BMC Bioinformatics

RESEARCH

Open Access

Small protein complex prediction algorithm based on protein–protein interaction network segmentation

Jiaqing Lyu[1], Zhen Yao[2], Bing Liang[3*], Yiwei Liu[3] and Yijia Zhang[4*]

*Correspondence:
liangbing@dlut.edu.cn;
zhangyijia@dlmu.edu.cn

[1] School of Computer Science and Technology, Dalian University of Technology, Dalian, China
[2] School of Chemical Engineering, Dalian University of Technology, Dalian, China
[3] School of Innovation and Entrepreneurship, Dalian University of Technology, Dalian, China
[4] School of Information Science and Technology, Dalian Maritime University, Dalian, China

Abstract

Background: Identifying protein complexes from protein-protein interaction network is one of significant tasks in the postgenome era. Protein complexes, none of which exceeds 10 in size play an irreplaceable role in life activities and are also a hotspot of scientific research, such as PSD-95, CD44, PKM2 and BRD4. And in MIPS, CYC2008, SGD, Aloy and TAP06 datasets, the proportion of small protein complexes is over 75%. But up to now, protein complex identification methods do not perform well in the field of small protein complexes.

Results: In this paper, we propose a novel method, called BOPS. It is a three-step procedure. Firstly, it calculates the balanced weights to replace the original weights. Secondly, it divides the graphs larger than MAXP until the original PPIN is divided into small PPINs. Thirdly, it enumerates the connected subset of each small PPINs, identifies potential protein complexes based on cohesion and removes those that are similar.

Conclusions: In four yeast PPINs, experimental results have shown that BOPS has an improvement of about 5% compared with the SOTA model. In addition, we constructed a weighted Homo sapiens PPIN based on STRINGdb and BioGRID, and BOPS gets the best result in it. These results give new insights into the identification of small protein complexes, and the weighted Homo sapiens PPIN provides more data for related research.

Keywords: Protein complex identification, Small protein complex, Protein–protein interaction, Graph segmentation

Introduction

Since the launch of the Human Genome Project in 1990, massive amounts of genomic data have emerged, and bioinformatics appeared. With the advent of the post-genomic era, the focus of life science research has shifted from genomics to proteomics [1]. Protein complexes take part in a variety of biological processes including: cell cycle regulation, differentiation and protein folding [2]. With the development of Biotechnology, a great number of ways to get protein-protein interaction network (PPIN) appeared, such as X-ray crystallography, Nuclear magnetic resonance (NMR) [2, 3] tandem affinity

purification [4, 5] (TAP), various massspectrometry techniques such as native, cross-linked [6] (CX or XL), ion mobility [7, 8] two-hybrid system [9] and protein micro array. Therefore, predicting protein complexes in PPI networks has gradually become a research hotspot [10].

Protein complexes are groups of proteins that interact with each other at the same time and place, forming a single multimolecular machine [11]. Due to its essential role in the understanding of cellular organizations and functions, such as replication, transcription and the control of gene expression, etc [4, 12, 13]. One of the purposes of studying PPIN is to obtain protein complexes or functional modules in the network. However, experimentally determining protein complex data are still somewhat limited as they are largely obtained through small-scale experimental techniques, which are time-consuming and tedious [14]. At the same time, many large-scale PPIN have been constructed with the advances of high-throughput technologies. Therefore, predicting protein complexes in PPIN through computational algorithms can provide reliable guidance and help for biological experiments.

A protein-protein interaction network can be modeled as an undirected graph. The vertices in the graph represent proteins, and the edges represent the interactions between proteins. Therefore, the problem of protein complex prediction can be approximated as a graph theory problem. The predecessors proposed some computational algorithms to predict protein complexes in PPI networks. Most of these protein complexes identification methods are based on the principle that densely linked regions in the PPI network correspond to actual protein complexes [15]. Therefore, the protein complex prediction problem can be further regarded as the problem of detecting densely linked regions in PPIN [16, 17].

Subject to biological technology, researchers usually conduct in-depth research on smaller proteins. At the same time, small protein complexes also play an irreplaceable role in life activities. For example, PSD-95 consists of 6 proteins and plays an important role in synaptic plasticity and the stabilization of synaptic changes during long-term potentiation [18]. CD44 consists of 8 proteins and participates in a wide variety of cellular functions including lymphocyte activation [19], recirculation and homing [20], hematopoiesis [21], and tumor metastasis [22]. PKM2 consists of 8 proteins and is expressed in most human tumors [23]. BRD4 consists of 5 proteins and most cases of NUT midline carcinoma involve translocation of the BRD4 with NUT genes [24]. So, predicting smaller protein complexes may provide more help for biological research. But up to now, there is not a specifical method to identify the complex whose size is no more than ten effectively from PPIN. And the performances of traditional methods are not so satisfying and promising.

Therefore, we designed the BOPS algorithm to specially predict smaller protein complexes. The BOPS algorithm means "Based On PPIN Segmentation". The basic idea of BOPS is to divide the PPIN according to the reliability of the interaction. The BOPS algorithm divides the original graph into some small networks and enumerates connected subsets of small PPINs to check if subsets are protein complexes. Finally, BOPS successfully transforms the problem of predicting protein complexes into a problem of judging whether a subgraph is a protein complex, thereby greatly improving the prediction effect. We evaluate BOPS compared to the state-of-the-art methods. The experimental

results show that the BOPS algorithm has achieved very great results for complexes with not exceeding ten.

Related work

Generally, the computational methods for protein complex prediction can be divided into three main categories: network-based, biological-context-aware, and specialized methods [2, 25]. Network-based approaches exploit the network structure to detect protein complexes and Biological-context-aware approaches combine topo-logical and gene information as functional information to detect complexes. However, all of them try to predict protein complexes of various sizes. Therefore, the approaches developed to predict small complexes are summarized as "specialized methods".

Among the three main categories, there are many studies in this field of network-based algorithms. These algorithms are based solely on PPIN. The network-based algorithms can be further divided into agglomerative methods and divisive methods. For example, CFinder [26], PEWCC [27] and ClusterONE [28] are all classic network-based algorithms which use agglomerative methods. CFinder is based on subgraph merge to predict protein complexes. The ClusterONE first selects the protein with the highest degree from the PPI network as the seed node and uses the greedy algorithm to add or remove protein to form a highly aggregated subgraph. The PEWCC assesses the reliability of the interaction data, then predicts protein complexes based on the concept of the weighted clustering coefficient. MCL [29] is the representative algorithm using divisive method. This method detects dense subgraphs as predicted complexes in a given PPIN by simulating random walks. To simulate the random walk (flow), MCL uses "expansion" (controls the spread of the flow) and "inflation" (controls the spread of the flow) operation iteratively.

Some methods are based on PPIN and some additional biological insights [30]. The number of these methods is not so large, and the most famous algorithm is COACH [31]. The protein complex has a combination feature, and the protein complex is composed of a core and some attachments. The proteins in the core part have high levels of co-expression and functional similarity [4]. Therefore, the COACH is based on this theory and has two steps. First, the core structures of the proteins are determined according to the neighboring relationships of the proteins, and then the proteins in the core structures are expanded to get attachments according to the biological significance. Kouhasr et al [32] improved COACH to be compatible with weighted PPI networks for protein complex detection. They proposed a new method WCOACH based on Gene Ontology structure as an optimized version of COACH. Recently a new method called GANE based on Gene Ontology attributed network embedding was proposed to predict protein complexes [33]. This method learns the vector representation for each protein from a GO attributed PPI network. Then, it uses the clique mining method to generate candidate cores. For each seed core, its attachments are the proteins with a correlation score that is larger than a given threshold.

