have.

As the database ITFP includes just the transcription factors for mammalian, here we only analyze the instances for *H.sapiens.* In order to keep consistent with TFs, we only extract essential genes of *H.sapiens*from the database DEG [54] for the following analysis. As previously mentioned, there may be slightly difference in the aspects of related percentages, topological centralities and biological significance between the subcategories in those four PPI networks, however this does not affect the final conclusions about various subcategories at all. As shown in Figs. 7(a-b), we illustrate the average numbers of contained TFs in CT1, CT2, CT3, CT4, SC1 and SC2 with corresponding standard variances in *Hsa*HPRD and *Hsa*DIP. As shown in Figs. 7(c-d), we illustrate the average numbers of contained essential genes in CT1, CT2, CT3, CT4, SC1 and SC2 with corresponding standard variances in *Hsa*HPRD and *Hsa*DIP. CT4 and SC1 contain more TFs and essential genes than other subcategories, leading to the largest numbers of TFs and essential genes contained in SC1. Although the TFs and essential genes possess their own distinct biological connotations, their distributed numbers within CT4 as well as SC1 are striking similar. Thus the evidence of containing TFs and essential genes can verify that CT4 and SC1 have the functional potential to play the central functional modules in entire PPI networks.

In this part we are going to assess the relative importance of various subcategories by comparing the numbers of their own matched benchmark functional modules like the instances of contained TFs and essential genes. This idea is illuminated by the methods to identify essential genes based on the known information of functional modules. Within those methods the genes distributed in functional modules are regarded more important than the genes that do not belong to any functional modules. From this we can obtain the assumption that the more functional modules a gene participates in, the more important it is natural. As shown in Figs. 7(e-h), we describe the average numbers of matched protein complexes or GO terms with corresponding standard variances for the six subcategories in *Hsa*HPRD, *Hsa*DIP, *Sce*DIP and *Sce*Krogan respectively. The average number of matched benchmark protein complexes or GO terms for CT4 and SC1 are larger than those of other subcategories. Although the benchmark functional modules cannot cover all functional modules in PPI networks but just can cover a fraction of them, the numbers of matched benchmark functional modules distributed in CT4 and SC1 are still similar to the numbers of TFs and essential genes which are mainly distributed within the whole functional modules. Thus the integrated supporting evidences in the numbers of involved TFs and essential genes, as well as the numbers of matched benchmark protein complexes or GO terms verify the fact that CT4 and SC1 may really have

