

RESEARCH

Open Access



Finding low-conductance sets with dense interactions (FLCD) for better protein complex prediction

Yijie Wang and Xiaoning Qian*

From Third International Workshop on Computational Network Biology: Modeling, Analysis, and Control (CNB-MAC 2016) Seattle, WA, USA. 10 February 2016

Abstract

Background: Intuitively, proteins in the same protein complexes should highly interact with each other but rarely interact with the other proteins in protein-protein interaction (PPI) networks. Surprisingly, many existing computational algorithms do not directly detect protein complexes based on both of these topological properties. Most of them, depending on mathematical definitions of either “modularity” or “conductance”, have their own limitations: Modularity has the inherent resolution problem ignoring small protein complexes; and conductance characterizes the separability of complexes but fails to capture the interaction density within complexes.

Results: In this paper, we propose a two-step algorithm FLCD (**F**inding **L**ow-**C**onductance sets with **D**ense interactions) to predict overlapping protein complexes with the desired topological structure, which is densely connected inside and well separated from the rest of the networks. First, FLCD detects well-separated subnetworks based on approximating a potential low-conductance set through a personalized PageRank vector from a protein and then solving a mixed integer programming (MIP) problem to find the minimum-conductance set within the identified low-conductance set. At the second step, the densely connected parts in those subnetworks are discovered as the protein complexes by solving another MIP problem that aims to find the dense subnetwork in the minimum-conductance set.

Conclusion: Experiments on four large-scale yeast PPI networks from different public databases demonstrate that the complexes predicted by FLCD have better correspondence with the yeast protein complex gold standards than other three state-of-the-art algorithms (ClusterONE, LinkComm, and SR-MCL). Additionally, results of FLCD show higher biological relevance with respect to Gene Ontology (GO) terms by GO enrichment analysis.

Keywords: Protein complex identification, Low conductance set, Dense subnetwork, Mixed integer programming

Background

Recent developments of high-throughput profiling techniques, such as yeast two-hybrid (Y2H) and tandem affinity purification (TAP) with mass spectrometry (MS), allow scientists to generate large-scale protein-protein interaction (PPI) datasets for different species [1–5]. These interactome data have enabled us to discover biological

insights from a systematic point of view through PPI networks, where nodes represent proteins and edges denote biological relationships (either physical binding or statistical association) between two proteins. In this paper, we focus on predicting protein complexes in derived PPI networks from high-throughput profiling.

Based on the inherent topological structures of protein complexes [6], prediction of protein complexes can be formulated as searching for subnetworks that are densely connected inside and well separated from the rest of the

*Correspondence: xqian@ece.tamu.edu
Department of Electrical & Computer Engineering, Texas A and M University,
MS 3128, TAMU, College Station, TX, USA

PPI networks. Many algorithms have been developed and applied for this purpose of detecting protein complexes.

These existing algorithms can be grouped into three categories. The first category includes the algorithms that mimic Markovian random walk on graphs, pioneered by MCL [7]. MCL does not have explicit mathematical definitions for the desired properties of subnetworks to detect as protein complexes. Similar to random walk, it iteratively implements “Expand” and “Inflation” operations to generate non-overlapping complexes. R-MCL [8] and SR-MCL [9] are improved versions of MCL. R-MCL penalizes the large complexes at each iteration in order to obtain more size-balanced complexes with a similar number of nodes within them. SR-MCL executes R-MCL many times to yield overlapping complexes. All those algorithms have shown good empirical performance, despite the mystery of parameter tuning and the lack of theoretic understanding of their working mechanisms.

Algorithms in the second category do not directly predict complexes according to the topological structure of subnetworks but resemble traditional clustering methods based on derived similarity measures between nodes or edges. For example, MCODE [1], CFinder [10], and RRW [11] grow complexes from single nodes by iteratively adding similar nodes in terms of different similarity criteria that help form local dense subnetworks. However, they only concentrate on the internal connectivity of the subnetworks and neglect the connectivity between the subnetworks and the rest of the networks. LinkComm [12] represents networks with edge graphs, whose nodes are interactions and edges reflect the similarity between interactions, and derives potential complexes by hierarchical clustering to partition the edge graphs.

Algorithms in the third category detect complexes based on explicit topological definitions of protein complexes. For example, modularity [13] and conductance [6, 14] are two widely used definitions. Algorithms based on modularity [15] aim to detect subnetworks that have higher than expected internal connections. And algorithms, such as ClusterONE [6], based on finding low-conductance sets, focus on the separability of the subnetworks, which can be quantified by the ratios between the external connections of subnetworks and the total number of interactions of the proteins within the subnetworks. However, these methods have their own limitations. Modularity-based methods have the inherent resolution problem [16], which leads to ignorance of small-size protein complexes. Algorithms based on conductance minimization [6, 17] consider the relationships between the internal connections and the external connections of subnetworks, but neglect the density of the interactions within the subnetworks.

In this paper, we propose a two-step algorithm FLCD (Finding Low-Conductance sets with Dense interactions) to detect protein complexes that have dense interactions inside and sparse interactions outside in a given PPI network. FLCD explicitly takes care of both the internal and external connectivity of protein complexes in two steps. FLCD first identifies a low-conductance set around a protein, which is locally well separated from the rest of the network. Then a densely connected subnetwork within the low-conductance set is detected based on the definition of the edge density of a subnetwork proposed in [18]. We compare our FLCD with three state-of-the-art overlapping complex prediction algorithms, which are ClusterONE [6], LinkComm [12], and SR-MCL [9], respectively. Experimental results on four different yeast PPI networks from different publicly accessible databases demonstrate that our FLCD outperforms all competing algorithms for biological significance in terms of yeast protein complex gold standards and Gene Ontology (GO) term annotations [19].

Results and discussion

We first introduce the implementation details of the algorithms that we take for comparison; the information of the PPI networks, the reference protein complex datasets as our gold standards, and the GO terms we use for evaluation; and the criteria for the performance comparison. In order to demonstrate the robust performance of FLCD, we then compare predicted protein complexes from *three* selected state-of-the-art protein complex prediction algorithms based on *two* golden standard protein complex datasets on *four* public yeast PPI networks. What's more, we apply GO enrichment analysis to the entire set of detected complexes by all the competing algorithms. At the end, we illustrate differences between protein complexes predicted by all competing algorithms corresponding to specific reference complexes to further demonstrate the superiority of our FLCD.