The smaller protein complex contains fewer proteins, so the topology in the PPI network is not obvious. All of the aforementioned methods try to predict protein complexes of various sizes and densities. Those general algorithms cannot efficiently find specific types of complexes, particularly sparse and small ones [2]. These complexes are riddled

with various challenges in the course of prediction, particularly when only topological information of the PPIN is available. Therefore, the special-purpose strategies developed to address this problem are classified as "specialized methods". CPredictor2.0 is a method to detect "very small complexes" (size not exceeding three) [34]. The method groups proteins of similar functions, then uses the Markov clustering algorithm to discover clusters in each group and merge some of them. The merged clusters as well as the remaining clusters constitute the set of detected complexes.

Method

Problem statement and notation

A PPIN can be represented as a graph $G = (V, E, W)$. The PPIN has $|V|$ proteins which are indicated by vertices. And the PPIN has $|E|$ interactions which are indicated by edges. Additionally, W reflects the weights of the interactions. The protein complex can be represented as a subset of proteins with high cohesion in the graph. As a result, the protein complex prediction problem can be regarded as a graph theory problem. Therefore, in "Method", we will mainly use graph theory to describe BOPS algorithm, thereby enhancing rigor of the paper.

An undirected edge can be represented as $e = (x_e, y_e, w_e)$, where x_e and y_e are endpoints of e, and w_e represents the weight of e. The cnt_v represents the number of edges from vertex v, which is the degree of vertex v without weight. The sum_v represents the sum of the weights of all edges from vertex v, which is the degree of vertex v with weight.

$$cnt_v = \sum_{x_e = v} 1 \tag{1}$$

$$sum_v = \sum_{x_e = v} w_e \tag{2}$$

Algorithm overview

The BOPS algorithm for protein complex prediction is a three-step procedure. First, the BOPS algorithm calculates the balanced weights, and replaces the original weights with balanced weights (3.3). Second, the BOPS algorithm divides the graphs larger than MAXP until the original PPIN is divided into small networks (3.4.1) and the details is described in (3.4.2). Third, the BOPS algorithm enumerates every connected subset of each small network (3.5.1), calculates the cohesion of each connected subset (3.5.2), identifies potential protein complexes based on cohesion and removes those that are similar (3.5.3). Figure 1 shows the overall flow of the algorithm to identify complexes in a PPIN.

Calculation of balanced weights

PPIN is obtained by many biological experiment methods. The weight reflects the reliability of interactions. Based on previous studies in yeast, each complex is composed of a core and attachments [4]. Proteins in the core interact with each other closely, which decides the main biological function of the complex. Some proteins are bound to the core to complete their function. These proteins are called attachments.

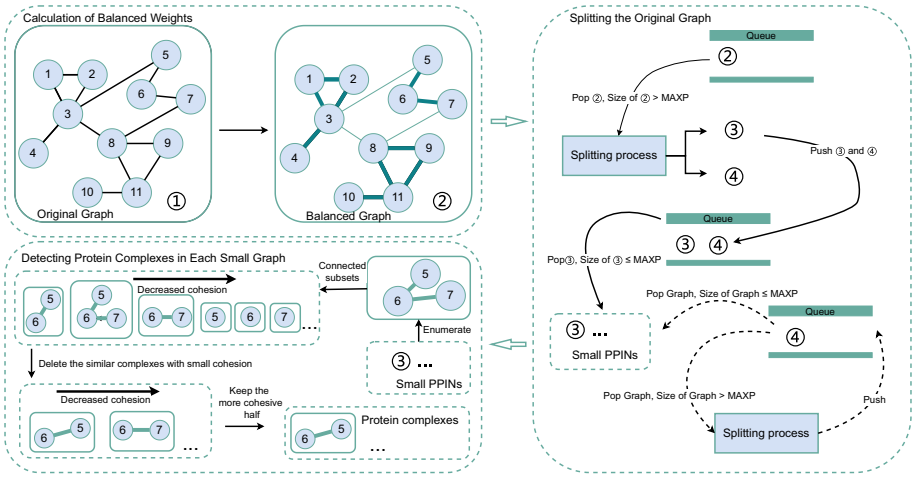

Fig. 1 The description of the BOPS algorithm

The proteins in the core usually have a high level of interaction with a large number of proteins in the same core. But the attachment usually only interacts with a small number of proteins in the core, and the level of interaction is low. Because the BOPS algorithm only uses cohesion to determine whether a set of proteins is a complex, the BOPS algorithm adjusts the weights of edges according to the core-attachment biological structure. Indeed, the BOPS algorithm calculates balanced weights to balance the importance of the original edge weight and core-attachment structure. For an edge e, we define its balanced weight bw_e as follows:

$$bw_e = \frac{1}{2}\left(\frac{{w_e}^{\beta}}{{sum_{x_e}}^{\beta-1}} + \frac{{w_e}^{\beta}}{{sum_{y_e}}^{\beta-1}}\right) \tag{3}$$

Parameter β is used in the calculation of balanced weights. The default value of β is 1.5. When the β is 1, the BOPS algorithm will not change original weights. When β is large, biological structure information will affect the performance of our method significantly. The value range of this parameter is from 1.0 to 2.0.

When the interaction e is between the core and an attachment, one of sum_{x_e} and sum_{y_e} will be small, and since the value is used as a denominator, the balanced weight of e becomes larger than the original weight. As a result, the BOPS algorithm indirectly considers the core-attachment structure in the balanced weights.

Splitting the original graph

An overview of the graph segmentation process

The BOPS algorithm constructs an empty queue, and pushes the original graph into it. As long as the queue is not empty, the algorithm pops the head element, a graph every time, and then checks whether its size is greater than MAXP. If the size of the graph does not exceed MAXP, the BOPS will no longer split this graph. If the size of the graph is greater than MAXP, the BOPS algorithm will split it and push the subgraphs into the back of queue. When the queue is empty, the size of all graphs are all not greater than MAXP.

The MAXP represents the maximum number of proteins in each graph after splitting the original graph into small graphs. The larger MAXP is, the better result is. The time complexity is $O(2^{\mathrm{MAXP}})$ approximately. We recommend set MAXP to 20 when detecting small complexes.

The details of segmentation

The details of the splitting process is described below. The algorithm deletes some edges to make the graph disconnected, and simultaneously makes sure the highest weight of the deleted edges are minimized. The cost is defined as follow:

$$cost = max\{bw_e\}(e \,\epsilon\, deleted\ edges) \tag{4}$$

Removing some edges is equivalent to deleting all the edges and adding back some edges. In the same way, deleting some edges to make the graph split into two parts, is equivalent to removing all edges and adding back some edges to make the graph join into two parts.

