

Extracting active modules from multilayer PPI network: a continuous optimization approach

Dong Li, Zexuan Zhu, Zhisong Pan, Guyu Hu, Shan He*

June 2017

Abstract: Active modules identification has received much attention due to its ability to reveal regulatory and signaling mechanisms of a given cellular response. Most existing algorithms identify active modules by extracting connected nodes with high activity scores from a graph. These algorithms do not consider other topological properties such as community structure, which may correspond to functional units. In this paper, we propose an active module identification algorithm based on a novel objective function, which considers both and network topology and nodes activity. This objective is formulated as a constrained quadratic programming problem, which is convex and can be solved by iterative methods. Furthermore, the framework is extended to the multilayer dynamic PPI networks. Empirical results on the single layer and multilayer PPI networks show the effectiveness of proposed algorithms.

Availability: The package and code for reproducing all results and figures are available at <https://github.com/fairmiracle/ModuleExtraction>.

1 Introduction

Identifying functional modules from protein-protein interaction (PPI) networks is one of the central challenge in network biology [1]. The modules (or subnetworks) represent certain interacted components of the whole network to exert some functions, which may shed light to the complex mechanisms behind living organisms. Modular structure in various PPI networks has been reported in [2, 3, 4, 5], and community detection techniques [6, 7] have been applied in a wide range of biological networks, including PPI networks.

Though the topology of a biological network does not always precisely reflects the function or even disease-determined regions [8], which are the real concerns in biology. The topological modules and functional modules may have some overlapped components, but still vary from definition. To better address these concerns, the **active module** identification, which aims to find connected regions of the network showing striking changes in molecular activity or phenotypic signatures that are associated with a given cellular response in biological networks [9], has become an important issue in network biology.

Ideker et al. [10] proposed a general method to discover active connected regions by given network with scores on each node, which reflects the changes over particular conditions. The seminal work [10] not only defined an optimization problem on a weighted molecular interaction networks, which incorporates basic network structure with high-throughput omics data, but also proposed simulated annealing algorithm to solve it. Some following researches improved this basic active module identification from many aspects, resulting into a series of heuristic based methods [11, 12, 13, 14] and integer programming based methods [15, 16]. A comprehensive review [9] listed several advances in this area.

However, in the existing works of active modules identification, node [15] or edge weights [13] (or both [17]) weight too much in original active module identification objective, which leads to modules without topological properties other than connectivity. In other words, current active modules identification algorithms tend to categorize several unrelated nodes in terms of their interactions pattern.

But some earlier results derived from pure topological structure, such as communities [6] or motifs [18], also make sense of biological meaning [19]. The communities are important since they may correspond to functional units [20]. As a recent example of 2016 Disease Module Identification DREAM Challenge [21], the participants were asked only to use network structural information to find disease modules, which were evaluated by pre-defined genome-wide association study (GWAS) sets. Quite a few teams only adopted basic communities detection methods which maximize the modularity [22, 23] also achieved good performance. As a community effort, basic modules detection algorithms which only consider the topology of biological networks are proven to be effective as well.

In this work, we try to combine these two parallel lines of research, active modules identification and communities detection, to extract modules from PPI networks. Specifically, we consider the objective of representing an active module is composed of two parts: the topological part and the active part. By combining these two parts, we can reveal functional change and regulatory or signaling mechanisms. The topological property of a module is derived from graph partitioning, and the active property is highlighted by the higher expected average node score. As a result, we formulate the active modules identification on connected PPI network as a constrained quadratic programming problem, which can be solved efficiently by iterative methods. We also show that this method can be extended to multilayer networks, by conducting the same algorithm on an aggregation of multiple networks. The package and code for reproducing all results and figures are available at <https://github.com/fairmiracle/ModuleExtraction>.

