A method for identifying protein complexes with the features of joint co-localization and joint co-expression in static PPI networks

Zhang, Jinxiong; Zhong, Cheng; Huang, Yiran; Lin, Hai Xiang; Wang, Mian

DOI
10.1016/j.compbiomed.2019.103333

Publication date
2019

Document Version
Accepted author manuscript

Published in
Computers in Biology and Medicine

Citation (APA)
Zhang, J., Zhong, C., Huang, Y., Lin, H. X., & Wang, M. (2019). A method for identifying protein complexes with the features of joint co-localization and joint co-expression in static PPI networks. *Computers in Biology and Medicine*, *111*, 1-19. Article 103333. https://doi.org/10.1016/j.compbiomed.2019.103333

A method for identifying protein complexes with the features of joint co-localization and joint co-expression in static PPI networks

Jinxiong Zhang[*a, b], Cheng Zhong[*b], Yiran Huang[b], Hai Xiang Lin[c] and Mian Wang[d]

[a]School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

[b]School of Computer, Electronics and Information, Guangxi University, Nanning, China

[c]Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, the Netherlands

[d]College of Life Science and Technology, Guangxi University, Nanning, China

Email:

Jinxiong Zhang* - zhangjx@gxu.edu.cn;

Cheng Zhong* - chzhong@gxu.edu.cn;

Yiran Huang - hyr@gxu.edu.cn;

Hai Xiang Lin - h.x.lin@tudelft.nl;

Mian Wang - mianwang@gxu.edu.cn

[*]Corresponding author

Abstract

Identifying protein complexes in static protein-protein interaction (PPI) networks is essential for understanding the underlying mechanism of biological processes. Proteins in a complex are co-localized at the same place and co-expressed at the same time. We propose a novel method to identify protein complexes with the features of joint co-localization and joint co-expression in static PPI networks. To achieve this goal, we define a joint localization vector to construct a joint co-localization criterion of a protein group, and define a joint gene expression to construct a joint co-expression criterion of a gene group. Moreover, the functional similarity of proteins in a complex is an important characteristic. Thus, we use the CC-based, MF-based, and BP-based protein similarities to devise functional similarity criterion to determine whether a protein is functionally similar to a protein cluster. Based on the core-attachment structure and following to seed expanding strategy, we use four types of biological data including PPI data with reliability score, protein localization data, gene expression data, and gene ontology annotations, to identify protein complexes. The experimental results on yeast data show that comparing with existing methods our proposed method can efficiently and exactly identify more protein complexes, especially more protein complexes of sizes from 2 to 6. Furthermore, the enrichment analysis demonstrates that the protein complexes identified by our method have significant biological meaning.

Keywords: protein complexes; static PPI networks; joint co-localization; joint co-expression; core-attachment structure; seed expanding strategy.

1. Introduction

Protein complexes are fundamental functional units in biological processes. A protein complex is a group of proteins that form a single macromolecular entity in performing a biological function. For instance, the RNA polymerase II complex, containing 12 proteins, is responsible for unwinding the DNA double helix,

polymerizing RNA, and proofreading the nascent transcript [1]. The translation initiation complex, composed of IF1, IF2, IF3, S30, and initiator tRNA, is in charge of starting the process of mRNA translation [2]. The anaphase-promoting complex containing 15 proteins is a large E3 ubiquitin ligase which controls the cell cycle process [3]. Therefore, identifying protein complexes is essential for understanding the mechanism of specific biological process in cell. Tandem Affinity Purification with Mass Spectrometry (TAP-MS) [4] is a widely used method for identifying protein complexes in wet lab. Some drawbacks remain in this experimental method. For instance, transient low affinity complexes are prone to be removed during TAP, and the protein-tag may influence protein function in the experiment [5]. Moreover, TAP-MS can only identify a limited number of known yeast complex subunits [6]. Hence, developing alternative methods to identify protein complexes remains an important issue. High-throughput experiments, such as yeast-two-hybrid (Y2H) [7, 8], protein-fragment complementation assays (PCA) [9], and TAP-MS, have produced a large number of protein-protein interaction (PPI) data from various model organisms. These PPI data can be modeled as static PPI networks whose nodes and edges represent proteins and interactions respectively. As the PPI data increase, it becomes a computational challenge to identify protein complexes in the large-scale static PPI networks.

Over the past decade, a number of research groups have studied various computational methods to identify protein complexes in static PPI networks. These computational methods identifying protein complexes can be mainly classified into two categories. The first one is solely based on the topology of static PPI networks. The topology-based methods mine highly dense sub-graphs in static PPI networks to identify complexes. In [10-12], the concept of clique is used to detect complexes in PPI networks. Instead of enumerating clique in the dense PPI networks, the complexes are identified by searching local cliques in [13-14]. Some clustering methods, such as APCluster [15] and MCL [16], are applied to find complexes in PPI networks. The seed expanding based strategy is also employed to predict complexes in PPI networks in

MCODE [17], DPCLus [18], ClusterONE [19], SPICi [20], and NEOComplex [21]. The above-mentioned methods detect dense sub-graphs to identify protein complexes by using only topology of static PPI networks.

To further improve detection accuracy, based on the integration of the topology of static PPI networks and biological information, another type of computational method for identifying protein complexes has emerged. In this type of method, biological findings, such as core-attachment structure and available data including protein Gene Ontology term annotations and gene expression data, have been integrated into computational identification of protein complexes in static PPI networks.

The study in [6] reports that the yeast complexes exhibit core-attachment structure. The protein core is the key functional unit of a protein complex. The protein attachment assists the protein core to implement the specific function. Hasin et al. [22] pointed out that a protein complex typically has two regions, viz., core and periphery. The core part is a highly dense central region where proteins are strongly connected with each other, and periphery region is a part of the complex where proteins are weakly connected with the core [22]. In fact, the concept of the core/periphery structure is originated from the core-attachment structure. In this paper, the core/periphery structure is synonymous with the core-attachment structure [22]. In [23-27], the idea of the core-attachment structure is exploited to detect protein complexes. However, these methods [23-27], which identify protein complexes by mining protein core and adding protein attachment, are still based on the topology of PPI networks.

Gene ontology (GO) project aims at standardizing the annotation of genes across species and databases by an expert-curated mechanism [28]. The GO project is divided into three ontologies: biological process (BP), molecular function (MF), and cellular component (CC). BP is referred to as a biological objective to which the gene or gene product contributes. MF is defined as the biochemical activity of a gene product. And CC is referred to as the place in the cell where a gene product is active [28]. Some protein complex identification

methods using GO annotations have been developed. In RNSC [29], the GO-based functional homogeneity, cluster size, and density are used to filter out the partitioned sub-networks to predict complexes. Price et al. [30] weighted the PPI networks with the GO-based protein similarity, and compared six existing complex prediction algorithms. To identify complexes effectively, Yang et al. [31] also used the GO-based protein similarity to weight PPI networks. Based on the organism-specific GO Slims and the GO term semantics, the similarity between two proteins are calculated to rank and predict PPI pairs [32]. Subsequently, the PPI network is reconstructed for identifying protein complexes. In PCE-FR [33], the PPI network is weighted with the GO-based protein similarity, and the pseudo-cliques are greedily extended to identify the overlapping protein complexes rapidly and effectively. The aforementioned methods [30,31,33] measure the functional similarity between two proteins by the GO-based protein similarity, but do not measure the functional homogeneity among all proteins in a complex.

Gene expression data are also widely used to analyze PPI pairs [32,34-36]. The method in [32] calculates Pearson correlation coefficient between two proteins to verify true PPI pairs through a machine learning approach using microarray gene expression data series. Feng et al. [34] weighted appropriately each node with microarray gene expression data in PPI networks, and utilized the density information to identify complexes from PPI networks. Tang et al. [35] used the gene expression data to calculate Pearson correlation coefficient between two proteins to predict complexes. WEC [36] identifies protein complexes based on the weight defined by the edge clustering coefficient and the gene expression correlation between the interacting proteins. These methods [34-36] measure the co-expression between two proteins, but do not evaluate the group co-expression among proteins in a protein complex.

Recently, some researchers used network embedding method [37] to extract the topological features of proteins in PPI network and learn protein feature vector representation. PC-SENE [38] combines node

embedding similarity with seed-extension method to detect protein complexes. In PC-SENE, the node embedding vectors generated by Node2Vec [39], are used to represent features of protein nodes in PPI network, and the node embedding similarity between interacting proteins is calculated by the generated node embedding vectors. GLONE [40] uses a global network embedding method to learn protein vector representation to preserve both high-order structure proximity and biological attribute proximity. Furthermore, based on the calculation of the cosine similarity of the protein vector representation, GLONE applies a seed-extension clustering method to detect the overlapping protein complexes. CPredictor 5.0 [41] also uses the network embedding method Node2Vec to learn node feature vector representation, and further calculates and combines the vector-based topological similarity and the GO-based functional similarity to weight PPI networks. These methods [38,40,41] employ the network embedding method to boost the performance of complexes identification.

A protein complex consists of proteins that interact with each other at the same time and place [10]. In other words, the proteins in a protein complex are jointly co-localized, jointly co-expressed, and functionally similar in biology, and they are densely connected in static PPI networks. In this paper, we propose a joint co-localization criterion, a joint co-expression criterion, and a functional similarity criterion, and design a novel method to identify protein complexes from static PPI networks. In addition, we use the yeast data sets including PPI data with reliability score, protein localization data, gene expression data, and gene ontology annotations to compute the statistical matching based metrics, and analyze the BP-based significant enrichment to evaluate our proposed method and ten other competing methods.

The main contributions of this paper are as follows. We define the localization vector and propose the joint co-localization criterion to judge whether the members of a protein group are jointly co-localized. We further define the joint gene expression and present the joint co-expression criterion to determine whether the

members of a gene group are jointly co-expressed. We calculate the CC-based, MF-based, and BP-based protein similarities and present the functional similarity criterion to ensure that the identified complexes are of functional homogeneity. Based on the three above-mentioned criteria, we design a novel method to identify protein complexes with the features of joint co-localization and joint co-expression.

The rest of this paper is organized as follows. Section 2 describes our proposed identifying method in detail. Section 3 evaluates experimental results. Section 4 discusses the characteristics of our proposed method. Section 5 concludes this paper and discusses the potential improvement for the future work.

2. Methods

In this section, we introduce three criteria and describe our proposed method in detail.

2.1. Joint co-localization criterion

A protein performs specific function in certain subcellular localization. The subcellular localization category can be classified into 22 categories listed in **Table 1** [42]. Based on the subcellular localization categories and protein localization data, we now introduce the joint localization vector to depict the co-localization of a protein group.

Table 1

Subcellular localization category

No.	subcellular localization category	No.	subcellular localization category	No.	subcellular localization category	No.	subcellular localization category
1	mitochondrion	7	ER	13	late Golgi	19	early Golgi
2	vacuole	8	nuclear periphery	14	peroxisome	20	lipid particle
3	spindle pole	9	endosome	15	actin	21	nucleus
4	cell periphery	10	bud neck	16	nucleolus	22	bud
5	punctate composite	11	microtubule	17	cytoplasm		
6	vacuolar membrane	12	Golgi	18	ER to Golgi		

NOTE: No. is the subcellular localization category number.

Definition 1. **L**ocalization **V**ector (***LV***). Given a protein P, $LV(P)$ is defined as the localization vector of P.

$LV(P)$ is a 22-dimension 0-1 vector. Let $LV_i(P)$ denote the i-th element of $LV(P)$. If protein P is localized at the i-th subcellular localization category during a cell cycle, $LV_i(P)=1$; otherwise, $LV_i(P)=0$, $i=1,\ldots, 22$.

Definition 2. **J**oint **L**ocalization **V**ector (***JLV***). Given a set of k proteins $PS=\{P_1, P_2,\ldots, P_k\}$ and $LV(P_j)$ is the localization vector of P_j, $j=1,\ldots, k$, $JLV(PS)$ is defined as the joint localization vector of PS, and $JLV_i(PS)=\bigwedge_{j=1}^{k} LV_i(P_j)$, $i=1,\ldots, 22$, where "$\wedge$" is the logical AND operation of the corresponding elements among localization vectors of proteins in PS.

If all proteins in PS perform a specific function in the i-th subcellular localization category, then $JLV_i(PS)=1$; otherwise $JLV_i(PS)=0$, $i=1,\ldots, 22$. Obviously, $JLV(PS)$ is also a 22-dimension 0-1 vector.

Definition 3. **J**oint co-localization **C**ount (***JC***). Given a set of proteins PS and its $JLV(PS)$, $JC(PS)=\sum_{i=1}^{22} JLV_i(PS)$ is defined as the joint co-localization count of PS.

If $JC(PS)>0$, we will call that all proteins in PS are jointly co-localized. If $JC(PS)=0$, we will call that all proteins in PS are not jointly co-localized. When $PS=\{P\}$, $JC(PS)=JC(\{P\})$ measures the localization count of protein P. The conditional expression "$JC(PS)>0$" is used to denote the joint co-localization criterion.

Given a joint co-localization protein set PL, there is $JC(PL)>0$. For any protein $P\notin PL$, if $JC(PL\cup\{P\})>0$, the protein P is jointly co-localized with PL; if $JC(PL\cup\{P\})=0$, the protein P is not jointly co-localized with PL.

Given a complex C and a protein group S, the complex C is *gamma-tubulin* complex comprised of YHR172W, YLR212C, and YNL126W [43], and the protein group S is composed of YBL021C, YPL246C, and YPL242C. **Table 2** shows the LV, JLV, and JC for C and S.

Table 2

LV, JLV, and JC for the complex C and the protein group S

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	JC

LV_i(YHR172W)	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
LV_i(YLR212C)	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
LV_i(YNL126W)	0	0	**1**	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**
$JLV_i(C)$	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
LV_i(YBL021C)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	**1**	0	**2**
LV_i(YPL246C)	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	**1**	0	0	0	**2**
LV_i(YPL242C)	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	**1**
$JLV_i(S)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

NOTE: All nonzero values in $LV(\cdot)$ and $JLV(\cdot)$ are marked in boxed presentation. All $JC(\cdot)$s are also marked in boxed presentation.

As can be seen in **Table 2**, the joint co-localization count of the complex C is 1, i.e. $JC(C)=1$. It means that proteins YHR172W, YLR212C, and YNL126W in the complex C are jointly co-localized. In addition, the joint co-localization count of the protein group S is 0, i.e. $JC(S)=0$. This indicates that proteins YBL021C, YPL246C, and YPL242C in the protein group S are not jointly co-localized.