Indeed, the BOPS algorithm firstly removes all the edges from the graph into the array and sorts them from largest to smallest by edge weight. Then, BOPS adds the edges back to the graph, one at a time. The BOPS maintains a classic data structure called disjoint set union to support queries of the connectivity of the graph. If the edge that is currently being added causes the graph to become connect, It is proved that this edge must be removed if the graph is to be divided at minimum cost. In this case, BOPS removes this edge completely.

Detecting possible protein complexes

Enumerating connected subsets

The connected subsets of each graph may be protein complexes. After splitting the original graph into small graphs, all graphs do not exceed MAXP. Therefore, the BOPS algorithm can enumerate all connected subsets of each graph to calculate their cohesion, which is used for filtering the potential complexes. In this way, we convert a generative problem into a decision problem.

The BOPS algorithm uses breadth-first search to obtain all connected subsets in graph $G = (V, E, BW)$. Then for each connected subgraph, the algorithm will compute its cohesion.

Calculation of sets' cohesion

In a protein complex, a protein should interact with most of other proteins. As a result, the BOPS algorithm uses $(cnt_x + 1)/|V|$ to reflect the number of proteins which interact with protein x. $(cnt_x + 1)/|V|$ reflects the proportion of proteins interacting with proteins x in the current complex. The BOPS adds 1 to the numerator because the protein is always interacting with itself.

At the same time, a protein should have high level of interactions with other proteins. Therefore, the BOPS algorithm uses sum_x to reflect the level of interaction of protein x.

As a result, the algorithm uses $sum_x \times (cnt_x + 1)/|V|$ to measure the denote the possibility of protein x in the complex.

Finally, the algorithm averages the possibility of each protein in the set to reflect the cohesion of the entire set.

$$Cohension(V) = \frac{1}{|V|} \sum_{x \in V} sum_x \times \frac{cnt_x + 1}{|V|} \tag{5}$$

V, $|V|$, sum_x and cnt_x here only consider vertices in the set, and edges whose endpoints are all in the set.

Detecting protein complexes

The BOPS algorithm calculates the cohesion of every complexes in each graph and ranks the complexes from most cohesive to least cohesive. Then the BOPS iterates through all the complexes, if one complex is similar to complexes which have larger cohesion, it will be deleted. Finally, the algorithm takes the most cohesive half of the candidates as the final result.

Time complexity analysis

The bottleneck of the algorithm is enumerating the connected subset of all graphs and calculating the cohesion. The graph $G = (V, E, W)$ can be divided into at most $|V|$/MAXP graphs (e.g. The size of each graph is MAXP). The maximum size of each graph is MAXP (e.g. Each graph reaches an upper limit in size). The number of connected subset of each graph is 2^{MAXP} (e.g. Any two points are connected to each other).

As a result, the time complexity of calculating cohesion is O(MAXP), the time complexity of calculating cohesion of all connected subsets of one graph is O(MAXP $\times$ 2^{MAXP}), the time complexity of calculating all connected subset of all graphs is O($V \times 2^{\mathrm{MAXP}}$) and the time complexity of BOPS is O($V \times 2^{\mathrm{MAXP}}$).

According to our experiments in "Experimental result and analysis" section, the results of conventional data sets can be finished in less than 20 minutes, which is consistent with the time complexity analysis. The hardware environment is Intel Core i5-9500 @ 3.00GHz.

The intention of algorithm design

Detecting protein complexes in PPIN is a generative problem. However, judging whether a connected set of protein is a protein complex is a decision problem. Usually, a decision problem is easier than a generative problem. Therefore, we convert a generative problem into a decision problem by splitting PPIN into small networks and enumerating connected subsets. The algorithm is summarized as the pseudo-code shown in Algorithm 1.

Algorithm 1 The BOPS Algorithm

Input: The original graph $G_o = (V_o, E_o, W_o)$, the balance-index $/beta$ and the maximum number of vertices in one small graph MAXP
Output: A set of protein complexes Res

```
Calculating balanced weights BW_o to replace W_o
Constructing a queue Q with the initial element G_o
Initializing a set of graphs SG to be ∅
while Q ≠ ∅ do
    Pop the head element of Q as the G_c = (V_c, E_c, BW_c)
    if |V_c| ≤ MAXP then
        Add G_c to SG
    else
        Split G_c to two small graphs G_1 and G_2
        Push G_1 and G_2 in the back of Q
    end if
end while
Initialize a set of complexes TmpRes to be ∅
for G in SG do
    BFS to get the set of complexes ComSet from G
    Rank complexes in ComSet in order of cohesion
    Delete similar complexes with small cohesion in ComSet
    Add ComSet into TmpRes
end for
Take the most cohesive half of TmpRes as Res
Return Res
```

Experimental result and analysis

First, we present the details of baseline methods : GANE [33], WCOACH [32], ClusterONE [28] , PEWCC [27], CPredictor [34], MCL [29] and CFinder [26]. Second, We introduce the evaluation metrics. Third, we systematically evaluate the performance of our method compared to 7 baseline algorithms in yeast PPINs. Fourth, we discuss the effect of parameters and the effect of graph segmentation. Fifth, we test the adaptability of BOPS with various sizes and other species. Sixth, we evaluate the reasonability and validity of the predicted complexes by their p-values under GO terms of biological process.

The details of baseline methods

- *MCL* Markov clustering is a representative graph-based clustering algorithm. It utilizes the random walk theory to discover the cluster core nodes and Markov chains rule to translate between within-cluster and across-cluster. MCL does not require the number of clusters to be known in advance.
- *CFinder* CFinder is an approach to analyzing the main statistical features of the interwoven sets of overlapping communities. Unlike the BOPS split the subgraph by the greed method, CFinder uses the greedy method to form clusters for finding maximal cliques with at least k vertices.
- *ClusterONE* ClusterONE is a clustering method with overlapping neighborhood expansion. It is a greedy method to predict protein complexes. In each iteration, it selects a node as the core node and extends it through the other node to increase the density of the cluster. Differing from the up-bottom method BOPS, ClusterONE is a bottom-up clustering method.

- *PEWCC* PEWCC is a kind of graph mining algorithm. Firstly, PEWCC assesses the reliability of the interaction data, then predicts protein complexes based on the concept of weighted clustering coefficient. BOPS and PEWCC methods are considered the reliability for the interaction of proteins.
- *WCOACH* WCOACH proposes a semantic similarity measure between proteins, based on Gene Ontology structure, which is applied to weigh PPI networks. It improved the well-known method COACH, which has been improved to be compatible with weighted PPI networks for protein complex detection.
- *CPredictor* CPredictor is a method to detect "very small complexes" (size not exceeding three). The method groups proteins of similar functions and then uses the Markov clustering algorithm to discover clusters in each group and merge some of them. The merged clusters, as well as the remaining clusters, constitute the set of detected complexes. BOPS predicts that the number of proteins in the small complex does not exceed ten. So BOPS is more universal than CPredictor.
- *GANE* GANE is a method to predict protein complexes based on Gene Ontology. First, it learns the vector representation for each protein from a GO attributed PPI network. Then, it uses the clique mining method to generate candidate cores. Similar to BOPS, it selects the proteins with more significant correlation scores as predicted proteins.

