Accepted Manuscript

Journal of Bioinformatics and Computational Biology

Article Title:	Integrating network topology, gene expression data and GO annotation information for protein complex prediction
Author(s):	Wei Zhang, Jia Xu, Yuanyuan Li, Xiufen Zou
DOI:	10.1142/S021972001950001X
Received:	28 November 2017
Accepted:	24 October 2018
To be cited as:	Wei Zhang *et al.*, Integrating network topology, gene expression data and GO annotation information for protein complex prediction, *Journal of Bioinformatics and Computational Biology*, doi: 10.1142/S021972001950001X
Link to final version:	https://doi.org/10.1142/S021972001950001X

World Scientific
www.worldscientific.com

Journal of Bioinformatics and Computational Biology

Integrating network topology, gene expression data and GO annotation information for protein complex prediction

Wei Zhang
School of Science, East China Jiaotong University
Nanchang 330013, *China*
wzhang_math@whu.edu.cn

Jia Xu
School of Mechatronic Engineering, East China Jiaotong University
Nanchang 330013, *China*

Yuanyuan Li
School of Mathematics and Statistics, Wuhan Institute of Technology in Wuhan, Wuhan, 430072, *China*

Xiufen Zou
School of Mathematics and Statistics, Wuhan University
Wuhan 430072, *China*

The prediction of protein complexes based on the protein interaction network is a fundamental task for the understanding of cellular life as well as the mechanisms underlying complex disease. A great number of methods have been developed to predict protein complexes based on protein protein interaction(PPI) networks in recent years. However, because the high throughput data obtained from experimental biotechnology are incomplete, and usually contain a large number of spurious interactions, most of the network-based protein complex identification methods are sensitive to the reliability of the PPI network. In this paper, we propose a new method, IPC-RPIN(Identification of Protein Complex based on Refined Protein Interaction Network), which integrates the topology, gene expression profiles and GO functional annotation information to predict protein complexes from the reconstructed networks. To demonstrate the performance of the IPC-RPIN method, we evaluated the IPC-RPIN on three PPI networks of *Saccharomycescerevisiae* and compared it with four state-of-the-art methods. The simulation results show that the IPC-RPIN achieved a better result than the other methods on most of the measurements and is able to discover small protein complexes which have traditionally been neglected.

Keywords: PPI Network, Protein Complexes, GO Ontology, Gene expression profile.

1. Introduction

Protein complexes are formed by groups of physically interacting proteins into functional units that are integral to maintaining normal physiological activities [1]. Discovering protein complexes is one of the key areas of research for understanding the

cellular function and the dynamic mechanisms underlying complex disease. Experimentally, the identification of protein complexes is typically performed by the use of Tandem Affinity Purification with mass spectrometry(TAP-MS) [2]. However, these biological experiments are expensive and time-consuming.

Due to the rapid development of modern high-throughput technologies such as yeast two-hybrid(Y2H) screens [3, 4], tandem affinity purification(TAP) [5], and mass spectrometric protein complex identification(MS-PCI) [6], large scale protein-protein interaction(PPI) data are available for many organisms. PPI networks provide a comprehensive view of the global interaction structure of an organism's proteome, as well as detailed information on specific interactions, which make it possible and feasible to use computational methods to predict protein complexes from the perspective of large molecular network.

The past decade has witnessed the rapid development of computational methods for identifying protein complexes from PPI network datasets. Many approaches have been developed to predict protein complexes [7–13, 18–28], such as MCL [7], MCODE [8], COACH [11], CMC [12], ClusterOne [13], PEWCC [18], WPNCA [19], and WCOACH [22]. These methods basically fall into three categories: seed expanding-based methods, clustering-based methods and dynamic PPI-based methods. The first category identifies protein complexes based on seed expanding or core expanding by capturing the density of subgraphs in PPI networks [8,11–13]. For example, Liu G. et al. [12] developed an iterative scoring method(AdjstCD) to evaluate the reliability of interaction between protein pairs by assigning weight to protein pairs and predicting protein complexes in weighted PPI networks. Wu et al. [11] proposed a core-attachment based method, COACH, which first identifies the core of protein complexes and then incorporates the attachments associated with the core proteins. Nepusz and Paccanaro et al. [13] proposed a seed growth method for predicting protein complexes, named ClusterOne, by defining a cohesiveness score for a subnetwork.

It has been widely accepted that the efficient and effective integration of multiple data sources yields better results [29,30]. Han Yong et al. [31] proposed a supervised maximum-likelihood method to weight edges in a composite network constructed by heterogeneous data and adapted clustering algorithms for protein complex prediction. Wu et al. [32] developed an integrative approach, named InteHC, which utilized the Support Vector Machine(SVM) to weight features for heterogeneous data and generated protein complexes by using the hierarchical clustering algorithm. Comparative analysis showed that InteHC outperformed the other 14 methods. Recently, methods have incorporated additional information including Gene Ontology(GO) [33, 34], gene expression data [35] and structural interface data of protein domains [40] to help predict protein complexes. The simulation results of these methods show that integrating multiple sources of data can boost the detection of protein complexes.

The second category consists of clustering the proteins based on clustering and evolutional algorithms [14–17,36]. For example, Ou-Yang et al. [36] proposed a new

protein prediction method based on nonnegative matrix factorization. Sanghamitra B. et al. [15] proposed a multi-objective framework for predicting protein complexes based on both topological properties of PPI network and Gene Ontology semantic similarity information.

The methods in the third category predict protein complexes by constructing dynamic protein-protein interaction networks. Zhang et al. [37] identify protein complexes by constructing dynamic probabilistic protein networks. Ou-Yang et al. [38] developed a novel time smooth overlapping complex detection model(TS-OCD) for predicting temporal protein complexes and tracking the evolutionary process of the temporal complexes. Li M. et al. [39] proposed a model-based scheme for construction of the spatial and temporal protein interaction network by combing gene expression and subcellular location information, and the algorithm MCL is used for predicting protein complexes based on the temporal network.

However, most of the methods described above only focus on the original network obtained from high-throughput techniques, and few of them consider the inherent incompleteness and noise of the PPI networks [41].

Due to the fact that PPI data are inherently noisy and incomplete, the incompleteness and noise underlying the PPI networks severely hindered the prediction accuracy of these methods.

In this research, we try to introduce a novel method(IPC-RPIN) for predicting protein complexes by constructing a refined PPI network based on integrating gene expression data and GO ontology information. To validate the performance of the IPC-RPIN, we compare it with four other state-of-the-art methods(ClusterOne, PEWCC, COACH, and TNC([24])) on three benchmark yeast PPI datasets. The simulation results show that the new method performs well in identifying protein complexes.