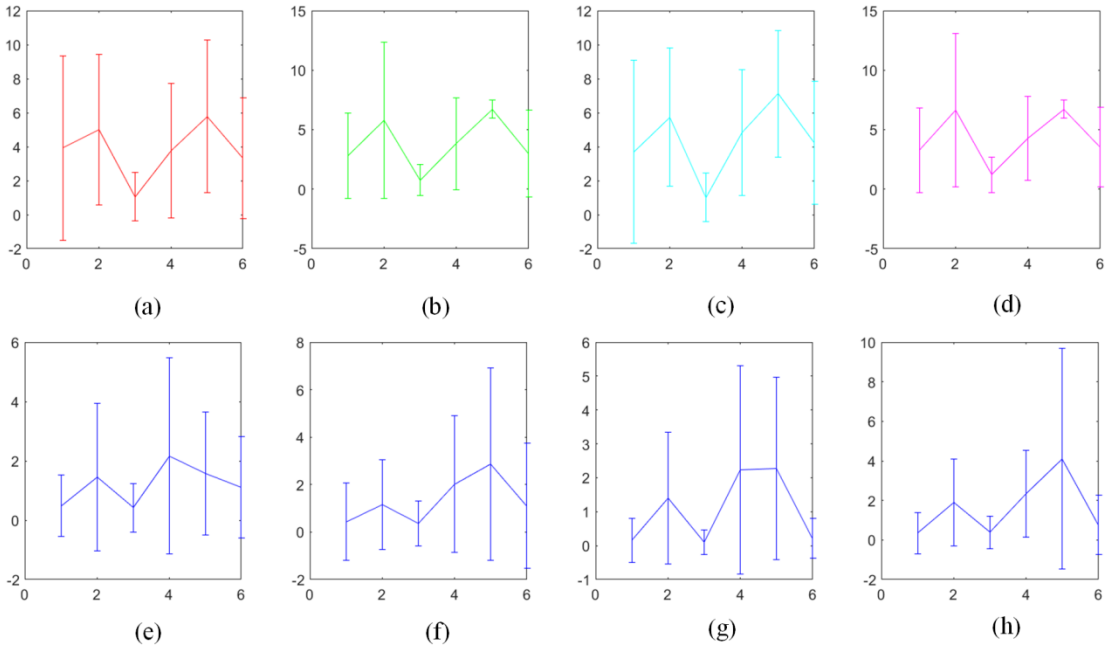


FIGURE 7. Illustrating the average numbers of contained TFs and essential genes, matched protein complexes or GO terms. (a-b) describing the average numbers of TFs involved in CT1, CT2, CT3, CT4, SC1 and SC2 with corresponding standard variances in *Hsa*HPRD and *Hsa*DIP; (c-d) describing the average numbers of essential genes involved in CT1, CT2, CT3, CT4, SC1 and SC2 with corresponding standard variances in *Hsa*HPRD and *Hsa* DIP; (e-h) describing the average numbers of matched protein complexes or GO terms with corresponding standard variances for the six subcategories in *Hsa*HPRD, *Hsa*DIP, *Sce*DIP and *Sce*Krogan respectively.

the functional potential to play the roles of central functional modules with various perspectives.

B. SC2 COORDINATING CENTRAL FUNCTIONAL MODULES

SC2 are the 2-clubs with the property that they have smallest spanning trees of diameter 3 and containing 4-cycle smallest cycles. As shown in Fig. 4(h), the smallest spanning trees of SC2 consists of four-layer vertices. It deserve noting the following key conditions: (i) there must be at least one edge to connect the vertices of layer L_1 and L_4 since SC2 containing 4-cycle smallest cycles; (ii) there are no edges from E_2 or E_3 since if there is any one edge from E_2 or E_3 it may not be considered as SC2, instead it may be considered SC1. The connections from E_4 between the highest layer L_1 and lowest layer L_4 imply that SC2 have very strong ability in coordinating the functional modules surrounding themselves. There are two reasons for the strong ability of SC2 in coordinating. Firstly, the connections from E_4 imply that there are cycles among those four layers and the 4-cycle smallest cycles which connect the vertices from different layers together, and can take a shape of a whole self-contained unit to transmit information efficiently. Secondly, unlike those vertices of SC1 which take shape 3-cycle smallest cycles only tend to communicate with the vertices also within themselves, the vertices of SC2 do not tend to communicate with their inner vertices since they prohibit the edges from E_2 or E_3 existing, thus SC2 are regarded as those vertices tend to communicate with the outside vertices.

As shown in Figs. 5(a) and 5(c), Figs. 5(d) and 5(f), Figs. 5(g) and 5(i), as well as in Figs. 5(j) and 5(l), the percentages of SC2 in total detected functional modules and in total matched functional modules look the same. They do not show significant change, such as the percentages of SC2 in *Hsa*HPRD from 4.65% to 5.89%, in *Hsa*DIP from 3.36% to 4.59%, in *Sce*DIP from 1.24% to 0.94% and in *Sce*Krogan from 0.61% to 0.51%. This consistent percentages indicate that there are no preferences for SC2 to match the benchmark functional modules. In other words, SC2 themselves do not tend to reveal individual functions.