Algorithms, data, and evaluation metrics

Algorithms

We compare our FLCD algorithm with other three state-of-the-art overlapping complex prediction algorithms, which are ClusterONE [6], LinkComm [12], and SR-MCL [9]. The JAVA implementation of ClusterONE does not require any tuning parameters. For LinkComm, we set the tuning parameter t (the threshold to cut the dendrogram for hierarchical clustering) to 0.2 that achieves the best performance empirically in our experiments. For SR-MCL, we set the inflation parameter $I = 3$ and other parameters to their default settings since they yield the best results in our experiments. We set the only parameter k of our FLCD, the size of local neighbors based on personalized PageRank computation, to 20.

Data

We take four yeast PPI networks for performance evaluation: SceDIP, SceBG, SceIntAct, and SceMINT, extracted respectively from the Database of Interacting Proteins (DIP) [2], the Biological General Repository for Interaction Datasets (BioGRID) [3], the IntAct Molecular Interaction Database (IntAct) [4], and the Molecular INteraction database (MINT) [5]. We note that we only consider protein-protein interactions by removing all genetic interactions from SceBG. We download the protein complex gold standards from the supplementary data in [6], which are obtained from the Saccharomyces Genome Database (SGD) [20] and the Munich Information Center for Protein Sequences (MIPS) [21] databases. For each PPI network, we remove reference protein complexes if their size smaller than 3 or half of the proteins of them are not in the network. The detailed information of four PPI networks and the gold standard reference complex datasets are provided in Table 1.

Due to the possible incompleteness of the reference protein complexes, we further examine the biological relevance of every predicted complex by GO enrichment analysis. We download the mappings of yeast genes and proteins to GO terms according to [20] (version 20150411).

Evaluation metrics for protein complex prediction

For the protein complex prediction, we assess the performance of all competing algorithms by a composite score consisting of three quality measures: F-measure [9, 14]; the geometric accuracy (Acc) score [14]; and the maximum matching ratio (MMR) [6]. For fair comparison, we remove predicted complexes of two or fewer proteins by all competing algorithms.

For a gold standard reference protein complex set $C = \{c_1, c_2, \dots, c_n\}$ and a set of predicted complexes $S = \{s_1, s_2, \dots, s_m\}$, the F-measure is defined as the harmonic mean of precision and recall defined as follows:

$$\text{precision} = \frac{|N_{cs}|}{|S|}; \quad \text{recall} = \frac{|N_{cp}|}{|C|}, \quad (1)$$

in which $N_{cs} = \{s_i \in S | NA(c_j, s_i) \geq 0.25, \exists c_j \in C\}$ is the set of the complexes that match to one or more reference protein complexes; $|N_{cs}|$ is the size of the set N_{cs} .

$N_{cp} = \{c_i \in C | NA(c_i, s_j) \geq 0.25, \exists s_j \in S\}$ is a set of reference protein complexes that are matched by predicted complexes. We consider a reference protein complex c_j is matched by a predicted complex s_j if $NA(c_i, s_j) \geq 0.25$ [9, 22], where $NA(c_i, s_j) = \frac{|c_i \cap s_j|^2}{|c_i| \times |s_j|}$ is called neighborhood affinity. Finally, the F-measure is

$$\text{F-measure} = 2 \times \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \quad (2)$$

The geometric accuracy (Acc) score is the geometric mean of two other measures — the cluster-wise sensitivity (Sn) and cluster-wise positive predictive value (PPV) [6]. Given m predicted and n reference complexes, let t_{ij} denote the number of proteins that exist in both predicted complex s_i and reference complex c_j , and w_j represent the number of proteins in reference complex c_j . Then Sn and PPV can be computed as

$$\text{Sn} = \frac{\sum_{j=1}^n \max_{i=1, \dots, m} t_{ij}}{\sum_{j=1}^n w_j}; \quad \text{PPV} = \frac{\sum_{i=1}^m \max_{j=1, \dots, n} t_{ij}}{\sum_{i=1}^m \sum_{j=1}^n t_{ij}}. \quad (3)$$

The Acc score provides a balanced measure of Sn and PPV: $\text{Acc} = \sqrt{\text{Sn} \times \text{PPV}}$.

The maximum matching ratio (MMR) is the ratio of the weight of maximum weight matching to the size of the reference set.

GO enrichment analysis

Suppose that a given PPI network has N proteins with M proteins annotated with one GO term and the predicted complex has n proteins with m proteins annotated with the same GO term. The p -value of the complex enriched with that GO term can be calculated as similarly done in [23]:

$$p\text{-value} = \sum_{i=m}^n \frac{\binom{m}{i} \binom{N-M}{N-i}}{\binom{N}{n}}. \quad (4)$$

We choose the lowest p -value of all its enriched GO terms for a predicted complex as its final p -value. A GO term is statistically significantly enriched when the p -value of any complex corresponding to this GO term is lower than $1e-3$.

Comparison on protein complex prediction

We apply all competing algorithms to search for potential protein complexes in four yeast PPI networks and compare them in terms of the composite score, consisting of F-measure, Acc score and MMR based on both the SGD and MIPS reference protein complex datasets.

We note that the different sizes and different numbers of detected complexes would affect the scores for the metrics that we have employed. However, in the context of complex prediction, there is no universal gold-standard

Table 1 The detailed information of four yeast PPI networks and the numbers of covered SGD and MIPS reference complexes

Network	#. proteins	#. interactions	SGD	MIPS
SceDIP	5136	22491	224	184
SceBG	6438	80577	234	189
SceIntAct	5453	54134	231	187
SceMINT	5414	27316	230	188

metric. Hence, we apply three aforementioned metrics that have been commonly adopted in many other related works [6, 9]. We also note that the average sizes of the complexes generated by FLCD in our experiments are from 6 to 8 for four networks under study. The average complex sizes are indeed comparable to the average sizes of detected complexes by other algorithms. For example, the average sizes of complexes produced by LinkCommunity are from 5 to 6; The average sizes of complexes produced by ClusterONE are from 7 to 9; The average sizes of complexes produced by SR-MCL are from 8 to 10. Furthermore, the total numbers of predicted complexes yielded by FLCD, LinkCommunity and SR-MCL are much larger than that of ClusterONE. The reason is that the post-processing procedure of ClusterONE filters out complexes with lower scores but FLCD and LinkCommunity output all complexes without filtering.