2. Method

2.1. *Protein complex prediction from refined network*

In this section, we first present a new strategy for obtaining a refined PPI network by link prediction based on gene expression profiles, and then we introduce the new proposed method based on the definition of protein complexes.

Given a PPI network with N proteins, usually represented by an undirected graph $G = (V, E)$, the vertex set V represents the proteins and the edge set E represents the set of interactions between pairs of proteins.

Previous works have demonstrated that integrating multiple data sources could improve the prediction accuracy. In addition, constructing new networks by PPI prediction based on gene expression profiles demonstrates more reliable for mining key information such as essential proteins [42, 43]. So we firstly constructed a new PPI network by PPI prediction based on gene expression profiles and then incorporated GO annotation information to qualify the compactness of the subnetwork.

The new proposed algorithm is as follows:

Algorithm 1 Prediction protein complexes

Input: The PPI nework G=(V,E); GE: gene expression data; GO:Gene ontology annotation data.

Output: Predicted protein complexes (PPC)

for each unlinked protein pairs (u,v) $\notin$ E **do** Compute the PCC(u,v) between two proteins u and v.
  if PCC(u,v)$\geq$ 0.98 **then** Add the link (u,v) to the original network G
Obtain new network G'
Compute the GO semantic similarity matrix G_simgo based on GO annotation information under Biological Process term.
for each protein u $\in$ V **do**
  Create a rough group Vg by adding the neighbors of u.
  Evaluate the compactness of the rough group $CVgt$(Topology compactness of rough group),$CVgo$(GO compactness of rough group).
  Add vertices on the outer boundary of the Vg iteratively to increase the compactness $CVgt$ and $CVgo$ until they reaches the maximum.
  Remove vertices in the Vg if the compactness of the remaining group $CVgt$ and $CVgo$ are larger than original, and iteratively increase the compactness until it reaches the maximum.
Obtain initial protein complex set PCS
Merge the pairs of candidate connected protein complexes with OS $\geq$ 0.8 and obtain refined protein complexes sets.
for each protein complex pc $\in$ PCS **do**
  if size(pc)<3 or density(pc)<0.5 **then**
    delete the pc from PCS
Obtain the last predicted protein complex PPC
return Predicted protein complexes PPC

The topology compactness of a group ($CVgt$) is adapted from our previous work [24]

$$CVgt(G') = \frac{Nl^{in}(G')}{Nl^{in}(G') + Nl^{out}(G') + p|G'|} \tag{1}$$

where the $Nl^{in}(G')$ denotes the total number of 3-cliques in the subgraph G', and the $Nl^{out}(G')$ denotes the total number of 3-cliques that connect the subgraph with the rest of the network, $p|G'|$ is a penalty term which measures the inaccuracy of the network interaction. The value of p is set to 0.1 in this work and the effect of parameter p on the results is discussed in section 3.4.

The new compactness after the addition of protein i (protein i is a member of Vb') can be calculated as follows:

$$CVgt(G' \cup \{i\}) = \frac{Nl^{in}(G') + Nl^{in}(i)}{Nl^{in}(G') + Nl^{out}(G') + Nl^{out}(i) + p(|G'| + 1)} \tag{2}$$

where the $Nl^{in}(i)$ denotes the total number of 3-cliques that connect protein i with proteins in subgraph G', and $Nl^{out}(i)$ denotes the total number of 3-cliques that connect protein i with nonmembers of G'.

Similarly, the new compactness after removal of protein i (in subgraph G') can be calculated as follows:

$$CVgt(G'\backslash\{i\}) = \frac{Nl^{in}(G') - Nl^{in}(i)}{Nl^{in}(G') + Nl^{out}(G') - Nl^{out}(i) + p(|G'| - 1)} \tag{3}$$

The GO compactness of rough group $G'(V', E')$ is defined as

$$CVgo(G') = \frac{\sum\limits_{u,v \in E'} GO\_sim(X, Y)}{|G'|(|G'| - 1)/2} \tag{4}$$

The $GO\_sim(X, Y)$ is the GO functional similarity between two proteins a and b, which is calculated by using the GO semantic similarity. To evaluate the gene functional similarity between GO terms annotated to proteins in the interaction networks, the method developed by Wang et al. [44] is applied to calculate the semantic similarity between proteins. In this work, we only consider the BP subontology term in evaluating the two considered proteins. The GO similarity between two connected proteins is defined as:

$$GO\_sim(X, Y) = \frac{\sum\limits_{1 \le i \le m} s(go_{x_i}, Y) + \sum\limits_{1 \le j \le n} s(go_{y_j}, X)}{m + n} \tag{5}$$

where $s(go_{x_i}, Y) = \max\limits_{1 \le j \le n} (S\_GO(go_{x_i}, go_{y_j}))$, $s(go_{y_j}, X) = \max\limits_{1 \le i \le n} (S\_GO(go_{x_i}, go_{y_j}))$, and $S\_GO(go_{x_i}, go_{y_j})$ is the semantic similarity between terms go_{x_i} and go_{y_j}. The semantic similarity between two considered terms A and B is defined as:

$$S\_GO(A, B) = \frac{\sum\limits_{\mathrm{r} \in T_A \cap T_B} (S_A(r) + S_B(r))}{\sum\limits_{\mathrm{r} \in T_A} S_A(r) + \sum\limits_{\mathrm{r} \in T_B} S_B(r)} \tag{6}$$

where $S_A(r)$ is the S-value of GO term r related to term A and $S_B(r)$ is the S-value of GO term r related to term B.

2.2. *Experimental Data*

Genome-wide protein-protein interaction networks are publicly available in several open databases. The yeast *Saccharomycescerevisiae* PPI networks are widely used as benchmarks for evaluating the performance of a newly proposed algorithm, as they have been well characterized by biological experiments.

In our proposed method, we used three different yeast PPI datasets including the collins2007 dataset, which contains 9074 interactions and 1622 proteins, the gavin2006 dataset, which contains 1855 proteins and 7669 interactions, and the krogan_extended dataset, which consists 14317 interactions and 3672 proteins. The three datasets are obtained from the published work in [13]. In the following statement, we use collins, gavin and krogan_ extended to represent these three networks.

Three benchmark reference complexes are used to evaluate the performance of the new method: the CYC2008 dataset [45], which contains 408 manually curated heteromeric protein complexes, the Mips dataset [46], which contains 203 protein complexes and the Aloy dataset [47],which contains 101 protein complexes.