2 Method

2.1 Optimization on single-layer PPI network

We aim to extract a module from a weighted network with both topological and active features. The problem is formally defined as:

Problem 1. *Given a weighted graph $G = (V, E, W)$ where w_{ij} is the weight between vertex i and j , vertices weights z_i for each $i \in V$, find a connected subnetwork $S = (V_S, E_S)$ of G with significant separation from the rest in both edges and vertices weight.*

The separation from topological perspective has been intensively studied in graph theory. We thus start with the logic of spectral clustering (partitioning) on the graph (network) G . The *cut* for two disjoint subsets A and its complement $\bar{A}$ on G , defined as

$$cut(A, \bar{A}) = \sum_{i \in A, j \in \bar{A}} w_{ij}. \quad (1)$$

The basic criteria to measure the quality of a partition on graph G is to minimize $cut(A, \bar{A})$ over all possibilities of A . But there exists a trivial case when $A = V$ or $A = \emptyset$. Even when there is only one node or very few nodes in A , the $cut(A, \bar{A})$ is small, which is not preferred. Several modifications based on *cut* were proposed, such as the *ratio cut* [24]:

$$R = \frac{cut(A, \bar{A})}{|A||\bar{A}|}. \quad (2)$$

Zhao et al. [25] proposed the community extraction framework based on *ratio cut*, incorporating with the assumption that the interactions in the extracted module should be denser. The same intuition had been formulated as the well-known *modularity* [22], which also measures the quality of a partition instead of certain extraction module. Adopting the actual-minus-expected paradigm on the average node score, we define the following objective to be minimized:

$$f = \frac{\sum_{i \in A, j \in \bar{A}} w_{ij}}{|A||\bar{A}|} - \lambda \sqrt{|A||\bar{A}|} \left(\frac{\sum_{i \in A} z_i}{|A|} - \frac{\sum_{i \in \bar{A}} z_i}{|\bar{A}|} \right). \quad (3)$$

The first part is the same as *ratio cut*, which is supposed to be minimized from the topological perspective. And $(\sum_{i \in A} z_i/|A| - \sum_{i \in \bar{A}} z_i/|\bar{A}|)$ in the second term is the gap between the expected node score in the module and in the rest, which is supposed to be maximized (equivalent to minimize the negative) from the active consideration. We also add the penalty $\sqrt{|A||\bar{A}|}$ to this part, when $|A| = |\bar{A}|$ the penalty has a maximal value, which is consistent with the denominator of *ratio cut*. Parameter λ is the used to balance the consideration between edge part and node part. Without prior knowledge or preference on the two parts, we suggest the default value $\lambda = 1$.

Being similar to Luxburg [26] (section 5, Graph cut point of view), we define the vector $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ with entries:

$$x_i = \begin{cases} \sqrt{|\bar{A}|/|A|} & \text{if } i \in A \\ -\sqrt{|A|/|\bar{A}|} & \text{if } i \in \bar{A}. \end{cases} \quad (4)$$

Note the following two facts which are the same as in [26]:

$$\mathbf{x}^\top \mathbf{1} = \sum_i x_i = \sum_{i \in A} \sqrt{|\bar{A}|/|A|} - \sum_{i \in \bar{A}} \sqrt{|A|/|\bar{A}|} = 0. \quad (5)$$

and

$$\mathbf{x}^\top \mathbf{x} = \sum_i x_i^2 = \sum_{i \in A} (|\bar{A}|/|A|) + \sum_{i \in \bar{A}} (|A|/|\bar{A}|) = |V|. \quad (6)$$

We also have

$$\begin{aligned} \mathbf{x}^\top L \mathbf{x} &= \sum_{i,j} w_{ij} (x_i - x_j)^2 \\ &= \sum_{i \in A, j \in \bar{A}} w_{ij} \left(\sqrt{|\bar{A}|/|A|} + \sqrt{|A|/|\bar{A}|} \right)^2 + \sum_{i \in \bar{A}, j \in A} w_{ij} \left(-\sqrt{|\bar{A}|/|A|} - \sqrt{|A|/|\bar{A}|} \right)^2 \\ &= 2 \text{cut}(A, \bar{A}) \left(\frac{|\bar{A}|}{|A|} + \frac{|A|}{|\bar{A}|} + 2 \right) = 2 \text{cut}(A, \bar{A}) \left(\frac{|\bar{A}|+|A|}{|A|} + \frac{|A|+|\bar{A}|}{|\bar{A}|} \right) \\ &= 2|V|R \end{aligned} \quad (7)$$