2.2. Joint co-expression criterion

The subunits in a permanent complex are co-expressed [44]. It means that the protein co-expression is a prerequisite for forming a permanent complex. Because there is relation between gene expression level and protein activity [45], there must exist co-expression between genes whose products are assembled to form a permanent complex. To reveal the potential co-complex of interacting proteins, we need to deeply analyze gene co-expression. Hence, we introduce a joint co-expression criterion to judge whether a gene group is of co-expression in the following.

Under a certain condition, the gene expression profile depicts the varying pattern of RNA abundance over time. During the observed period, the gene expression value of gene g at time t is represented by $gev_g(t)$, $t=1,\ldots, T$, where T is the number of time point.

Definition 4. **g**ene **e**xpression **p**attern (***gep***). Given a gene g and its expression profile $gev_g=\{gev_g(t) \mid t=1,\ldots,T\}$, $gep_g=\{gep_g(t) \mid t=1,\ldots,T\}$ is defined as the gene expression pattern of gene g, where $gep_g(t)=$

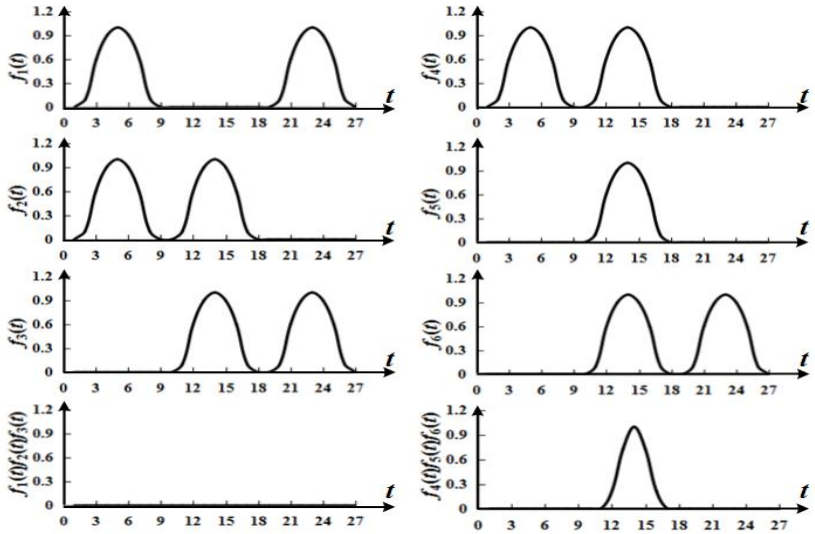
$\frac{gev_g(t) - gev_{min}}{gev_{max} - gev_{min}}$, t=1,…,T, and $gev_{min}= \min_{t=1}^{T} gev_g(t)$, $gev_{max}= \max_{t=1}^{T} gev_g(t)$. In fact, gep_g is the normalized gev_g and consists of T temporal real values in [0, 1].

Pearson correlation coefficient can be used to measure the similarity between two gene expression patterns [46]. Given two gene expression patterns $x=\{x(t) \mid t=1,\ldots,T\}$ and $y=\{y(t) \mid t=1,\ldots,T\}$, the expression pattern similarity between x and y, $pcc(x,y)$, can be calculated by formula (1).

$$pcc(x,y) = \frac{\sum_{t=1}^{T}(x(t)-\bar{x})(y(t)-\bar{y})}{\sqrt{\sum_{t=1}^{T}(x(t)-\bar{x})^2}\sqrt{\sum_{t=1}^{T}(y(t)-\bar{y})^2}} \quad (1)$$

where $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x(t)$, $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y(t)$.

Hence, for any two genes g_i and g_j, as well as their expression patterns $gep_i=\{gep_i(t) \mid t=1,\ldots,T\}$ and $gep_j=\{gep_j(t) \mid t=1,\ldots,T\}$, $pcc(gep_i, gep_j)$ can be calculated by formula (1). The higher the value of $pcc(gep_i, gep_j)$, the more similar the gene expression patterns of g_i and g_j are. If $pcc(gep_i, gep_j)\geq\delta$, we call that the gene expression patterns of g_i and g_j are similar and co-expressed, where δ is a given similarity threshold of gene expression pattern.



(a) non-collective but pairwise co-expression (b) collective co-expression

Fig. 1. Two cases that all three expression patterns in a group of hypothetical normalized expression patterns are pairwise co-expressed. (a) All pairs of the patterns $f_1(t)$, $f_2(t)$, and $f_3(t)$ are co-expressed, but $f_1(t)$, $f_2(t)$, and $f_3(t)$ are not co-expressed

together, because $f_1(t)\ f_2(t)\ f_3(t)=0$. (b) All pairs of the patterns $f_4(t)$, $f_5(t)$, and $f_6(t)$ are co-expressed, and $f_4(t)$, $f_5(t)$, and $f_6(t)$ are also co-expressed together, since $f_4(t)\ f_5(t)\ f_6(t)>0$ for $11<t<17$.

Even when every pair of genes in a gene group are co-expressed, all genes in a gene group are not necessary co-expressed all together. **Fig.1** shows two cases that all three expression patterns in a group of hypothetical normalized expression patterns are pairwise co-expressed.

To measure the joint co-expression of a gene group, we next introduce the notions of joint gene expression and joint gene expression pattern.

Definition 5. **J**oint **G**ene **E**xpression (***JGE***). Given a set of k genes $GS=\{g_1, g_2,..., g_k\}$ and $gep_j(t)$, $t=1,\ldots,T$ is the gene expression pattern of gene g_j, $j=1,\ldots,k$, let $JGE_{GS}=\{JGE_{GS}(t) \mid t=1,\ldots,T\}$ denote the joint gene expression of GS, where $JGE_{GS}(t)=\prod_{j=1}^{k} gep_j(t)$, $t=1,...,T$, "Π" is the multiplication operation of the expression pattern values among genes in GS. In fact, $JGE_{GS}(t)$ is generated by calculating the product of those normalized expression values of k genes at time t, $t=1,...,T$.

Definition 6. **J**oint co-expression **Q**uantity (***JQ***). Let $JQ(GS)=\frac{1}{T}\sum_{t=1}^{T} JGE_{GS}(t)$ denote the joint co-expression quantity of GS. If $JQ(GS)\geq\gamma$, all genes in GS are considered to be jointly co-expressed, γ is the given threshold of the joint co-expression quantity. The conditional expression "$JQ(GS)\geq\gamma$" is used to denote the joint co-expression criterion.

Definition 7. **J**oint **G**ene **E**xpression **P**attern (***JGEP***). Given a set of genes GS and its $JGE_{GS}(t)$, $t=1,...,T$, $JGEP_{GS}=\{JGEP_{GS}(t)|t=1,...,T\}$ is defined as the joint gene expression pattern of GS, where $JGEP_{GS}(t)=\frac{JGE_{GS}(t)-JGE_{min}}{JGE_{max}-JGE_{min}}$, $t=1,...,T$, and $JGE_{min}=\min_{t=1}^{T} JGE_{GS}(t)$ and $JGE_{max}=\max_{t=1}^{T} JGE_{GS}(t)$.

In fact, $JGEP_{GS}$ is the normalized joint gene expression of GS. Similar to gep, $JGEP_{GS}$ is also composed of T temporal real values in [0,1].

Given a joint co-expression gene set GS, there is $JQ(GS)\geq\gamma$. For gene $g \notin GS$, let $JGE_{GS\cup\{g\}}(t)= JGEP_{GS}(t)\times gep_g(t)$. We have $JQ(GS\cup\{g\})=\frac{1}{T}\sum_{t=1}^{T} JGE_{GS\cup\{g\}}(t)$. If $JQ(GS\cup\{g\})\geq\gamma$, we call that the gene g is jointly co-expressed with GS.

Given a gene set GS and gene g, together with their $JGEP_{GS}$ and gep_g, $pcc(JGEP_{GS}, gep_g)$ can also be computed by formula (1). If $pcc(JGEP_{GS}, gep_g)\geq\delta$, we call that gene g is similar to GS on expression pattern and jointly co-expressed with GS.

2.3. GO-based protein similarity

To express quantitatively the relationship between GO terms, term semantic similarity (SS) measures have been widely studied in the past decade. In [47], SS measures are grouped according to the following characteristics: (i) Term Information Content, (ii) Term Depth, (iii) based on a common ancestor, (iv) based on all common ancestors, (v) Path Length and (vi) Vector Space Models (*VSM*). Being one of the most popular SS measures, Resnik's measure [48] is based on the Maximum Informative Common Ancestor (*MICA*). Based on *MICA*, Zhang et al. [49] redefined three concepts used in [50]: the path length (l) between two terms, the depth (h) of a term, and local semantic density (d), to calculate SS measure between terms t_1 andt_2 by formula (2)[49].

$$sim(t_1,t_2)=e^{-c_1 l}\cdot\frac{e^{c_2 h}-e^{-c_2 h}}{e^{c_2 h}+e^{-c_2 h}}\cdot\frac{e^{c_3 d}-e^{-c_3 d}}{e^{c_3 d}+e^{-c_3 d}} \quad (2)$$

where $c_1\geq 0$, $c_2>0$, $c_3>0$. In our study, formula (2) is used to compute GO term SS measure, where $c_1=0.2$, $c_2=0.3$, $c_3=30$ [49].

In biological field, term similarity measures have been extended to objects (such as gene products and proteins) that are annotated with terms belonging to the ontology, allowing to draw a conclusion on the relationship of two proteins relying on the similarity of GO terms [47]. To accurately measure the functional similarity between two proteins, we must consider the contributions from the semantically similar terms that

annotate the two proteins respectively [51]. Thus we define the semantic similarity between a GO term and a set of GO terms. Given GO term *go* and GO term set $ST=\{t_1, t_2, ..., t_k\}$, let $Sim(go, ST)$ denote the maximum semantic similarity between term *go* and any of the terms in set *ST*. So $Sim(go, ST)$ can be represented by formula (3) [51].

$$Sim(go, ST) = \max_{j=1}^{k} sim(go, t_j) \tag{3}$$

Furthermore, given two proteins P_1 and P_2 annotated with two GO term sets $ST_1=\{t_{11}, t_{12}, ..., t_{1m}\}$ and $ST_2=\{t_{21}, t_{22}, ..., t_{2n}\}$ respectively, we can define the GO-based similarity between proteins P_1 and P_2 as $sim_{go}(P_1, P_2)$ represented by formula (4) [51].

$$sim_{go}(P_1, P_2) = \frac{\sum_{i=1}^{m} Sim(t_{1i}, ST_2) + \sum_{j=1}^{n} Sim(t_{2j}, ST_1)}{m+n} \tag{4}$$

Taking molecule function, cellular component, and biological process into consideration, we use formulas (2), (3), and (4) to calculate the MF-based protein similarity $sim_{mf}(P_1, P_2)$, the CC-based protein similarity $sim_{cc}(P_1,P_2)$, and the BP-based protein similarity $sim_{bp}(P_1, P_2)$ between proteins P_1 and P_2 respectively. The value of $sim_{mf}(P_1, P_2)$ is in [0, 1], and so are the values of $sim_{cc}(P_1, P_2)$ and $sim_{bp}(P_1, P_2)$. The larger these values are, the more similar proteins P_1 and P_2. If $sim_{mf}(P_1, P_2)\geq\omega$, proteins P_1 and P_2 are similar to each other based on MF terms, where ω is a given threshold for the MF-based protein similarity. Similarly, if $sim_{cc}(P_1, P_2)\geq\sigma$, proteins P_1 and P_2 are similar to each other based on CC terms, and if $sim_{bp}(P_1, P_2)>\theta$, proteins P_1 and P_2 are similar to each other based on BP terms, where σ and θ are given thresholds for the CC-based protein similarity and the BP-based protein similarity.

2.4. Other used terminologies

A PPI set with reliability score can be represented by a 2-tuple(I, s), where I is a set of protein-protein interactions with reliability scores s. For an interaction $(u, v) \in I$, $s(u, v)$ denotes the reliability score of the interaction (u, v), where u and v denote two interacting proteins respectively, and $s(u, v) \in [1, 2,\ldots, 999]$ [52].

A PPI network can be represented by an undirected and weighted graph $GW=(V, E, W)$, where V is a set of nodes (proteins), E is a set of edges (protein-protein interactions). For nodes $u, v \in V$, $W(u, v)$ denotes the weight of the edge(u, v) between nodes u and v. Given a PPI set (I, s) with reliability score$s$, $W(u, v)$ is computed by formula (5).

$$W(u,v) = \begin{cases} 0 & , \quad (u,v) \notin E \\ 1 & , \quad (u,v) \in E - I \\ s(u,v) & , \quad (u,v) \in E \bigcap I \end{cases} \tag{5}$$

If $W(u, v) \geq r$, we will call that there is a r-reliable link between nodes u and v, and the edge (u, v) is referred as a r-reliable edge, where r is a given reliability threshold and $r \in [1, 2,...,999]$ [52].

Let $N_r(v)=\{u|W(u,v) \geq r, u \in V\}$ denote the r-reliable neighborhood of node v, and $deg_r(v)=|N_r(v)|$ denote the r-reliable degree of node v.

Given a PPI sub-network $SN=(V', E', W')$, let $RE=\{(u, v) \mid W'(u, v) \geq r, (u, v) \in E'$, and $u, v \in V'\}$ be a set of r-reliable edges. We define the r-reliable density of SN, $d_r(SN)$, as follows:

$$d_r(SN) = \frac{2 \times |RE|}{|V'| \times (|V'| - 1)} \tag{6}$$

If $d_r(SN) \geq \rho$, SN is called a densely and r-reliably linked sub-network, where ρ is a given threshold of r-reliable density.

2.5. Finding protein cores

According to core-attachment structure, our method first finds protein cores. To find a protein core, our method initializes a protein core by seeding a protein. To add the jointly co-localized, jointly co-expressed, densely and r-reliably linked proteins into a protein core, our method dynamically constructs the jointly co-localized, jointly co-expressed, and densely r-reliable neighborhood of a protein core.