The gene expression data used in our experiment are from GSE3431 dataset [48], which collects the data of 12 time points during three successive metabolic cycles and approximately 25 min per time interval.

2.3. *Evaluation measures*

To evaluate the efficiency of the proposed method, we compare the IPC-RPIN algorithm to the following state-of-the-art algorithms: ClusterOne, PEWCC and TNC. All methods are performed on the refined networks. There are four different quantity measures, namely, the number of matched protein complexes (Match) in the reference protein complexes, the fraction of reference protein complexes predicted (Frac), the prediction geometric accuracy (ACC) [49], and maximum matching ratio (MMR) [13].

The Overlapping Score(OS) [8] between a predicted protein complex P and a known protein complex R is defined as following.

$$OS(R,P) = \frac{|R \cap P|^2}{|R||P|} \tag{7}$$

where R and P represent the benchmark reference protein complexes and predicted protein complexes respectively. We assume the reference complex has been predicted when the predicted protein complex and reference complexes with an OS ≥ 0.25.

The MMR is based on a maximal one-to-one mapping between predicted and reference complexes.

$$MMR(R,P) = \frac{\sum_{i=1}^{n} \max_{j=1}^{m} OS(R_i, P_j)}{n} \tag{8}$$

where the OS refers to the overlap score between two protein sets.

The Acc, is the geometric mean of the clustering-wise sensitivity (Sn) and the clustering-wise positive predictive value (PPV). Given m predicted and n reference protein complexes, the two measures are based on the confusion matrix $T = [t_{ij}]$ of the complexes, where t_{ij} denotes the number of proteins that are found both in reference complex i and predicted complex j. The Sn and PPV are defined as:

$$S_n = \frac{\sum_{i=1}^{n} \max_{j=1}^{m} t_{ij}}{\sum_{i=1}^{n} n_i} \tag{9}$$

where n_i is the number of proteins in reference complex i.

$$PPV = \frac{\sum_{j=1}^{m} \max_{i=1}^{n} t_{ij}}{\sum_{j=1}^{m} \sum_{i=1}^{n} t_{ij}} \tag{10}$$

The geometry accuracy (Acc) represents a tradeoff between sensitivity and positive predictive value and is defined as:

$$Acc = \sqrt{S_n * PPV} \tag{11}$$

The high geometric accuracy (Acc) indicates that the two criteria of the Sn and the PPV metric are indicative of high performance.

3. Results

In this section, we first systematically evaluate the performance of the new method and against four other existing methods (ClusterOne, PEWCC, TNC, and COACH) on the three test PPI networks using four evaluation measurements. Next, we compare the evaluation results under different combination strategies in section 3.3; Last, the effects of the parameters on the evaluation results are analyzed. The parameters in ClusterOne, PEWCC and TNC are set as default. The minimum number of proteins in protein complexes is 3, the overlap score is set to 0.25, and the density threshold of the predicted protein complexes is set to 0.5 as default for the new method.

3.1. *Comparison with the benchmark protein complexes*

Table 1 shows the comparative results of the new method and four existing methods on the three datasets using four evaluating measurements under the CYC2008 reference protein complex dataset. We can see that the new proposed method IPC-RPIN performs better than the other considered methods in most of these measurements. Especially, for the collins dataset, the predicted protein complexes could match 112 reference protein complexes, which is better than other methods.

Table 1. Result of the three methods on the original network and new constructed network under CYC2008 reference protein complex dataset

PPI data	Method	Matched	Frac	Acc	MMR
collins	IPC-RPIN	**112**	**0.572**	0.599	**0.416**
	TNC	97	0.538	0.608	0.388
	ClusterOne	96	0.504	**0.640**	0.377
	PEWCC	91	0.492	0.591	0.365
	COACH	98	0.513	0.607	0.383
krogan_extended	IPC-RPIN	**96**	**0.551**	0.584	**0.356**
	TNC	88	0.496	**0.595**	0.331
	ClusterOne	91	0.487	0.563	0.342
	PEWCC	81	0.462	0.518	0.308
	COACH	93	0.496	0.530	0.338
gavin	IPC-RPIN	**95**	**0.521**	0.576	**0.355**
	TNC	84	0.521	**0.598**	0.344
	ClusterOne	69	0.381	0.502	0.262
	PEWCC	88	0.492	0.590	0.340
	COACH	92	0.504	0.591	0.346

Similarly, we applied the methods on the three test datasets with respect to two other reference complex datasets. The performance comparison of the three datasets under the Mips reference complexes and Aloy reference complexes are presented in Tables 2 and 3, respectively.

As shown in the Table 2, in addition to the ACC measurement, the proposed method shows superiority to other methods in the other three measurements. Our method always achieves the highest Match number and MMR score.

Similar results were obtained using the Aloy reference protein complex dataset, as seen from Table 3, where the new method performs better than the other methods in most of the four measurements under all three considered datasets, suggesting that the new method is capable of detecting more true complexes and provides a strong argument in favor of the proposed method.

3.2. *Protein complexes identified by our method*

To further exhibit the performance of the new proposed method, we collected the matched protein complexes that could only be detected by the new method under the three PPI networks with respect to the three reference protein complex datasets.

The protein complexes that could only be predicted by the new method under the collins PPI network are listed in Figure 1. Complexes A, B ,D and E in Figure 1 are evaluated under CYC2008 reference protein complexes. The three small protein complexes, A, B and C were identified accurately by the new method using the Aloy reference protein

Table 2. Result of the three methods on the original network and new constructed network under Mips reference protein complex dataset

PPI data	Method	Matched	Frac	Acc	MMR
collins	IPC-RPIN	**74**	**0.537**	0.448	**0.360**
	TNC	66	0.512	**0.456**	0.341
	ClusterOne	67	0.512	0.397	0.332
	PEWCC	61	0.468	0.394	0.322
	COACH	66	0.507	0.437	0.339
krogan_extended	IPC-RPIN	**58**	**0.443**	0.335	**0.288**
	TNC	53	0.399	**0.338**	0.265
	ClusterOne	54	0.404	0.336	0.266
	PEWCC	53	0.414	0.312	0.256
	COACH	53	0.429	0.318	0.271
gavin	IPC-RPIN	**67**	0.488	0.368	**0.321**
	TNC	58	0.458	0.366	0.303
	ClusterOne	45	0.374	0.287	0.243
	PEWCC	62	0.468	0.374	0.317
	COACH	62	0.458	**0.375**	0.315

Table 3. Result of the three methods on the original network and new constructed network under Aloy reference protein complex dataset

