BMC Bioinformatics

RESEARCH ARTICLE

Open Access

Protein complexes detection based on node local properties and gene expression in PPI weighted networks

Yang Yu* and Dezhou Kong

*Correspondence:
yuyangsd1204@126.com
Software College, Shenyang Normal University, Shenyang 110034, People's Republic of China

Abstract

Background: Identifying protein complexes from protein–protein interaction (PPI) networks is a crucial task, and many related algorithms have been developed. Most algorithms usually employ direct neighbors of nodes and ignore resource allocation and second-order neighbors. The effective use of such information is crucial to protein complex detection.

Result: Based on this observation, we propose a new way by combining node resource allocation and gene expression information to weight protein network (NRAGE-WPN), in which protein complexes are detected based on core-attachment and second-order neighbors.

Conclusions: Through comparison with eleven methods in Yeast and Human PPI network, the experimental results demonstrate that this algorithm not only performs better than other methods on 75% in terms of f-measure+, but also can achieve an ideal overall performance in terms of a composite score consisting of five performance measures. This identification method is simple and can accurately identify more complexes.

Keywords: Protein complex, Protein–protein interaction (PPI), Resource allocation, Weighted graph construction

Background

Proteins are the basis of biological activities, and their functions are generally expressed by the interactions between proteins [1]. In organisms, protein–protein interaction (PPI) networks consist of proteins and protein interactions. PPI networks provide an elegant means for expressing gene regulation and metabolic pathways in complex biological systems [2]. Protein complexes are the locally dense regions of PPI networks and possess graph-like structures in which a node represents a protein and an edge represents interaction between two proteins [3].

Complexes take part in many diverse biochemical activities that are fundamental to all kinds of functions, such as cell homeostasis, cell cycle control, growth, and proliferation. Moreover, specific functional modules usually are related to certain diseases.

Although great progress has been made in identifying protein complexes, laboratory-based methods are expensive, ineffective and sometimes even infeasible, and only parts of protein complexes are located. In addition, experiments in the laboratory are often incomplete because of the constraints of experimental conditions. As it is necessary to overcome the lacking of laboratory-based methods, a large number of computational algorithms have been designed as alternative methods to identify protein clusters, such as density-based clustering [4–8], hierarchical clustering[8–10], partition-based clustering [11, 12], flow simulation-based clustering[13–16] and other methods with integrating biological and topological multiple information [17–20]. Although methods of protein complexes detection have achieved some effective results, how to reasonably integrate PPI node local data and gene expression biological information to construct weighted graphs, and how to define effective detection methods to identify complexes from the weighted network still need further study. Only direct neighbors are applied to PPI network clustering problems, which is not sufficient. In fact, node resource allocation information and second-order neighbors often contain some important potential information in PPI networks.

Aiming at the solution for the above-mentioned problems, we introduce a novel method based on resource allocation and gene expression in weighted PPI networks (called NRAGE-WPN) with based on core-attachment structure and second-order neighbors searching. First, based on the resource allocation and gene expression of the PPI network, a new weight metric is designed to accurately describe the interaction between proteins. Then our method detects a series of dense complex cores based on density and network diameter constraints and the final complexes are recognized by expanding the second-order neighbors of nodes in core complexes. This identification method is simple and can accurately identify more complexes.

Methods

Protein complex detection with a computational approach from PPI data is useful as the useful supplement to the limited experimental methods. Besides the enhancement in graph clustering techniques, successful and accurate methods for protein complex prediction depends more on the construction of weighted graphs. Therefore, constructing weighted graph for protein interactions is essential. In this section, we introduce a novel method based on resource allocation and gene expression in weighted PPI networks with two main steps. First, a method is proposed to evaluate the reliability of the protein interaction data considering both the common neighbor information and gene expression profiles through the weighted graph construction. Second, protein complexes are detected based on core-attachment and second-order neighbors in this new weighted graph. The workflow of our method is shown in Fig. 1.

Assessing the reliability of protein interaction

To represent a PPI network, a 3-element tuple $G = (\mathrm{V}, E, \mathrm{W})$ is employed, where $V = (V_i)(1 \leq i \leq N)$ is a set of N proteins, and $E = \{e_{ij}\}$ is the set of PPI edges whose values are stored in matrix W. For each pair of nodes, $i, j \in V$ and the edge e_{ij} is assigned a score as w_{ij}. Inspired by the reference [21], resource allocation index (RA), is