and

$$\begin{aligned} \mathbf{x}^\top \mathbf{z} &= \sum_i x_i z_i = \sum_{i \in A} x_i z_i + \sum_{i \in \bar{A}} x_i z_i \\ &= \sum_{i \in A} z_i \sqrt{|\bar{A}|/|A|} - \sum_{i \in \bar{A}} z_i \sqrt{|A|/|\bar{A}|} \\ &= \sqrt{|A||\bar{A}|} \left(\frac{\sum_{i \in A} z_i}{|A|} - \frac{\sum_{i \in \bar{A}} z_i}{|\bar{A}|} \right) \end{aligned} \quad (8)$$

Combining equation (7), (8), and conditions (5) and (6), we can rewrite the objective (3) into the matrix form:

$$\text{minimize } f = \frac{\mathbf{x}^\top L \mathbf{x}}{2|V|} - \lambda \mathbf{z}^\top \mathbf{x} \quad \text{subject to} \quad \mathbf{x}^\top \mathbf{1} = 0, \mathbf{x}^\top \mathbf{x} = n. \quad (9)$$

Since the objective function is smooth and differentiable, we use a projected gradient [27] approach to find the local optimum, which works as follow. The gradient is used to determine the primary direction, followed by orthogonalization to satisfy the condition (5) and normalization step to satisfy the condition (6). Following [28] we claim that any limit point of sequence $\{\mathbf{x}^{(k)}\}$ generated by this procedure is a stationary point of (3). When L is positive semi-definite, the problem is convex thus a local optimum is also the the global optimum.

After getting the target set A , it may comes to case that the candidates nodes are not connected, due to the relaxation on $\mathbf{x}$, which was supposed to be 0-1 vector. We conduct a connected components finding (CCF) operation on A as in [29] to extract large component. The existence of giant component in

Algorithm 1: Module extraction via continuous optimization

Input: The adjacency matrix W , node score vector $\mathbf{z}$, parameter λ , maximal candidate set size N_{max} , step-size $\alpha^{(1)}$, maximal iteration T , tolerate τ

Output: The extracted module node set A .

```

1 Laplacian: Compute diagonal degree matrix  $D$  and Laplacian
    $L = D - W$ ;
2 for  $k=1,2,\dots,T$  do
3    $\alpha^{(k)}$  is picked by a non-monotone line search;
4    $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha^{(k)} \nabla f(\mathbf{x}^{(k)})$ ;
5   Orthogonalization:  $\mathbf{x}^{(k+1)} = \mathbf{1}^\top \mathbf{1} \cdot \mathbf{x}^{(k+1)} - \mathbf{x}^{(k+1)\top} \mathbf{1} \cdot \mathbf{1}$ ;
6   Normalization:  $\mathbf{x}^{(k+1)} = \sqrt{n} \cdot \mathbf{x}^{(k+1)} / \sqrt{\mathbf{x}^{(k+1)\top} \mathbf{x}^{(k+1)}}$ ;
7   if  $\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2 < \tau$  then
8     break;
9   end
10 end
11 Determine the primary partition:  $i \in A$  if  $x_i > 0$  otherwise  $i \in \bar{A}$ ;
12 if  $|A| > N_{max}$  then
13   Further division: Recursively the algorithm;
14 end
15 else
16   CCF: Conduct connected component finding on  $A$  if it is not
       connected;
17   Output the module vertices set.
18 end

```

a random set of vertices from a graph is supported by the Erdős-Rényi model [30]. In summary, the procedure is formally described as Algorithm 1.

For a large network, we may need to execute Algorithm 1 multiple times, to get proper number of modules. There are two strategies: recursively extraction and components finding. The former works for dense interacted networks, which mines module one by one and remove the identified module from background network. And the latter aims to leverage the relatively small modules from the candidate set, which avoids wasting isolated nodes when the interactions are not enough.