PPI data	Method	Matched	Frac	Acc	MMR
collins	IPC-RPIN	**68**	**0.936**	0.793	**0.743**
	TNC	65	**0.936**	0.804	0.719
	ClusterOne	64	0.897	**0.842**	0.702
	PEWCC	60	0.846	0.800	0.654
	COACH	66	0.910	0.827	0.729
krogan_extended	IPC-RPIN	**50**	**0.782**	**0.764**	0.500
	TNC	48	0.718	0.769	0.475
	ClusterOne	34	0.590	0.689	0.383
	PEWCC	44	0.679	0.714	0.467
	COACH	49	0.756	0.729	**0.507**
gavin	IPC-RPIN	**67**	**0.974**	0.826	**0.685**
	TNC	65	**0.974**	**0.847**	0.684
	ClusterOne	53	0.756	0.720	0.550
	PEWCC	64	0.897	0.831	0.667
	COACH	**67**	0.949	0.841	0.682

complexes dataset. Complexes A and E predicted by the new method were matched by using the Mips reference protein complex dataset.

Figure 2 shows the protein complexes identified by the new method for the gavin PPI dataset. Both complexes A and B could only be detected when evaluated by using the CYC2008 and Aloy reference protein complex datasets, however, complex A could only be detected by the new method for the gavin dataset under the Mips reference protein complex dataset.

The protein complexes that could only be predicted by the new method for the krogan_extended PPI network under the CYC2008 reference protein complex dataset are shown in Figure 3. The small protein complexes that were accurately identified by using the new algorithm indicates that constructing refined network and adding GO information when measuring compactness of subnetworks will help in the detection of small protein

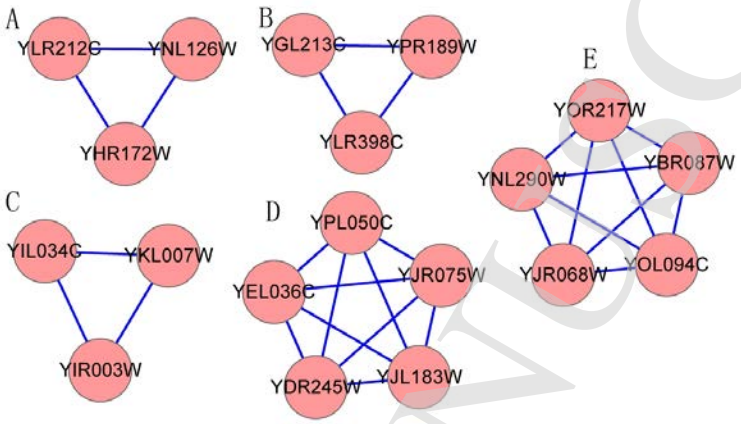

Fig. 1. The protein complexes could only be detected by using our method under the collins PPI dataset with respect to the CYC2008, Aloy and Mips reference protein complexes datasets.

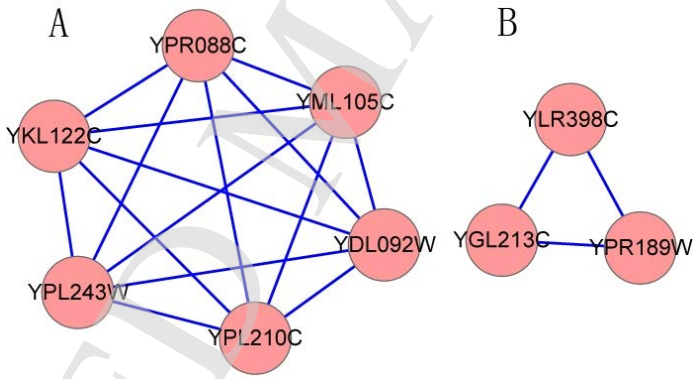

Fig. 2. The protein complexes could only be detected by using our method under the gavin PPI dataset with respect to the CYC2008, Aloy and Mips reference protein complexes datasets.

complexes that were ignored by all four state-of-the-art methods.

3.3. *Comparing results under different types of strategies and threshold parameters*

The new proposed algorithm involved three types of biological data, and different strategies were used to evaluate the cohesiveness of protein complexes. It was necessary to analyze the performance under different combination of data and strategies. To test the performance of the new proposed strategy under different scenarios of additional data combination, we compared the results with different PCC thresholds under which both topological information and GO information are considered, only topological information is considered, and only GO information is considered in measuring the compactness of subnetwork. For simplicity, we set the threshold to 0.98 and 0.96, respectively.

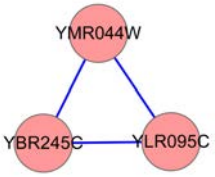

Fig. 3. The protein complexes could only be detected by using our method under the krogan_extended PPI dataset with respect to the CYC2008, Aloy and Mips reference protein complexes datasets.

As shown in Table 4, we list the evaluation results under different scenarios of information used in measuring the compactness of the subnetwork with different PCC thresholds. The only TNC in the bracket of the method column means that we only consider the topological compactness when performing the simulation of the IPC-RPIN algorithm, and only GO denotes that only GO compactness is calculated in step 7 of the algorithm. Comparing Table 4 with results listed in Table 1 for IPC-RPIN, performance is slightly worse when only TNC is considered for three PPI networks under the CYC2008 reference protein complex dataset. When only the GO compactness is considered in the algorithm, the results are the worst. The comparison results indicated that the appropriate integration of different data for evaluation compactness will increase the performance of protein complex prediction and that the topological information is more useful than GO information when only one dataset is considered.

The evaluation results of the Mips and Aloy benchmark reference protein complex datasets are shown in Table 5 and Table 6, respectively. Similar results can be seen from the two tables, in which IPC-RPIN algorithm performs better than the original TNC-based method and when only one type of data information is considered in measuring the compactness of the subgraph. These results suggest that the addition of GO annotation information could improve the performance of predicting protein complexes.

In addition, the results under different PCC thresholds show that when the PCC is set to be relatively large (PCC=0.98), the performance of the four considered measurements is better than when the PCC is set to be relatively small(PCC=0.96), indicating that when the PCC threshold is set to be relatively large, the added edges may include the real missed interactions between proteins and the new obtained network will be more reliable and complete for protein complex detection.