As shown in Figs. 6(b), 6(g), 6(l) and 6(q), other than the cluster coefficient, SC2 have slightly lower average vertex betweenness, closeness centrality, degree and eccentricity than the central functional modules CT4 and SC1. As shown in Figs. 6(h), 6(r), 6(j) and 6(t), other than the closeness centrality and eccentricity of SC2 from *Hsa*DIP and*Sce*Krogan appear lowest and are a bit abnormal comparing to those of SC2 in *Hsa*HPRD and *Sce*DIP. Especially lower cluster coefficient indicates SC2 do not tend to be individual functional modules again. The reason of being abnormal in *Hsa*DIP and*Sce*Krogan is not very clear, may be also due to the unconnectedness and relatively small numbers themselves. As shown in Figs. 7(a-d), the numbers of TFs and essential genes contained in SC2 are lower than those of the central functional modules CT4 and SC1. As shown in Figs. 7(e-h), the average number of matched benchmark protein complexes or GO terms for SC2 is also lower than

those of CT4 and SC1 like the instances of TFs and essential genes.

Summarizing, the structural properties indicate SC2 have strong coordinating ability. The related centralities reveal that they are distributed around the central functional modules CT4 and SC1, while for SC2 themselves do not tend to reveal clear individual biological functions based on the related analyses of IFs, essential genes, protein complexes and GO terms. Thus the aforementioned multi-angle evidences show that SC2 tend to coordinate nothing but the central functional modules CT4 and SC1.

C. CT2 PLAYING THE ROLES OF REGIONAL CENTRAL FUNCTIONAL MODULES

CT2 are the 2-clubs with the property that each 2-club has a star-like vertex and a smallest spanning tree of diameter 2, further if the star-like vertex is removed, the remaining vertices neither compose an entire connected component nor turn into all isolated vertices overall. As shown in Fig. 4(g), the property of the remaining vertices neither compose an entire connected component nor turn into all isolated vertices overall after the star-like vertex is removed means the set of edges E_1 is neither empty nor complete. Thus the strength of robustness falls in between that of CT4 and CT1 since CT4 are central functional modules while CT1 consist of a special star-like vertex and other isolated vertices.

The percentages of CT2 in total matched functional modules increase observably comparing to those in total detected functional modules. For example, the specific percentages in *Hsa*HPRD increase from 19.8% to 27.24% as shown in the light blue areas of Figs. 5(a) and 5(c), those in *Hsa*DIP increase from 14.97% to 25.57% as shown in Figs. 5(d) and 5(f),while those in *Sce*DIP from 16.49% to 41.98% as shown in Figs. 5(g) and 5(i), those in *Sce*Krogan from 15.36% to 31.12% as shown in Figs. 5(j) and 5(l). The significant increase of CT2 is very similar to the instances of the central functional modules CT4 and SC1. Although their common increase show that CT2 tend to match at least one of the benchmark protein complexes or GO terms at a certain extent, CT2 cannot play the roles of central functional modules like CT4 and SC1 since there are quite a few CT2 in total detected functional modules.

As shown in Figs. 6, other than those in Figs. 6(h), 6(j), 6(r) and 6(t), the average centralities of CT2 are similar to the central functional modules CT4 and SC1, but they just reveal local hump in vertex betweenness, clustering coefficient, closeness centrality, degree and eccentricity.

As shown in Fig. 7, the instances of contained TFs, essential genes and numbers of matched protein complexes or GO terms in CT2 are surprisingly similar to their tendency of related topological centralities. That is CT2 reveal local hump in the numbers of contained TFs, essential genes and numbers of matched protein complexes or GO terms comparing to those of CT1 and CT3.

In summary, the structural properties indicate CT2 have local hump in robustness comparing to CT1 and CT3. The related percentages show that CT2 can be regarded as neither central functional modules nor as ordinary functional modules since they are in the right amounts for regional central roles. The related centralities reveal that CT2 are distributed in regional central areas surrounded by CT1 and CT3. The related analyses of IFs, essential genes, protein complexes and GO terms indicate the biological importance of CT2 is in the middle level among the total detected functional modules in PPI networks. Thus the aforementioned multi-angle evidences show that CT2 tend to be the regional central functional modules surrounded with CT1 and CT3.