For a protein core PC, let $N_r(PC)=(\bigcup_{v\in PC} N_r(v))$-$PC$ denote the r-reliable neighborhood of PC. Thus, the jointly co-localized, jointly co-expressed, and densely r-reliable neighborhood of PC, $N_{led}(PC)$, is defined by formula (7).

$$N_{led}(PC)=\{u \mid JC(PC\cup\{u\})>0,\ pcc(JGEP_{PC}, gep_u)\geq\delta,\ d_r(PC\cup\{u\})\geq\rho,\text{and } u\in N_r(PC)\} \quad (7)$$

For the added node u, let $T_1(PC)=\{w \mid w\in V\text{-}(PC\cup N_{led}(PC)),\ W(w,u)\geq r,\ JC(PC\cup\{w\})>0,\ pcc(JGEP_{PC}, gep_w)\geq\delta$, and $d_r(PC\cup\{w\})\geq\rho\}$, and $T_2(PC)=\{w \mid w\in N_{led}(PC),(W(w,u)<r$ or $JC(PC\cup\{w\})=0$ or $pcc(JGEP_{PC}, gep_w)<\delta$ or $d_r(PC\cup\{w\})<\rho)\}$. After node u is added to PC, the proteins in $T_1(PC)$ are first added to $N_{led}(PC)$, and the proteins in $T_2(PC)$ are then removed from $N_{led}(PC)$. As a result, $N_{led}(PC)$ is updated by formula (8).

$$N_{led}(PC)\leftarrow(N_{led}(PC)\cup T_1(PC))\text{-}T_2(PC) \quad (8)$$

Our method expands a protein core PC by adding nodes in $N_{led}(PC)$ till $N_{led}(PC)$ becomes empty.

Each protein core is initialized by seeding a protein which does not belong to any found protein cores, and is expanded by adding proteins in $N_{led}(PC)$. Also, the added proteins do not belong to any found protein cores. Following this way of finding protein cores, any two protein cores are not allowed to be overlapped with each other.

2.6. Functional similarity criterion

To add the proteins with similar function to a protein core, now we discuss how to judge whether a protein is functionally similar to the protein core.

Given a protein core PC and node $u\in N_{led}(PC)$, the CC-based minimal similarity $CC(PC,u)$, the MF-based minimal similarity $MF(PC,u)$, and the BP-based minimal similarity $BP(PC,u)$ between PC and u are defined by formulas (9), (10), and (11) respectively.

$$CC(PC,u)=\min\{sim_{cc}(u,v) \mid W(u,v)\geq r, v\in PC\} \quad (9)$$

$$MF(PC,u)=\min\{sim_{mf}(u,v) \mid W(u,v)\geq r, v\in PC\} \quad (10)$$

$$BP(PC,u)=\min\{sim_{bp}(u,v) \mid W(u,v)\geq r, v\in PC\} \quad (11)$$

The boolean variables *bcc*, *bmf*, and *bbp* in formulas (12), (13), and (14) determine whether $CC(PC,u)$, $MF(PC,u)$, and $BP(PC,u)$ have reached their specified thresholds σ, ω, and θ respectively.

$$bcc=\begin{cases} true & , \quad CC(PC,u)\geq\sigma \\ false & , \quad otherwise \end{cases} \quad (12)$$

$$bmf=\begin{cases} true & , \quad MF(PC,u)\geq\omega \\ false & , \quad otherwise \end{cases} \quad (13)$$

$$bbp=\begin{cases} true & , \quad BP(PC,u)\geq\theta \\ false & , \quad otherwise \end{cases} \quad (14)$$

If at least 2 out of 3 boolean variables *bcc*, *bmf*, and *bbp* are "true" at the same time, the value of $B(PC,u)$ in formula (15) will become "true". It means that the protein node u is sufficiently similar to the protein core PC when at least 2 out of 3 boolean variables *bcc*, *bmf*, and *bbp* are "true" at the same time. The conditional expression "$B(PC,u)$=true" is used to denote functional similarity criterion.

$$B(PC,u)=(bcc\wedge bmf\vee bcc\wedge bbp\vee bmf\wedge bbp) \quad (15)$$

2.7. Adding attachment proteins

After finding all protein cores in static PPI networks, our method adds the attachment proteins around each protein core PC in the way similar to finding protein core to generate candidate protein complexes CPC.

Given a protein core PC and a protein node $u\in N_r(PC)$, $RC(PC, u)=\frac{|\{v \mid W(u,v)\geq r, v\in PC\}|}{|PC|}$ is defined as the r-reliable connectivity between PC and u. If $RC(PC, u)\geq\eta$, we will term that there is a adequately r-reliable connectivity between PC and u,where η is a given threshold of r-reliable connectivity. Thus, we define $N(PC)=\{u \mid JC(PC\cup\{u\})>0,\ JQ(PC\cup\{u\})\geq\gamma,\ RC(PC,\ u)\geq\eta,$ and $u\in N_r(PC)\}$ as the jointly co-localized, jointly co-expressed, and adequately r-reliable neighborhood of PC.

For a protein core *PC*, the corresponding *N*(*PC*) is first constructed. Subsequently, according to the aforementioned functional similarity criterion, attachment proteins in the *N*(*PC*) are added to the *PC* till *N*(*PC*) becomes empty. Finally, the *PC* with the added attachment proteins becomes a *CPC*. For all protein cores, this procedure is repeated to produce all candidate protein complexes. Regardless of whether the added attachment proteins belong to any produced candidate protein complexes, the identified candidate protein complexes are allowed to overlap with each other.

2.8. Algorithm

In this subsection, we describe in detail our method ICJointLE (Identifying protein Complexes with the features of Joint co-Localization and joint co-Expression). **Fig.2** shows the flow-chart ofICJointLE.

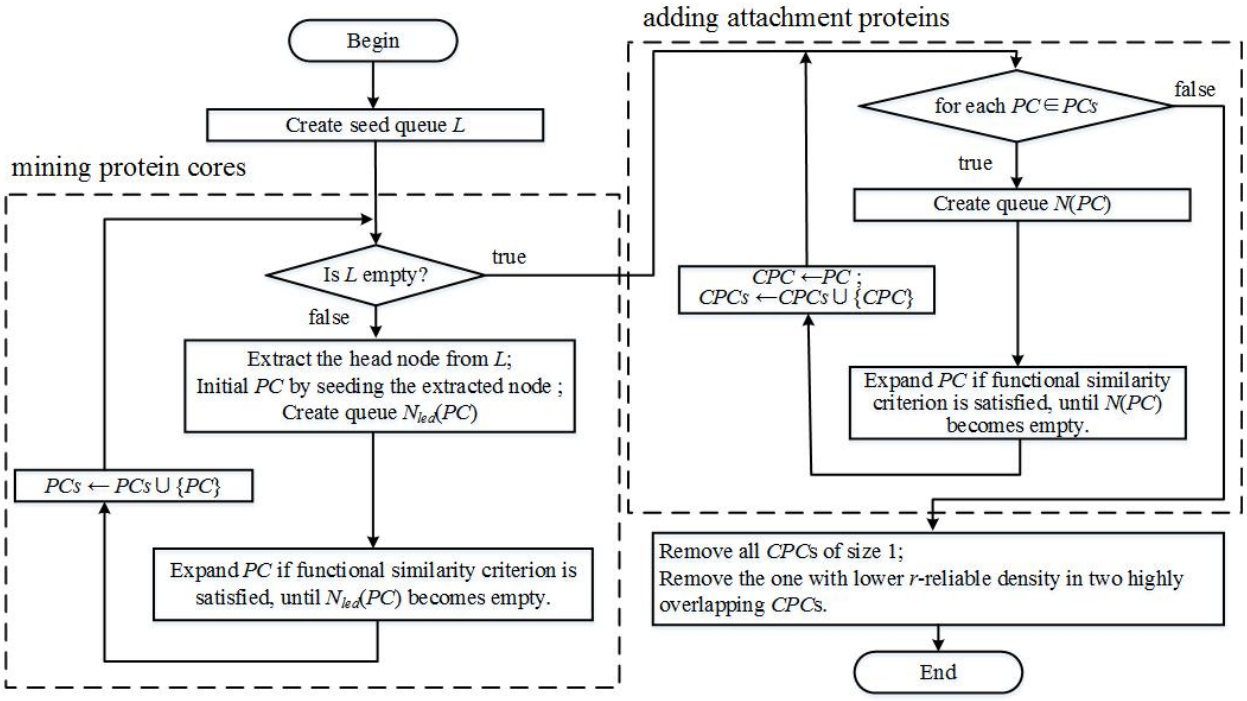


Fig.2. Flow-chart of ICJointLE.

Algorithm 1 is the pseudo-code of our method ICJointLE.

Algorithm 1: ICJointLE

Input: Three GO-based protein similarity matrices SM_{cc}, SM_{mf}, and SM_{bp}; reliability-marked PPI data set *GW*, protein localization matrix M_{locus}, gene

expression data matrix M_{ge}.

Output: Complexes set *CPC*s.

Begin

1. Create priority queue L in non-descending order of $deg_r(v)$ for each $v \in V$;
2. $PC \leftarrow \Phi$; $N_{led}(PC) \leftarrow \Phi$; $PCs \leftarrow \Phi$;
3. **while** (L is not empty) **do**
4. $seed \leftarrow$the first node of L;
5. $PC \leftarrow PC \cup \{seed\}$;
6. Create priority queue $N_{led}(PC)$ in non-ascending order of $pcc(JGEP_{PC}, gep_u)$, $u \in N_{led}(PC)$;
7. **while** ($N_{led}(PC)$ is not empty) **do**
8. $v \leftarrow$the first node of $N_{led}(PC)$;
9. **if** ($B(PC, v)$ is true) **then**
10. $PC \leftarrow PC \cup \{v\}$;
11. Update $N_{led}(PC)$;
12. **end if**
13. **end while**
14. $PCs \leftarrow PCs \cup \{PC\}$;
15. $PC \leftarrow \Phi$;
16. **end while**
17. $CPCs \leftarrow \Phi$;
18. **for each** $PC \in PCs$ **do**
19. Create priority queue $N(PC)$ in non-ascending order of $JQ(PC \cup \{u\})$, $u \in N(PC)$;
20. **while** ($N(PC)$ is not empty) **do**
21. $v \leftarrow$the first node of $N(PC)$;
22. **if** ($B(PC, v)$ is true) **then**
23. $PC \leftarrow PC \cup \{v\}$;
24. Update $N(PC)$;
25. **end if**
26. **end while**
27. $CPC \leftarrow PC$;
28. *CPC*s$\leftarrow$*CPC*s$\cup \{CPC\}$;
29. **end for**
30. Remove all *CPC*s of size 1 from *CPC*s;
31. Rearrange *CPC*s in non-ascending order of $d_r(CPC)$, $CPC \in$ *CPC*s;
32. Remove the one with lower r-reliable density in two highly overlapping *CPCs*.

End.

Algorithm ICJointLE includes three main stages: finding protein cores, adding attachment proteins, and filtering candidate protein complexes. After creating seed priority queue (line 1), ICJointLE enters the first stage(lines 2-16). In first stage, ICJointLE finds all protein cores with the characteristics of joint co-localization, joint co-expression, densely *r*-reliable link, and biologically functional homogeneity. Firstly, ICJointLE selects a seed to initialize a protein core *PC* (lines 4-5). Secondly, ICJointLE extracts the proteins from $N_{led}(PC)$ and adds the proteins, which satisfy the functional similarity criterion, into *PC* until $N_{led}(PC)$ becomes empty (lines 7-13). This extracting-adding procedure is repeated until all protein cores are found. In this procedure, any two *PC*s found by ICJointLE are not allowed to overlap with each other.

In the second stage, ICJointLE adds attachment proteins to each *PC*. First, ICJointLE selects a *PC* and creates the corresponding neighborhood $N(PC)$ (line 19). Then ICJointLE extracts the proteins from $N(PC)$ and adds the proteins satisfying the functional similarity criterion to the *PC* by the expanding strategy (lines 20-26). This extracting-expanding procedure is repeated until $N(PC)$ becomes empty. In the second stage, the added attachment proteins may be or not be the proteins belonging to other *CPC*s. It indicates that any two *CPC*s can overlap with each other.

In the final stage, ICJointLE filters *CPC*s. First, ICJointLE eliminates the *CPC*s containing only one protein (line 30), and sorts the remainder *CPC*s in non-ascending order of *r*-reliable density (line 31). Then ICJointLE removes the one with lower *r*-reliable density in any two *CPC*s whose overlapping score is not less than α (line 32), where α=0.8 [19]. Finally, ICJointLE outputs the final remaining *CPC*s as the resulting protein complexes.

3. Results

3.1. Evaluation metrics

There are two kinds of evaluation metrics to assess the quality of identified complexes. One is the statistical matching based metrics. The other is the biological relevance based metrics.

Here some notations are introduced to represent the statistical matching based metrics. Symbol *ic* denotes an identified complex, V_{ic} is the set of proteins in *ic*, *kc* denotes a known complex, V_{kc} represents the set of proteins in *kc*, *IC* is a set of identified complexes and $m=|IC|$, and *KC* is a set of known complexes and $n=|KC|$.

3.1.1. Statistical matching based metrics

The overlapping score between identified complex *ic* and known complex *kc*, $OS(ic, kc)$, is computed by the following formula:

$$OS(ic,kc)=\frac{|V_{ic}\bigcap V_{kc}|^2}{|V_{ic}|\times|V_{kc}|} \tag{16}$$

If $OS(ic, kc)\geq\lambda$, *ic* and *kc* are matched with each other, where λ usually is set to 0.2 [17,22].

Let N_{ci} denote the number of identified complexes which match with at least one known complex in *KC*, and N_{ck} denote the number of known complexes which match with at least one identified complex in *IC*. That is,

$$N_{ci}=|\{ic \mid ic\in IC, \exists kc\in KC, OS(ic, kc)\geq\lambda\}| \tag{17}$$

$$N_{ck}=|\{kc \mid kc\in KC, \exists ic\in IC, OS(ic, kc)\geq\lambda\}| \tag{18}$$

Precision (*prec*), Recall (*rec*), and F-measure (*fm*) are used to evaluate the quality of the identified complexes and are defined as follows [12].

$$prec=\frac{N_{ci}}{|IC|}=\frac{N_{ci}}{m} \tag{19}$$

$$rec=\frac{N_{ck}}{|KC|}=\frac{N_{ck}}{n} \tag{20}$$

$$fm=\frac{2\times prec\times rec}{prec+rec} \tag{21}$$

FRAC is the fraction of matched complexes, which calculates the percentage of known complexes that are matched with identified complexes [19]. In fact, *FRAC* is equal to *rec*.

The maximum matching ratio (*MMR*) [19] is based on a maximal one-to-one mapping between identified complex and known complex. And *MMR* is calculated by formula (22).

$$MMR=\frac{\sum_{i=1}^{n}\max\{OS(kc_i,ic_j)\mid j=1,...,m\}}{n} \tag{22}$$

Where kc_i is the i-th known complex, and ic_j is the j-th identified complex.