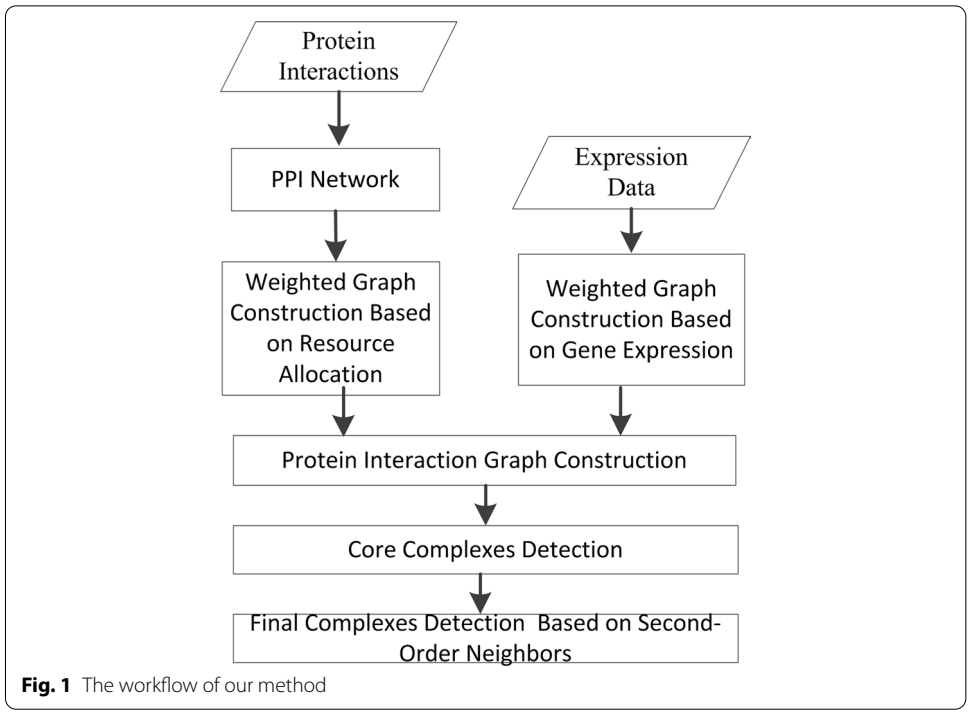


Fig. 1 The workflow of our method

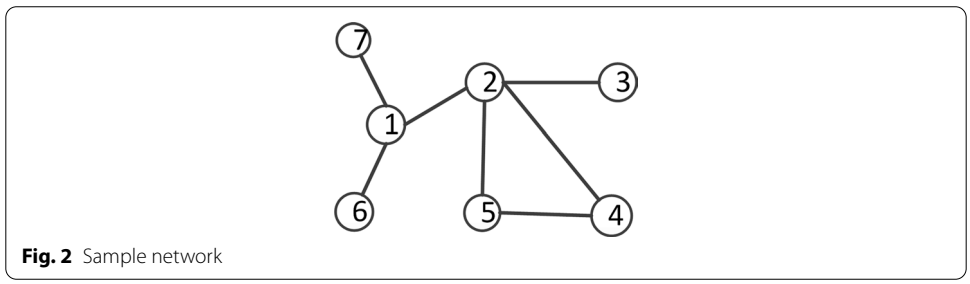


Fig. 2 Sample network

introduced to measure the similarity of interaction proteins in a network and a weighted graph based on resource allocation (WRA) is constructed in this step.

Taking Fig. 2 as an example, there is an edge between node 1 and node 2 and no common neighbors between them, but e_{12} is an important bridge for information transmission between node group{1, 2, 6, 7} and node group{1, 2, 3, 4, 5}. Simply, it is assumed that the transmitter 1 can carry resources, and will equally deliver it among all its neighbors. Based on this, the similarity of two nodes is shown in Eq. (1). We can consider node i and node j, which are directly connected without common neighbors and the node i can transmit the information to node j through edge e_{ij} to help the communication between two clusters {1, 2, 6, 7} and {1, 2, 3, 4, 5}. The value range of WRA belongs to [0 1]. This measure requires only the information of the nearest neighbors which therefore has very low computational complexity. $N(i)$ is the set of the neighbors of node i and node i, N(j) is the set of the neighbors of node j and node j.

$$WRA_{\mathrm{ij}} = \sum_{u \in N(i) \cap N(j)} \frac{1}{\mathrm{N}(u)} \tag{1}$$

$$W_N = \left\{ WRA_{\mathrm{ij}} \right\} \tag{2}$$

Pearson's correlation of expression levels

Co-expression genes tend to encode interacting proteins [22]. In this paper, we mainly concentrate on linear gene expression networks unless explicitly stated otherwise and Pearson's correlation coefficient of expression levels (PCC) is employed as biological information for interacting protein pair p and q. According to GBA principle (i.e. genes with similar expression spectrums have similar biological functions) [23], a higher correlation suggests a higher confidence in their interaction. PCC is generally used to measure the strength of the linear relationship between two variables and is also commonly used to measure the linear relationship between two sets of gene expression values. Suppose there are two columns of gene expression profiles $\mathrm{X} = (x_1, \ldots, x_n)$ and $\mathrm{Y} = (\mathrm{y}_1, \ldots, \mathrm{y}_n)$. Matrix W_{p} is formed by the PCC calculation formula, which is defined in Eq. (3). The value range of PCC belongs to [−1 1]. If PCC (X,Y) <0, it means that gene X and Y show a negative correlation; if PCC (X,Y) >0, it means gene X and Y show a positive correlation, PCC (X, Y)=0 means that there is no correlation between genes X and Y. If PCC(X, Y)<0, protein pairs will be removed from PPI network in order to reduce the negative effect of low noise data on the detection results of mining protein complexes. The value range [0 1] of PCC is employed in this step.

$$\mathrm{PCC}_{\mathrm{ij}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{3}$$

$$W_{\mathrm{p}} = \left\{ PCC_{ij} \right\} \tag{4}$$

where $\bar{x}$ denotes the average value of the expression value of gene X at 36 different times and $\bar{y}$ denotes the average value of the expression value of gene Y at 36 different times.

Weighted graph construction

In this part, we first describe how to compute the weighted value by combining gene expression information (GEI) based on PCC and RA information between two interaction proteins. The final weighted construction formula is proposed in Eq. (5).

$$W = \alpha W_{\mathrm{P}} + (1 - \alpha) W_{\mathrm{N}} \tag{5}$$

Matrix W_P is constructed based on Pearson correlation coefficient and matrix W_N is constructed based on RA, respectively. After a simple calculation, the range of values can be known from 0 to 2. The final values are normalized to [0 1]. $\alpha (0 \leq \alpha \leq 1)$ is a constant, where a smaller α indicates that the importance of the modules is dependent more on RA information of the network, and a bigger α indicates that the importance of the modules depends more on gene expression information. When $\alpha = 0$, the weighted

method only considers RA information. When $\alpha = 1$, the weighted method only considers gene expression information. Therefore the Eq. (5) can measure the differential importance of interaction in protein networks by integrating node local information and biological information.