Evaluation metrics

To formally evaluate the performance of our method, we use the same evaluation metrics as other methods [27, 35]. In the beginning, we need to assess the quality of one predicted protein complex by comparing it with the protein complexes in the reference set. P denotes the set of predicted protein complexes from one method, and B denotes the set of gold standard protein complexes. And $p \in P$ is an identified protein complex; $b \in B$ is a known protein complex. The neighborhood affinity score $NA(p, b)$ is defined as:

$$\mathrm{NA}(p,b) = \frac{|V_p \cap V_b|^2}{|V_p| \times |V_b|} \tag{6}$$

where V_p is the set of proteins in the predicted protein complex p and V_b is the set of proteins in the reference protein complex b. Following the previous studies when NA(p,b) is not less than 0.25, we consider the p and b are matched [36].

Based on the neighborhood affinity score, N_{cp} is defined as the number of predicted protein complexes that match at least one reference protein complex, and N_{cb} is the number of the reference protein complexes that matches at least one predicted protein complex.

$$N_{cp} = \left|\left\{p | p \in P, \exists b \in B, \mathrm{NA}(p,b) \geq \omega\right\}\right| \tag{7}$$

$$N_{cb} = \left|\left\{b | b \in B, \exists p \in P, \mathrm{NA}(p,b) \geq \omega\right\}\right| \tag{8}$$

In Eqs. (7) and (8), ω is a threshold parameter, which is typically specified to be 0.25. The first three measures used in experiments for evaluating the performance of different methods are Precision, Recall, and F-score [37]. They can be defined as follows:

$$\text{Precision} = \frac{N_{cp}}{|V_p|}, \text{Recall} = \frac{N_{cb}}{|V_b|} \tag{9}$$

$$\text{F-score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

Precision is the rate of predicted protein complexes that match at least one reference complex, which is used to assess the quantity of matched predicted complexes Recall is the rate of reference protein complexes that match at least one predicted complex, which is used to assess the quantity of matched reference complexes. F-score is the harmonic mean of Recall and Precision, which is used to assess the overall performance for the quantity of matched complexes.

The other three measures used in experiments are clustering-wise sensitivity (Sn), clustering-wise positive predictive value (PPV) and geometric accuracy (ACC) [38, 39]. Given an identified complex p in predicted cluster P and a known complex b in gold reference cluster B. T_{pb} is defined as the number of proteins that can be found both in the reference set V_{band} predicted set V_p.

$$T_{pb} = \left|\left\{V_p \cap V_b\right\}\right| \tag{11}$$

Sn is the rate of the maximum-sum number of matched proteins to the total number of proteins in the set of the reference protein complex. PPV is the rate of the maximum-sum number of matched proteins to the total matched number of proteins in the set of the predicted protein complex. So, Sn and PPV are defined as follows:

$$\text{Sn} = \frac{\sum_{b=1}^{|B|} \max_{p=1}^{|P|} \left\{T_{pb}\right\}}{\sum_{b=1}^{|B|} |b|} \tag{12}$$

$$\text{PPV} = \frac{\sum_{p=1}^{|P|} \max_{b=1}^{|B|} \left\{T_{pb}\right\}}{\sum_{p=1}^{|P|} \sum_{b=1}^{|B|} T_{pb}} \tag{13}$$

ACC is the geometric mean of Sn and PPV, which is used to assess the overall performance for the quality of matched complexes.

$$\text{ACC} = \sqrt{\text{Sn} \times \text{PPV}} \tag{14}$$

The third metric we used is the maximum matching ratio (MMR) [28], which is based on a maximal matching between gold standard complexes and predicted complexes in a bipartite graph. The bipartite graph is the two sets of nodes representing the reference and predicted complexes, respectively, and an edge connecting a reference complex with a predicted one is weighted by the overlap score between the two. MMR offers a natural, intuitive way to compare predicted complexes with a gold standard and it explicitly

penalizes cases when a reference complex is split into two or more parts in the predicted set, as only one of its parts is allowed to match the correct reference complex.

F-score, ACC and MMR are useful in the sense that they assess how well a protein complex detection method is able to rediscover the known complexes. They are measured in the range [0, 1] and a high value indicates a good quality of detection [40].

Performance comparison in the yeast PPINs

Data sets and gold standard

We conduct experiments on four PPI networks: Krogan-core [5], Krogan-extended [5], Gavin [41], Collins [42]. The detailed information of these four datasets is shown in Table 1. To compare the identified complexes with the known complexes, we have constructed a benchmarking set as the gold standard by selecting the protein complexes which have at most ten proteins from MIPS, CYC2008, SGD, Aloy and TAP06. Therefore, there are 596 protein complexes in the reference set [25].

Performance comparison

To evaluate the effectiveness of our proposed mothed, we compare it with other seven protein identification methods: GANE [33], WCOACH [32], ClusterONE [28], PEWCC [27], CPredictor [34], MCL [29] and CFinder [26]. The parameters of these methods are set as the recommended values as mentioned in their original papers [43]. For our method, we set the β to 1.5. For fairness, we filter out the predicted protein complexes whose sizes are not more than 10 in all methods. All experimental results are listed in Table 2. According to the section of evaluation metrics, considering both F-score, ACC and MMR are overall evaluation metrics, so the best scores of F-score, ACC and MMR are highlighted in bold for easy comparison.

According to the data in Table 2, we observed that BOPS obtains the best scores for F-score, ACC and MMR in all datasets. The result of F-score is 3.2% higher than that of the second method on average. The result of ACC is 7.5% higher than that of the second method on average. The result of MMR is 40.6% higher than that of the second method on average. It illustrates that the overall accuracy of protein complexes identified by BOPS is better than prevalent algorithms in the field of small protein complexes. What's more, both Recall, Sn and MMR rank first in all datasets. So, BOPS covered more real protein complexes relatively. In other words, it has a high quantity and quality for matched complexes with respect to the reference set.

But BOPS do not achieve the highest Precision and PPV. For these datasets, WCOACH is the best for Precision. WCOACH is a semantic similarity measure

Table 1 The Yeast PPIN datasets used in the experiment

PPIN	#Proteins	#Interactions	Edge weight average	Edge weight variance	PPIN density
Krogan-core	2708	7123	0.67978	0.06407	0.00194
Krogan-extended	3672	14317	0.41552	0.10200	0.00212
Gavin	1855	7669	0.35643	0.01996	0.00446
Collins	1622	9074	0.78214	0.03310	0.00690