2.2 Aggregation on multilayer PPI network

The community structure has been explored on multilayer connected graphs. Various algorithms have been proposed to detect the communities on multilayer graphs, see the review [31]. One of the nature idea is to extend the concept of modularity to multilayer case, as [32] did. Taking the layer into consideration, the matrix calculation in original modularity was transformed into tensor computation. The underlying assumption of this extension is that each layer shares similar structure, and the word "multiscale" in the title of [32] also indicates that each network shares similar structure.

However, in some real applications, the multiple layers can be diverse. The sub-challenge 2 of Disease Module Identification DREAM Challenge [21] required find modules across multiple networks, including two protein-protein interaction networks, one signaling network, one co-expression network, one cancer network and one homology network. In such a multilayer network system, the size and topology are diverse with all nodes aligned. Due to the fact that there be very less shared edges among different networks, even no single edge shared by all networks, we can highlight the majority of presence of edges by adopting the idea of **aggregated graph** or **consensus graph**. An aggregated graph is simply defined by the sum of all edge weight matrices, followed by proper cut-off to remove low valued entries. The intuition behind this aggregating is to enhance the concurrent edges across more layers and remove layer-specific edges. As a result, the aggregated graph is supposed to encode the conservation information of multiple layers. And the module (defined by the solution of (3)) extracted from this graph represents conserved subnetwork with maximal node activities across multiple layers. When each layer is a snapshot of a dynamic system, the extracted modules can be viewed as time-invariant subnetworks, which may be related to persistent biological processes.

The complete procedure to mine modules from multiple networks $G_1, G_2, \dots, G_n$ is describe as Algorithm 2.

Algorithm 2: Module extraction via continuous optimization on multi-layer network

Input: The adjacency matrices $W_1, W_2, \dots, W_n$, node score vectors $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$ of n layers, edge weight cut-off threshold τ_1 and node weight cut-off threshold τ_2

Output: The extracted module node set A .

- 1 **Normalization:** Make sure $W_1, W_2, \dots, W_n$ and $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$ are normalized respectively;
 - 2 **Aggregation:** $W = \sum_{i=1}^n W_i$, $\mathbf{z} = \sum_{i=1}^n \mathbf{z}_i$;
 - 3 **Cut-off:** $W(W < \tau_1) = 0$, $\mathbf{z}(\mathbf{z} < \tau_2) = 0$;
 - 4 **Module extraction:** Call Algorithm 1;
 - 5 Output the module vertices set;
-

3 Results and Discussion

3.1 Toy example

We first compare the proposed method with *ratio cut* minimization, which actually solves the graph partitioning without using the node score:

$$\text{minimize } f = \frac{\mathbf{s}^\top L \mathbf{s}}{2|V|} \quad \text{subject to} \quad \mathbf{s}^\top \mathbf{1} = 0, \mathbf{s}^\top \mathbf{s} = n. \quad (10)$$

According to Rayleigh-Ritz theorem [26, 33], the solution is directly given by the eigenvector corresponding to the second smallest eigenvalue of L , denoted as $\mathbf{u}$. The nodes of extracted module are selected according to the sign of entries in $\mathbf{u}$, i.e. $i \in A$ if $u_i > 0$. We denote this method as spectral algorithm in the following discussion.

We compare the spectral algorithm with Algorithm 1 on the widely used Zachary karate club network, as Figure 1 (a) shows. Red nodes mean the members in the target module. The simulated node score of all nodes are randomly generated via $\mathbf{z} \sim [0, 1]$, but with higher expectation in target module members as $\mathbf{z} \sim [0.5, 1]$. We can see that spectral algorithm (see Figure 1c) not only failed to find the important node '34' in the target module, but also wrongly include two nodes '1' and '17' from the opposite module. Furthermore, node '1' is considered to play a core role in the opposite module. The proposed method (see Figure 1c) successfully found all the nodes but introduced additional node '3', which also interacts with other nodes in the target module. In Figure 1c the node size indicates the assigned score, and the result validates the purposed of (9), which combines modular structure and high scored nodes. In other words, not all high scored nodes are included, such as the connected and high-scored nodes '20', '1', '13' and '22'.