D. CT1 PLAYING THE ROLES OF SATELLITE FUNCTIONAL MODULES AROUND CENTRAL MODULES AND REGIONAL CENTRAL ONES

CT1 are the 2-clubs with the property that each 2-club has a star-like vertex and a smallest spanning tree of diameter 2, further if the star-like vertex is removed, the remaining vertices will break up into all isolated vertices. As shown in Fig. 4(g), the property of the remaining vertices will break up into all isolated vertices implying that there are no any edges such as those edges of E_1 which can connect the vertices from the remaining layers except the special layer L'. Obviously, the robustness of CT1 is extremely weak since if the star-like vertex is under attack, there may be nothing but just all isolated vertices. Moreover, it is reported that the star-like structure possesses the property that can maintain the star-like structure themselves with consuming the least energy [64]. The least energy needed for maintaining their own structures again indicates their weak robustness from the viewpoints of energy. The weak structure decides that CT1 cannot play the roles such as central functional modules or regional central ones. However, the star-like structure tend to possess relatively independent functions since the star-like structure owns very strong cohesiveness.

The percentages of CT1 in total matched functional modules decrease observably comparing to those in total detected functional modules. For example, the specific percentages in *Hsa*HPRD decrease from 36.31% to 28.83% as shown in the dark blue areas of Figs. 5(a) and 5(c), those in *Hsa*DIP decrease from 36.75% to 25.25% as shown in Figs. 5(d) and 5(f), those in *Sce*DIP decrease from 39.63% to 16.04% as shown in Figs. 5(g) and 5(i), and those in *Sce*Krogan decrease from 37.94% to 18.37% as shown in Figs. 5(j) and 5(l). The significant decrease of CT1 in total matched functional modules is one of the two sole subcategories whose percentages appear decreasing instances. There are two reasons for the clear decrease. One is that CT1 themselves do not tend to reveal strong ability in specific biological functions, the other is that most of CT1 belong to the unknown groups of the whole PPI network so far.

As shown in Fig. 6, the average centralities of CT1 are rather consistent like the instances of CT4 and CT2. Specifically, the lower average vertex betweenness,

clustering coefficient, closeness centrality, degree and eccentricity than those of the regional central functional modules CT2. The related centrality analyses show that although the tendency of CT1 is similar to CT2, the status of CT1 may be weaker than CT2.

As shown in Fig. 7, the instances of contained TFs, essential genes and numbers of matched protein complexes or GO terms in CT1 are surprisingly similar to the tendency of their related topological centralities. In specific, CT1 have lower numbers of contained TFs, essential genes and numbers of matched protein complexes or GO terms comparing to those of CT2. And this again shows that CT1 do not tend to reveal more clear specific biological functions or there are still much unknown biological knowledge of CT1 needed to explore further.

Summarizing, the extremely weak robustness of CT1, the decreasing percentages among total matched functional modules, as well as the related centrality evidence and the biological analyses together show that CT1 play the roles of satellite functional modules around central functional modules CT4 and SC1 as well as regional central functional modules CT2.

Finally, we specifically mention the subcategory CT3, which just consists of two vertices which are connected with one edge trivially. Although we make the same related analyses about CT3 like the other subcategories, we cannot obtain clear conclusions about their potential roles in PPI networks. There are about one third of CT3 in the all subcategories but they are trivially distributed around other subcategories.

VI. CONCLUSIONS

In this paper, we try to analyze the roles of functional modules in PPI networks based on the refined inner topological structural properties of 2-clubs. We skillfully take advantage of 2-club as the meso-scale structure which is used for bridging local scale structures and global scale structures. By integrating the analyses of structural properties, related percentages, centralities and biological factors such as IFs, essential genes, matched protein complexes or GO terms, we can infer that CT4 and SC1 play the roles of central functional modules, SC2 play the roles of coordinating central functional modules, CT2 play the roles of regional central functional modules and CT1 play the roles of satellite functional modules around CT4, SC1 and CT2. Knowing clear topological structural properties of functional modules and their roles in the whole PPI network is helpful for controlling and examining PPI networks precisely, as PPI networks are usually used as skeleton networks for further research in systems biology such as prioritizing causal genes for diseases, identifying drug targets and other applications, as examples.