As shown in Figs. 1 and 2, FLCD clearly outperforms other state-of-the-art algorithms for all four networks on both SGD and MIPS reference datasets. Therefore, the complexes detected by FLCD have the best correspondence with the reference datasets. The detailed evaluation scores in Figs. 1 and 2 are displayed in Tables 2 and 3, respectively.

When we take SGD reference dataset as our gold standard protein complexes, from Table 2, we find that FLCD consistently achieves the best MMR scores among all competing algorithms because FLCD is the only algorithm that can capture the desired network structure of protein complexes. In the table, we also compare F-measure and the precision and recall scores that are used to compute F-measure. We observe that for all four PPI networks, FLCD predicts the largest number of matched

reference protein complexes, and therefore FLCD attains the best recall scores for all PPI networks. With respect to the precision score, FLCD is the best for SceMINT but ClusterONE performs the best for the rest. However, since the post-processing step in ClusterONE only keeps the dense complexes, ClusterONE has low coverage. Based on the precision and recall scores, we find that FLCD attains the best F-measures for SceDIP and SceMINT PPI networks and ClusterONE obtains the best scores for SceBG and SceIntAct PPI networks. In addition to MMR and F-measure, we show comparison on the cluster-wise sensitivity (S_n), the cluster-wise positive predictive value (PPV) and the Acc score. We notice that FLCD has the best Acc scores for SceBG and SceIntAct. LinkComm obtains the best Acc scores for SceDIP and SceMINT, since LinkComm detects several large-size and many small-size complexes, which favors both the S_n and PPV scores [6]. We also compare the coverage of the competing algorithms and notice that SR-MCL has the largest coverage and FLCD has competitive coverage to SR-MCL. Here, the coverage is defined as the number of proteins covered by all predicted complexes, which is typically used to evaluate whether complex prediction algorithms can help comprehensively predict functionalities for all the proteins in a given network.

For MIPS reference dataset, we notice the similar trend for the evaluation scores in Table 3. FLCD finds the largest number of matched reference complexes in MIPS and attains the best recall scores, F-measures and MMR scores for all four PPI networks. The Acc scores of FLCD are competitive to LinkComm, which achieves the best Acc scores for all four yeast PPI networks. FLCD covers the competitive number of proteins to SR-MCL, which covers

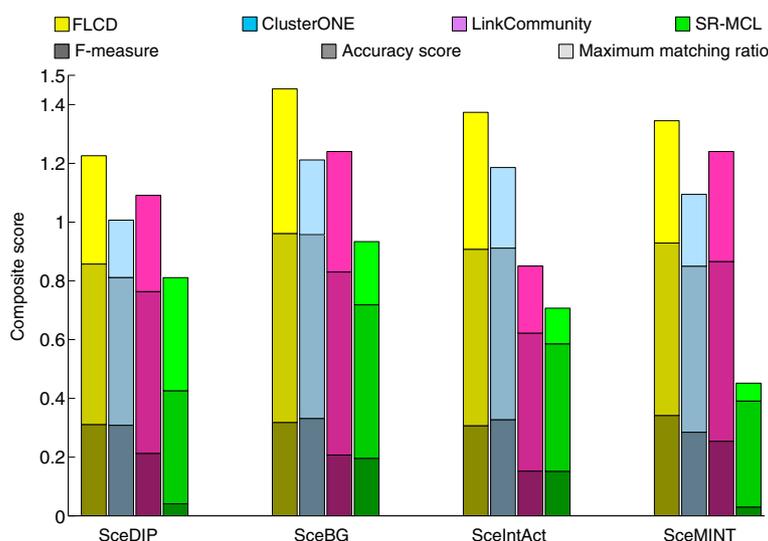


Fig. 1 Comparison of all competing algorithms by SGD reference dataset in terms of the composite scores. *Shades of the same color* indicate different evaluating scores. Each bar height reflects the value of the composite score

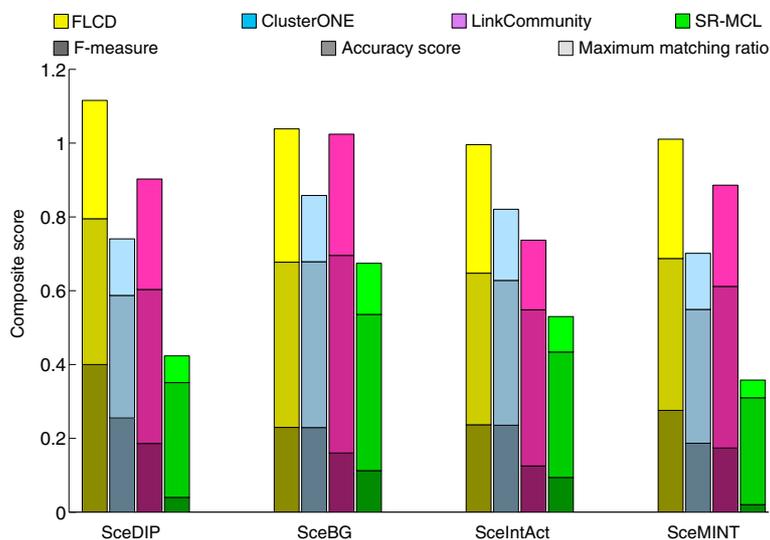


Fig. 2 Comparison of all competing algorithms by MIPS reference dataset in terms of the composite scores. Shades of the same color indicate different evaluating scores. Each bar height reflects the value of the composite score

the largest number of proteins in all four yeast PPI networks. However, by the overall performance, which is represented by the composite score, FLCD is superior to other competing algorithms as shown in Fig. 2.

In summary, considering the composite score based on three metrics, our FLCD outperforms the other algorithms. To further validate all competing algorithms, we

perform GO enrichment analysis in the next section to see whether all predicted complexes by different algorithms have significant biological meaning.