Detecting protein complexes in weighted graphs

The proposed algorithm, NRAGE-WPN, consists of two phases: weighted graph construction and core-attachment protein complex detection based on second-order neighbors searching. In the weighted graph construction phase, gene expression information and common neighbor information are integrated. A detailed description of the algorithm is outlined in Algorithm 1. Line 1 is for constructing matrix W_N with the given PPI datasets. Line 2 is for constructing matrix W_{p} with the gene expression data. Line 3 is for constructing the new matrix W based on W_N and W_{p}, and the protein interaction confidence is the sum of the weights of W_N and W_{p}. Lines 4–8 are for identifying core clusters. Lines 9–11 are for enlarging core clusters based on second-order neighbors of nodes in each core.

Algorithm 1 Prediction of Protein Complexes

Input: The PPI network G=(V, E, W)

PPI: protein-protein interaction network

GE: gene expression data; parameter: α

Output: Detected protein complexes(DP)

Description:

1. Constructing matrix W_N based on RA using Eqs. (1) and (2).

2. Constructing matrix W_{p} based on PCC using Eqs. (3) and (4).

3. Constructing final matrix W and forming final graph G= (V, E, W) using Eq. (5).

4. Forming nodes set Q according to the descending degree of node in weighted Graph G.

5. Selecting a node in Q as a new cluster.

6. Traversing every neighbor of this node.

7. If a neighbor node meets the condition (7) and it is added to the current cluster and updating this new cluster. Go to step 6 until traversing all neighbors to form a core cluster C_i.

8. Determining the next node to be processed in Q, repeat 5-7 until traversing all nodes in Q and forming core complex set $C=\{C_1, C_2, \ldots\ldots, C_{\mathrm{m}}\}$.

9. Enlarging the core complexes based on second-order neighbors searching of node in C_i,

10. If a node in core complex has second-order neighbors and meets the condition (7) , enlarge this core complex by adding second-neighbors and form final complexes

11. Repeating step 9, 10, enlarging every core complex and forming final protein clusters $DP=\{DP_1, DP_2, \ldots\ldots, DP_m\}$.

In this algorithm, density and diameter are employed as the condition for complex detection.

If a node meets the two constraints in condition (7), it is added to the current cluster (subgraph). Generally, λ is usually set to 0.7 and δ is set to 2, according to the references [12, 24].

(1) Density: The degree of a node V is the sum of the weights for each edge connecting to this node. Density in the weighted subgraph G = (V, E) is defined in (6). |N| is the number of nodes in G and w(e) is the weight of the edge e_{ij} in G.

$$\begin{aligned} \mathrm{m} &= \sum_{e_{ij} \in E} w(e) \\ density(G) &= \frac{2 * m}{(|\mathrm{N}| * (|N| - 1))} \end{aligned} \tag{6}$$

(2) Network Diameter: Diameter is the shortest path in a cluster.

$$diameter \leq \delta \; and \; density \geq \lambda \tag{7}$$

Results

Datasets

The effectiveness of our method is evaluated using PPI networks and gold standards of protein complexes from yeast and human and the detail information is shown in Table 1 and relative detail information can be find in reference [25]. GSE3431 dataset [26] is employed in our paper which records the data of 36 time points during three successive metabolic cycles.

Evaluation criteria

To evaluate our method on benchmark datasets and compare NRAGE-WPN with other methods, evaluation measures are given in this section, such as sensitivity (SN), positive predictive value (PPV), accuracy (ACC), separation (SEP), fraction match (FRM), maximum matching ratio(MMR), precision (Prec), recall (Rec) and f-measure, precision+, recall+, f-measure+, the sum (F_MMR) of MMR and f-measure+, the composite score(CS) of MMR, FRM, SEP, ACC and f-measure [25]. Given a set of benchmark protein complexes $R = \{R_1, R_2, \ldots, R_n\}$ and a set of predicted clusters $P = \{P_1, P_2, \ldots, P_n\}$, two protein complexes, namely, R_i and P_j, are generated from benchmark complex datasets R and predicted protein complex sets P, respectively. T_{ij} is the number of proteins in common between ith benchmark complex R_i and jth predicted complex P_j. SN, PPV and ACC are defined as follows.

Table 1 PPI networks and gold standards in our experiments

	Yeast	Human
PPI networks	Collins [36],Gavin [37], Krogan core [38], Krogan extended [29]	STRING [39], PIPS [40]
Gold standards	CYC2008 [41], MIPS [42]	Corum [43]

N_i presents the size of proteins in the ith benchmark module. Here, n is the number of benchmark complexes and m is the number of predicted complexes.

$$SN = \frac{\sum_{i=1}^{n} \max_{j}\{T_{ij}\}}{\sum_{i=1}^{n} N_i} \quad PPV = \frac{\sum_{j=1}^{m} \max_{i}\{T_{ij}\}}{\sum_{j=1}^{m}\sum_{i=1}^{n} T_{ij}} \quad ACC = \sqrt{Sn \times PPV} \tag{8}$$

To evaluate protein complex prediction in terms of precision and recall, the Jaccard index is employed. The located complex P_j is defined to match the real complex R_i if the Jacquard similarity is greater than 0.5.