Let n_i denote the number of proteins in the i-th known complex, t_{ij} denote the number of common proteins between the i-th known complex and the j-th identified complex, i =1,2,…,n and j=1,2,…,m. Sensitivity (*Sn*), positive predictive value (*PPV*), and the geometric mean of *Sn* and *PPV* (*Acc*) which measures the accuracy of identification method, are computed by formulas (23-25) respectively [53].

$$Sn=\frac{\sum_{i=1}^{n}\max\{t_{ij}\mid j=1,2,...,m\}}{\sum_{i=1}^{n}n_i} \tag{23}$$

$$PPV=\frac{\sum_{j=1}^{m}\max\{t_{ij}\mid i=1,2,...,n\}}{\sum_{j=1}^{m}\sum_{i=1}^{n}t_{ij}} \tag{24}$$

$$Acc=\sqrt{Sn\times PPV} \tag{25}$$

To take *FRAC*, *Acc*, and *MMR* into consideration at the same time, we used the comprehensive score *FAM* by formula (26) to measure the performance of various methods [19].

$$FAM=FRAC+Acc+MMR \tag{26}$$

3.1.2. Biological relevance based metrics

The statistical matching based metrics depends on the known complexes. But the known complexes are generally incomplete [54]. Although an identified complex does not match with any known complexes, it may be an uncharacterized but valid complex. A complex tends to be responsible for a specific biological function or molecular process[55]. Hence, it is necessary to perform over-expression score based biological relevance evaluation.

The GO-based over-expression analysis on biological process and molecular function can be used to reveal functional homogeneity of proteins in a complex to some extent [19]. Let N_s be the total number of proteins and K be the total number of the proteins annotated by term X in PPI network. For a given complex containing n_s proteins, if there are k_s term X-annotated proteins in this complex, the *p-value* of this complex is computed as follows [56]:

$$p = 1 - \sum_{i=0}^{k_s-1} \frac{\binom{N_s - K}{n_s - i}\binom{K}{i}}{\binom{N_s}{n_s}} = \sum_{i=k_s}^{n_s} \frac{\binom{N_s - K}{n_s - i}\binom{K}{i}}{\binom{N_s}{n_s}} \tag{27}$$

If $p<\psi$, we call that the term X-annotated proteins enrich the complex at ψ-level, where ψ is a given threshold. If the term X-annotated proteins enrich a complex at the level of ψ=0.01, this complex will have significantly biological function and be called significant complex [19]. The over-expression score of a set of complexes is generally the proportion of the significant complexes enriching the proteins annotated at least one functional term[19]. We used the software GO::TermFinder [57] to calculate the *p*-value of an identified complex.

3.2. Experiment materials

S. cerevisiae as a model organism has been well studied. A great number of biological data on S. cerevisiae have been produced. Hence we used the S. cerevisiae data sets including protein localization data and gene expression data to conduct the experiments. We selected six yeast PPI data sets to conduct the comparison experiment. The first yeast PPI data set is downloaded from the STRING database V10 version [52]. This yeast PPI data set consists of 6418 proteins and 939998 interactions with reliability score, and it is also used as the scoring data set. The second yeast PPI data set, which consists of 5811 proteins and 256516 interactions, is downloaded from the BioGrid database 3.4.128 version [58]. The third yeast PPI data set, which contains 5022 proteins and 22381 interactions, is downloaded from the DIP database with the release

date 2015/07/01[59]. The other three PPI data sets Uetz [7], Ito [8], and Yu [60] are the yeast binary interactome derived by Y2H. Uetz data set contains 910 proteins and 823 interactions. Ito data set is composed of 765 proteins and 733 interactions. Yu data set is comprised of 1203 proteins and 1610 interactions. Three data sets Uetz, Ito, and Yu can be extracted in file interaction_data.tab downloaded from https://downloads.yeastgenome.org/curation/literature/ respectively. Apparently, the first three PPI data sets correspond to dense PPI networks while the corresponding PPI networks of the three other PPI data sets are sparse.

The known complex set CYC2008 is obtained from http://wodaklab.org/cyc2008/[43], which comprises 408 manually curated heterometric protein complexes.The gene expression data [61] are obtained from http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2267. We extracted the gene expression data from the file GDS2267_full.soft.GSE3431is downloaded from ftp://ftp.ncb.nlm.nih.gov/geo/series/GSE3nnn/GSE3431/soft/, which contains not only yeast affymetrix gene expression data over three successive metabolic cycles but also three kinds of GO term annotations of expressed genes. The yeast's protein localization data [42] are obtained from http://yeastgfp.yeastgenome.org. We noted that a few proteins in CYC2008 and a PPI data set have no protein localization data available. In order to accurately identify as many protein complexes in CYC2008 as possible, for proteins without localization data, we set their localization vectors to all "1" to calculate the joint co-localization count of the protein group containing these proteins. By doing so, our method ICJointLE is able to identify the protein complexes containing proteins without localization data in CYC2008.

3.3. Results

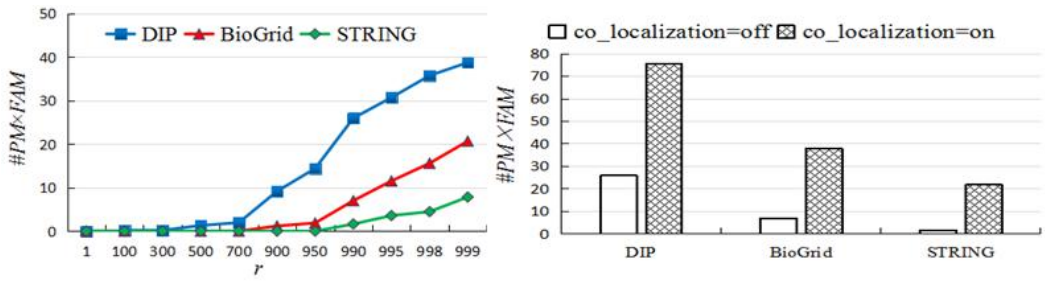
Firstly, we conducted experiments to test the effect of threshold *r* and the co-localization constraint on the quality of complexes identified by ICJointLE. Secondly, we presented two examples to illustrate the procedure of identifying complexes using ICJointLE. Finally, we compared the performance of ICJointLE and the existing methods.

By analyzing the experimental results, we found that the *r*-reliable density of a protein core declines with the increasing number of proteins in a protein core. Thus, the threshold of *r*-reliable density ρ is calculated by $e^{-\mu c}$, namely $\rho=e^{-\mu c}$, where c is the number of proteins in a protein core and the decaying coefficient μ controls the declining rate of ρ with the increase of c.

3.3.1. The effect of threshold r and co-localization constraint

In order to evaluate the number of precisely identified complexes, we used *#PM* to denote the number of identified complexes that are matched exactly with known complexes in CYC2008. Furthermore we adopted the product *#PM×FAM* to comprehensively estimate the quality of identified complexes.

For the setting of μ=0, δ=0, σ=0, ω=0, θ=0, γ=0, and η=1, we conducted experiments to investigate the influence of threshold *r* and the co-localization constraint on the value of *#PM×FAM* produced by ICJointLE on the three data sets STRING, BioGrid, and DIP respectively. The experimental results are shown in **Fig.3**.



(a) (b)

Fig.3. The values of *#PM×FAM* for different value of *r* and different setting of *co_localization* variable. (a) The variation curves between *#PM×FAM* and *r*. (b) Comparison of *#PM×FAM* between *co_localization*=off and *co_localization*=on.

From **Fig.3**(a), we can see that, for the setting of μ=0, δ=0, σ=0, ω=0, θ=0, γ=0, η=1, and *co_localization* =off, the value of *#PM×FAM* increases gradually as threshold *r* increases. In order to get a non-zero value of *#PM×FAM*, we made threshold *r* larger than or equal to 500, 900, and 990 for three data sets DIP, BioGrid, and STRING respectively. When the threshold *r* is set to 990 or larger, the value of *#PM×FAM* is always larger than zero for all three data sets.

co_localization is an on-off variable. In ICJointLE, If the joint co-localization criterion functions, *co_localization*=on; otherwise *co_localization*=off. As can be seen in **Fig.3** (b), for the setting of r=990, μ=0, δ=0, σ=0, ω=0, θ=0, γ=0, and η=1, the protein complexes identified by ICJointLE have higher quality under *co_localization*=on than under *co_localization*=off. For the DIP data set, if *co_localization*=off, the value of #*PM*×*FAM* is less than 30; otherwise, exceeds 70. For the BioGrid data set, if *co_localization*=off, the value of #*PM*×*FAM* is less than 10; otherwise, exceeds 30. And for the STRING data set, if *co_localization*=off, the value of #*PM*×*FAM* is less than 5; otherwise, exceeds 20. Hence, *co_localization* is set to "on".

3.3.2. Identification results

We determined the value of threshold r by the experiment in the previous subsection. For three given Y2H PPI data sets, due to their sparse interactions, we set threshold r=1. Meanwhile, we experimentally determined the values of other seven thresholds μ, δ, σ, ω, θ, γ, and η. The setting of eight thresholds for six data sets are shown in **Table 3**.

Table 3

The setting of eight thresholds for ICJointLE on the six data sets

data sets	r	δ	μ	σ	ω	θ	γ	η
STRING	999	0.3	0.08	0.7	0.75	0.3	0.01	0.9
BioGrid	999	0.3	0.1	0.7	0.75	0.3	0.01	0.7
DIP	990	0.3	0.4	0.6	0.8	0.1	0.01	0.7
Uetz	1	0.3	0.4	0.8	0.3	0.2	0.01	0.6
Ito	1	0.3	0.4	0.7	0.3	0.2	0.01	0.6
Yu	1	0.3	0.4	0.7	0.4	0.3	0.01	0.6

In the following, we give two examples to demonstrate how ICJointLE identifies complexes on the DIP data set. The first example shown in **Fig.4** is to illustrate how to identify the tRNA-intron endonuclease complex by ICJointLE.

As illustrated in **Fig.4 (b)**, ICJointLE uses YAR008W to initialize protein core (*PC*), and $N_{led}(PC)$= {YPL083C,YLR105C}. Let u denote YMR059W and v denote YBL051C, because $pcc(JGEP_{PC}, gep_u)<\delta$ and

$pcc(JGEP_{PC},gep_v)<\delta$, YMR059W and YBL051C are not added to $N_{led}(PC)$. **Fig.4 (c)** shows that YPL083C and YLR105C are successively added to PC. Since $pcc(JGEP_{PC}, gep_u)<\delta$,YMR059W is not added to $N_{led}(PC)$ during identifying protein core. In **Fig.4 (d)**, because $JC(PC \cup \{YBL051C\})=0$, YBL051C is not inserted into $N(PC)$. From **Fig.4 (e)** we can see that after YMR059C is added to PC, $N(PC)$ becomes empty. Hence, the PC, namely tRNA-intron endonuclease complex, is the complex identified by ICJointLE.

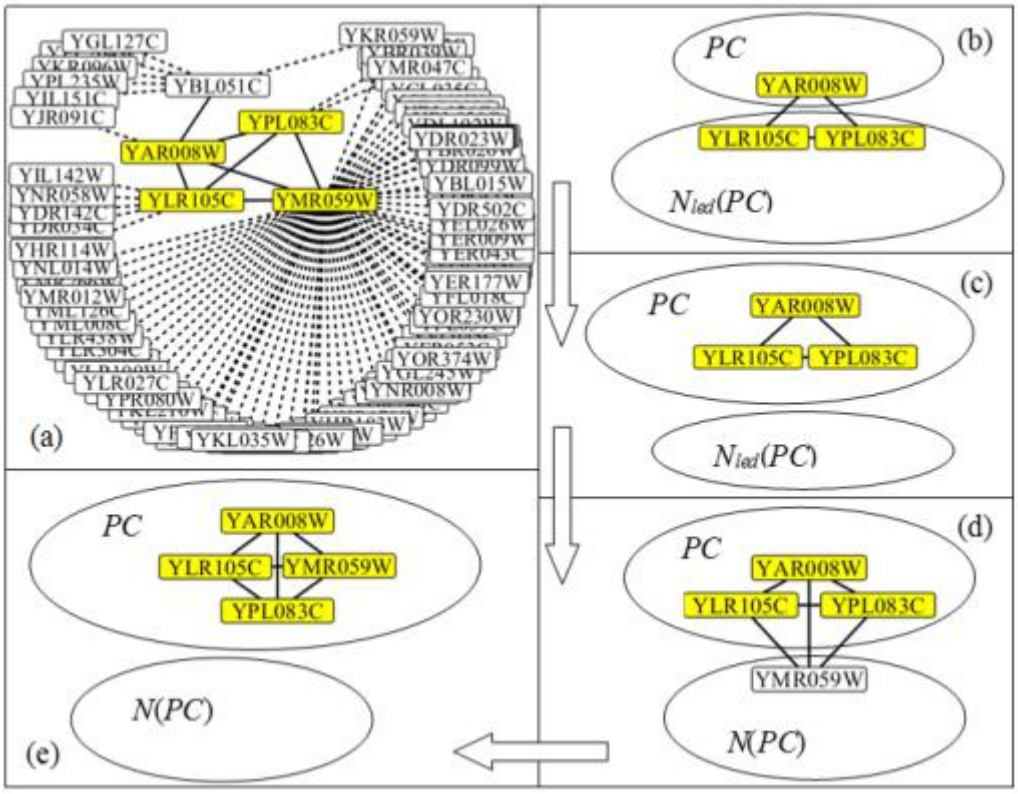


Fig.4. The process of identifying the tRNA-intron endonuclease complex in ICJointLE. (a) A sub-graph including the tRNA-intron endonuclease complex and its neighborhood in DIP network, where the tRNA-intron endonuclease complex is comprised of four yellow-colored proteins. Dash line denotes the edge with score$<r$, and solid line denotes the edge with score$\geq r$. (b) After YAR008W is seeded into PC. Then PC={YAR008W}. Thus $N_{led}(PC)$ is composed of YPL083C and YLR105C. Let u denote YMR059W and v denote YBL051C, because $pcc(JGEP_{PC},gep_u)<\delta$ and $pcc(JGEP_{PC},gep_v)<\delta$, YMR059W and YBL051C are not contained in $N_{led}(PC)$. (c) YPL083C and YLR105C are successively added to PC. Because $pcc(JGEP_{PC},gep_u)<\delta$, YMR059W is still not inserted into $N_{led}(PC)$ during the process of identifying protein core. (d) Because $JC(PC \cup \{YBL051C\})=0$,

YBL051C is not inserted into $N(PC)$. (e) After YMR059C is added into PC, $N(PC)$ becomes empty. Thus PC is a complex identified by ICJointLE.