ACKNOWLEDGMENT

The authors would like to thank the editors and referees for their suggestions that improve the paper.

REFERENCES

[1] P. Sun *et al.*, "Protein function prediction using function associations in protein–protein interaction network," *IEEE Access*, vol. 6, pp. 30892–30902, 2018, doi: 10.1109/ACCESS.2018.2806478.
[2] H. Chen, W. Guo, J. Shen, L. Wang, and J. Song, "Structural principles analysis of host-pathogen protein–protein interactions: A structural bioinformatics survey," *IEEE Access*, vol. 6, pp. 11760–11771, 2018.
[3] X. Cao, W. Zhang, and Y. Yu, "A bootstrapping framework with interactive information modeling for network alignment," *IEEE Access*, vol. 6, pp. 13685–13696, 2018.
[4] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Mol. Syst. Biol.*, vol. 4, no. 1, p. 189, 2008.
[5] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, 2010.
[6] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Res.*, vol. 18, no. 4, pp. 644–652, 2008.
[7] D. Ghersi and M. Singh, "Interaction-based discovery of functionally important genes in cancers," *Nucleic Acids Res.*, vol. 42, no. 3, p. e18, 2013.
[8] M. A. Yıldırım, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal, "Drug—Target network," *Nature Biotechnol.*, vol. 25, no. 10, p. 1119, 2007.
[9] M. Zhu *et al.*, "The analysis of the drug–targets based on the topological properties in the human protein–protein interaction network," *J. Drug Target*, vol. 17, no. 7, pp. 524–532, 2009.
[10] R. Liu, N. Singh, G. J. Tawa, A. Wallqvist, and J. Reifman, "Exploiting large-scale drug-protein interaction information for computational drug repurposing," *BMC Bioinformat.*, vol. 15, no. 1, p. 210, 2014.
[11] W.-P. Lee, J.-Y. Huang, H.-H. Chang, K.-T. Lee, and C.-T. Lai, "Predicting drug side effects using data analytics and the integration of multiple data sources," *IEEE Access*, vol. 5, pp. 20449–20462, 2017.
[12] G. Liu, L. Wong, and H. N. Chua, "Complex discovery from weighted PPI networks," *Bioinformatics*, vol. 25, no. 15, pp. 1891–1897, 2009.
[13] K. Sun, J. P. Gonçalves, C. Larminie, and N. Pržulj, "Predicting disease associations via biological network analysis," *BMC Bioinformat.*, vol. 15, no. 1, p. 304, 2014.
[14] J. Woodsmith and U. Stelzl, "Studying post-translational modifications with protein interaction networks," *Current Opinion Struct. Biol.*, vol. 24, pp. 34–44, 2014.
[15] J. Woodsmith, A. Kamburov, and U. Stelzl, "Dual coordination of post translational modifications in human protein networks," *PLoS Comput. Biol.*, vol. 9, no. 3, p. e1002933, 2013.
[16] T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein–protein interaction networks," *Nature Methods*, vol. 9, no. 5, pp. 471–472, 2012.
[17] T. He and K. C. Chan, "Evolutionary graph clustering for protein complex identification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 3, pp. 892–904, May/Jun. 2018, doi: 10.1109/TCBB.2016.2642107.
[18] T. Will and V. Helms, "Identifying transcription factor complexes and their roles," *Bioinformatics*, vol. 30, no. 17, pp. i415–i421, Sep. 2014.
[19] C.-H. Huang, T.-H. Chen, and K.-L. Ng, "Graph theory and stability analysis of protein complex interaction networks," *IET Syst. Biol.*, vol. 10, no. 2, pp. 64–75, 2016.
[20] R. D. Alba, "A graph-theoretic definition of a sociometric clique," *J. Math. Sociol.*, vol. 3, no. 1, pp. 113–126, Jul. 1973.
[21] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.
[22] R. D. Luce, "Connectivity and generalized cliques in sociometric group structure," *Psychometrika*, vol. 15, no. 2, pp. 169–190, Jun. 1950.
[23] Y. Xiaoyan, "On fuzzy cliques in fuzzy networks," *J. Math. Sociol.*, vol. 13, no. 4, pp. 359–389, Jun. 1988.
[24] L. C. Freeman, "Cliques, Galois lattices, and the structure of human social groups," *Social Netw.*, vol. 18, no. 3, pp. 173–187, Aug. 1996.
[25] A. Veremyev and V. Boginski, "Identifying large robust network clusters via new compact formulations of maximum k-club problems," *Eur. J. Oper. Res.*, vol. 218, no. 2, pp. 316–326, Apr. 2012.
[26] J. Pattillo, N. Youssef, and S. Butenko, "On clique relaxation models in network analysis," *Eur. J. Oper. Res.*, vol. 226, no. 1, pp. 9–18, Apr. 2013.
[27] J. Nastos and Y. Gao, "Familial groups in social networks," *Soc. Netw.*, vol. 35, no. 3, pp. 439–450, Jul. 2013.
[28] S. Jia *et al.*, "Defining and identifying cograph communities in complex networks," *New J. Phys.*, vol. 17, no. 1, p. 013044, 2015.