Comparison on GO enrichment analysis

We perform GO enrichment analysis for all protein complexes predicted by the competing algorithms and report

Table 2 Comparison of protein complex prediction by SGD reference dataset

Network	Method	# complex	#. matched	coverage	Recall	Precision	F-measure	Sn	PPV	Acc	MMR
SceDIP	FLCD	2134	152	3921	0.6786	0.2020	0.3113	0.5964	0.5003	0.5462	0.3685
	CONE	380	86	1503	0.3839	0.2579	0.3085	0.4082	0.6203	0.5032	0.1950
	LinkC	1839	137	3735	0.6116	0.1289	0.2130	0.6290	0.4820	0.5506	0.3276
	SR-MCL	3216	44	4678	0.2228	0.0221	0.0412	0.5120	0.2893	0.3489	0.0708
SceBG	FLCD	4027	183	5836	0.7821	0.2000	0.3181	0.7363	0.5621	0.6433	0.4920
	CONE	522	122	2735	0.5214	0.2433	0.3318	0.6488	0.6035	0.6257	0.2542
	LinkC	5382	164	6076	0.7008	0.1217	0.2072	0.8880	0.4373	0.6231	0.4100
	SR-MCL	1862	108	5889	0.4615	0.1245	0.1961	0.8999	0.3034	0.5225	0.2151
SceIntAct	FLCD	3394	172	4678	0.7446	0.1933	0.3069	0.6699	0.5391	0.6009	0.4661
	CONE	496	117	1994	0.5065	0.2419	0.3275	0.5742	0.5944	0.5842	0.2742
	LinkC	1297	93	5290	0.4026	0.0941	0.1525	0.9223	0.2393	0.4698	0.2285
	SR-MCL	1079	68	5342	0.2294	0.0437	0.1517	0.7784	0.2402	0.4341	0.1213
SceMINT	FLCD	2483	157	4210	0.6826	0.2280	0.3418	0.6524	0.5284	0.5871	0.4163
	CONE	513	110	2335	0.4783	0.2027	0.2848	0.5370	0.5954	0.5654	0.2442
	LinkC	2201	144	4068	0.6261	0.1595	0.2542	0.6757	0.5540	0.6119	0.3743
	SR-MCL	3698	33	4976	0.1435	0.0169	0.0302	0.5013	0.2597	0.3608	0.0609

CONE and LinkC are short for ClusterONE and LinkComm, respectively
 Bold values denote the best scores corresponding to specific criteria

Table 3 Comparison of protein complex prediction by MIPS reference dataset

Network	Method	# complex	#. matched	Coverage	Recall	Precision	F-measure	Sn	PPV	Acc	MMR
SceDIP	FLCD	2134	120	3921	0.6522	0.1603	0.2573	0.4001	0.3901	0.3951	0.3206
	CONE	380	74	1503	0.4022	0.1868	0.2551	0.2749	0.4015	0.3322	0.1533
	LinkC	1839	109	3735	0.5924	0.1104	0.1862	0.4775	0.3646	0.4173	0.2993
	SR-MCL	2851	41	4687	0.1964	0.0230	0.0402	0.4592	0.2104	0.3108	0.0726
SceBG	FLCD	4027	124	5836	0.6561	0.1393	0.2298	0.4643	0.4315	0.4476	0.3611
	CONE	522	86	2735	0.4450	0.1533	0.2293	0.4537	0.4452	0.4494	0.1795
	LinkC	5382	109	6076	0.6349	0.0918	0.1604	0.8179	0.3504	0.5354	0.3285
	SR-MCL	1862	65	5889	0.3439	0.0673	0.1126	0.7360	0.2436	0.4234	0.1384
SceIntAct	FLCD	3394	120	4678	0.6417	0.1452	0.2368	0.4183	0.4034	0.4108	0.3482
	CONE	496	79	1994	0.4225	0.1633	0.2356	0.3587	0.4296	0.3925	0.1927
	LinkC	1297	80	5290	0.4278	0.0732	0.1251	0.9028	0.1986	0.4234	0.1886
	SR-MCL	1079	45	5342	0.1337	0.0190	0.0941	0.6246	0.1850	0.3399	0.0960
SceMINT	FLCD	2483	111	4210	0.5904	0.1800	0.2759	0.4147	0.4086	0.4116	0.3231
	CONE	513	67	2335	0.3564	0.1267	0.1869	0.3274	0.4017	0.3626	0.1519
	LinkC	2201	100	4068	0.5319	0.1040	0.1740	0.4744	0.4038	0.4377	0.2744
	SR-MCL	3698	24	4976	0.1277	0.0112	0.0205	0.4192	0.1999	0.2894	0.0481

CONE and LinkC are short for ClusterONE and LinkComm, respectively
 Bold values denote the best scores corresponding to specific criteria

the percentages of the predicted protein complexes that are significantly enriched with at least one GO term and the total number of GO terms that are enriched in the predicted complexes in Table 4. We find that our FLCD achieves the best percentages of the enriched predicted protein complexes in SceDIP and SceIntAct PPI networks. ClusterONE obtains the best percentages for SceBG and SceMINT PPI networks but with the smaller number of GO terms enriched in the detected complexes because ClusterONE may remove meaningful functional modules in its post-processing step. Furthermore, the protein complexes detected by FLCD are significantly associated with the largest number of GO terms over all competing algorithms on all four PPI networks.