The second example is shown in **Fig.5** to demonstrate how ICJointLE identifies a candidate protein complex containing only one protein. From **Fig.5 (b)** we can see that by seeding YOR281C, ICJointLE adds YOR281C into PC. Let w denote YJL152W, because $pcc(JGEP_{PC}, gep_w)<\eta$, $N_{led}(PC)$ is empty. **Fig.5 (c)** shows that YJL152W is contained in $N(PC)$. We know that YJL152W is not sufficiently similar to PC in function, i.e. $B(PC, w)$=false. So, YJL152W is not added to PC, and $N(PC)$ becomes empty. At this time, the PC, containing only YOR281C, is the complex identified by ICJointLE. Since the size of this PC is 1, this PC is discarded in the final stage.

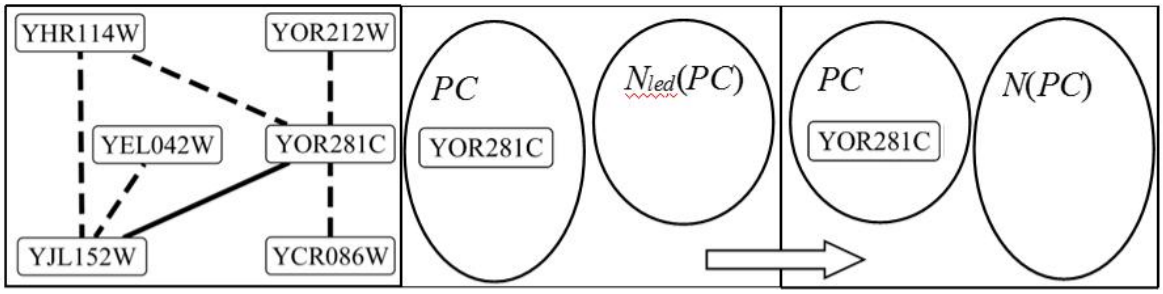


Fig.5. The process of identifying a complex containing only one protein by seed YOR281C in ICJointLE. (a) The sub-graph including YOR281C and its neighborhood in DIP network. The dash line denotes the edge with score<r, and solid line denotes the edge with score≥r. (b) After YOR281C is seeded into PC. PC={YOR281C}. Let w denote YJL152W, because $pcc(JGEP_{PC}, gep_w)<\delta$, YJL152W is not contained in $N_{led}(PC)$. $N_{led}(PC)$ is empty. (c) In adding attachment stage, YJL152W appears in $N(PC)$ once. Because it is not sufficiently similar to PC, i.e. $B(PC,w)$=false, YJL152W is not added to PC. Thus $N(PC)$ becomes empty. So PC, only containing YOR281C, becomes the identified complex of size 1, and it will be discarded.

In order to evaluate the effectiveness of ICJointLE, we compared ICJointLE with ten other existing methods PCP*, MCL [16], WPNCA [27], APcluster [15], SPICi [20], ClusterONE [19], WEC [36], RNSC [29], CORE [22], and MCODE [17]. PCP* is the extension of PCP [12], where the threshold FS is added to filter

out the interactions with reliability score lower than *FS*. In APcluster method, parameter *preference* is set to the value of node index. No parameters need to be set in both SPICi and CORE methods. **Table 4** lists the parameter setting of seven other existing methods PCP*, WPNCA, ClusterONE, WEC, RNSC, MCL, and MCODE on the six data sets.

Table 4

The setting of parameters for seven methods on the six data sets

Methods	Parameters	STRING	BioGrid	DIP	Uetz	Ito	Yu
PCP*	ICD threshold	0.6	0.6	0.6	0.6	0.6	0.6
	Minclique size	2	2	2	2	2	2
	FS	0.999	0.999	0.99	0	0	0
WPNCA	λ	0.6	0.7	0.8	0.4	0.4	0.5
	min size	2	2	2	2	2	2
ClusterONE	*d*	0.8	0.9	0.6	0.6	0.6	0.6
	min size	2	2	2	2	2	2
WEC	Balance Factor	0.8	0.9	0.9	0.8	0.8	0.8
	Edge weight	0.7	0.5	0.2	0.4	0.5	0.5
	Enrich	0.8	0.9	0.9	0.8	0.8	0.8
	Filter	0.8	0.9	0.9	0.9	0.9	0.9
RNSC	size	2	2	2	2	2	2
	density	0.2	0.2	0.2	0.2	0.2	0.2
	p-value	0.01	0.01	0.01	0.01	0.01	0.01
MCL	inflation	3	3	2.5	2.5	2.5	2.5
MCODE	Include Loops	false	false	false	false	false	false
	Degree cutoff	2	2	2	2	2	2
	Node Score	0.1	0.1	0.1	0.2	0.2	0.2
	Haircut	true	true	true	true	true	true
	Fluff	false	false	false	false	false	false
	K-Core	2	2	2	2	2	2
	Max. Depth	100	100	100	100	100	100

By analyzing the known complexes in CYC2008, we found that the number of the complexes of sizes from 2 to 6 exceeds 84% of the total in CYC2008. Therefore,it is necessary to identify complexes of small size indeed. For the six PPI data sets, **Tables 5** and **6** list the distribution of the size of the complexes identified exactly by eleven methods respectively.

Table 5

Distribution of the size of the exactly identified complexes on three PPI data sets STRING, BioGrid, and DIP

Data sets	Methods	Number of the exactly identified complexes of different size											Total
		size	2	3	4	5	6	7	8	9	10	12	14	
STRING	ICJointLE		92	34	10	5	3	0	2	1	0	1	1	149
	PCP*		52	11	1	5	6	1	3	0	0	1	1	81
	WPNCA		0	0	0	1	0	0	0	0	0	0	0	1
	APcluster		0	0	0	0	0	0	0	0	0	0	0	0
	SPICi		1	1	0	0	0	0	0	0	0	0	0	2
	ClusterONE		0	0	0	0	0	0	0	0	0	0	0	0
	WEC		0	0	0	0	0	0	0	0	0	0	0	0
	RNSC		0	0	0	0	0	0	0	0	0	0	0	0
	CORE		0	0	0	0	0	0	0	0	0	0	0	0
	MCL		0	0	0	0	0	0	0	0	0	0	0	0
	MCODE		0	0	0	0	0	0	0	0	0	0	0	0
BioGrid	ICJointLE		94	34	8	4	3	0	0	1	0	1	0	145
	PCP*		54	12	1	6	6	1	3	0	1	1	0	85
	WPNCA		13	3	3	1	5	0	0	0	0	0	0	25
	APcluster		1	0	1	1	1	0	0	0	0	0	0	4
	SPICi		18	3	6	2	4	0	0	0	0	0	0	33
	ClusterONE		5	1	0	0	1	0	0	0	0	0	0	7
	WEC		0	0	0	0	0	0	0	0	0	0	0	0
	RNSC		2	1	0	0	0	0	0	0	0	0	0	3
	CORE		0	0	0	0	0	0	0	0	0	0	0	0
	MCL		1	0	0	0	0	0	0	0	0	0	0	1
	MCODE		0	0	0	0	0	0	0	0	0	0	0	0
DIP	ICJointLE		97	22	8	1	0	1	1	1	0	1	0	132
	PCP*		62	9	7	1	2	1	1	0	0	0	0	83
	WPNCA		8	3	6	2	1	0	1	1	1	0	0	23
	APcluster		20	3	2	2	0	0	0	0	0	0	0	27
	SPICi		14	8	4	1	0	1	0	0	1	0	0	29
	ClusterONE		4	3	1	1	0	0	0	0	1	0	0	10
	WEC		0	14	11	5	0	0	8	0	7	0	0	45
	RNSC		14	8	4	2	0	1	1	1	0	0	0	31
	CORE		16	6	4	2	1	0	1	1	0	0	0	31
	MCL		11	3	1	1	1	0	1	0	0	0	0	18
	MCODE		0	1	2	0	0	0	0	0	0	0	0	3

Note: The number of complexes of sizes from 2 to 6 and the total number of the complexes identified exactly by ICJointLE are marked in boxed presentation.

Table 6

Distribution of the size of the exactly identified complexes on three PPI data sets Uetz, Ito, and Yu

Data sets	Methods	Number of the exactly identified complexes of different size											Total
		size	2	3	4	5	6	7	8	9	10	12	14	
Uetz	ICJointLE		18	1	0	0	0	0	0	0	0	0	0	19
	PCP*		17	0	0	0	0	0	0	0	0	0	0	17
	WPNCA		5	1	0	0	0	0	0	0	0	0	0	6
	APcluster		11	0	0	0	0	0	0	0	0	0	0	11

	SPICi	5	0	0	0	0	0	0	0	0	0	0	5
	ClusterONE	5	1	0	0	0	0	0	0	0	0	0	6
	WEC	0	2	0	0	0	0	0	0	0	0	0	5
	RNSC	7	2	0	0	0	0	0	0	0	0	0	9
	CORE	8	2	0	0	0	0	0	0	0	0	0	10
	MCL	8	1	0	0	0	0	0	0	0	0	0	9
	MCODE	0	0	0	0	0	0	0	0	0	0	0	0
	ICJointLE	14	2	0	0	0	0	0	0	0	0	0	16
	PCP*	12	1	0	0	0	0	0	0	0	0	0	13
	WPNCA	5	4	1	1	0	0	0	0	0	0	0	11
	APcluster	6	3	0	0	0	0	0	0	0	0	0	9
	SPICi	3	1	0	0	0	0	0	0	0	0	0	4
Ito	ClusterONE	5	4	0	0	0	0	0	0	0	0	0	9
	WEC	0	5	0	0	0	0	0	0	0	0	0	5
	RNSC	7	5	0	0	0	0	0	0	0	0	0	12
	CORE	7	5	1	0	0	0	0	0	0	0	0	13
	MCL	6	5	1	0	0	0	0	0	0	0	0	12
	MCODE	1	0	0	0	0	0	0	0	0	0	0	1
	ICJointLE	30	2	0	0	0	0	0	0	0	0	0	32
	PCP*	28	2	0	0	0	0	0	0	0	0	0	30
	WPNCA	14	3	0	0	0	0	0	0	0	0	0	17
	APcluster	20	3	0	0	0	0	0	0	0	0	0	23
	SPICi	7	2	0	0	0	0	0	0	0	0	0	9
Yu	ClusterONE	16	3	0	0	0	0	0	0	0	0	0	19
	WEC	0	5	0	0	0	0	0	0	0	0	0	5
	RNSC	19	4	0	0	0	0	0	0	0	0	0	23
	CORE	20	3	0	0	0	0	0	0	0	0	0	23
	MCL	19	3	0	0	0	0	0	0	0	0	0	22
	MCODE	0	0	0	0	0	0	0	0	0	0	0	0

Note: The number of complexes of sizes from 2 to 6 and the total number of the complexes identified exactly by ICJointLE are marked in boxed presentation.

We can see from **Tables 5** and **6** that ICJointLE can exactly identify 149, 145, 132, 19, 16, and 32 complexes on six data sets STRING, BioGrid, DIP, Uetz, Ito, and Yu respectively. PCP* can exactly identify 81, 85,83, 17, 13, and 30 complexes on data sets STRING, BioGrid, DIP, Uetz, Ito, and Yu respectively by filtering out interactions with low reliability. It indicates that scoring PPI data with reliability and filtering out interactions with low reliability are helpful for exactly identifying more complexes. We can also find that APcluster, ClusterONE,WEC, RNSC, CORE, MCL and MCODE are unable to exactly identify any complexes on data set STRING. Furthermore, WEC, CORE, and MCODE fail to exactly identify any complexes on data

set BioGrid, and MCODE does not exactly identify any complexes on data sets Uetz and Yu.In summary, our method ICJointLE can exactly identify more complexes than ten other existing methods. Furthermore, the results in **Tables 5** and **6** show that our method ICJointLE is capable of exactly identifying complexes with small size.

Tables 7 and **8** show the values of *#PM*, *prec*, *rec*, *fm*, *Sn*, *PPV*, *Acc*, *MMR*, and *FAM* of the identified complexes on the six data sets for eleven methods respectively.

Table 7

Comparison of identified results for eleven methods on three PPI data sets STRING, BioGrid, and DIP.

Data sets	Methods	*#PM*	*prec*	*rec*	*fm*	*Sn*	*PPV*	*Acc*	*MMR*	*FAM*
STRING	ICJointLE	149	0.47	0.91	0.62	0.72	0.93	0.82	0.67	2.40
	PCP*	81	0.36	0.84	0.50	0.70	0.86	0.77	0.53	2.14
	WPNCA	1	0.19	0.23	0.21	0.77	0.49	0.61	0.14	0.98
	APcluster	0	0.17	0.20	0.18	0.73	0.52	0.61	0.13	0.94
	SPICi	2	0.12	0.19	0.15	0.79	0.45	0.59	0.13	0.91
	ClusterONE	0	0.08	0.14	0.10	0.86	0.36	0.55	0.10	0.79
	WEC	0	0.02	0.02	0.02	0.96	0.11	0.32	0.03	0.37
	RNSC	0	0.08	0.03	0.04	0.78	0.25	0.44	0.03	0.50
	CORE	0	0.02	0.04	0.02	0.80	0.19	0.39	0.06	0.49
	MCL	0	0.02	0.01	0.01	0.96	0.07	0.26	0.01	0.28
	MCODE	0	0	0	N/A	0.40	0.14	0.24	0.01	0.25
BioGrid	ICJointLE	145	0.46	0.89	0.61	0.67	0.92	0.78	0.66	2.33
	PCP*	85	0.37	0.83	0.52	0.63	0.92	0.76	0.53	2.12
	WPNCA	25	0.36	0.61	0.45	0.89	0.53	0.68	0.36	1.65
	APcluster	4	0.17	0.39	0.24	0.62	0.68	0.65	0.22	1.26
	SPICi	33	0.22	0.48	0.30	0.73	0.64	0.68	0.31	1.47
	ClusterONE	7	0.29	0.48	0.36	0.72	0.61	0.66	0.25	1.39
	WEC	0	0.15	0.17	0.16	0.92	0.13	0.34	0.12	0.63
	RNSC	3	0.23	0.26	0.24	0.68	0.54	0.61	0.16	1.03
	CORE	0	0.03	0.16	0.06	0.71	0.25	0.43	0.13	0.72
	MCL	1	0.12	0.12	0.12	0.43	0.31	0.36	0.07	0.55
	MCODE	0	0.08	0.02	0.03	0.31	0.17	0.23	0.02	0.27
DIP	ICJointLE	132	0.57	0.83	0.67	0.54	0.94	0.72	0.59	2.14
	PCP*	83	0.47	0.78	0.59	0.47	0.95	0.67	0.48	1.93
	WPNCA	23	0.66	0.45	0.53	0.56	0.75	0.65	0.27	1.37
	APcluster	27	0.22	0.59	0.32	0.48	0.77	0.60	0.32	1.52
	SPICi	29	0.43	0.60	0.5	0.54	0.86	0.68	0.36	1.64
	ClusterONE	10	0.27	0.39	0.32	0.40	0.83	0.56	0.23	1.19
	WEC	45	0.53	0.54	0.53	0.65	0.57	0.61	0.32	1.47
	RNSC	31	0.45	0.58	0.50	0.47	0.88	0.64	0.34	1.56

	CORE	31	0.16	0.64	0.25	0.60	0.68	0.64	0.37	1.65
	MCL	18	0.16	0.53	0.25	0.48	0.85	0.64	0.29	1.46
	MCODE	3	0.42	0.08	0.13	0.23	0.49	0.33	0.07	0.48

Note: The best performers for the relative item are marked in boxed presentation.