$$\begin{aligned} \text{Jaccard}(P_j, R_i) &= \frac{|P_j \cap R_i|}{|P_j \cup R_i|} \quad precision = \frac{|\{P_j \in P | \exists R_i \in R, P_j matches\, R_i\}|}{m} \\ recall &= \frac{|\{R_i \in R | \exists P_j \in P, P_j matches R_i\}|}{n} \quad f - measure = \frac{2 * recall * precision}{recall + precision} \end{aligned} \tag{9}$$

In terms of precision+, recall+ and f-measure+, neighborhood affinity score $NA(P_j, R_i)$ between P_j and R_i, as defined in Eq. (10) can be used to determine whether they match with each other. If $NA(P_j, R_i) = \omega$, $\omega \geq t$, ω is greater than 0.2, P_j and R_i are considered to be matching. In this paper, t is usually set as 0.20. $|\mathrm{P_i}|$ $|P_i|$ and $|\mathrm{R_j}|$ $|R_j|$ are the numbers of proteins in $\mathrm{P_i}$ P_i and R_j, respectively.

$$NA(P_j, R_i) = \frac{|P_j \cap R_i|^2}{|P_j| * |R_i|} \tag{10}$$

$$\begin{aligned} recall^{+} &= \frac{|\{R_j | R_j \in R \wedge P_i \in P, P_i matches R_j\}|}{n} \\ precision^{+} &= \frac{|\{P_i | P_i \wedge R_j \in R, R_j matches P_i\}|}{m} \\ f - measure^{+} &= \frac{2 * recall^{+} * precision^{+}}{recall^{+} + precision^{+}} \end{aligned} \tag{11}$$

Comparison with other methods

To inspect the performance of our proposed algorithm, we compare our algorithm with MCODE [6], Cfinder [4], ClusterOne [20], ProRank+ [27], MCL [28], PC2P [25], CLE [7], CW [8], CLP [29], CI [13], DPCT [30] in different measures as shown in Additional file 1: Table S1 and all the weighted graphs are constructed based on Eq. (5). Comparison results about CS measure in four PPI networks of Yeast on CYC2008 are shown in Fig. 3.

Comparative analysis is performed with the sum score of MMR, FMR, SEP, ACC and f-measure. Performances among different methods are compared for yeast and human with the corresponding complex datasets and PPI networks. First, as is illustrated in Fig. 3 that NRAGE-WPN can achieve best performance in Collins, Gavin and KroganExt network and perform better than other ten methods except PC2P in KroganCore in terms of CS on CYC2008. On MIPS, NRAGE-WPN outperforms all methods on MIPS in network Collins and ten methods in network Gavin, KroganExt and KroganCore except PC2P in Additional file 1: Table S1. On CORUM in

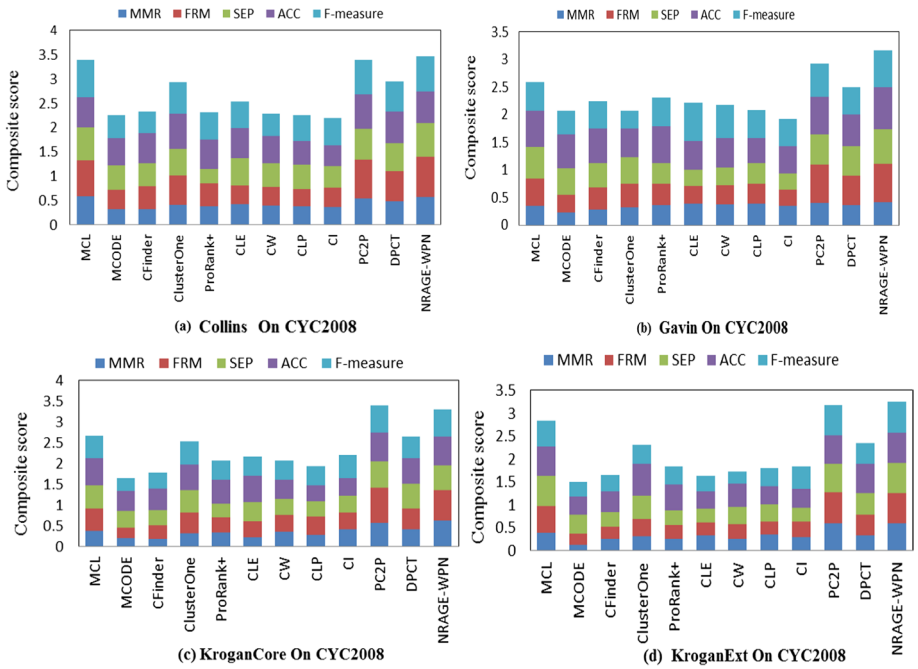
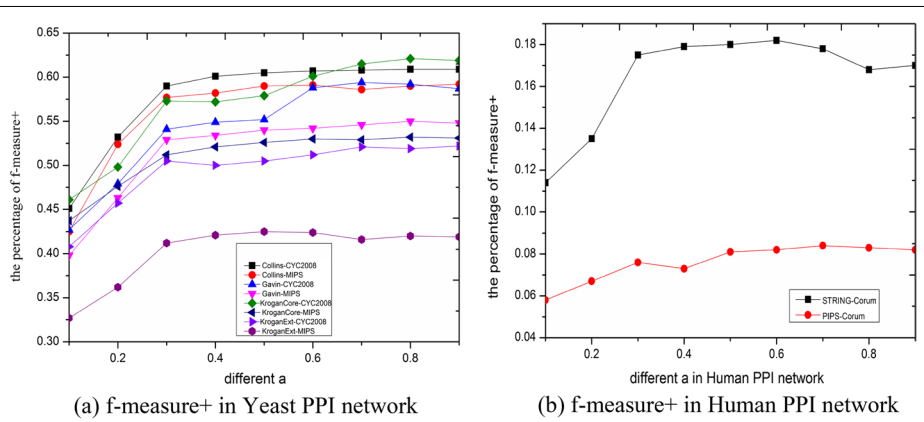