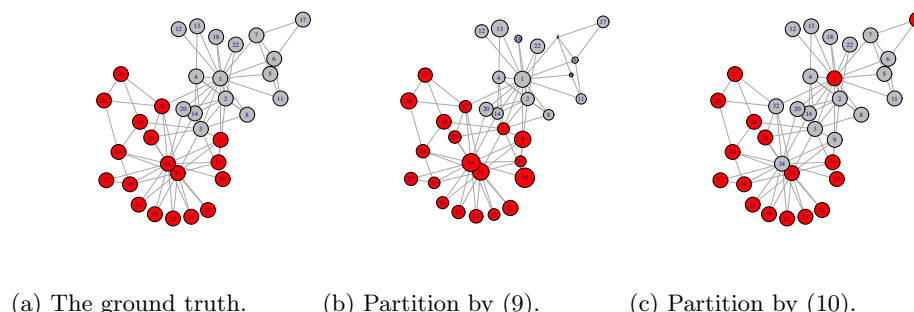


Figure 1: Comparison of the ground truth with partition by proposed method (9) and spectral algorithm (10), on the widely used Zachary karate club network. Red nodes mean the members in the target module. And node size indicates the simulated node score.

3.2 Single layer PPI network

In order to check the biological relevance of identified modules by the proposed algorithm, we apply it on the real world protein-protein interactions (PPI) network. The human PPI network and p-values derived from differential expression and survival analysis [34] come from package **BioNet** [35], which implements the integer programming based active module identification algorithm [15] as well as a heuristics algorithm. The heuristics algorithm can be viewed as an alternative of the Cytoscape plugin **jActiveModules** [10]. Being different from the objective (3), both the exact approach [15] and the heuristics method [10] aim to collect as more high-scored and connected nodes but the exact approach relies on external library CPLEX.

We compare the heuristics method from **BioNet** and Algorithm 1 (using default parameters: $\lambda = 1$, $\alpha = 0.01$, $\tau = 1E - 9$, $T = 100$, $N_{max} = 100$) on the same PPI network with 2559 nodes and 7788 edges, and the nodes are scored by the Beta-Uniform-Mixture (BUM) model [15] based on expression p-values. **BioNet** identifies a module with 37 nodes and 44 edges, and the Algorithm 1 identifies a module with 54 nodes and 54 edges. And there are 20 nodes in common, most of which are high-scored nodes. In order to show the difference directly, we plot both modules in one figure, with different colors, as Figure 2 shows. We can see that in the both modules, there are some low-scored nodes such as SMAD2, LYN, CDC2, NME1, serving as bridge nodes which connect server component. But the module identified by Algorithm 1 includes more low-scored nodes and their interactions.

The identified modules are enriched by biological progresses and pathways related to the experimental settings. Since these two modules have different size, we use the false discovery rate (FDR) of enriched KEGG pathways to show the significance of enrichment analysis. Table shows the top 5 pathways enriched

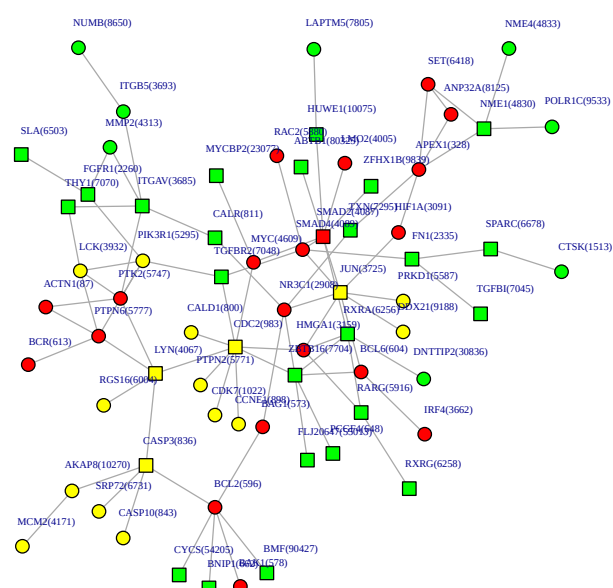


Figure 2: The identified modules by BioNet and Algorithm 1. The red nodes are shared by both methods, yellow for BioNet and green for Algorithm 1. The shape of nodes indicates score, squares indicate negative scores and circles for positive.