To further examine the statistical significance of the complexes detected by the competing algorithms, we compare the p -values of the complexes under GO terms of biological process, molecular function, and cellular component domains. We use the lowest p -value for each predicted complex and show the comparison of the statistical significance of the complexes detected by all competing algorithms in Fig. 3. The y -axis of Fig. 3 represents the negative \log - p -values while the x -axis is the ordered list of the complexes detected by all competing algorithms in terms of their negative \log - p -values. Since complexes with significant biological relevance have lower p -values, higher values in Fig. 3 represent the higher quality of the detected complexes. As shown in Fig. 3, for all four

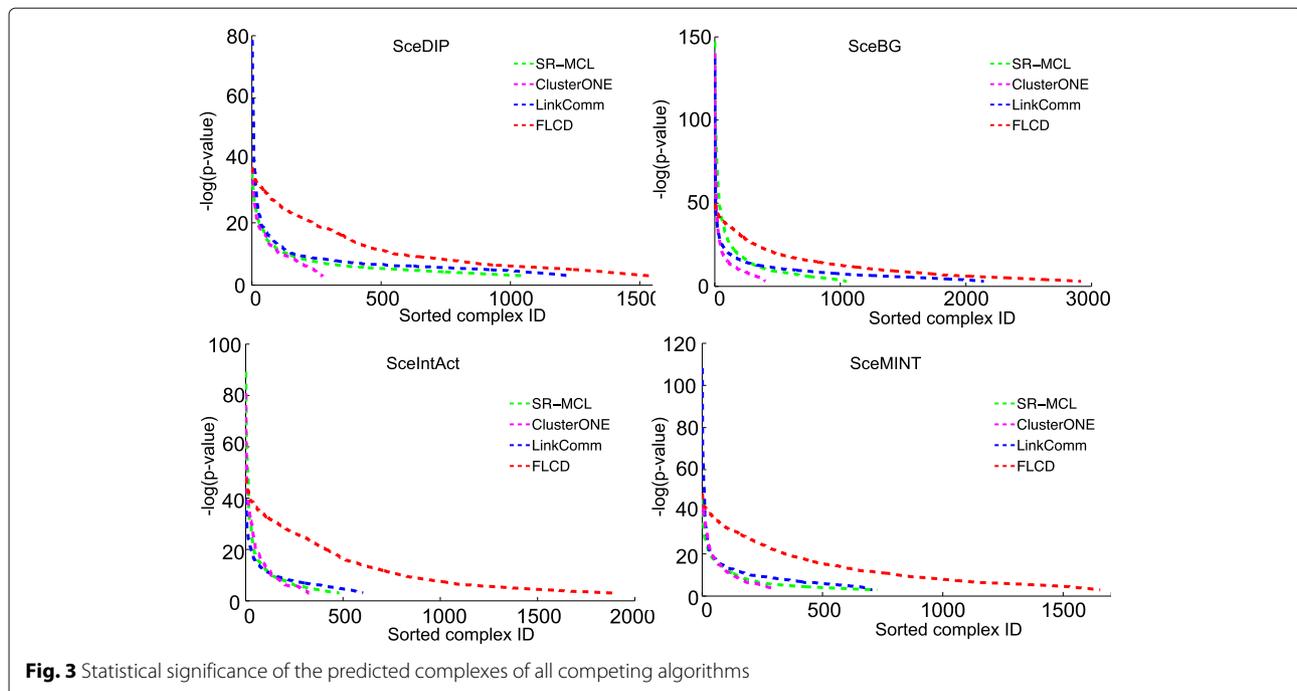
Table 4 Comparison by GO enrichment analysis

Network	Method	# complex	% enriched	# GO
SceDIP	FLCD	2134	72.2	1442
	CONE	380	71.8	852
	LinkC	1839	67.4	1273
	SR-MCL	2851	23.5	957
SceBG	FLCD	4027	72.4	1800
	CONE	522	77.4	1282
	LinkC	5382	39.8	1554
	SR-MCL	1862	56.4	1702
SceIntAct	FLCD	3394	62.4	1414
	CONE	496	65.6	1031
	LinkC	1297	46.5	1129
	SR-MCL	1079	44.7	888
SceMINT	FLCD	2483	62.3	1416
	CONE	513	59.4	954
	LinkC	2201	32.1	1123
	SR-MCL	3698	19.7	856

"% enriched" presents the percentage of complexes that are enriched with at least one GO term.

"# GO" denotes the number of enriched GO terms

Bold values denote the best scores corresponding to specific criteria



yeast PPI networks, in addition to the fact that FLCD detects significantly more GO-enriched complexes, FLCD clearly outperforms other competing algorithms because the curves of FLCD are consistently on top of the others. The outperformance of FLCD further demonstrates that network structure that has dense internal connectivity and sparse external connectivity can better depict complexes of biological significance and FLCD provides an effective way to predict complexes with the desired network structure through explicitly taking care of internal and external connectivity of potential subnetworks.