Table 2 Performance comparision

PPIN	Method	#Predicated	F-score	Precision	Recall	ACC	Sn	PPV	MMR
Krogan-core	BOPS	704	0.558^{1st}	0.463	0.701	0.528^{1st}	0.610	0.457	0.332^{1st}
	GANE	140	0.539^{2nd}	0.636	0.467	0.442	0.485	0.403	0.182
	WCOACH	70	0.382	0.729	0.259	0.320	0.317	0.322	0.084
	ClusterONE	551	0.407	0.332	0.524	0.489^{2nd}	0.513	0.466	0.223^{2nd}
	PEWCC	177	0.468	0.599	0.385	0.387	0.414	0.361	0.165
	CPredictor	155	0.520^{3rd}	0.703	0.413	0.412	0.400	0.425	0.188^{3rd}
	MCL	337	0.349	0.279	0.464	0.480^{3rd}	0.478	0.482	0.164
	CFinder	108	0.372	0.528	0.288	0.364	0.289	0.457	0.120
Krogan-extended	BOPS	778	0.538^{1st}	0.476	0.620	0.482^{1st}	0.533	0.420	0.288^{1st}
	GANE	183	0.496^{3rd}	0.579	0.434	0.421^{3rd}	0.453	0.390	0.174
	WCOACH	97	0.390	0.701	0.270	0.342	0.325	0.361	0.095
	ClusterONE	910	0.398	0.374	0.427	0.433^{2nd}	0.456	0.411	0.197^{2nd}
	PEWCC	225	0.436	0.524	0.373	0.367	0.407	0.331	0.143
	CPredictor	180	0.507^{2nd}	0.689	0.401	0.395	0.401	0.390	0.175^{3rd}
	MCL	419	0.250	0.203	0.326	0.418	0.368	0.475	0.111
	CFinder	118	0.261	0.39	0.196	0.302	0.209	0.436	0.071
Gavin	BOPS	832	0.668^{1st}	0.585	0.777	0.560^{1st}	0.727	0.431	0.435^{1st}
	GANE	182	0.593	0.604	0.582	0.480	0.500	0.461	0.211
	WCOACH	199	0.654^{3rd}	0.859	0.528	0.465	0.531	0.407	0.258^{3rd}
	ClusterONE	200	0.653	0.770	0.567	0.514^{3rd}	0.568	0.465	0.249
	PEWCC	203	0.656^{2nd}	0.768	0.573	0.489	0.592	0.404	0.404^{2nd}
	CPredictor	180	0.527	0.722	0.415	0.417	0.408	0.427	0.195
	MCL	231	0.516	0.463	0.582	0.523^{2nd}	0.570	0.479	0.205
	CFinder	115	0.550	0.713	0.448	0.457	0.468	0.447	0.175
Collins	BOPS	794	0.614^{1st}	0.506	0.779	0.550^{1st}	0.707	0.428	0.422^{1st}
	GANE	126	0.607^{2nd}	0.675	0.552	0.493	0.528	0.461	0.221
	WCOACH	65	0.465	0.800	0.328	0.362	0.325	0.403	0.121
	ClusterONE	178	0.604^{3rd}	0.607	0.602	0.531^{2nd}	0.563	0.501	0.264^{2nd}
	PEWCC	97	0.548	0.732	0.438	0.449	0.459	0.440	0.167
	CPredictor	150	0.506	0.640	0.418	0.426	0.402	0.452	0.196
	MCL	160	0.591	0.594	0.589	0.520^{3rd}	0.525	0.516	0.250^{3rd}
	CFinder	102	0.529	0.676	0.435	0.450	0.401	0.506	0.195

between proteins, based on Gene Ontology structure. The complexes detected by WCOACH generally had more proteins, so the number of small complexes is very rare, which leads to the high "hitting accuracy" relatively.

As for PPV, MCL algorithm utilizes random walk theory and Markov chains rule. It divides PPI network into many dense subgraphs; thus, every protein only belongs to one specific complex. So, the PPV is higher than our method. But the Precision and Recall of ours are all over than MCL.

BOPS achieved the highest MMR, indicating that the predictions matched the gold standard quite naturally. This shows that BOPS does not rely on increasing similar prediction results to improve the values of F-value and ACC, and BOPS has high biological experimental significance. And For comparing these methods more visually,

we plot Fig. 2 to show the F-score of each method. BOPS always obtains the highestF-score. Overall, the performance on the task of small protein complex identification is very promising. It obtains better results in both F-score, ACC and MMR in all datasets.

We recommend the applicability of each algorithm. If you need protein complex prediction to guide biological experiments, BOPS will be the best choice. But if you are short on funds and can only detect fewer protein complexes, then WCOACH will be the best choice. WCOACH has few predictions, but a high hit rate. When funding is replenished, the BOPS predictions can be used to detect more complexes.

The effects of parameter settings

Accuracy of graph segmentation algorithm

According to "Method", our method can be summarized as three steps. Firstly, it divides the PPI network into many subgraphs. This step will make the original network unrecoverable. Therefore, the quality of the segmented subgraphs affects the final performances directly. For evaluating the accuracy of this step, we define the expected regression ratio(ERR)to represent the accuracy of segmentation. Analogous to the definition of neighborhood affinity score $\mathrm{NA}(p, b)$, we define regression degree score $\mathrm{RD}(b, s)$ as follows

$$\mathrm{RD}(b, s) = \frac{|V_s \cap V_b|}{|V_b|} \tag{15}$$

where V_b is the set of proteins in the reference protein complex B and V_s is the set of proteins in the PPI sub-network. When $\mathrm{RD}(p, b)$ is not less than 0.25, we consider the b is recalled. Based of regression degree score, we define expected regression ratio (ERR) in Eps. 16 to evaluate the accuracy of segmentation:

$$ERR = \frac{\left|\left\{b | b \in B, \exists n \in N, RD(b, n) \geq \omega\right\}\right|}{\left|\left\{b | b \in B, RD(b, M) \geq \omega\right\}\right|} \tag{16}$$

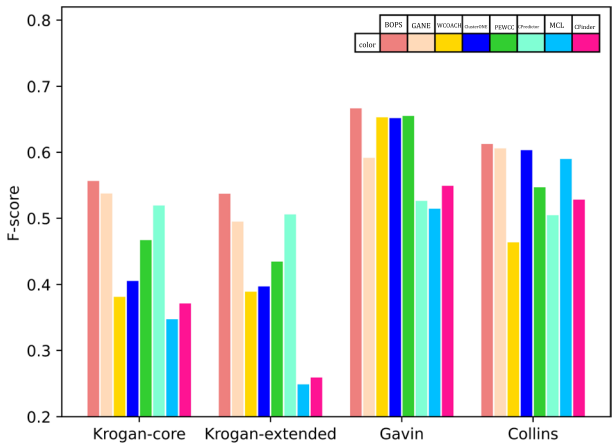

Fig. 2 Comparison with six protein complex identification algorithms in terms of F-score. Each bar height reflects the value of the F-score

where δ is specified to be 0.25 typically. b is a reference complex in the reference complex set B. n is a segmented PPI subgraph in the PPI subgraph set N. M is the original PPI network. We set the β change from 1.0 to 2.0 using a 0.2 increment, and get the result of ERR in four databases. At the same time, we execute the segmentation randomly and get the result as a reference. These results are listed in Table 3.

As shown in Table 3, the subgraphs divided by our method can obtain a much higher ERR than random. The expected regression ratio first increases and then decreases as the β increases and peaks at 1.4 or 1.6. Therefore, we set 1.6 as the default value and range from 1.4 to 1.6 as recommended an interval of β. When set default, the values are even more than 0.96 in the Gavin and Collins. It indicates our method achieves a high accuracy in these PPI networks. More than 96% of expected reference complexes are reserved in sub-networks. At the same time, we found the Recall in Gavin and Collins are more than Krogan-core and Krogan-extended. These results also confirmed the reliability of the data in Table 3. However, the ERR in Krogan-extended is 0.88. It is the least value in all networks. The reason why about 12% of the expected recall complexes are destroyed may be due to the large edge weight variance and the low PPIN density in the Krogan-extended. Considering in one graph, if the distribution of edge weights is discrete there are more edges will be deleted and if the graph is relatively sparse the possibility of deleting the correct edge will be increased. But compared with random segmentation, our method takes advantage of the weight and topology and achieves superior performance (0.880 vs 0.149).