Table 8

Comparison of identified results for eleven methods on three PPI data sets Uetz, Ito, and Yu.

Data sets	Methods	*#PM*	*prec*	*rec*	*fm*	*Sn*	*PPV*	*Acc*	*MMR*	*FAM*
	ICJointLE	19	0.41	0.22	0.29	0.12	0.97	0.34	0.13	0.69
	PCP*	17	0.26	0.24	0.25	0.14	0.95	0.36	0.14	0.74
	WPNCA	6	0.28	0.16	0.20	0.17	0.69	0.34	0.10	0.60
	APcluster	11	0.24	0.20	0.22	0.15	0.79	0.35	0.12	0.67
	SPICi	5	0.36	0.13	0.19	0.09	0.95	0.29	0.07	0.49
Uetz	ClusterONE	6	0.24	0.12	0.16	0.10	0.87	0.29	0.08	0.49
	WEC	2	0.29	0.005	0.01	0.01	0.79	0.11	0.006	0.12
	RNSC	9	0.46	0.16	0.24	0.10	0.93	0.30	0.09	0.55
	CORE	10	0.21	0.18	0.19	0.16	0.86	0.36	0.12	0.67
	MCL	9	0.21	0.17	0.19	0.17	0.82	0.37	0.12	0.66
	MCODE	0	0.13	0.005	0.01	0.01	0.86	0.09	0.004	0.10
	ICJointLE	16	0.34	0.20	0.25	0.11	0.96	0.33	0.12	0.65
	PCP*	13	0.28	0.22	0.24	0.12	0.96	0.34	0.12	0.67
	WPNCA	11	0.38	0.16	0.23	0.15	0.81	0.34	0.10	0.61
	APcluster	9	0.28	0.18	0.22	0.14	0.78	0.33	0.11	0.62
	SPICi	4	0.43	0.12	0.19	0.08	0.93	0.26	0.07	0.45
Ito	ClusterONE	9	0.26	0.12	0.16	0.09	0.90	0.28	0.07	0.47
	WEC	5	0.65	0.02	0.03	0.01	0.97	0.11	0.01	0.14
	RNSC	12	0.42	0.16	0.24	0.11	0.92	0.32	0.10	0.58
	CORE	13	0.26	0.19	0.22	0.14	0.87	0.35	0.12	0.66
	MCL	12	0.26	0.18	0.21	0.15	0.81	0.35	0.11	0.64
	MCODE	1	0.47	0.02	0.03	0.01	0.93	0.11	0.01	0.14
	ICJointLE	32	0.41	0.28	0.33	0.15	0.97	0.38	0.18	0.84
	PCP*	30	0.29	0.30	0.30	0.17	0.96	0.40	0.18	0.88
	WPNCA	17	0.29	0.25	0.27	0.20	0.72	0.38	0.14	0.78
	APcluster	23	0.27	0.26	0.27	0.20	0.78	0.39	0.17	0.82
	SPICi	9	0.32	0.17	0.22	0.11	0.93	0.32	0.09	0.58
Yu	ClusterONE	19	0.30	0.18	0.22	0.12	0.92	0.34	0.12	0.63
	WEC	5	0.47	0.03	0.06	0.03	0.86	0.15	0.02	0.20
	RNSC	23	0.44	0.24	0.31	0.14	0.95	0.36	0.15	0.75
	CORE	23	0.25	0.26	0.25	0.18	0.87	0.40	0.17	0.83
	MCL	22	0.29	0.26	0.28	0.21	0.80	0.41	0.17	0.84
	MCODE	0	0.27	0.007	0.01	0.02	0.70	0.11	0.006	0.12

Note: The best performers for the relative item are marked in boxed presentation.

From **Tables 7** and **8,** we noticed that WEC obtains high value of *prec* and but low value of *rec* on data sets Uetz, Ito, and Yu. The reason is that many complexes identified by WEC match with one known complex. That is, there exists many-to-one matching between the identified complexes and known complexes. In addition, PCP* gets the highest value of *rec* on data sets Uetz, Ito, and Yu due to filtering out the interactions with lower reliability score. We can also see from **Tables 7** and **8** that with regard to *prec*, ICJointLE is inferior to RNSC on Uetz, Ito, and Yu, ICJointLE is inferior to SPICi on Ito, and ICJointLE is inferior to WPNCA on DIP, but ICJointLE performs better than other competing methods on STRING and BioGrid. Meanwhile, ICJointLE gains the highest value of *rec* on data sets STRING, BioGrid, and DIP. Furthermore, ICJointLE obtains the highest value of *fm* on all six data sets among eleven methods. It indicates that in overall, ICJointLE can accurately identify complexes in static PPI networks.

We also noticed that for the known complexes in CYC2008, the denominator item of the formula (23) is a definite value. Hence the value of *Sn* depends on the numerator item of the formula (23). The greater the numerator item of the formula (23) is, the greater the value of *Sn*. It means that the greater the number of common proteins between the identified complex and known complex is, the higher the value of *Sn*. From **Tables 7** and **8**, we can see that, with regard to *Sn*, WEC performs well on STRING, BioGrid, and DIP. Considering both *max size* and *average size* in Table 7, we found that WEC might identify a number of complexes of large size and achieves the highest value of *Sn* among eleven competing methods in dense PPI networks STRING, BioGrid, and DIP. Because these identified complexes of large size share many common proteins with known complexes, WEC obtains higher value of *Sn* than our method ICJointLE. Conversely, ICJointLE gets lower value of *Sn* because ICJointLE identifies complexes which share relatively few common proteins with known complexes. In terms of #*PM*, *rec* (*Frac*), *Acc*, *MMR*, and *FAM* (*Frac*+*Acc*+*MMR*), WEC performs poorly on STRING, BioGrid, and DIP. This implies that for STRING, BioGrid, and DIP, a lot of

complexes of large size identified by WEC share many common proteins with known complexes, but do not accurately match with known complexes.

We can see from **Tables 7** and **8** that because of identifying more complexes of small size, ICJointLE obtains larger value of *PPV* than ten other existing methods. With respect to *Acc*, ICJointLE performs well on data sets STRING, BioGrid, and DIP, but ICJointLE performs poorly on data sets Uetz, Ito, and Yu. In addition, because ICJointLE can exactly match the most known complexes with identified complexes, it achieves higher value of *MMR* than ten other existing methods on STRING, BioGrid, and DIP.

The obtained superb performance for *rec*, *Acc*, and *MMR* in ICJointLE leads to the highest corresponding comprehensive score *FAM* on data sets STRING, BioGrid, and DIP. However, for the sparse PPI sets Ito, Uetz, and Yu, PCP* performs better than ICJointLE in terms of *rec*, *Sn*, *Acc*, *MMR*, and *FAM*, and MCL gains the highest value of *Acc* because it obtains the highest value of *Sn* and the relatively stable value of *PPV*. In addition, CORE is slightly superior to ICJointLE on data set Ito in terms of *FAM*. In summary, ICJointLE performs better than ten other existing methods on the dense PPI networks, and in most cases ICJointLE performs well but is inferior to PCP* on three sparse PPI networks.

To compare the biological significance of identified complexes, we listed the proportion of identified complexes that significantly enrich the BP term-annotated proteins on six data sets STRING, BioGrid, DIP, Uetz, Ito, and Yu respectively in **Tables 9** and **10**, where #*IC* is the total number of identified complexes, #*SC* is the number of identified complexes with significant biological function, % *of significant* denotes the percentage of identified complexes with significant biological function, *Max size* represents the maximal size of identified complexes, and *Average size* is the mean of the sizes of identified complexes. The BP-based enrichment analysis and the statistic result of significant complexes identified by ICJointLE are available in Additional file 1.

Table 9

Proportion of the complexes that significantly enrich the BP term-annotated proteins on three PPI data sets STRING, BioGrid, and DIP.

Data sets	Methods	*#IC*	*#SC*	*% of significant*	*Max size*	*Average size*	*% of significant* (*size*≤6)	*% of significant* (6<*size*<20)	*% of significant* (*size*≥20)
STRING	ICJointLE	1118	1040	93.02%	68	3.37	92.51%	100.00%	100.00%
	PCP*	857	699	81.56%	116	3.11	80.27%	100.00%	100.00%
	WPNCA	1235	895	72.47%	85	17.86	34.62%	74.46%	96.28%
	APcluster	401	326	81.30%	554	16.01	48.28%	87.77%	97.65%
	SPICi	585	338	57.78%	355	9.46	42.23%	84.72%	96.36%
	ClusterONE	269	187	69.52%	402	17.87	21.43%	74.26%	89.29%
	WEC	1484	1457	98.18%	1864	270.4	73.23%	89.54%	100.00%
	RNSC	130	130	100.00%	607	38.96	14.38%	60.81%	83.67%
	CORE	845	549	64.97%	384	40.14	10.56%	63.83%	85.95%
	MCL	161	34	21.12%	5373	37.77	13.70%	90.91%	100.00%
	MCODE	80	45	56.25%	267	59.83	35.48%	44.44%	75.00%
BioGrid	ICJointLE	1101	987	89.65%	67	3.26	88.67%	100.00%	100.00%
	PCP*	856	690	80.61%	23	2.88	79.25%	100.00%	100.00%
	WPNCA	2278	2086	91.57%	78	14.92	77.95%	99.13%	99.69%
	APcluster	756	532	70.37%	709	7.63	56.25%	92.34%	96.77%
	SPICi	760	540	71.05%	123	5.00	66.00%	100.00%	100.00%
	ClusterONE	1057	725	68.59%	94	8.48	51.67%	88.68%	100.00%
	WEC	1534	1514	98.70%	1986	250.7	74.19%	90.91%	100.00%
	RNSC	377	377	100.00%	260	9.08	26.77%	95.92%	100.00%
	CORE	2098	1204	57.39%	180	18.24	19.44%	54.99%	81.10%
	MCL	322	132	40.99%	1568	10.65	35.74%	87.50%	100.00%
	MCODE	60	42	70%	141	32.88	71.43%	70.00%	68.97%
DIP	ICJointLE	917	796	86.80%	78	2.92	86.04%	100.00%	100.00%
	PCP*	662	541	81.72%	12	2.49	81.18%	100.00%	0.00%
	WPNCA	301	277	92.03%	47	8.71	84.00%	100.00%	100.00%
	APcluster	1071	513	47.90%	739	4.62	43.28%	80.65%	77.78%
	SPICi	491	359	73.12%	24	3.82	69.37%	100.00%	100.00%
	ClusterONE	1036	452	43.63%	19	3.75	32.85%	87.30%	100.00%
	WEC	1654	1547	93.53%	126	16.8	83.40%	96.72%	98.95%
	RNSC	453	453	100.00%	40	3.69	27.86%	100.00%	100.00%
	CORE	1632	424	25.98%	79	3.46	21.69%	94.37%	92.31%
	MCL	1240	395	31.85%	59	3.63	29.10%	63.74%	46.67%
	MCODE	66	59	89.39%	70	9.17	89.58%	90.00%	87.50%

Note: For the relative item, the performers better than ICJointLE are marked in boxed presentation.

Table 10

Proportion of the complexes that significantly enrich the BP term-annotated proteins on three PPI data sets Uetz, Ito, and Yu.

Data sets	Methods	*#IC*	*#SC*	*% of significant*	*Max size*	*Average size*	*% of significant* (*size*≤6)	*% of significant* (6<*size*<20)	*% of significant* (*size*≥20)
Uetz	ICJointLE	251	145	57.77%	5	2.04	57.77%	0.00%	0.00%
	PCP*	261	142	54.41%	4	2.01	54.41%	0.00%	0.00%
	WPNCA	274	146	53.28%	18	4.39	48.68%	76.09%	0.00%
	APcluster	306	116	37.91%	141	2.85	38.00%	40.00%	0.00%
	SPICi	122	70	57.38%	8	2.27	57.02%	100.00%	0.00%
	ClusterONE	178	71	39.89%	8	2.52	39.20%	100.00%	0.00%
	WEC	21	14	66.67%	9	4	61.11%	100.00%	0.00%
	RNSC	211	211	100.00%	6	2.43	100.00%	0.00%	0.00%
	CORE	324	117	36.11%	10	2.63	35.85%	50.00%	0.00%
	MCL	301	116	38.54%	15	3.01	37.59%	63.64%	0.00%
	MCODE	8	5	62.50%	4	3.25	62.50%	0.00%	0.00%
Ito	ICJointLE	255	152	59.61%	4	2.03	59.61%	0.00%	0.00%
	PCP*	317	135	42.59%	3	2.03	42.59%	0.00%	0.00%
	WPNCA	201	152	75.62%	15	3.92	74.01%	87.50%	0.00%
	APcluster	247	140	56.68%	111	3.01	57.08%	40.00%	50.00%
	SPICi	83	57	68.67%	5	2.37	68.67%	0.00%	0.00%
	ClusterONE	159	95	59.75%	5	2.43	59.75%	0.00%	0.00%
	WEC	34	31	91.18%	5	3.3	91.18%	0.00%	0.00%
	RNSC	141	141	100.00%	5	2.44	100.00%	0.00%	0.00%
	CORE	270	137	50.74%	10	2.58	51.70%	0.00%	0.00%
	MCL	254	140	55.12%	45	3.02	55.51%	37.50%	100.00%
	MCODE	15	13	86.67%	4	3.33	86.67%	0.00%	0.00%
Yu	ICJointLE	346	220	63.58%	5	2.05	63.58%	0.00%	0.00%
	PCP*	383	201	52.48%	6	2.05	52.48%	0.00%	0.00%
	WPNCA	358	244	68.16%	22	4.69	67.14%	71.43%	100.00%
	APcluster	369	200	54.20%	169	3.16	54.24%	57.14%	0.00%
	SPICi	177	111	62.71%	6	2.36	62.71%	0.00%	0.00%
	ClusterONE	214	136	63.55%	7	2.54	63.38%	100.00%	0.00%
	WEC	60	52	86.67%	11	3.8	86.21%	100.00%	0.00%
	RNSC	194	194	100.00%	6	2.43	100.00%	0.00%	0.00%
	CORE	405	198	48.89%	22	2.61	49.75%	0.00%	0.00%
	MCL	344	194	56.40%	47	3.48	56.04%	56.25%	80.00%
	MCODE	11	10	90.91%	27	7.27	100.00%	50.00%	100.00%

Note: For the relative item, the performers better than ICJointLE are marked in boxed presentation.