[29] S. Jia *et al.*, "Exploring triad-rich substructures by graph-theoretic characterizations in complex networks," *Phys. Stat. Mech. Appl.*, vol. 468, pp. 53–69, Feb. 2017.
[30] B. Balasundaram, S. Butenko, and S. Trukhanov, "Novel approaches for analyzing biological networks," *J. Combinatorial Optim.*, vol. 10, no. 1, pp. 23–39, Aug. 2005.
[31] B. Balasundaram, S. Butenko, and I. V. Hicks, "Clique relaxations in social network analysis: The maximum k-Plex problem," *Oper. Res.*, vol. 59, no. 1, pp. 133–142, Feb. 2011.
[32] Y. T. Huang, K. H. Lin, and B. Y. Wu, "A structural approach for finding real-friend links in Internet social networks," in *Proc. Int. Conf. Internet Things 4th Int. Conf. Cyber, Phys. Social Comput.*, 2011, pp. 305–312.
[33] S. Bruckner, F. Hüffner, and C. Komusiewicz, "A graph modification approach for finding core-periphery structures in protein interaction networks," *Algorithms Mol. Biol.*, vol. 10, p. 16, May 2015.
[34] C.-P. Yang, C.-Y. Liu, and B. Y. Wu, "Influence clubs in social networks," in *Computational Collective Intelligence. Technologies and Applications*. Berlin, Germany: Springer, 2010, pp. 1–10.
[35] M. C. Golumbic, "Trivially perfect graphs," *Discrete Math.*, vol. 24, no. 1, pp. 105–107, 1978.
[36] Y. Jing-Ho, C. Jer-Jeong, and G. J. Chang, "Quasi-threshold graphs," *Discrete Appl. Math.*, vol. 69, no. 3, pp. 247–255, Aug. 1996.
[37] M. Hellmuth and N. Wieseke, "On symbolic ultrametrics, cotree representations, and cograph edge decompositions and partitions," in *Computing and Combinatorics*. Cham, Switzerland: Springer, 2015, pp. 609–623.
[38] M. Hellmuth and N. Wieseke, "On tree representations of relations and graphs: Symbolic ultrametrics and cograph edge decompositions," *J. Combinatorial Optim.*, vol. 36, no. 2, pp. 591–616, 2017.
[39] D. G. Corneil, H. Lerchs, and L. S. Burlingham, "Complement reducible graphs," *Discrete Appl. Math.*, vol. 3, no. 3, pp. 163–174, 1981.
[40] B. Andreas and S. Spinrad, *Graph Classes: A Survey*, vol. 3. Philadelphia, PA, USA: SIAM, 1999.
[41] R. J. Mokken. (Mar. 2012). "Coteries, social circles and hamlets. Close communities: A study of acquaintance networks." [Online]. Available: https://arxiv.org/abs/1203.5218
[42] X. Li, M. Wu, C.-K. Kwoh, and S.-K. Ng, "Computational approaches for detecting protein complexes from protein interaction networks: A survey," *BMC Genomics*, vol. 11, no. 1, p. S3, 2010.
[43] T. S. Keshava Prasad *et al.*, "Human protein reference database–2009 update," *Nucleic Acids Res.*, vol. 37, no. 1, pp. D767–D772, 2008.
[44] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic Acids Res.*, vol. 32, no. 1, pp. D449–D451, 2004.
[45] N. J. Krogan *et al.*, "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, p. 637, 2006.
[46] Y. Wang and X. Qian, "Functional module identification in protein interaction networks by interaction patterns," *Bioinformatics*, vol. 30, no. 1, pp. 81–93, Jan. 2014.