The effect of β on performance of BOPS

As described before, there is only one parameter in BOPS: β. In order to investigate how the different parameters affect the performance of the protein complex identification [1, 44]. We try β changing from 1.0 to 2.0 to detect complexes in four datasets respectively. Considering that F-score generally reflect the accuracy of prediction sets, in Fig. 3, we plot the F-score with different parameters.

As the Fig. 3 shows, the value of the F-score shows a trend of increasing first and then decreasing. For Krogan-core and Krogan-extended datasets, the F-score reaches a peak when the balanced index is 1.4. For Gavin and Collins datasets, the F-score reaches a peak when the balanced index is about 1.8. That is because the values of edge weight variance in Krogan-core and Krogan-extended are higher than Collins and Gavin. The higher the balanced index is, the more discrete of the modified edge weights are. For Collins and Gavin datasets with more concentrated weights, a larger β is conducive to make the modified weights more decentralized, which is convenient for subsequent graph segmentation. Overall, when the balanced index is 1.5, it has a

Table 3 ERR on the four datasets

PPIN	Random	$\beta=1.0$	$\beta=1.2$	$\beta=1.4$	$\beta=1.6$	$\beta=1.8$	$\beta=2.0$
Krogan-core	0.444	0.764	0.901	0.917	0.921	0.903	0.877
Krogan-extended	0.149	0.733	0.856	0.870	0.888	0.861	0.863
Gavin	0.770	0.928	0.953	0.973	0.966	0.959	0.957
Collins	0.751	0.922	0.967	0.969	0.964	0.964	0.955

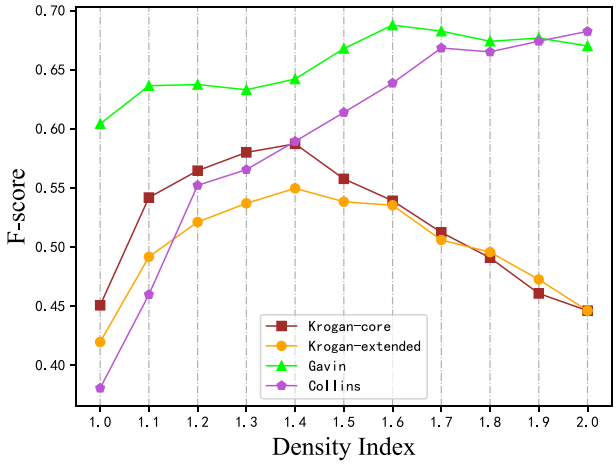

Fig. 3 The performances and quality of BOPS with different setting of β

Table 4 Performance comparision on four datasets(include large protein complexes)

	Krogan-core		Krogan-extended		Gavin		Collins	
	#Predicated	*F*-value	#Predicated	*F*-value	#Predicated	*F*-value	#Predicated	*F*-value
BOPS	710	0.596^{3rd}	793	0.582^{2nd}	851	0.729	813	0.737^{2nd}
GANE	208	0.674^{1st}	251	0.603^{1st}	182	0.594	202	0.759^{1st}
WCOACH	308	0.510	528	0.466	406	0.739^{3rd}	247	0.649
ClusterONE	600	0.476	972	0.456	240	0.769^{2nd}	207	0.701
PEWCC	283	0.600^{2nd}	464	0.548^{3rd}	401	0.772^{1st}	277	0.705^{3rd}
CPredictor	168	0.577	190	0.534	207	0.637	172	0.614
MCL	376	0.412	483	0.311	253	0.587	183	0.686
CFinder	114	0.412	120	0.234	137	0.628	113	0.575

good performance on the four datasets. For different datasets, we encourage to use a suitable balanced index according to edge weight variance in BOPS.

Adaptability of BOPS

The performance of BOPS on complexes of all sizes

In the previous section, we evaluate the performance of BOPS with small protein complexes identification. In order to make the experimental data more comprehensive, we compare it with other methods in the field of total protein complexes. Considering ACC is used to evaluate the overall performance in the field of quality, when one small identified complex is matched to a large reference complex, although all of the predicted proteins can be found in the reference complex, the Sn will be very low. This situation is inconsistent with our original intention. Thus, ACC is not a fair reference standard. Therefore, we evaluate F-score with a set of gold standard protein complexes (789 protein complexes totally). The results are listed in Table 4. Overall, the performance of BOPS is better than most algorithms with respect to the whole protein complexes.

Table 5 The Homo sapiens PPIN datasets used in the experiment

PPIN	#Proteins	#Interactions	Edge weight average	Edge weight variance	PPIN density
HomoSTRING	8654	97674	0.84259	0.03680	0.00261

Table 6 Performance comparision in Homo sapiens PPIN

Method	#Predicated	F-score	Precision	Recall	ACC	Sn	PPV	F-score + ACC
BOPS	2140	0.307	0.296	0.318	0.274	0.395	0.190	0.581^{1st}
PEWCC	1584	0.346^{1st}	0.321	0.374	0.211	0.567	0.079	0.557
ClusterONE	798	0.141	0.175	0.118	0.298^{1st}	0.268	0.331	0.439

Performance comparison in the Homo Sapiens PPIN

The STRING database [45] aims to integrate all known and predicted associations between proteins, including both physical interactions as well as functional associations. BioGRID [46] is a biomedical interaction repository with data compiled through comprehensive curation efforts. An unweighted Homo sapiens PPIN is obtained from BioGRID, which contains 206930 interactions. And interactions in PPIN are assigned weights based on STRING database. If an interaction can not be found in STRING database, it will be removed from PPIN. In that way, a weighted Homo sapiens PPIN is constructed, and the detail is shown in Table 5.

BOPS is compared to PEWCC and ClusterONE which are the second best in the Yeast experiment. The parameter β is still set to the default value of 1.5. And the gold standard is Corum [47]. The result shows in Table 6. In the Homo sapiens PPIN, BOPS shows the best overall performance. BOPS has strong adaptability between different species.

Biological significance of the identification protein complex

To assess the biological sense of the predicted protein complexes generated by BOPS, we calculate the Min *P*-value by the tool GOTermFinder [48]. *P*-value is defined as follows:

$$P{-}value = 1 - \sum_{i=0}^{k-1} \frac{\binom{|F|}{i}\binom{|V|-|F|}{|C|-i}}{\binom{V}{C}} \tag{17}$$

where a predicted complex C contains k proteins in the functional group F and the whole PPI network contains $|V|$ protein. The functional homogeneity of a predicted complex is the Min *P*-value overall of the possible functional groups. A predicted complex with a low functional homogeneity indicates it is enriched by proteins from the same function group [49]. So, the collective occurrence of these proteins in a complex does not occur merely by chance [50].