As can be seen in **Tables 9** and **10**, with regard to the proportion of significant complexes, ICJointLE is inferior to RNSC. This is because the post-processing stage in RNSC filters out the partitioned clusters with *p*-value≥0.01, the proportion of significant complexes identified by RNSC reaches 100%. For the sparse PPI data sets Uetz, Ito, and Yu, both of WEC and MCODE identify fewer complexes respectively. So it is relatively easy for WEC and MCODE to obtain a high proportion of significant identified complexes. For the dense PPI data sets STRING, BioGrid, and DIP, WEC identifies a lot of significant complexes of large size. A lot of significant complexes of large size and a few complexes of small or middle size contribute to a high proportion of significant complexes. From **Tables 9** and **10** we can see that WPNCA obtains higher proportion of significant identified complexes than ICJointLE on data sets BioGrid, DIP, Ito, and Yu respectively. In addition, SPICi attains higher proportion of significant identified complexes than ICJointLE on data set Ito. For the other cases, ICJointLE performs better than other competing methods in terms of the proportion of significant identified complexes.

The *p*-value of an identified complex has close association with the size of the identified complex [22]. In order to further compare the proportion of the identified significant complexes of different size, we partitioned the identified complexes into three groups. The size of identified complexes in the first group is less than or equal to 6, the size of identified complexes in the second group is greater than 6 and less than 20, and the size of identified complexes in the last group is greater than or equal to 20. The proportion of the significant complexes of these three groups is shown respectively in **Tables 9** and **10**. We can see from **Tables 9** and **10** that ICJointLE performs poorly on the three data sets Uetz, Ito, and Yu. In particular, for all three groups of different size, ICJointLE is inferior to WEC, RNSC, and MCODE on data sets Uetz and Yu, ICJointLE performs more poorly than WPNCA on data set Yu, and ICJointLE performs worse than WPNCA, SPICi, ClusterONE, WEC, RNSC, and MCODE on data set Ito. However, for the three other data sets STRING,

BioGrid, and DIP, regarding the proportion of significant complexes of the first group of size≤6, ICJointLE outperforms ten other existing methods except for the case of MCODE on DIP, and concerning the proportion of significant complexes of two other groups of size>6, ICJointLE outperforms or performs equally as ten other competing methods.

In order to further demonstrate the effectiveness of ICJointLE, we showed six examples of the complexes identified by ICJointLE and their CC-based enrichment analysis on data sets STRING, BioGrid and DIP respectively in **Tables 11-13**. The CC-based enrichment analysis of six examples of complexes identified by ICJointLE is available in Additional file 2. In **Tables 11-13**, the first column shows the name of the known complex matched with an identified complex, the second column is the number (*#kc*) of proteins in the matched known complex, the third column displays the proteins of the identified complex, the fourth column is the overlapping score (*OS*) between the identified complex and the known complex, the fifth column shows the CC term which annotates the proteins belonging to the identified complex, and the sixth column is the *p-value* of the identified complex enriching the proteins annotated with the CC term.

Table 11

Six complexes identified by ICJointLE on STRING and their CC term enrichment analyses

Matched known complex name	*#kc*	Proteins of the identified complex	*OS*	CC annotation	
				term	*p*-value
anaphase-promoting complex	15	YBL084C, YDR118W, YDR260C, YFR036W, YGL240W, YIR025W, YKL022C, YLR102C, YLR127C, YNL172W, YOR249C	0.73	anaphase-promoting complex	6.26e-30
20S proteasome	14	YBL041W, YER012W, YER094C, YFR050C, YGL011C, YGR135W, YGR253C,**YHL030W**, YJL001W, YML092C, YMR314W, YOL038W, YOR157C, YOR362C, YPR103W	0.93	proteasome core complex	7.94e-40
DASH complex	10	YDR016C, YDR201W, YGL061C, YKL052C, YKR037C,YKR083C	0.60	DASH complex	1.09e-16
SWI/SNF complex	12	YBR289W, YDR073W, YFL049W, YHL025W, YJL176C, YNR023W, YOR290C, YPL016W, YPR034W	0.75	SWI/SNF complex	1.31e-24

SAGA complex	20	YBR081C,YBR198C,YDR145W,YDR167W,YDR392W,YDR448W,YGL066W,YGL112C,YHR099W,YLR055C,YMR223W,YMR236W,YOL148C,YPL047W,YPL254W	0.75	SAGA complex	8.18e-40
U1 snRNP complex	17	YBR119W, YDR240C, YER029C, YGR013W, YGR074W, YIL061C, YKL012W, YLR147C, YLR298C	0.53	U1 snRNP	2.04e-22

NOTE: Systematic names in boldrepresent those proteins in the complex identified by ICJointLE, but do not appear in the matched known complex in the first column.

Table 12

Six complexes identified by ICJointLE on BioGrid and their CC term enrichment analyses

Matched known complex name	*#kc*	Proteins of the identified complex	*OS*	CC annotation	
				term	*p*-value
anaphase-promoting complex	15	YBL084C, YDR118W, YDR260C, YFR036W, YGL240W, YIR025W, YKL022C, YLR102C, YLR127C, YNL172W, YOR249C	0.73	anaphase-promoting complex	1.86e-29
20S proteasome	14	YBL041W, YER012W, YER094C, YFR050C, YGL011C, YGR135W, YGR253C, YJL001W, YML092C, YMR314W, YOL038W, YOR157C, YOR362C,YPR103W,**YHL030W**	0.93	proteasome core complex	1.11e-37
DASH complex	10	YDR016C, YDR201W, YGL061C, YGR113W, YKL052C, YKR037C, YKR083C	0.70	DASH complex	1.38e-19
SWI/SNF complex	12	YBR289W, YDR073W, YFL049W, YHL025W, YJL176C, YMR033W, YNR023W, YOR290C, YPL016W, YPL129W,YPR034W	0.92	SWI/SNF complex	1.90e-30
SAGA complex	20	YBR081C, YBR198C, YCL010C, YDR145W, YDR176W, YDR392W, YDR448W, YGL066W, YGL112C, YGR252W, YLR055C, YMR223W, YMR236W, YOL148C, YPL254W	0.75	SAGA complex	3.41e-39
U1 snRNP complex	17	YBR119W, YDR235W, YDR240C, YER029C, YGR013W, YGR074W, YIL061C, YKL012W, YLR147C, YLR275W, YLR298C	0.65	U1 snRNP	1.95e-27

NOTE: Systematic names in bold represent those proteins in the complex identified by ICJointLE, but do not appear in the matched known complex in the first column.

Table 13

Six complexes identified by ICJointLE on DIP and their CC term enrichment analyses

Matched known complex name	*#kc*	Proteins of the identified complex	*OS*	CC annotation	
				term	*p*-value
anaphase-promoting complex	15	YBL084C, YDR118W, YFR036W, YGL240W, YKL022C, YLR127C, YNL172W, YOR249C	0.53	anaphase-promoting complex	3.50e-20
20S proteasome	14	YER012W, YER094C, YGL011C, YML092C,YMR314W,YPR103W	0.43	proteasome core complex	1.23e-14
DASH complex	10	YDR016C, YDR201W, YGR113W, YKR037C, YKR083C	0.50	DASH complex	3.96e-14
SWI/SNF complex	12	YBR289W, YFL049W, YOR290C, YPL016W, YPR034W	0.42	SWI/SNF complex	2.49e-12
SAGA complex	20	YBR198C, YCL010C, YDR167W, YDR176W, YDR448W, YGL112C, YMR236W, YOL148C	0.4	SAGA complex	4.95e-19
U1 snRNP complex	17	YBR119W, YDL087C, YDR235W, YDR240C, YGR013W, YHR086W, YIL061C, YML046W	0.47	U1 snRNP complex	3.17e-19

From **Tables 11-13**, we can see that the complexes identified by ICJointLE are matched with the known complexes well and are enriched the proteins annotated with the corresponding CC term. This indicates that the complexes identified by ICJointLE have significantly biological meaning.

4. Discussion

Most existing methods for identifying complexes in static PPI networks are based on mining densely connected regions [10-20] and integrated gene expression data [32-35] and GO functional annotation [29-31]. These methods do not use both of gene expression data and protein localization data to identify complexes. So there is no guarantee that the complexes identified by these methods are of co-localization and co-expression. In this paper, we have proposed the method ICJointLE to identify jointly co-localized and jointly co-expressed protein complexes in static PPI networks.

On one hand, If proteins in the same functional module work together, they should have high chance to show up at the same physical location [62]. When a protein complex is assembled, its constituent proteins must be localized at the same subcellular localization category. To depict co-localization among members of a protein group, we defined the joint localization vector to construct the joint co-localization criterion of a protein group. Then we can use the joint co-localization criterion to guarantee that proteins in an identified

protein complex are jointly co-localized. Furthermore, we noticed that even if all proteins in a protein group are pairwise co-localized, they are not always necessary to be jointly co-localized. It is worthwhile pointing out that the joint co-localization of a protein group is a group relationship. Evidently, using the combination of joint co-localization criterion and protein localization data can judge whether the proteins in a protein group are jointly co-localized at the same subcellular localization category.

On the other hand, the methods [34-36] measure the co-expression between two proteins. However, when a protein complex is assembled, the genes coding members of a protein complex must be co-expressed at the same time. To describe co-expression among members of a gene group, we defined the joint gene expression to construct the joint co-expression criterion of a gene group. Then we can use the joint co-expression criterion to ensure that the genes coding constituent proteins of an identified protein complex are jointly co-expressed to some extent. We also found that even if all genes in a gene group are pairwise co-expressed, they are not always necessary to be jointly co-expressed, namely, they are partially pairwise co-expression. Likewise, it is worth noting that the joint co-expression of a gene group is also a group relationship. Obviously, using the integration of joint co-expression criterion and gene expression data can determine whether the genes in gene group are jointly co-expressed at the same time.

Moreover, members of a protein complex tend to be functionally similar. Inspired by [30,31,33], we integrated the CC-based, MF-based, and BP-based protein similarities into functional similarity criterion to identify biologically significant complexes. As found in **Tables 9** and **10**, the use of functional similarity criterion can ensure that the majority of identified complexes have significantly biological meaning.

In addition, PPI data produced by high throughput technology usually contain some amount of noises or spurious interactions. The method in [32] calculates Pearson correlation coefficient between two proteins to verify true PPI pairs. While ICJointLE uses reliability score to mark PPI data and filtered out PPIs with low

reliability score. The experimental results have indicated that marking PPI data with reliability score and filtering out interactions with low reliability are helpful for precisely identifying more protein complexes.

Furthermore, according to core-attachment structure [6, 22-27], ICJointLE finds densely and *r*-reliably connected region as protein cores and adds the attachment proteins adequately and *r*-reliably connected with protein core to identify protein complexes. To some extent, our method ICJointLE is able to avoid missing attachment proteins which are not densely but adequately connected to protein core.

Considering complexes of small size (consisting of two or three distinct proteins), Xu et al.[63] found that there are 156 size-two and 66 size-three complexes in CYC2008. They proposed a method CPredictor 2.0 which achieves better performance of detecting complexes with small size in terms of F-measure. Yong et al.[64] exploited size-specific supervised weighting (SSS) to weight each edge in PPI network, and predicted and scored candidate small complexes. We evaluated the distribution of sizes of perfectly matched protein complexes identified by various methods. The evaluated results confirm that our proposed method ICJointLE has stronger ability of identifying complexes with small size.

5. Conclusions

In this paper, first, we have introduced the joint co-localization criterion, the joint co-expression criterion, and functional similarity criterion. Then we proposed a novel method ICJointLE which uses four types of biological data including PPI data with reliability score, protein localization data, gene expression data, and gene ontology annotations to identify protein complexes in static PPI networks. The experimental results on yeast showed that our method can precisely identify more complexes, especially more complexes of sizes from 2 to 6. Besides, the complexes identified by ICJointLE have significantly biological meaning.

Despite of having the advantage in precisely identifying protein complexes of small size, ICJointLE yet remains failure to precisely identify some protein complexes with small size in CYC2008. Integrating more

biological information such as PTM-dependent PPIs [65] and domain-domain interactions [66, 67] may be helpful for identifying more protein complexes exactly. In addition, PPI networks are dynamic in nature [68]. Dynamic PPI networks modeling could reveal the mechanisms of protein complex formation and contribute to identification of protein complexes. Our future work will focus on modeling dynamic PPI networks and integrating more biological information to identify more protein complexes in dynamic PPI networks.

Conflict of interest

The authors hereby declare to have no conflict of interests.

Availability of software and materials

Algorithm ICJointLE is implemented in C++. The software suite of our method and the results produced by ICJointLE from six yeast PPI data sets STRING, BioGrid, DIP, Uetz, Ito, and Yu are available at http://dx.doi.org/10.6084/m9.figshare.7719296. Or please contact to zhangjx@gxu.edu.cn.

Acknowledgement

The authors thank the editor and anonymous reviewers for their constructive comments and suggestions, which greatly help us improve our manuscript. We are grateful to Juan Shi, Li Wang, Chun yan Tang and Xi Qin for their discussion.