In this experiment, we evaluate the effect of the density threshold parameters of our method on the considered three PPI networks. Table 7 shows the performance of the new method under the three benchmark reference protein complexes; when the density threshold in the last step of IPC-PRIN is set to 0.3, the Match number of predicted protein complexes and fractions and MMR measurements are higher than the results obtained by setting the density to 0.5 as default. Especially for the collins PPI network, the number of matched protein complexes under the density threshold of 0.3 is 121, which is larger than the number of matched protein complexes (112) under the density threshold of 0.5. This finding indicated that the benchmark protein complexes are sparse; however, when the density threshold is set to be relatively large, some protein complexes with sparse density may be ignored. Almost the same results can be obtained by setting the density threshold to 0, indicating that the density of the predicted protein complexes is equal or larger than 0.3.

Table 4. Evaluation results under different threshold or strategy on corresponding PPI networks with respect to the CYC2008 reference protein complexes dataset.

PPI data	Method	PCC threshold	Matched	Frac	Acc	MMR
collins	IPC-RPIN(only TNC)	PCC=0.98	97	0.534	0.595	0.384
	IPC-RPIN(only GO)		70	0.407	0.502	0.262
	IPC-RPIN(only TNC)	PCC=0.96	95	0.521	0.555	0.378
	IPC-RPIN(only GO)		70	0.407	0.504	0.262
krogan_extended	IPC-RPIN(only TNC)	PCC=0.98	81	0.479	0.575	0.318
	IPC-RPIN(only GO)		74	0.386	0.462	0.268
	IPC-RPIN(only TNC)	PCC=0.96	76	0.462	0.479	0.303
	IPC-RPIN(only GO)		73	0.386	0.461	0.264
gavin	IPC-RPIN(only TNC)	PCC=0.98	82	0.508	0.584	0.332
	IPC-RPIN(only GO)		72	0.419	0.486	0.263
	IPC-RPIN(only TNC)	PCC=0.96	80	0.492	0.489	0.324
	IPC-RPIN(only GO)		71	0.407	0.484	0.258

Table 5. Evaluation results under different threshold or strategy on corresponding PPI networks with respect to the Mips reference protein complexes dataset.

PPI data	Method	PCC threshold	Matched	Frac	Acc	MMR
collins	IPC-RPIN(only TNC)	PCC=0.98	66	0.512	0.445	0.339
	IPC-RPIN(only GO)		46	0.374	0.346	0.225
	IPC-RPIN(only TNC)	PCC=0.96	64	0.507	0.421	0.338
	IPC-RPIN(only GO)		46	0.374	0.347	0.225
krogan_extended	IPC-RPIN(only TNC)	PCC=0.98	55	0.404	0.334	0.268
	IPC-RPIN(only GO)		48	0.360	0.289	0.224
	IPC-RPIN(only TNC)	PCC=0.96	48	0.409	0.306	0.263
	IPC-RPIN(only GO)		48	0.369	0.295	0.223
gavin	IPC-RPIN(only TNC)	PCC=0.98	58	0.443	0.363	0.294
	IPC-RPIN(only GO)		47	0.394	0.308	0.240
	IPC-RPIN(only TNC)	PCC=0.96	56	0.448	0.332	0.294
	IPC-RPIN(only GO)		47	0.384	0.307	0.236

Table 6. Evaluation results under different threshold or strategy on corresponding PPI networks with respect to the Aloy reference protein complexes dataset.

PPI data	Method	PCC threshold	Matched	Frac	Acc	MMR
collins	IPC-RPIN(only TNC)	PCC=0.98	65	0.936	0.778	0.711
	IPC-RPIN(only GO)		46	0.705	0.672	0.452
	IPC-RPIN(only TNC)	PCC=0.96	62	0.897	0.694	0.692
	IPC-RPIN(only GO)		47	0.705	0.666	0.458
krogan_extended	IPC-RPIN(only TNC)	PCC=0.98	47	0.731	0.758	0.479
	IPC-RPIN(only GO)		42	0.615	0.640	0.421
	IPC-RPIN(only TNC)	PCC=0.96	42	0.654	0.654	0.445
	IPC-RPIN(only GO)		42	0.615	0.636	0.415
gavin	IPC-RPIN(only TNC)	PCC=0.98	65	0.962	0.836	0.665
	IPC-RPIN(only GO)		51	0.821	0.702	0.512
	IPC-RPIN(only TNC)	PCC=0.96	63	0.923	0.708	0.643
	IPC-RPIN(only GO)		50	0.795	0.692	0.497

Table 7. Performance of new method on three PPI datasets with the density threshold of 0.3.

Reference date	PPI date	Matched	Frac	Acc	MMR
Mips	collins	**77**	**0.557**	**0.449**	**0.373**
	krogan_extended	64	0.498	0.339	0.318
	gavin	68	0.502	0.371	0.328
CYC2008	collins	**121**	**0.614**	**0.597**	**0.447**
	krogan_extended	101	0.610	0.569	0.387
	gavin	97	0.525	0.552	0.360
Aloy	collins	**69**	**0.974**	0.777	**0.761**
	krogan_extended	51	0.821	0.754	0.525
	gavin	67	**0.974**	**0.801**	0.685

3.4. *Influence of Parameter p*

The penalty term p in the formulas of measuring the compactness of protein complexes is used to fine-tune the inaccuracy of the network. Hence, it is necessary to analyze the effects of parameter p on the performance of the new method. Figure 4, Figure 5 and Figure 6 show the three core measurements as a function of parameter p when parameter p is varied from 0 to 0.3. We can observe that parameter p has slight effect on the core evaluation measurements under different PPI networks and benchmark reference protein complex datasets, and when parameter p varies in the interval [0, 0.3], the new proposed method always achieves good results under the tested PPI networks. Hence, we set parameter p to 0.1 in performing simulation under all of the considered PPI datasets.

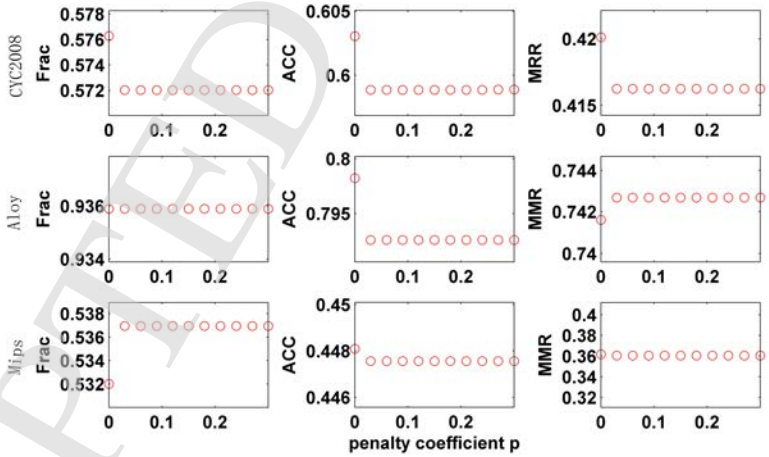

Fig. 4. The three core measurement plots as a function of penalty term p under the collins PPI network. The three rows of the subfigures (on the left, middle and right) are corresponding to the CYC2008, Aloy and Mips reference protein complexes datasets.