Fig. 3 Comparisons of four yeast PPI networks on CYC2008



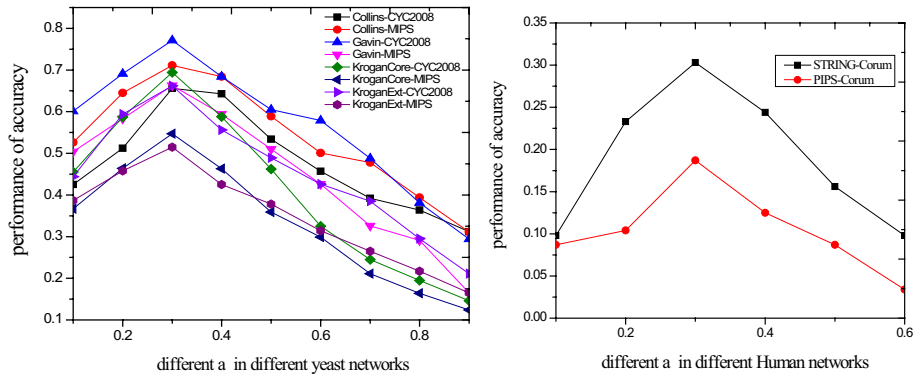
(a) f-measure+ in Yeast PPI network (b) f-measure+ in Human PPI network

Fig. 4 f-measure+ in yeast and human for different parameter *α*

2 combinations, NRAGE-WPN can achieve best performances in terms of CS. Second, in terms of f-measure+, NRAGE-WPN results the best performance except in Collins on MIPS. Third, in the rest measures, NRAGE-WPN performs better than most other methods and the all detail information can be shown in Additional file 1: Table S1.

Assessment performances of f-measure+ and accuracy with parameter α

By evaluating the importance of parameter α, we can more intuitively observe the influence of a certain parameter on the experimental results, and it is helpful to understand the advantages and disadvantages of the algorithm and enhance it. The critical parameter



(a) accuracy in Yeast PPI network (b) accuracy in Human PPI network

Fig. 5 Accuracy yeast and human for different parameter *a*

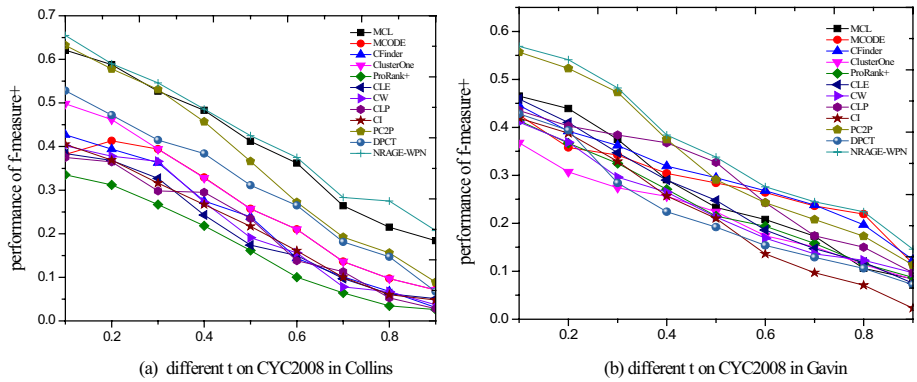


Fig. 6 f-measure+ in yeast for different threshold t

α in our method is mainly employed to show the effectiveness of information fusion from local neighbors and gene expression information and to affect the detection results of protein complexes. This experiment investigates the effects of different parameters α from 0.1 to 0.9 at interval of 0.1 on complex detection performance. Using f-measure+ and accuracy as our experimental evaluation criterion, the performances with different α are evaluated as shown in the Figs. 4 and 5, respectively. In Fig. 4,when the parameter α is greater than or equal to 0.3, the f-measure+ tends to stable. In In Fig. 5, when $\alpha = 0.3$, the best performance of accuracy can be achieved. In this article, we take $\alpha = 0.3$.

Robustness to the different thresholds (t)

In order to illustrate the comprehensive performance of NRAGE-WPN, we demonstrate f-measure+ performances with nine thresholds $t = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ among different methods in Fig. 6. Figure 6a shows the comparisons of f-measure+ performances on the CYC2008 benchmark dataset in Collins. It can be illustrated that NRAGE-WPN outperforms other eleven methods. Similar results can also be found on the CYC2008 benchmark in Gavin in Fig. 6b. Other comparisons are shown in Additional file 1: Fig. S1, which illustrates that NRAGE-WPN performs better than other

combinations on 50%. This further demonstrates the effectiveness of the fusion information from local node and gene expression data.

Discussion

Functional analysis

For the protein complexes identified by the NRAGE-WPN algorithm, we measure the effectiveness of the algorithm quantitatively and qualitatively. We analyze the biological significance of the identified protein complexes. Real protein complexes often present high functional homogeneity, so the function enrichment test is employed to demonstrate the biological significances of detected protein complexes [31]. The function enrichment analysis of protein complexes identified from yeast PPI network is carried out to further verify the effectiveness of NRAGE-WPN algorithm. The analysis and comparison of *P* value are shown in Table 2. *P* value of each complex can be divided into one of four intervals from small to large: < E-15, [E-15, E-10], [E-10, E-5], [E-5, 0.001]. When *P* value is greater than 0.001, it is generally considered that the function of the complex is very likely to be randomly assigned and has no biological significance. The percentages in brackets in Table 2 indicate the ratio of the number of complexes in a certain interval to the number of complexes in all intervals. For example, a total of 325 complexes are predicted by NRAGE-WPN on CYC2008 in Collins and effective percentage of NRAGE-WPN is greater than other eleven algorithms. Further, with respect to the biological relevance, the enrichment score of the annotations are employed to evaluate the performance of predicted complex. The average of detected complexes with at least one enriched annotation over all clusters among eleven approaches on six datasets is compared in Additional file 1: Table S2. The results illustrate that NRAGE-WPN predicts biologically relevant clusters with enrichment scores with the top 70% of other methods in terms of the different GO categories.