We counted the distribution of the negative logarithm of the *P*-value of the unmatched protein complexes predicted and plotted it into Fig. 4. The heatmap shows that most *P*-values of BOPS are less than 1e-5, which indicates that these unmatched complexes in krogan-extended also have high biological significance. In addition, the *P*-values of GANE and WCOACH are small compared to BOPS, which may be because these

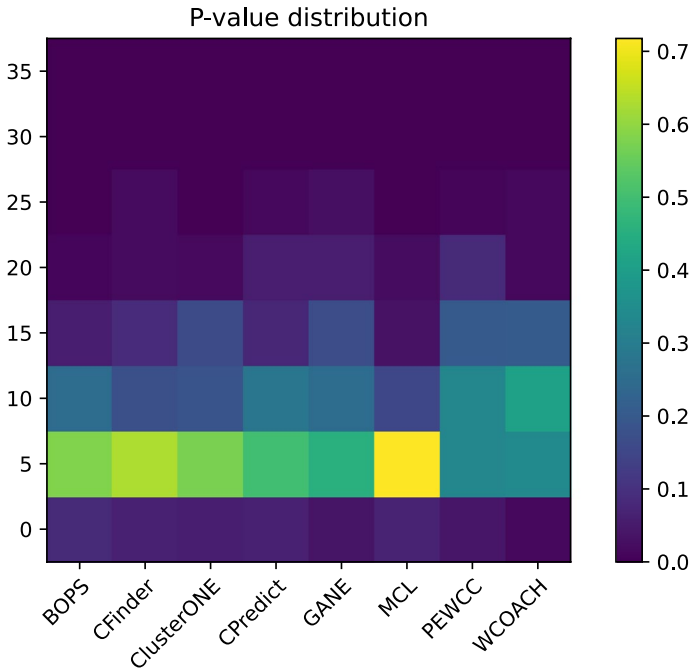

Fig. 4 The heatmap of unmatched complexes' *P*-values

methods consider GO information in the process of the complex prediction. In a word, the predicted results of BOPS showed a higher biological function.

Conclusion

In this paper, a protein complex prediction algorithm based on graph segmentation, BOPS is proposed. Firstly, the BOPS algorithm calculates the balanced weight. Secondly, the BOPS algorithm divides the original PPIN into small networks. Thirdly, the BOPS algorithm enumerates the connected subset of each small network and determines whether it is a protein complex based on the cohesion of the subset.

The experimental performance proves that the BOPS algorithm can obtain the best results when identifying small protein complexes. And the performance of BOPS is better than most algorithms for the whole protein complexes. In addition, we constructed a weighted Homo sapiens PPIN based on STRINGdb and BioGRID, and provided more data for related research.

At the same time, we convert a generative problem into a decision problem by splitting PPIN into small PPINs and enumerating connected subsets. We have succeeded in segmenting PPIN, retaining most of the protein complexes. We believe that graph segmentation can be combined with many other algorithms to make better results in the future. And the way to solve problems by converting a generative problem into a decision problem is firstly introduced into protein complex prediction. We believe that this method will have greater application prospects in the future.

In the future, we will attempt to improve the performance of BOPS, develop a better graph segmentation algorithm, apply the convert way to more problems, and focus on combining the identified proteins' structural information with BOPS to assess the structural compatibility of predicted protein complexes.

Abbreviations

PPI Protein–protein interaction
GO Gene Ontology

Acknowledgements
The authors wish to thank the reviewers for their helpful suggestions.

Author contributions
JL designed and developed the BOPS algorithm. ZY designed the evaluation algorithm and evaluated the result. Both JL and ZY made substantial contributions to acquisition of data, analysis and interpretation of data. Both BL, YL and YZ provided suggestions and support for the research process. All authors read and approved the final manuscript.

Funding
This work is supported by grant from the Natural Science Foundation of China (No. 62072070) and College Student Innovation and Entrepreneurship Training Program Support Project of China (No. 2019101411600010093, No. 202010141160001012 9).

Availability of data and materials
All datasets and the source code of BOPS are available at https://github.com/jiaqinglv2000/BOPS.

Declaration

Competing interests
The authors declare that they have no competing interests.

Received: 5 May 2022 Accepted: 19 September 2022
Published online: 30 September 2022

References

1. Zhang X-F, Dai D-Q, Li X-X. Protein complexes discovery based on protein–protein interaction data via a regularized sparse generative network model. IEEE/ACM Trans Comput Biol Bioinform. 2012;9(3):857–70. https://doi.org/10.1109/TCBB.2012.20.
2. Zahiri J, Emamjomeh A, Bagheri S, Ivazeh A, Mahdevar G, Tehrani HS, Mirzaie M, Fakheri BA, Mohammad-Noori M. Protein complex prediction: a survey. Genomics. 2020;112(1):174–83. https://doi.org/10.1016/j.ygeno.2019.01.011.
3. Dias DM, Ciulli A. NMR approaches in structure-based lead discovery: recent developments and new frontiers for targeting multi-protein complexes. Prog Biophys Mol Biol. 2014;116(2–3):101–12. https://doi.org/10.1016/j.pbiomolbio.2014.08.012.
4. Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B. Proteome survey reveals modularity of the yeast cell machinery. Nature. 2006;440(7084):631–6. https://doi.org/10.1038/nature04532.
5. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. Nature. 2006;440(7084):637–43. https://doi.org/10.1038/nature04670.
6. Hopper JT, Robinson CV. Mass spectrometry of intact protein complexes. Proteomics Biol Discov. 2019. https://doi.org/10.1002/9781119081661.ch6.
7. Hale OJ, Illes-Toth E, Mize TH, Cooper HJ. High-field asymmetric waveform ion mobility spectrometry and native mass spectrometry: analysis of intact protein assemblies and protein complexes. Anal Chem. 2020;92(10):6811–6. https://doi.org/10.1021/acs.analchem.0c00649.
8. Hale OJ, Sisley EK, Griffiths RL, Styles IB, Cooper HJ. Native LESA TWIMS-MSI: spatial, conformational, and mass analysis of proteins and protein complexes. J Am Soc Mass Spectrom. 2020;31(4):873–9. https://doi.org/10.1021/jasms.9b00122.
9. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci. 2001;98(8):4569–74. https://doi.org/10.1073/pnas.061034498.
10. Marx V. Proteomics: taking on protein complexes. Nat Methods. 2016;13(9):721–7. https://doi.org/10.1038/nmeth.3966.
11. Guo Y, Shang X, Zhu Q, Huang M, Li Z. Identification of protein complexes and functional modules in integrated ppi networks. In: IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2014. p. 8–15. https://doi.org/10.1109/BIBM.2014.6999291.
12. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402(6761):47–52. https://doi.org/10.1038/35011540.
13. Barabasi A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101–13. https://doi.org/10.1038/nrg1272.
14. Wu M, Li X-L, Kwoh C-K, Ng S-K, Wong L. Discovery of protein complexes with core-attachment structures from tandem affinity purification (tap) data. J Comput Biol. 2012;19(9):1027–42. https://doi.org/10.1089/cmb.2010.0293.
15. Wang J, Li M, Deng Y, Pan Y. Recent advances in clustering methods for protein interaction networks. BMC Genomics. 2010;11(3):1–19. https://doi.org/10.1186/1471-2164-11-S3-S10.