This work is supported by the National Natural Science Foundation of China under Grant No. 61462005 and No. 61862006, and Natural Science Foundation of Guangxi under Grant No. 2014GXNSFAA118396.

References

1. P. Cramer, D.A. Bushnel, J. Fu, et al. **Architecture of RNA polymerase II and implications for the transcription mechanism**. *Science*.288(5466) (2000) 640–649.

2. S. Clancy, W. Brown. **Translation: DNA to mRNA to protein**. *Nature Education*. 1(1) (2008) 101.

3. A. Schreiber, F. Stengel, Z.G. Zhang, et al. **Structural basis for the subunit assembly of the anaphase-promoting complex**. *Nature*. 470(7333) (2011) 270-232.

4. A.C. Gavin, M. Bösche, R. Krause, et al. **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature*. 415(6868) (2002) 141-147.

5. A. Bauer, B. Kuster. **Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes**. *FEBS*. 270(4) (2003) 570-578. http://dx.doi.org/10.1046/j.1432-1033.2003.03428.x

6. A.C. Gavin, P. Aloy, P. Grandi, et al. **Proteome survey reveals modularity of the yeast cell machinery**. *Nature*. 440(7084) (2006) 631-636.

7. P. Uetz, L. Giot, G. Cagney, et al. **A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae**. *Nature*. 403(6770) (2000) 623–627.

8. T. Ito, T. Chiba, R. Ozawa, et al. **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *PNAS*. 98(8) (2001) 4569-4574.

9. S.W. Michnick, P.H. Ear, C. Landry, et al. **Protein-fragment complementation assays for large-scale analysis, functional dissection and dynamic studies of protein-protein interactions in living cells**. *Methods in Molecular Biology*. 756 (2011) 395-425.

10. V. Spirin, L. Mirny. **Protein complexes and functional modules in molecular networks**. *PNAS*. 100(21) (2003) 12123-12128.

11. G. Liu, L. Wong, H.N. Chua. **Complex discovery from weighted PPI networks**. *Bioinformatics*. 25(15) (2009) 1891-1897. http://dx.doi.org/10.1093/bioinformatics/btp311.

12. H.N. Chua, K. Ning, W.K. Sung, et al. **Using indirect protein-protein interactions for protein complex prediction**. *Bioinformatics&Computational Biology*. 6(3) (2008) 435-466.

13. X.L. Li, S.H. Tan, C.S. Foo, et al. **Interaction graph mining for protein complexes using local clique merging**. *Genome Informatics*. 16(2) (2005) 260-269.

14. X.L. Li, C.S. Foo, S.K. Ng. **Discovering protein complexes in dense reliable neighborhoods of protein interaction networks**. *Proc*. *Comput*. *Syst*. *Bioinform*. Conf. 6 (2007) 157-168.

15. B.J. Frey, D. Dueck. **Clustering by passing messages between data points**. *Science*. 315(5814) (2007) 972-976. http://dx.doi.org/10.1126/science.1136800.

16. S.M. Van Dongen. **Graph clustering by flow simulation**. Ph.D. Thesis, University of Utrecht, Utrecht, The Netherlands. (2000).

17. G.D. Bader, C.W.V. Hogue. **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics*. 4:2(2003).

18. M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, et al. **Development and implementation of an algorithm for detection of protein complexes in large interaction networks**. *BMC bioinformatics*. 7:207 (2006). http://dx.doi.org/10.1186/1471-2105-7-207.

19. T. Nepusz, H. Yu, A. Paccanaro. **Detecting overlapping protein complexes in protein-protein interaction networks**. *Nature Methods*. 9(5) (2012) 471-472.

20. P. Jiang, M. Singh. **SPICi: a fast clustering algorithm for large biological networks**. *Bioinformatics*. 26(8) (2010) 1105-1111. http://dx.doi.org/10.1093/bioinformatics/btq078.

21. C. Ma, Y.P. Chen, B.B.C. Liao, et al. **Identification of Protein Complexes by Integrating multiple alignment of protein interaction networks**. *Bioinformatics*. 33(11) (2017) 1681-1688. http://dx.doi.org/10.1093/bioinformatics/btx043.

22. A.A. Hasin, K.B. Dhruba, K.K. Jugal. **Core and peripheral connectivity based cluster analysis over PPI network**. *Computational Biology and Chemistry*. 59 (2015) 32-41. http://dx.doi.org/10.1016/

j.compbiolchem.2015.08.008.

23. H.C.M. Leung, Q. Xiang, S.M. Yiu, et al. **Predicting protein complexes from PPI data: a core-attachment approach**. *Computational Biology*. 16(2) (2009) 133-144.

24. M. Wu, X.L. Li, C.K. Kwoh, et al. **A core-attachment based method to detect protein complexes in PPI networks**. *BMC Bioinformatics*. 10(1) (2009) 169.

25. S. Srihari, K. Ning, H.W. Leong. **Refining Markov clustering for complex detection by incorporating core-attachment structure**. *Genome Inform*. 23(1) (2009) 159-168. http://dx.doi.org/10.1142/9781848165632_0015.

26. S. Srihari, K. Ning, H.W. Leong. **MCL-CAw: a refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure**. *BMC Bioinformatics*. 11:504 (2010).

27. W. Peng, J.X. Wang, B.H. Zhao, et al. **Identification of protein complexes using weighted pagerank-nibble algorithm and core-attachment structure**. *IEEE/ACM Transactions on computational biology and bioinformatics*. 12(1) (2015) 179-192. http://dx.doi.org/10.1109/TCBB.2014.2343954.

28. The Gene Ontology Consortium. **The Gene Ontology (GO) project in 2006**. *Nucleic Acids Research*. 34(1) (2006) 322-326.

29. A.D. King, N. Pržulj,I. Jurisica. **Protein complex prediction via cost-based clustering**. *Bioinformatics*. 2004; 20(17): 3013-3020.

30. T. Price, F.I. Peña, Y.R. Cho. **Survey: Enhancing Protein Complex Prediction in PPI Networks with GO Similarity Weighting**. *Interdisciplinary Sciences*: *Computational Life Sciences*. 5(3) (2013) 196 -210.

31. Z.H. Yang, H.F. Lin, B Xu. **Ontology integration to identify protein complex in protein interaction networks**. *Proteome Science*. 9(Suppl 1):S7 (2011). http://dx.Doi.org/10.1186/1477-5956-9-S1-S7.

32. B. Xu, H. Lin, Y. Chen, et al. **Protein complex identification by integrating protein-protein interaction evidence from multiple sources**. *PLoS ONE*. 8(12) (2013) e83841. http://dx.doi.org/10.1371/journal. pone.0083841.

33. B. Cao, J. Luo,C. Liang, et al. **Pce-fr: A novel method for identifying overlapping protein complexes in weighted protein-protein interaction networks using pseudo-clique extension based on fuzzy relation.** *IEEE Transactions on Nanobioscience*. 15(7) (2016) 728-738. http://dx.doi.org/10.1109/TNB.2016.2611683.

34. J. Feng, R. Jiang, T. Jiang. **A max-flow based approach to the identification of protein complexes using protein interaction and microarray data**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 8(3) (2011) 621-634. http://dx.doi.org/10.1109/TCBB.2010.78.

35. X.W. Tang, J.X. Wang, Y. Pan. **Predicting protein complexes via the integration of multiple biological information**. *IEEE 6th International Conference on Systems Biology*. (2012) 174-179. http://dx.doi.org/10.1109/ISB.2012.6314.132.

36. S. Keretsu, R. Sarmah. **Weighted edge based clustering to identify protein complexes in protein-protein interaction networks incorporating gene expression profile**. *Computational Biology and Chemistry*. (65) (2016) 69-79. http://dx.doi.org/10.1016/j.compbiolchem.2016.10.001.

37. P. Cui, X. Wang, J. Pei, et al. **A survey on network embedding**. *IEEE Transactions on Knowledge and Data Engineering*.31(5)(2019)833-852. http://dx.doi.org/10.1109/TKDE.2018.2849727.

38. X. Liu, Z. Yang, S. Sang, et al. **PC-SENE: A node embedding based method for protein complex detection**. *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*). (2018) 191-192. http://dx.doi.org/10.1109/BIBM.2018.8621338.

39. A. Grover, J. Leskovec. **node2vec: Scalable feature learning for networks**. *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, (2016) 855-864. http://dx.doi.org/10.1145/2939672.2939754.

40. B. Xu, K.Li, X. Liu, et al. **Protein Complexes Detection Based on Global Network Representation Learning**. *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*).(2018) 210-213. http://dx.doi.org/10.1109/BIBM.2018.8621541.

41. H. Yao, Y. Shi, J. Guan, et al. **Accurately Detecting Protein Complexes by Graph Embedding and Combining Functions with Interactions.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 11(1) (2019). http://dx.doi.org/10.1109/TCBB.2019.2897769.

42. W.K. Huh, J.V. Falvo, L.C. Gerke, et al.**Global analysis of protein localization in budding yeast**. *Nature*. 425(6959) (2003) 686-691. http://dx.doi.org/10.1038/nature02026.

43. S. Pu, J. Wong, B. Turner, et al. **Up-to-date catalogues of yeast protein complexes**. *Nucleic Acids Research*. 37(3) (2009) 825-831. http://dx.doi.org/10.1093/nar/gkn1005.

44. R. Jansen, D. Greenbaum, M. Gerstein. **Relating whole-genome expression data with protein-protein interactions**. *Genome Research*. 12(1) (2002) 37-46. http://dx.doi.org/10.1101/gr.205602.

45. B. Futcher, G.I. Latter, P. Monardo, et al. **A sampling of the yeast proteome**. *Molecular and Cellular Biology*. 19(11) (1999) 7357-7368. http://dx.doi.org/10.1128/MCB.19.11.7357.

46. K.I. Goh, M.E. Cusick, D. Valle, et al. **The human disease network**. *PNAS*. 104(21) (2007) 8685-8690.

47. P.H. Guzzi, M. Mina, C. Guerra, et al. **Semantic similarity analysis of protein data: assessment with**

biological features and issues. *Briefings in Bioinformatics*. 13(5) (2011) 569-585. http://dx.doi.org/10.1093/bib/bbr066.

48. P. Resnik. **Using information content to evaluate semantic similarity in a taxonomy**. arXiv preprint cmp-lg/9511007 1(1995). *In Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAI 95*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1995.
49. Y.F.Zhang, A.L. Xu. **Improved computation method for semantic similarity between gene ontology terms** (in Chinese). *Journal of Computer Applications*. 32(5) (2012) 1329-1331.
50. Y. Li, Z.A. Bandar, D. McLean. **An approach for measuring semantic similarity between words using multiple information sources**. *IEEE Transactions on Knowledge and Data Engineering*. 15(4) (2003) 871-882. http://dx.doi.org/10.1109/TKDE.2003.1209005.
51. J.Z. Wang, Z.D. Du, R. Payattakool, et al. **A new method to measure the semantic similarity of go terms**. *Bioinformatics*. 23(10) (2007) 1274-1281.
52. STRING http://string-db.org/ cgi/download_page.pl
53. S. Brohée, J. van Helden. **Evaluation of clustering algorithms for protein-protein interaction networks**. *BMC Bioinformatics*. 7(1) (2006) 488. http://dx.doi.org/10.1186/1471-2105-7-488.
54. R. Jansen, M. Gerstein. **Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction**. *Curr.Opin.Microbiol*. 7(5) (2004) 535-545. http://dx.doi.org/10.1016/j.mib.2004.08.012.
55. H. Ge, Z. Liu, G.M. Church, et al. **Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae**. *Nature Genetics*. 29(4) (2001) 482-486. http://dx.doi.org/10.1038/ng776.
56. B. Zhang, B.H. Park, T. Karpinets, et al. **From pull-down data to protein interaction networks and**

complexes with biological relevance. *Bioinformatics*. 24(7) (2008) 979-986. http://dx.doi.org/10.1093/bioinformatics/btn036.

57. E.I. Boyle, S. Weng, J. Gollub, et al. **GO::TermFinder-open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes**. *Bioinformatics*. 20(18) (2004) 3710-3715. http://dx.doi.org/10.1093/bioinformatics/bth456.
58. C. Stark, B.J. Breitkreutz, T. Reguly, et al. **BioGRID: A general repository for interaction datasets**. *Nucleic Acids Research*. 34(Suppl 1) (2006) 535-539. http://dx.doi.org/10.1093/nar/gkj109.
59. L. Salwinski, C.S. Miller, A.J. Smith, et al. **The database of interacting proteins: 2004 update**. *Nucleic Acids Research*. 32 (2004) 449-451.
60. H. Yu, P. Braun, M.A. Yildirim, et al. **High-quality binary protein interaction map of the yeast interactome network**. *Science*.322(5898) (2008) 104-10.
61. B.P. Tu, A. Kudlicki, M. Rowicka, et al. **Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes**. *Science*.310(5751) (2005)1152-1158.
62. A. Kumar, K.H. Cheung, P Ross-Macdonald, et al. **TRIPLES: a database of gene function in Saccharomyces cerevisiae**. *Nucleic Acids Res*. 28(1)(2000)81－84.
63. B. Xu, Y. Wang, Z. Wang, et al. **An Effective Approach to Detecting Both Small and large complexes from protein-protein interaction networks**. *BMC Bioinformatics*. 18(Suppl 12)(2017)419. http://dx.doi.org/10.1186/s12859-017-1820-8.
64. C.H. Yong, O. Maruyama, L. Wong. **Discovery of small protein complexes from PPI networks with size-specific supervised weighting**. *BMC Syst. Biol*. 8(S5)(2014)1-15.
65. J.X. Wang, X.Q. Peng, W. Peng, et al. **Dynamic protein interaction network construction and applications**. *Proteomics.* 14 (2014) 338–352. http://dx.doi.org/10.1002/pmic.201300257.

66. Y. Ozawa, R. Saito, S. Fujimori, et al. **Protein complex prediction via verifying and reconstructing the topology of domain-domain interactions**. *BMC Bioinformatics*. 11: 350 (2010).

67. L. Ou-Yang, H. Yan, X. Zhang. **A multi-network clustering method for detecting protein complexes from multiple heterogeneous networks**. *BMC Bioinformatics.* 18(Suppl 13)(2017) 463. http://dx.doi.org/10.1186/s12859-017-1877-4.

68. E.D. Levy, J.B. Pereira-Leal. **Evolution and dynamics of protein interactions and networks**. *Curr.Opin.Struct. Biol*. 18(3) (2008) 349-357. http://dx.doi.org/10.1016/j.sbi.2008.03.003.