Examples of predicted complexes

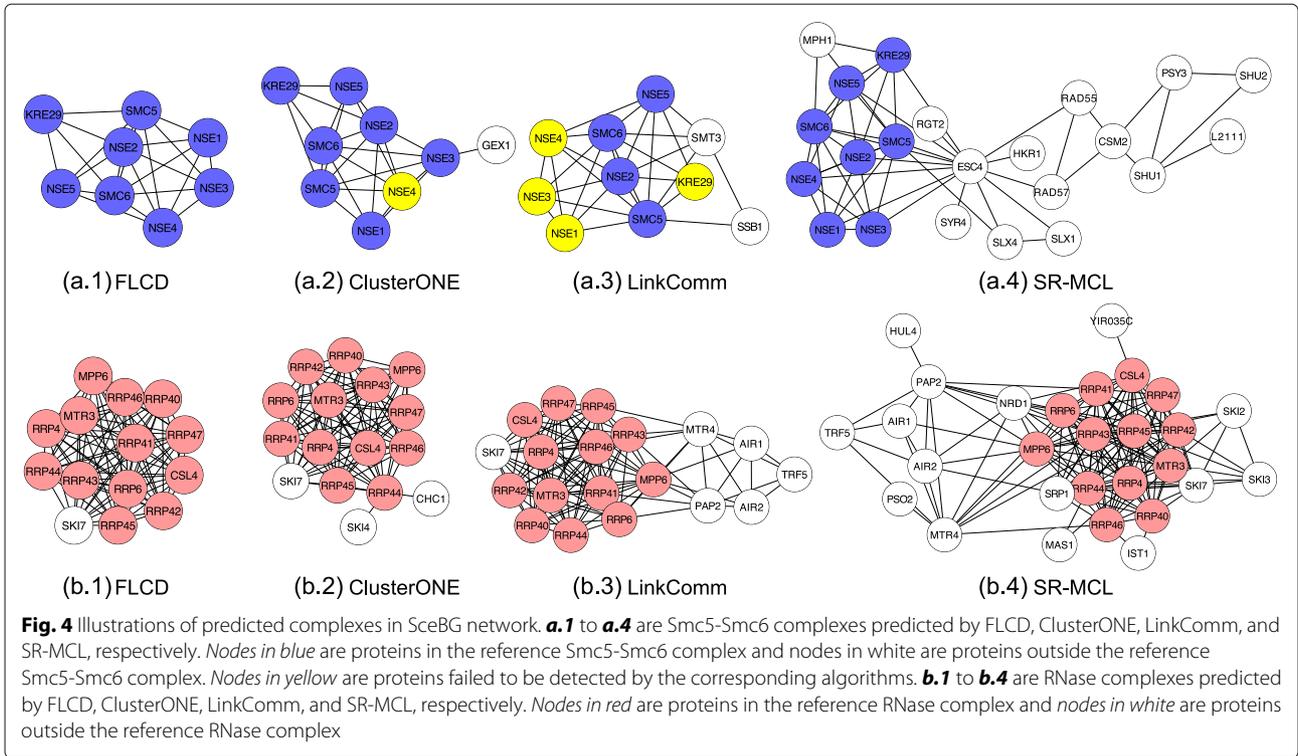
We further show the differences between the competing algorithms by illustrating the predicted complexes corresponding to two specific reference protein complexes. The first reference protein complex is the Smc5-Smc6 complex. In Fig. 4, the Smc5-Smc6 complexes predicted by FLCD, ClusterONE, LinkComm, and SR-MCL are displayed from (a.1) to (a.4), respectively. We notice that FLCD successfully identifies the Smc5-Smc6 complex as shown in Fig. 4(a.1). ClusterONE fails to detect the protein annotated as NSE4, probably due to the inaccuracy of the greedy algorithm used in ClusterONE. Also, we find that the protein annotated as GEX1 only interacts with the protein NSE3 but it is falsely added to the Smc5-Smc6 complex by ClusterONE. Because ClusterONE focuses on the separability of a complex but does not directly consider the internal density of the complex, it may mistakenly add proteins with small degrees into the final

result. The complex in Fig. 4(a.3) predicted by LinkComm contains false positives and false negatives since the similarities between interactions used in LinkComm can not describe the topological structure of protein complexes. In Fig 4(a.4), we find out that the Smc5-Smc6 complex predicted by SR-MCL consists of many false positives. However, it is hard to explain the performance of SR-MCL on predicting the Smc5-Smc6 complex due to the unclear working mechanism of SR-MCL.

Similarly, we show the predicted RNase complexes by all competing algorithms in Fig. 4 from (b.1) to (b.4). In (b.1), we observe that FLCD detects all proteins in the reference RNase complex but mistakenly includes the protein SKI7 due to the existence of false positive interactions between SKI7 and proteins in RNase complex. In addition to SKI7, the predicted complex by ClusterONE (shown in Fig. 4(b.2)) contains two false positive proteins with very small degrees due to the ignorance of the internal density. Because LinkComm does not explicitly characterize the separability of the complexes, it also recruits some false positive proteins as clearly shown in Fig. 4(b.3). For the complex obtained by SR-MCL, we note that it has lots of false positive proteins and the topological property of the predicted complex is not clear.

Conclusions

We propose an algorithm FLCD to predict protein complexes in protein-protein interaction networks. FLCD can better characterize the topological structure of a protein complex, which is densely connected inside and well



separated from the rest of the networks. We compare FLCD with other three state-of-the-art algorithms on protein complex prediction. The comparison results show that FLCD achieves superior performances. Furthermore, GO enrichment analysis of the results of the competing algorithms demonstrates that FLCD finds more biologically meaningful complexes, within which proteins tend to be in the same cellular components and have similar functions and/or participate in the same biological processes.

Methods

Terminologies and definitions

Let an undirected graph $G = (V, E)$ represent a PPI network, where V denotes the set of proteins in G and E is the interaction set. A is the adjacency matrix of G with $A_{ij} = A_{ji}$ and $A_{ij} = 1$ denoting node i interacts with node j and $A_{ij} = 0$ otherwise. The degree matrix D of G is a diagonal matrix with $D_{ii} = d_i$, where $d_i = \sum_j A_{ij}$ is the number of interactions connecting to protein i .

For a set S of proteins, the conductance of S in G is defined as [17]

$$\phi(S) = \frac{|E(S, \bar{S})|}{\min\{vol(S), vol(\bar{S})\}}, \quad S \cup \bar{S} = V, \quad (5)$$

where $E(S, \bar{S})$ denotes the edge cut, the set of edges between the set S and its complement set \bar{S} , $|\cdot|$ denotes the set size, and $vol(T) = \sum_{i \in T} d_i$ is the number of all

incident interactions of the set T . Here we make a mild assumption that $vol(S) \ll vol(V)$ for a small protein complex S in the large-scale PPI network G , which means $vol(S) = \min\{vol(S), vol(\bar{S})\}$. Hence, we have

$$\phi(S) = \frac{|E(S, \bar{S})|}{vol(S)} = \frac{\sum_i (D_{ii}^S - \sum_j A_{ij}^S)}{\sum_i D_{ii}^S}, \quad (6)$$

where A^S is the adjacency matrix of the induced subnetwork with respect to set S and D^S is the degree matrix for the nodes in S , where $D_{ii}^S = \sum_j A_{ij}^S = d_i$ for $i \in S$. For the same set S , the density of S is defined as [18]

$$\mathcal{D}(S) = \frac{|E(S, S)|}{|S|} = \frac{1}{2} \frac{\sum_{ij} A_{ij}^S}{\sum_i \mathbf{1}_{i \in S}}, \quad (7)$$

where $\mathbf{1}_{i \in S}$ is the indicator function depending on whether $i \in S$.