Effectiveness of RA

Due to the noise data in the PPI network, NRAGE-WPN uses gene expression and RA information to score a weight to each interaction of the PPI network. To assess the effect

Table 2 Performance of functional enrichment comparison and their *P* values in Collins on CYC2008

Methods	Clusters	Effective (%)	< E-15 (%)	E-15-E-5 (%)	E-5-0.001 (%)
MCL	212	86.46	13.25	34.00	39.21
MCODE	84	89.65	15.50	59	14.75
CFinder	73	83.50	24.25	42.45	16.80
ClusterOne	106	91.85	29.75	52.47	9.63
ProRank+	385	87.52	18.17	48.12	21.23
CLE	215	84.14	35.36	26.43	22.35
CW	164	92.79	19.24	54.31	19.24
CLP	207	94.44	8.79	60.28	25.37
CI	132	92.06	25.35	57.47	9.24
PC2P	283	93.20	43.80	32.65	16.75
DPCT	274	94.05	40.72	37.95	15.38
NRAGE-WPN	325	96.42	41.25	48.75	6.42

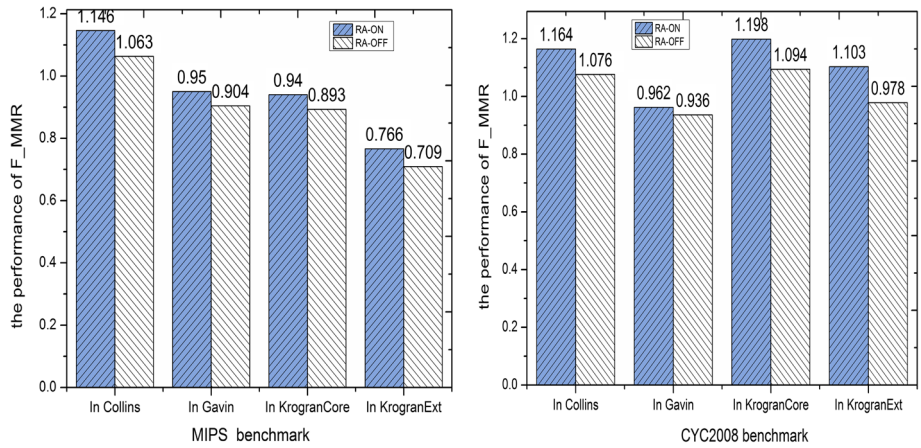


Fig. 7 The effectiveness of NRAGE-WPN when RA is off/on with CYC2008 and MIPS benchmarks

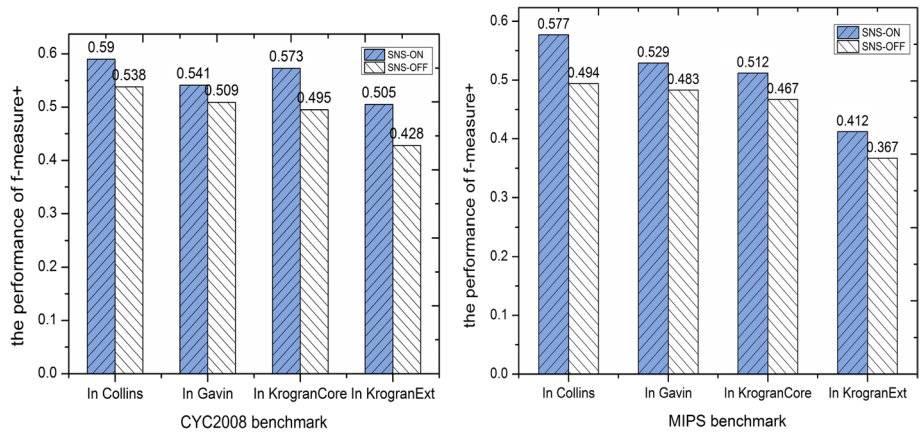


Fig. 8 The effectiveness of using SNS in NRAGE-WPN compare with NRAGE-WPN without using SNS in four PPI datasets with CYC2008 and MIPS benchmarks

of using RA in the f-measure+ for complexes detection, we conduct NRAGE-WPN without considering RA information and compare its results with normal the NRAGE-WPN which employs both gene expression and RA information. Without using RA situation, a weighted PPI network is constructed by gene expression only. Figure 7 shows the results of NRAGE-WPN in RA-OFF and RA-ON in Collins, Gavin, KroganCore and KroganExt datasets with CYC2008 and MIPS benchmarks, respectively. From Fig. 7, it can be shown that by introducing RA, the quality performance of F_MMR is enhanced. In term of RA-ON mode in Collins data, F_MMR increases 8.8% for the CYC2008 benchmark and 8.3% for the MIPS benchmark. According to Fig. 7, the same trend can also be shown in other three PPI datasets on two benchmarks, respectively. This experiment shows that using RA can reduce noise data and improve the overall performance of complexes detection.