16. Srihari S, Leong HW. A survey of computational methods for protein complex prediction from protein interaction networks. J Bioinform Comput Biol. 2013;11(02):1230002. https://doi.org/10.1142/S021972001230002X.
17. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol. 2007;3(1):88. https://doi.org/10.1038/msb4100129.
18. Meyer D, Bonhoeffer T, Scheuss V. Balance and stability of synaptic structures during synaptic plasticity. Neuron. 2014;82(2):430–43. https://doi.org/10.1016/j.neuron.2014.02.031.
19. Zohar R, Suzuki N, Suzuki K, Arora P, Glogauer M, McCulloch C, Sodek J. Intracellular osteopontin is an integral component of the CD44-ERM complex involved in cell migration. J Cell Physiol. 2000;184(1):118–30. https://doi.org/10.1002/(SICI)1097-4652(200007)184:1<118::AID-JCP13>3.0.CO;2-Y.
20. Sackstein R, Merzaban JS, Cain DW, Dagia NM, Spencer JA, Lin CP, Wohlgemuth R. Ex vivo glycan engineering of cd44 programs human multipotent mesenchymal stromal cell trafficking to bone. Nat Med. 2008;14(2):181–7. https://doi.org/10.1038/nm1703.
21. Alves CS, Burdick MM, Thomas SN, Pawar P, Konstantopoulos K. The dual role of CD44 as a functional P-selectin ligand and fibrin receptor in colon carcinoma cell adhesion. Am J Physiol Cell Physiol. 2008;294(4):907–16. https://doi.org/10.1152/ajpcell.00463.2007.
22. Li F, Tiede B, Massagué J, Kang Y. Beyond tumorigenesis: cancer stem cells in metastasis. Cell Res. 2007;17(1):3–14. https://doi.org/10.1038/sj.cr.7310118.
23. Reinacher M, Eigenbrodt E. Immunohistological demonstration of the same type of pyruvate kinase isoenzyme (M2-Pk) in tumors of chicken and rat. Virchows Archiv B. 1981;37(1):79–88. https://doi.org/10.1007/BF02892557.
24. French CA. Demystified molecular pathology of nut midline carcinomas. J Clin Pathol. 2010;63(6):492–6. https://doi.org/10.1136/jcp.2007.052902.
25. Li X, Wu M, Kwoh C-K, Ng S-K. Computational approaches for detecting protein complexes from protein interaction networks: a survey. BMC Genomics. 2010;11(1):1–19. https://doi.org/10.1186/1471-2164-11-S1-S3.
26. Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature. 2005;435(7043):814–8. https://doi.org/10.1038/nature03607.
27. Zaki N, Efimov D, Berengueres J. Protein complex detection using interaction reliability assessment and weighted clustering coefficient. BMC Bioinform. 2013;14(1):1–9. https://doi.org/10.1186/1471-2105-14-163.
28. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein–protein interaction networks. Nat Methods. 2012;9(5):471–2. https://doi.org/10.1038/nmeth.1938.
29. Pereira-Leal JB, Enright AJ, Ouzounis CA. Detection of functional modules from protein interaction networks. Proteins Struct Funct Bioinform. 2004;54(1):49–57. https://doi.org/10.1002/prot.10505.
30. Chen B, Fan W, Liu J, Wu F-X. Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks. Brief Bioinform. 2014;15(2):177–94. https://doi.org/10.1093/bib/bbt039.
31. Wu M, Li X, Kwoh C-K, Ng S-K. A core-attachment based method to detect protein complexes in PPI networks. BMC Bioinform. 2009;10(1):1–16. https://doi.org/10.1186/1471-2105-10-169.
32. Kouhsar M, Zare-Mirakabad F, Jamali Y. WCOACH: protein complex prediction in weighted PPI networks. Genes Genetic Syst. 2016. https://doi.org/10.1266/ggs.15-00032.
33. Xu B, Li K, Zheng W, Liu X, Zhang Y, Zhao Z, He Z. Protein complexes identification based on go attributed network embedding. BMC Bioinform. 2018;19(1):1–10. https://doi.org/10.1186/s12859-018-2555-x.
34. Xu B, Wang Y, Wang Z, Zhou J, Zhou S, Guan J. An effective approach to detecting both small and large complexes from protein–protein interaction networks. BMC Bioinform. 2017;18(12):19–28. https://doi.org/10.1186/s12859-017-1820-8.
35. Asur S, Ucar D, Parthasarathy S. An ensemble framework for clustering protein-protein interaction networks. Bioinformatics. 2007;23(13):29–40. https://doi.org/10.1093/bioinformatics/btm212.
36. Bhowmick SS, Seah BS. Clustering and summarizing protein–protein interaction networks: a survey. IEEE Trans Knowl Data Eng. 2015;28(3):638–58. https://doi.org/10.1109/TKDE.2015.2492559.
37. Wang X, Li J, Guo X, Ma Y, Qiao Q, Guo J. PLWRKY13: a transcription factor involved in abiotic and biotic stress responses in *Paeonia lactiflora*. Int J Mol Sci. 2019;20(23):5953. https://doi.org/10.3390/ijms20235953.
38. Wang R, Liu G, Wang C, Su L, Sun L. Predicting overlapping protein complexes based on core-attachment and a local modularity structure. BMC Bioinform. 2018;19(1):1–15. https://doi.org/10.1186/s12859-018-2309-9.
39. Wang R, Wang C, Liu G. A novel graph clustering method with a greedy heuristic search algorithm for mining protein complexes from dynamic and static PPI networks. Inf Sci. 2020;522:275–98. https://doi.org/10.1016/j.ins.2020.02.063.
40. Liu X, Yang Z, Sang S, Lin H, Wang J, Xu B. Detection of protein complexes from multiple protein interaction networks using graph embedding. Artif Intell Med. 2019;96:107–15. https://doi.org/10.1016/j.artmed.2019.04.001.
41. Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon A-M, Cruciat C-M. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002;415(6868):141–7. https://doi.org/10.1038/415141a.
42. Collins SR, Kemmeren P, Zhao X-C, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ. Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. Mol Cell Proteomics. 2007;6(3):439–50. https://doi.org/10.1074/mcp.M600381-MCP200.
43. A Maddi AM, Ahmadi Moughari F, Balouchi MM, Eslahchi C. CDAP: an online package for evaluation of complex detection methods. Sci Rep. 2019; 9(1), 1–13.
44. He T, Chan KC. Evolutionary graph clustering for protein complex identification. IEEE/ACM Trans Comput Biol Bioinform. 2016;15(3):892–904. https://doi.org/10.1109/TCBB.2016.2642107.
45. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49(D1):605–12.
46. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34(suppl_1):535–539

47. Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Ruepp A. CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. 2019;47(D1):559–63.
48. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. GO: TermFinder-open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. Bioinformatics. 2004;20(18):3710–5.
49. Xu B, Liu Y, Lin C, Dong J, Liu X, He Z. Reconstruction of the protein–protein interaction network for protein complexes identification by walking on the protein pair fingerprints similarity network. Front Genet. 2018;9:272. https://doi.org/10.3389/fgene.2018.00272.
50. Yan JY, Li CX, Sun L, Ren JY, Li GX, Ding ZJ, Zheng SJ. A WRKY transcription factor regulates Fe translocation under Fe deficiency. Plant Physiol. 2016;171(3):2017–27. https://doi.org/10.1104/pp.16.00252.

Publisher's Note