Motivation

FLCD is motivated by conductance minimization to identify well separated subnetworks in a given network. However, FLCD can overcome the problem of conductance minimization, which pays no attention to the internal connectivity within subnetworks as potential protein complexes. Figure 5 shows a motivating example: We can find two complexes enclosed in the red dotted lines in the network based on conductance minimization. The conductances of the complexes within red dotted lines are $\frac{2}{11}$

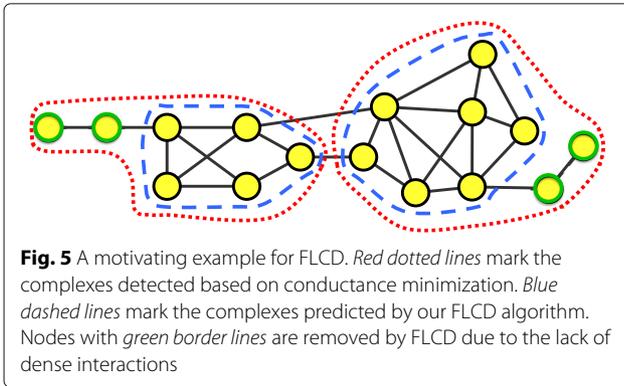


Fig. 5 A motivating example for FLCD. Red dotted lines mark the complexes detected based on conductance minimization. Blue dashed lines mark the complexes predicted by our FLCD algorithm. Nodes with green border lines are removed by FLCD due to the lack of dense interactions

and $\frac{2}{17}$ and the conductances of complexes within blue dashed lines are $\frac{3}{10}$ and $\frac{3}{16}$. Obviously, the conductances of the complexes within red dotted lines are lower than the complexes within blue dashed lines, indicating that the complexes within red dotted lines are topologically more separable than the complexes within blue dashed lines. However, the complexes within the blue dashed lines are more likely to be the desired complexes since the nodes with green border lines can not be confidently grouped into potential protein complexes due to their low degrees.

FLCD explicitly considers both the separability and internal edge density of complexes in two steps respectively. At the first step, it takes care of the separability of complexes by ensuring low conductance to hope for the complexes to have unique biological functions. At the second step, FLCD preserves the densely connected parts of the complexes identified in the first step. Because PPI networks are noisy and typically sparse, instead of finding cliques, we use the definition of internal density in (7) to search for dense subnetworks as final predicted complexes.

Searching for a low-conductance set H_v^*

Given a starting protein v , our goal is to find a protein set H_v^* with low conductance including v . We first apply the algorithm proposed in [17] to find a potential set H with low conductance, then the minimum-conductance set H_v^* in H is identified through solving a mixed integer programming (MIP) problem exactly.

Following [17], a low-conductance set including v can be efficiently approximated via the personalized PageRank vector of v . The personalized PageRank vector $p(\alpha, v)$ of v on G is the stationary distribution of the random walk on G , in which at every step, the random walker has the probability of α to restart the random walk at v and otherwise performs a lazy random walk. Mathematically, $p(\alpha, v)$ is the unique solution to

$$p(\alpha, v) = \alpha e_v + (1 - \alpha)p(\alpha, v)W, \quad (8)$$

where $\alpha \in (0, 1]$ is the “teleportation” constant, e_v is the indicator vector of v and $W = \frac{1}{2}(I + D^{-1}A)$ is the underlying probability transition matrix of the lazy random walk. We apply the local algorithm in [17] to efficiently approximate $\hat{p} \approx p(\alpha, v)$. Then we sort the nodes based on \hat{p} and attain an ordered set $\mathcal{H} = \{v_1, v_2, \dots, v_n\}$, whose elements satisfy $\hat{p}(v_i) > \hat{p}(v_{i+1})$. Inspired by PageRank-Nibble [17] that sweeps the ordered set \mathcal{H} to get the low-conductance set, we propose to find the minimum low-conductance set within a subnetwork of size k , which consists of the top k elements in \mathcal{H} , by solving a MIP problem. We take the top k elements out of \mathcal{H} , which are more likely to comprise a low-conductance set with v , and put them in H . The minimum-conductance set H_v^* in H can be derived by solving the following optimization problem based on (6):

$$\begin{aligned} \min: & \frac{x^T (D^H - A^H) x}{x^T d^H} \\ \text{s.t. } & x_v = 1, x_i \in \{0, 1\}, \end{aligned} \quad (9)$$

where x is a binary vector with $x_i = 1$ indicating that node i in H is assigned into H_v^* and $x_i = 0$ otherwise; and d^H is a vector containing the degrees of every node in H . We force node v to be in the low-conductance set by setting $x_v = 1$. By algebraic manipulations, (9) can be transformed into the following equivalent formulation:

$$\begin{aligned} \min: & z \\ \text{s.t. } & z \sum_i x_i d_i^H - \sum_i \sum_j (D_{ij}^H - A_{ij}^H) x_i x_j \geq 0, \\ & x_v = 1, x_i \in \{0, 1\}. \end{aligned} \quad (10)$$

After using standard techniques [24] to linearize $z x_i$ and $x_i x_j$, the optimization problem can be solved by any MIP solver, such as Gurobi [25]. Because the size of $|H| = k$ is much smaller than $|V| = n$ and we only focus on identifying one low-conductance set, we can efficiently obtain the minimum-conductance set H_v^* in H by solving (10) exactly.

If node v is in a connected component of size k' and we set $k > k'$, then we might have a trivial solution that the low-conductance set is the connected component with conductance 0. To avoid this, we apply the following procedure. We check every derived low-conductance set of size k' to see whether it has exactly 0 conductance, which implies that it is a connected component with size k' . If that is the case, we then set $k = k' - 1$, and re-solve the MIP to get a non-trivial solution.

Conservation of the densest subnetwork C_v^* in H_v^*

The induced subnetwork G_v with respect to the protein set H_v^* is well separated from the rest of the network; however, there may exist nodes with low degrees in H_v^* . As illustrated in Fig. 5, to remove low-degree nodes (nodes

with green border lines) as well as reserve densely connected subnetworks, we apply the definition of the internal density (7) to find the densest subnetwork in H_v^* . Because the problem size is small for such a local optimization problem, we can again take the full advantages of the power of MIP solvers. The node set $C_v^* \in H_v^*$ corresponding to the densest subnetwork can be identified based on (7) by deriving the exactly optimal solution to the following MIP problem:

$$\begin{aligned} \max: & \frac{r^T A_{ij}^{H_v^*} r}{r^T \mathbf{1}} \\ \text{s.t. } & r_i \in \{0, 1\}, \end{aligned} \quad (11)$$

where $\mathbf{1}$ is an all-one vector and r is the binary vector indicating the memberships of the nodes from H_v^* in the densest subnetwork. This optimization problem explicitly searches for the subnetwork with the highest internal density and it can be transformed into the equivalent problem, as similarly done in (10):

$$\begin{aligned} \max: & w \\ \text{s.t. } & w \sum_i r_i - \sum_i \sum_j A_{ij}^{H_v^*} r_i r_j \leq 0, \\ & r_i \in \{0, 1\}, \end{aligned} \quad (12)$$

which can also be cast into the MIP framework with the exactly optimal solution obtained by using standard MIP solvers after linearization [24].