[47] A. Ruepp *et al.*, "CORUM: The comprehensive resource of mammalian protein complexes," *Nucleic Acids Res.*, vol. 36, no. 1, pp. D646–D650, Jan. 2008.
[48] S. Kikugawa *et al.*, "PCDq: Human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from H-invitational protein-protein interactions integrative dataset," *BMC Syst. Biol.*, vol. 6, no. 2, p. S7, Dec. 2012.
[49] M. Ashburner *et al.*, "Gene Ontology: Tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, May 2000.
[50] H. W. Mewes *et al.*, "MIPS: Analysis and annotation of proteins from whole genomes," *Nucleic Acids Res.*, vol. 32, no. 1, pp. D41–D44, Jan. 2004.
[51] E. L. Hong *et al.*, "Gene ontology annotations at SGD: New data sources and annotation methods," *Nucleic Acids Res.*, vol. 36, no. 1, pp. D577–D581, Jan. 2008.
[52] D. S. Latchman, "Transcription factors: An overview," *Int. J. Biochem. Cell Biol.*, vol. 29, no. 12, pp. 1305–1312, Dec. 1997.
[53] S. Gerdes, R. Edwards, M. Kubal, M. Fonstein, R. Stevens, and A. Osterman, "Essential genes on metabolic maps," *Curr. Opin. Biotechnol.*, vol. 17, no. 5, pp. 448–456, Oct. 2006.
[54] H. Luo, Y. Lin, F. Gao, C.-T. Zhang, and R. Zhang, "DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D574–D580, Jan. 2014.
[55] N. Memon and H. L. Larsen, "Retracted: Structural analysis and mathematical methods for destabilizing terrorist networks using investigative data mining," in *Proc. Adv. Data Mining Appl.*, 2006, pp. 1037–1048.
[56] S. Pasupuleti, "Detection of protein complexes in protein interaction networks using n-clubs," in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Berlin, Germany: Springer, 2008, pp. 153–164.
[57] S. Hartung, C. Komusiewicz, and A. Nichterlein, "Parameterized algorithmics and computational experiments for finding 2-clubs," in *Proc. Int. Conf. Parameterized Exact Comput.*, 2012, pp. 231–241.
[58] F. D. Carvalho and M. T. Almeida, "Upper bounds and heuristics for the 2-club problem," *Eur. J. Oper. Res.*, vol. 210, no. 3, pp. 489–494, May 2011.
[59] S. Laan and Bachelor Opleiding Kunstmatige Intelligentie, "Fast finding of 2-clubs," Ph.D. dissertation, Fac. Sci., Univ. Amsterdam, Informat. Inst., Amsterdam, The Netherlands, 2012.
[60] G. Scardoni, M. Petterlini, and C. Laudanna, "Analyzing biological network parameters with CentiScaPe," *Bioinformatics*, vol. 25, no. 21, pp. 2857–2859, Nov. 2009.
[61] U. Brandes, "A faster algorithm for betweenness centrality," *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, 2001.
[62] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.
[63] M. E. J. Newman, "A measure of betweenness centrality based on random walks," *Social Netw.*, vol. 27, no. 1, pp. 39–54, 2005.
[64] M. O. Jackson, *Social and Economic Networks*. Princeton, NJ, USA: Princeton Univ. Press, 2010.