Effectiveness of second-order neighbors searching (SNS)

The second phase of the NRAGE-WPN method is to enlarge the core complexes by second-order neighbors. After detecting core protein complexes from weighted PPI network, due to the nature of complexes of core-attachment, there may be many attachment parts to be added to the cores. In this situation, the cores and attachment parts are combined to form final complexes. In order to assess the effect of introducing second-order neighbors searching(SNS), we conduct NRAGE-WPN without its second phase. Figure 8 shows the comparison between second-order neighbors searching-on (SNS-ON) and second-order neighbors searching-off (SNS-OFF) modes in terms of f-measure+. On the CYC2008 benchmark, when NRAGE-WPN uses the SNS phase, we can see a 5.2%, 3.2%, 7.8% and 7.7% rise in Collins, Gavin, KroganCore, KroganExt, respectively. As the results show, performance of f-measure+ can be improved by introducing the second-order neighbors searching.

Assessment of density in different weighted graphs

$$W = \alpha W_{\mathrm{P}} + (1-\alpha)\mathrm{KBRV} \tag{12}$$

Although PCC cannot identify whether gene variables are directly regulated or indirectly regulated [33–35], in this paper, we mainly focus on PCC as biological information to construct weighted graph network based on gene expression, which is one of the most commonly used methods for constructing gene regulatory networks. At the same time, we discuss the influence of nonlinear correlation of gene expression on the density of whole network. We construct another four weighted graphs based on KBRV [32] method and the density of networks are compared in Fig. 9. First, the results show that four weighted networks based on KBRV can increase the density of PPI network. The reason is that the weighted value of the protein pairs that can be increased by (12). Second, we can find that when α belongs to [0.3 0.5], the densities of four weighted graph by (5) decrease slow. In our experiment, $\alpha = 0.3$ is used. Lastly, in our future work,we will

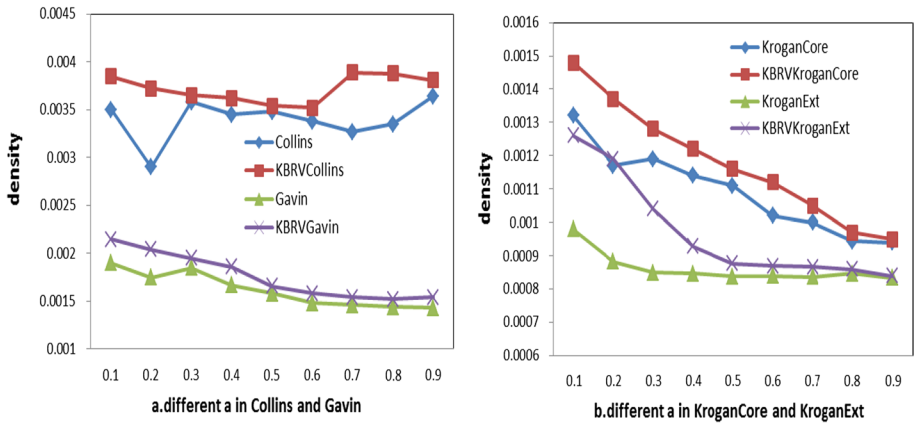


Fig. 9 The density of using different a by comparing with KVRB method for construction of four weighted graph in Yeast

focus on the nonlinear correlation of gene expression for weighted graph construction and complex detection.

Conclusions

The identification of protein complexes is important for discovering and understanding the cellular organizations and biological processes in PPI networks. In this paper a new approach named NRAGE-WPN is proposed for identifying protein complexes in protein–protein interaction networks. Based on the resource allocation and gene expression of the PPI network, we first design a new weight metric to accurately describe the interaction between proteins. Our method then constructs a series of dense complex cores based on density and network diameter constraints, and the final complexes are recognized by expanding the second-order neighbors of nodes in core complexes. Through comparison with eleven methods in Yeast and Human PPI network, the experimental results demonstrate that this algorithm not only performs better than other methods on 75% in terms of f-measure+, but also can achieve an ideal overall performance in terms of a composite score consisting of five performance measures. In the future work, we will focus on locating sparse and density protein complexes by integrating multiple information.

Abbreviations

PPI: Protein–protein interaction; NRAGE-WPN: A new way by combining node resource allocation and gene expression information to weight protein network; RA: Resource allocation; WRA: Weighted graph based on resource allocation; PCC: Pearson's correlation coefficient; GEI: Gene expression information; DP: Detected protein complexes; SN: Sensitivity; PPV: Positive predictive value; ACC: Accuracy; SEP: Separation; FRM: Fraction match; MMR: Maximum matching ratio; Prec: Precision; Rec: Recall; F_MMR: The sum of MMR and f-measure+; CS: The composite score of MMR, FRM, SEP, ACC and f-measure.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04543-4.

Additional fle 1. The Additional fle 1 contains Figures 1, Tables 1 and 2. **Figure 1** shows comparative analysis of approaches for prediction of protein complexes in Yeast on different threshold t. **Table 1** shows Comparative analysis of eleven algorithms with respect to different measures. **Table 2** shows the average of enrichment score of predicted complexes with at least one enrichedannotation over all clusters compared among eleven methods across six datasets.

Acknowledgements

We wish to thank the authors of the toolkits used in this paper and the reviewers of this paper.

Authors' contributions

YY designed the study and contributed manuscript preparation; DZK conducted the study. All authors read and approved the final manuscript.

Funding

This research is supported by the Liaoning Natural Science Foundation Project of China (20180550918). The funder YY took part in the formulation and development of methodology, and provided financial support for this study.