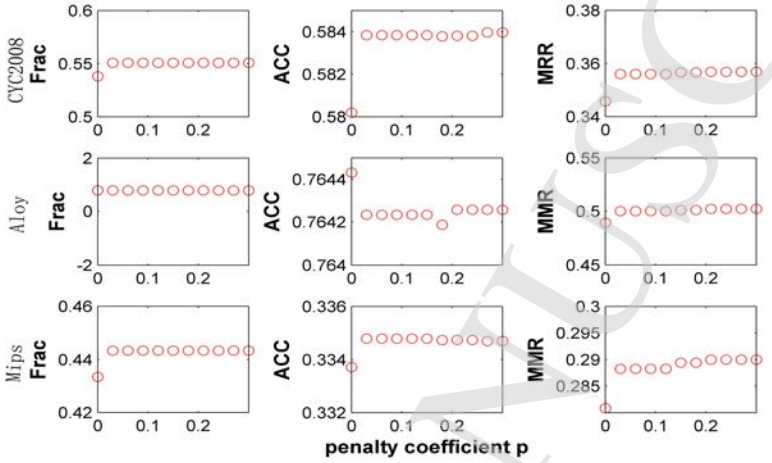

Fig. 5. The three core measurement plots as a function of penalty term p under the krogan_extended PPI network. The three rows of the subfigures (on the left, middle and right) are corresponding to the CYC2008, Aloy and Mips reference protein complexes datasets.

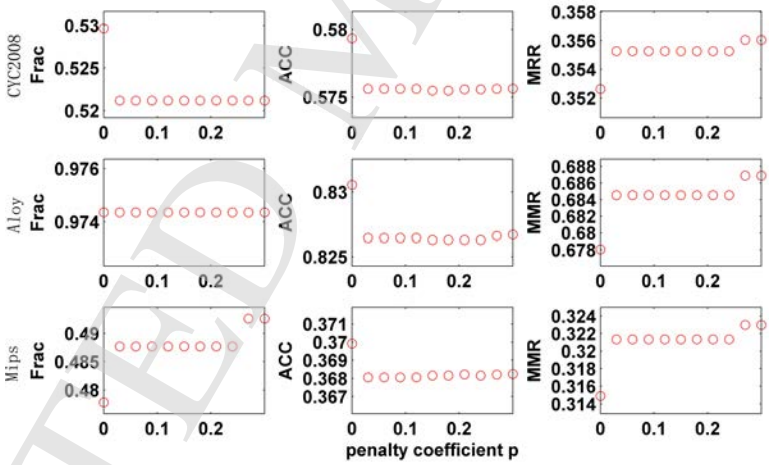

Fig. 6. The three core measurement plots as a function of penalty term p under the gavin PPI network. The three rows of the subfigures (on the left, middle and right) are corresponding to the CYC2008, Aloy and Mips reference protein complexes datasets.

3.5. *Effect of PCC threshold on prediction performance*

In the IPC-PRIC algorithm, the prediction accuracy of the method is closely related to the PCC threshold. To investigate the effect of PCC threshold on performance of our method, we evaluate the prediction measurements by setting the PCC threshold from 0.90 to 1 with a step 0.01. When the PCC threshold is set to 1, this corresponds to the original network. Figure 7, Figure 8 and Figure 9 show the three core measurement plots as a

function of the PCC threshold when the PCC threshold varies from 0.9 to 0.99 with a step 0.01 for the gavin, collins and krogan_extended networks, respectively. According to the results shown in the three figures, with an increase in the PCC threshold, the prediction accuracy is increased for both networks because with an increase in the PCC threshold, the added links in the constructed network will be more reliable, which contributes to the increase in ACC measurement. As for the MMR measurement, it also increases slightly with an increase in the PCC threshold. By contrast, the matched number of predicted protein complexes is not increased monotonically when the PCC threshold varies from 0.9 to 0.99; it depends on the PPI network and reference protein complexes we choose. Overall, the matched number is higher when the PCC threshold is set to be larger than 0.94.

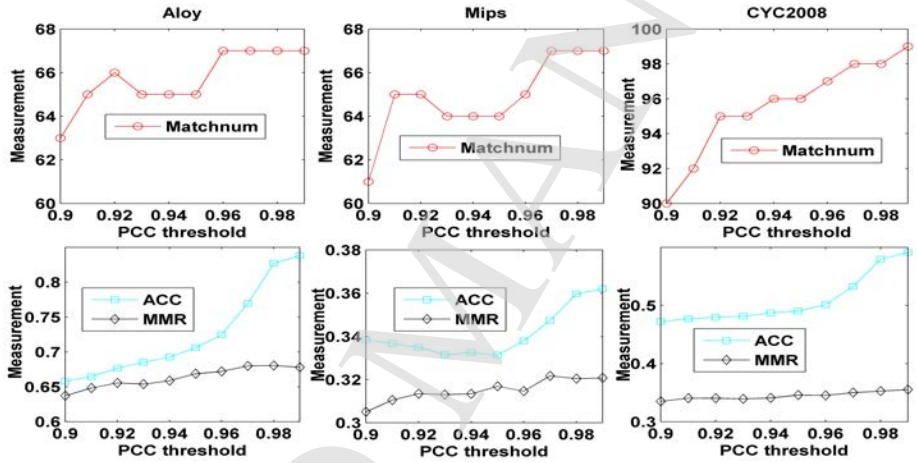

Fig. 7. The three core measurements as a function of PCC threshold under the gavin PPI network. The three columns of subfigures are corresponding to the Aloy, Mips and CYC2008 reference protein complexes datasets.

4. Conclusions

Predicting protein complexes from PPI networks is a hot topic in the post genome era and many computational methods have been developed to predict protein complexes for the PPI network. However, most of these methods are based on the original network only and the incomplete data hinder the prediction of protein complexes.

In this research, we constructed a refined PPI network based on time series gene expression data, and then introduced a new way to improve the prediction accuracy of the protein complexes by incorporating topology properties and GO annotation information. The compactness of the subnetwork is characterized by the number of three node cliques and GO functional semantic similarity under BP sub-annotation information. Based on the refined network and new evaluation of the compactness of the subnetwork, we developed a new method called IPC-RPIN for protein complex prediction.