SONGWEI JIA received the B.S. and M.S. degrees in mathematics from Northern Nationalities University, Yinchuan, Ningxia, China, in 2007 and 2010, respectively, and the Ph.D. degree in computer science and technology from Xidian University, Xi'an, Shaanxi, China, in 2016. Since 2016, he has been an Assistant Professor with the School of Software, Xidian University. He was commissioned by the National Natural Science Foundation of China to take charge of the Director Fund Project. He has published five research articles about network science. His research interests include the analysis about complex networks, graph theory, and computational biology. He was a recipient of the Foundation of China Postdoctoral Science and the Fundamental Research Funds for the Central Universities.

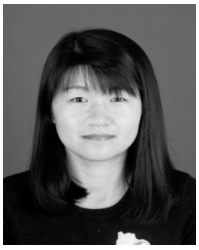
LIN GAO received the B.S. degree in computational mathematics from Xi'an Jiaotong University in 1987 and the M.S. degree in computational mathematics fromNorthwest University in 1990, and the Ph.D. degree in circuit and system from the Electronics Engineering Institute, Xidian University, in 2004. She was a Visiting Scholar with The University of Guelph, Canada, from 2004 to 2005. In 2016, she visited the University of Pennsylvania. She is currently a Professor with the School of Computer Science and Technology, Xidian University. Her research has been funded by NSFC, 863 Program, ROSC, and others. She has authored or co-authored over 100 publications in professional journals and conferences. Her research interests include data mining, bioinformatics, and graph theory and optimization. She has served on various conference program committees. She has also served as a reviewer for various journals.

YONG GAO received the B.S. degree in computational mathematics and the M.S. degree in probability and mathematical statistics from Xi'an Jiaotong University, in 1987 and 1990, respectively, and the M.S. and Ph. D. degrees in computer science and technology from the University of Alberta in 2000 and 2005, respectively. He is currently a Professor with the Department of Computer Science, The University of British Columbia, Okanagan Campus, Kelowna, Canada. His research has been funded over six times by the Natural Sciences and Engineering Research Council of Canada. He has published over 44 articles and they are cited 1274 times according to Google Scholar. He has served on various international conference program committees. He has also served as a reviewer for various journals.

JAMES NASTOS received the B.Math. degree from the University of Waterloo, the M.Sc. degree in computing science from the University of Alberta, and the Ph.D. degree from The University of British Columbia, Okanagan Campus. He is currently a Professor and the Chair of the Department of Computer Science, Okanagan College, Kelowna, Canada. He has published eight network science papers in refereed venues across mathematics, computer science, physics, social networking, and marketing venues. His research interests are graph theory and complex/social networks.

XIAO WEN received the B.S. degree in software engineering from Xidian University, Xi'an, Shaanxi, China, in 2014, where he is currently pursuing the Ph.D. degree. His research interests focus on computational biology.

XIAOTAI HUANG received the B.S. and M.S. degrees in bioengineering and bioinformatics from Northwest A&F University, China, in 2008 and 2011, respectively, and the Ph.D. degree in bioinformatics from the City University of Hong Kong in 2016. Since 2016, he has been an Assistant Professor with the School of Computer Science and Technology, Xidian University, China. He was commissioned by NSFC to take charge of the Director Fund Project. He has published five research articles. His research interests include bioinformatics, computational biology, and relation between biological networks and diseases. He was a recipient of the Foundation of China Postdoctoral Science and the Fundamental Research Funds for the Central Universities.



HAIYANG WANG received the B.S. degree in software engineering and the Ph.D. degree in computer science from Xidian University, XiâĂŹan, Shaanxi, China, in 2011 and 2018, respectively. She is currently an Assistant Professor with the School of Computer Science and Technology, Xi'an University of Technology, Shaanxi. Her research interests focus on model detection based on software and complex networks analysis.