Availability of data and materials

The datasets are available at https://github.com/graceyy000/dataset and NRAGE-WPN codes are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

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

Received: 14 October 2021 Accepted: 20 December 2021
Published online: 06 January 2022

References

1. Lei X, Yang X, Wu F. Artificial fish swarm optimization based method to identify essential proteins. IEEE/ACM Trans Comput Biol Bioinf. 2018;17(2):495–505.
2. Bo W, Pourshafeie A, Zitnik M, Zhu J, Bustamante CD, Batzoglou S, Leskovec J. Network enhancement as a general method to denoise weighted biological networks. Nat Commun. 2018;9:1–8.
3. Rehman ZU, Idris A, Khan A. Multi-dimensional scaling based grouping of known complexes and intelligent protein complex detection. Comput Biol Chem. 2018;74:149–56.
4. Adamcsek B, Palla G, Farkas IJ, Derényi I, Vicsek T. CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics. 2006;22:1021–3.
5. Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinform. 2006;7(1):1–13.
6. Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinform. 2003;4(1):2.
7. Newman M. Finding community structure in networks using the eigenvectors of matrices. PhRvE. 2006;74:036104.
8. Libraries M. Computing communities in large networks using random walks. In: Computer and information sciences—ISCIS 2005; 2005.
9. Ahn Y-Y, Bagrow J, Lehmann S. Link communities reveal multiscale complexity in networks. Nature. 2010;466:761–4.
10. Arnau V, Mars S, Marín I. Iterative cluster analysis of protein interaction data. Bioinformatics. 2004;21(3):364–78.
11. Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007;312:972–6.
12. King AD, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering. Bioinformatics. 2004;20(17):3013–20.
13. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci. 2008;105(4):1118–23.
14. Enright AJ, Dongen SV, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucl Acids Res. 2002;7(30):1575–84.
15. Pereira-Leal JB, Enright AJ, Ouzounis CA. Detection of functional modules from protein interaction networks. Proteins-Struct Funct Bioinform. 2004;54:49–57.
16. Cho YR, Hwang W, Ramanathan M, Zhang A. Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinform. 2007;8:1–13.
17. Hwang W, Cho YR, Zhang A, Ramanathan M. CASCADE: a novel quasi all paths-based network analysis algorithm for clustering biological interactions. BMC Bioinform. 2008;9(1):64.
18. Kentaro I, Weijiang L, Hiroyuki K, Ernberg IT. Diffusion model based spectral clustering for protein-protein interaction networks. PLoS ONE. 2010;5(9):e12623.
19. Lecca P, Re A. Detecting modules in biological networks by edge weight clustering and entropy significance. Front Genet. 2015;6:265.
20. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods. 2012;9(5):471–2.
21. Zhou T, Lü L, Zhang Y-C. Predicting missing links via local information. Eur Phys J B. 2009;71(4):623–30.
22. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. Comparative assessment of large-scale data sets of protein-protein interactions. Nature. 2002;417(6887):399–403.
23. Wolfe CJ, Kohane IS, Butte AJ. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinform. 2005;6(1):227–220.
24. Li M, Chen JE, Wang JX, Hu B, Chen G. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. BMC Bioinform. 2008;9:1–16.
25. Sara O, Angela A, Zoran N. PC2P: parameter-free network-based prediction of protein complexes. Bioinformatics. 2021;37:73–81.
26. Tu PB. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. Science. 2005;310(5751):1152–8.
27. Hanna EM, Zaki N. Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure. BMC Bioinform. 2014;15(1):204.
28. van Dongen S. Graph clustering by flow simulation. Ph.D. thesis, University of Utrecht, Utrecht, The Netherlands 2000.
29. Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E. 2007;76(3 Pt 2):036106.
30. SabziNezhad A, Jalili S. DPCT: a dynamic method for detecting protein complexes from TAP-aware weighted PPI network. Front Genet. 2020;11:567.
31. Ma J, Wang J, Ghoraie LS, Men X, Haibe-Kains B, Dai P. A comparative study of cluster detection algorithms in protein-protein interaction for drug target discovery and drug repurposing. Front Pharmacol. 2019;10:109.
32. Yao Z, Zhang J, Zou X. A general index for linear and nonlinear correlations for high dimensional genomic data. BMC Genomics. 2020;21(1):1–14.
33. Guo X, Zhang Y, Hu W, Tan H, Wang X. Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation. PLoS ONE. 2014;9(2):e87446.
34. Kontio JA, Rinta-Aho MJ, Sillanpää MJ. Estimating linear and nonlinear gene coexpression networks by semiparametric neighborhood selection. Genetics. 2020;215(3):597–607.

35. Piran M, Karbalaei R, Piran M, Aldahdooh J, Mirzaie M, Ansari-Pour N, Tang J, Jafari M. Can we assume the gene expression profile as a proxy for signaling network activity? Biomolecules. 2020;10(6):850.
36. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege F, Weissman JS, Krogan NJ. Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. Mol Cell Proteomics. 2007;6:439–50.
37. Gavin AC, 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.
38. Krogan N, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis A. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature. 2006;440(7084):637–43.
39. Damian S, Andrea F, Stefan W, Kristoffer F, Davide H, Jaime HC, Milan S, Alexander R, Alberto S, Tsafou KP. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(D1):D447–52.
40. McDowall MD, Scott MS, Barton GJ. PIPs: human protein-protein interaction prediction database. Nucl Acids Res. 2009;37:D651–6.
41. Pu S, Jessica W, Brian T, Emerson C, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucl Acids Res. 2009;37(3):825–31.
42. Mewes HW, Amid C, Arnold R, Frishman D, Güldener U, Mannhaupt G, Münsterkötter M, Pagel P, Strack N, Stümpflen V. MIPS: analysis and annotation of proteins from whole genomes. Nucl Acids Res. 2004;32(suppl_1):169–72.
43. Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G. CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. 2018;47:D559–63.

Publisher's Note