To evaluate the performance of the proposed strategy, we validated the new method on three test PPI networks under three reference protein complex datasets and compared

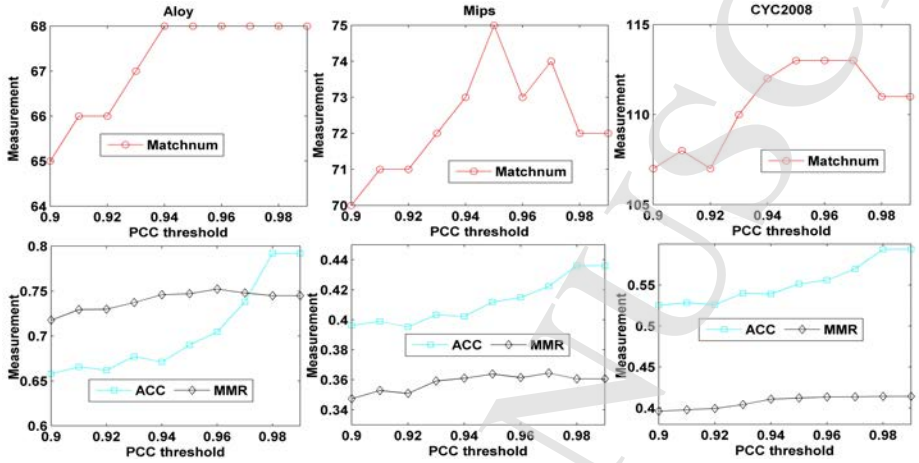

Fig. 8. The three core measurements as a function of PCC threshold under the collins PPI network. The three columns of subfigures are corresponding to the Aloy, Mips and CYC2008 reference protein complexes datasets.

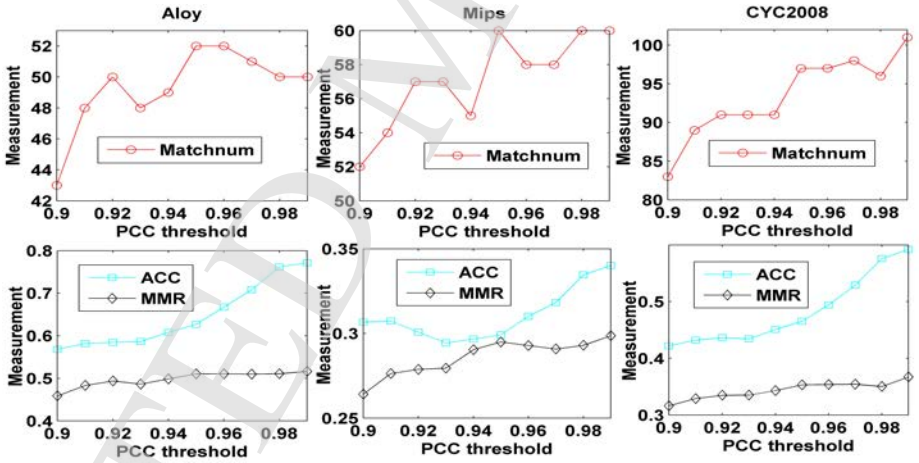

Fig. 9. The three core measurements as a function of PCC threshold under the krogan_extended PPI network. The three columns of subfigures are corresponding to the Aloy, Mips and CYC2008 reference protein complexes datasets.

it with four state-of-the-art methods. The simulation results demonstrated that the performance of the new proposed strategy is competitive in predicting protein complexes and that adding the GO annotation information will help to increase the prediction accuracy of protein complexes.

Although the new strategy performs well in the detection of protein complexes under the PPI network, the new network obtained by link prediction is still being refined. False-positive and negative links in the networks have not been considered, and the com-

putational complexity is relatively large. Therefore, in the future, we will try to design a new parallel method in predicting unrevealed links and filtering the noise underlying the PPI network.

Acknowledgment

This work was partly supported by the National Natural Science Foundation (No.61802125, No.11626102 and No.61672388), the Natural Science Foundation of Jiangxi Province (No.20181BAB202006 and No.20161BAB211022), and the Science Foundation of Wuhan Institute of Technology(Project No.K201746).

References

1. Spirin V, Mirny LA,Protein complexes and functional modules in molecular networks, *Proceedings of the National Academy of Sciences of the United States of America*, **100**(21):12123–12128, 2003.
2. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, & Séraphin B, A generic protein purification method for protein complex characterization and proteome exploration, *Nature biotechnology* **17**(10):1030–1032, 1999.
3. Fields S, Song O, A novel genetic system to detect protein-protein interactions, *Nature* **340**(6230): 245–246, 1989.
4. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, et al., A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae,*Nature* **403**(6770): 623–627, 2000.
5. Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature* **415**(6868): 141–147, 2002.
6. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y, A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. Sci. USA*. **98**(8):4569–4574, 2001.
7. Dongen SV, Graph Clustering by Flow Simulation, PhD thesis 2000, University of Utrecht.
8. Bader GD, Hogue CW, An automated method for finding molecular complexes in large protein interaction networks, *BMC Bioinformatics* **4**(1): 2, 2003.
9. Ulitsky I, & Shamir, R, Identification of functional modules using network topology and high-throughput data. *BMC systems biology*, **1**(1), 82007.
10. Chua HN, Ning K, Sung WK, Leong HW, & Wong L, Using indirect protein-protein interactions for protein complex prediction. *Journal of bioinformatics and computational biology*, **6**(03), 435-466, 2008.
11. Wu M, Li XL, Kwoh CK, Ng SK, A core-attachment based method to detect protein complexes in PPI networks, *BMC Bioinformatics* **10**(1): 169, 2009.
12. Liu GM, Wong L, Chua HN, Complex discovery from weighted PPI networks, *Bioinformatics* **25**(15)1891–1897, 2009.
13. Nepusz T, Yu HY, Paccanaro A, Detecting overlapping protein complexes in protein-protein interaction networks, *Nature Methods* **9**(5): 471–472, 2012.
14. Lei X, Ding Y, Fujita H, et al. Identification of dynamic protein complexes based on fruit fly optimization algorithm. *Knowledge-Based Systems* **105**: 270–277, 2016.
15. Bandyopadhyay S, Ray S, Mukhopadhyay A, & Maulik U, A multiobjective approach for identifying protein complexes and studying their association in multiple disorders, *Algorithms for Molecular Biology* **10**(1): 24, 2015.