The FLCD algorithm

The step-by-step procedure of FLCD algorithm is given in Table 5. The FLCD algorithm screens every protein with degree higher than two. For each selected protein, the FLCD algorithm first searches for the minimum-conductance set around it and then finds the densest subnetwork in the minimum-conductance set, which is

Table 5 The FLCD algorithm

Algorithm: The FLCD Algorithm

Input: $\mathcal{S} = V$ and $k = 20$.

Output: A set of predicted complexes R .

- 1 While ($\exists v \in \mathcal{S}$ and $d_v \geq 3$)
- 2 Estimate $\hat{p} \approx p(\alpha, v)$.
- 3 Sort nodes in V based on \hat{p} and collect the top k nodes in H_v .
- 4 Finding the lowest-conductance set $H_v^* \in H_v$ based on (10).
- 5 Identifying the node set C_v^* of the densest subnetwork in H_v^* based on (12).
- 6 Considering C_v^* as one predicted complex, let $R = \{R, C_v^*\}$ and $\mathcal{S} = \mathcal{S} - v$.
- 7 EndWhile
- 8 Remove duplicated complexes and complexes with size smaller than three in R .

considered as a predicted complex. After screening every possible proteins, we remove the duplicated complexes and complexes with size smaller than three. There is only one parameter k for the FLCD algorithm, where k can be considered as the upper bound of the sizes of the desired protein complexes. Also, the MIP problems (10) and (12) are both NP hard. The actual computational complexity of solving these MIP problems depends on the problem size of these local problems determined by k . The smaller k is, the less time it takes the FLCD algorithm to search for subnetworks as potential protein complexes. Throughout the experiments in this paper, we set $k = 20$.

Acknowledgements

XQ was partially supported by Awards CCF-1447235, CCF-1553281, IOS-1547557 from the National Science Foundation (NSF) and AFRI-2015-67013-22816 from the United States Department of Agriculture (USDA).

Funding

The publication costs of this article was funded by Award CCF-1447235 from the National Science Foundation.

Availability of data and materials

The Matlab source code and relevant network data can be found at <https://github.com/xqian37/FLCD>.

Authors' contributions

Conceived the algorithm: YW, XQ. Implemented the algorithm and performed the experiments: YW. Analyzed the results: YW, XQ. Wrote the paper: YW, XQ. Both authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

About this supplement

This article has been published as part of *BMC Systems Biology* Volume 11 Supplement 3, 2017: Selected original research articles from the Third International Workshop on Computational Network Biology: Modeling, Analysis, and Control (CNB-MAC 2016): systems biology. The full contents of the supplement are available online at <http://bmcscystbiol.biomedcentral.com/articles/supplements/volume-11-supplement-3>.

Published: 14 March 2017

References

1. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*. 2003;4:2.
2. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*. 2004;32:449–51.
3. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: A general repository for interaction datasets. *Nucleic Acids Res*. 2006;34:535–9.
4. Kerrien S, Aranda B, Breuza L, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res*. 2012;40(D1):841–6.
5. Licata L, Briganti L, Peluso D, et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res*. 2012;40:857–61.
6. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods*. 2012;9:471–2.

7. van Dongen S. A cluster algorithm for graphs. 2000. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam.
8. Satuluri V, Parthasarathy S. Scalable Graph Clustering Using Stochastic Flows: Applications to Community Discovery. In: 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09). Paris; 2009. p. 737–46.
9. Shih YK, Parthasarathy S. Identifying functional modules in interaction networks through overlapping Markov clustering. *Bioinformatics*. 2012;28(18):1473–9.
10. Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*. 2005;435:814–8.
11. Macropol K, Can T, Singh AK. RRW: Repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics*. 2009;10:283.
12. Ahn YY, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*. 2010;466:761–4.
13. Gavin AC, Aloy P, Grandi P, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*. 2006;440:631–6.
14. Wang Y, Qian X. Functional module identification in protein interaction networks by interaction patterns. *Bioinformatics*. 2014;30(1):81–93.
15. Newman MEJ. Modularity and community structure in networks. *Proc Nat Acad Sci USA*. 2006;103:8577–82.
16. Fortunato S, Barthélemy M. Resolution limit in community detection. *Proc Nat Acad Sci USA*. 2007;104(1):36–41.
17. Andersen R, Chung F, Lang K. Local Graph Partitioning Using PageRank Vectors. In: 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06). Berkeley; 2006. p. 475–86.
18. Corneil DG, Perl Y. Clustering and domination in perfect graphs. *Discrete Appl Math*. 1984;9(1):27–39.
19. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25–9.
20. Hong EL, et al. Gene ontology annotations at SGD: New data sources and annotation methods. *Nucleic Acids Res*. 2008;36:577–81.
21. Mewes HW, et al. MIPS: Analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*. 2004;32(Database issue):D41–44.
22. Wang Y, Qian X. Joint clustering of protein interaction networks through Markov random walk. *BMC Syst Biol*. 2014;8(suppl 1):9.
23. Shih Y, Parthasarathy S. Scalable global alignment for multiple biological networks. *BMC Bioinformatics*. 2012;13(Suppl 3):11.
24. Fan N, Pardalos P. Multi-way clustering and biclustering by the ratio cut and normalized cut in graphs. *J Combinatorial Optimization*. 2010;23(2): 224–51.
25. Gurobi Optimization Inc. Gurobi Optimizer Reference Manual. 2016. <http://www.gurobi.com>.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