16. Cao B, Luo J, Liang C, Wang S, & Song D. MOEPGA: A novel method to detect protein complexes in yeast proteinCprotein interaction networks based on MultiObjective Evolutionary Programming Genetic Algorithm. *Computational biology and chemistry* **58**:173–181, 2015.
17. Wu M, Ou-Yang L, Li XL. Protein complex detection via effective integration of base clustering solutions and co-complex affinity scores, *IEEE/ACM transactions on computational biology and bioinformatics* **14**(3): 733–739, 2017.
18. Zaki N, Efimov D, Berengueres J, Protein complex detection using interaction reliability assessment and weighted clustering coefficient,*BMC Bioinformatics*,**14**: 163, 2013.
19. Peng W, Wang J, Zhao B, Wang L, Identification of protein complexes using weighted PageRank-Nibble algorithm and core-attachment structure, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*,**12**(1):179–192, 2015.
20. Chen BL, Shi JH, Zhang SG, Wu FX, Identifying protein complexes in protein-protein interaction networks by using clique seeds and graph entropy, *Proteomics*,**13**(2): 269–277, 2012.
21. Shen X, Yi L, Jiang X, Zhao Y, Hu X, He T, & Yang J, Neighbor affinity based algorithm for discovering temporal protein complex from dynamic PPI network, *Methods*, **110**: 90–96, 2016.
22. Kouhsar M, Zare-Mirakabad F, & Jamali Y, WCOACH: Protein complex prediction in weighted PPI networks, *Genes & genetic systems*, **90**(5): 317–324, 2015.
23. Keretsu S, Sarmah R, Weighted edge based clustering to identify protein complexes in proteinCprotein interaction networks incorporating gene expression profile, *Computational biology and chemistry*, **65**:69–79, 2016.
24. Zhang W, Zou X, A new method for detecting protein complexes based on the three node cliques, *IEEE/ACM Trans Comput Biol Bioinform.*, **12**(4): 879–886, 2015.
25. Wang J, Li M, Deng Y, & Pan Y, Recent advances in clustering methods for protein interaction networks. *BMC genomics*, **11**(Suppl 3), S10, 2010.
26. Srihari S, & Leong, HW, A survey of computational methods for protein complex prediction from protein interaction networks. *Journal of bioinformatics and computational biology*, **11**(02), 1230002, 2013.
27. Ji J, Zhang A, Liu C, Quan X, & Liu Z, Survey: Functional module detection from protein-protein interaction networks. *IEEE Transactions on Knowledge and Data Engineering*, **26**(2), 261-277, 2014.
28. Ou-Yang L, Wu M, Zhang XF, Dai DQ, Li XL, & Yan H, A two-layer integration framework for protein complex detection. *BMC bioinformatics*, **17**(1), 100, 2016.
29. Meng J, Zhang X, Luan Y. Global Propagation Method for Predicting Protein Function by Integrating Multiple Data Sources, *Current Bioinformatics* **11**(2): 186–194(9), 2016
30. Zhang W, Xu J, Li Y, Zou X, Detecting Essential Proteins Based on Network Topology, Gene Expression Data and Gene Ontology Information, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **15**(1):109-116, 2018.
31. Yong CH, Liu G, Chua H N, et al. Supervised maximum-likelihood weighting of composite protein networks for complex prediction,*BMC systems biology BioMed Central*, **6**(2): S13, 2012.
32. Wu M, Xie Z, Li X, et al. Identifying protein complexes from heterogeneous biological data, *Proteins: Structure, Function, and Bioinformatics*, **81**(11): 2023-2033, 2013.
33. Zhang Y, Lin H, Yang Z, Wang J, Li Y, Xu B, Protein complex prediction in large ontology attributed protein-protein interaction networks, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **10**(3): 729–741, 2013.

34. Mukhopadhyay A, Ray S, De M, Detecting protein complexes in a PPI network: a gene ontology based multi-objective evolutionary approach, *Molecular Biosystems* **8**(11):3036–3048, 2012.
35. Hanna EM, Zaki N, Amin A, Detecting Protein Complexes in Protein Interaction Networks Modeled as Gene Expression Biclusters, *PLoS ONE* **10**(12): e0144163, 2015.
36. Ou-Yang L, Dai DQ, & Zhang XF, Protein complex detection via weighted ensemble clustering based on Bayesian nonnegative matrix factorization. *PloS one*, **8**(5), e62158, 2013.
37. Zhang Y, Lin H, Yang Z, Jian W, Liu Y, & Sang S, A method for predicting protein complex in dynamic ppi networks. *BMC Bioinformatics*, 2016, **17**(7), 229.
38. Ou-Yang L, Dai DQ, Li XL, Wu M, Zhang XF, & Yang P. Detecting temporal protein complexes from dynamic protein-protein interaction networks. *BMC bioinformatics*, **15**(1), 335, 2014.
39. Li M, Meng X, Zheng R, Wu FX, Li Y, Pan Y, & Wang J. Identification of protein complexes by using a spatial and temporal active protein interaction network, *IEEE/ACM transactions on computational biology and bioinformatics*, DOI 10.1109/TCBB.2017.2749571, 2017.
40. Ma W, McAnulla C, Wang L, Protein complex prediction based on maximum matching with domain-domain interaction, *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1824**(12): 1418–1424, 2012.
41. Sprinzak E, Sattath S, Margalit H, How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology* **327**(5):919–923, 2003.
42. Zhang W, Xu J, Li Y, Zou X, A new two-stage method for revealing missing parts of edges in protein-protein interaction networks, *PLoS One* **12**(5): e0177029, 2017.
43. Li M, Ni P, Chen X, Wang J, Wu F, & Pan Y, Construction of refined protein interaction network for predicting essential proteins, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2017, doi. 10.1109/TCBB.2017.2665482.
44. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF, A new method to measure the semantic similarity of GO terms, *Bioinformatics* **23**(10): 1274–1281, 2007.
45. Pu S, Wong J, Turner B, Cho E, Wodak SJ, Up-to-date catalogues of yeast protein complexes, *Nucleic Acids Research* **37**(3): 825–831, 2008.
46. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, et al.,The MIPS mammalian protein-protein interaction database, *Bioinformatics* **21**(6):832–834, 2004.
47. Aloy P, Böttcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, et al., Structure-based assembly of protein complexes in yeast, *Science* **303**(5666): 2026–2029, 2004.
48. Tu BP, Kudlicki A, Rowicka M, McKnight SL, Logic of the yeast metabolic cycle: temporal Compartmentalization of cellular processes, *Science* **310**(5751): 1152–1158, 2005.
49. Brohée S, Van Helden J, Evaluation of clustering algorithms for protein-protein interaction networks, *BMC Bioinformatics* **7**:488, 2006.