by both modules, and we can see find more significant results. Furthermore, all pathways are related to cancer, which is consistent with the data source [34].

3.3 Multilayer dynamic PPI networks

Multilayer dynamic PPI networks (DPPI or DPIN) are normally constructed by integrating static PPI network and a time-course gene expression data [36, 37, 38]. Each layer in a DPPI network represent a specific PPI of a time point, and the whole DPPI is supposed to model the dynamic properties of protein interactions. While the active modules across all layers can be viewed as complementary to those identified from static PPI [39], due to the time course data integration. Here we adopt the DPPI construction method by [38, 36], where the topological structure of each layer is determined by both the static yeast PPI network and the gene expression data at that time point. In addition, the node activity of each layer simply measured by the expression values. The processed gene expression data as well as DPPI construction can be found in [36].

First of all, we try to demonstrate the advantage of multilayer aggregation by comparing the result from the consensus graph and each single layer. We evaluate the result by f-measure [40], which assesses the similarity between identified modules (protein complexes) $C_i \in P$ and reference complexes $C_j \in B$. We claim C_i and C_j match if $v_{ij} \geq 0.25$:

$$v_{ij} = \frac{|C_i \cup C_j|^2}{|C_i \cap C_j|}. \quad (11)$$

F-measure is defined based on precision and recall:

$$\begin{aligned} \text{precision} &= \frac{|\{C_i \mid C_i \in P \wedge \exists C_j \in B, C_j \text{ matches } C_i\}|}{|P|} \\ \text{recall} &= \frac{|\{C_j \mid C_j \in B \wedge \exists C_i \in P, C_i \text{ matches } C_j\}|}{|B|} \\ \text{F-measure} &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \end{aligned} \quad (12)$$

Two widely used reference complexes CYC2008 [41] and MIPS [42] are chosen as gold standard sets B . Furthermore, we use SPICi [43] as baseline algo-

Table 1: The top 5 enriched KEGG pathways of the modules identified by Algorithm 1, and heuristics algorithm which targets at a group of high-scored and connected nodes, on single human PPI network.

Algorithm 1			Heuristics algorithm		
ID	Description	FDR	ID	Description	FDR
5200	Pathways in cancer	5.05E-16	5200	Pathways in cancer	1.23E-9
5222	Small cell lung cancer	5.36E-9	5210	Colorectal cancer	1.23E-9
5210	Colorectal cancer	1.03E-8	5161	Hepatitis B	4.77E-7
5205	Proteoglycans in cancer	1.18E-8	4110	Cell cycle	5.43E-6
5202	Transcriptional misregulation in cancer	1.72E-8	4520	Adherens junction	9.32E-6

rithm which was designed for detecting non-overlapping complexes from static PPI. Also note algorithms for detecting overlapping complexes such as ClusterONE [44] and TS-OCD [36] would achieve higher F-scores since part of proteins might appear frequently in different modules, which increases the chance to be captured by reference sets according to (12).

Figure 3 shows the overall result of modules from each layer, consensus graph and the baseline algorithm SPICi [43]. The purpose of this figure is twofold: identifying modules on the consensus graph is superior to each single one with the same algorithm and Algorithm 2 could achieve comparable result with other methods in terms of accuracy. Specifically, Algorithm 2 tends to achieve higher precision but lower recall compared with SPICi [43]. Possible reason for failing to recall as more complexes lies at the CCF operation in Algorithm 1, which may miss some proteins in the reference set.

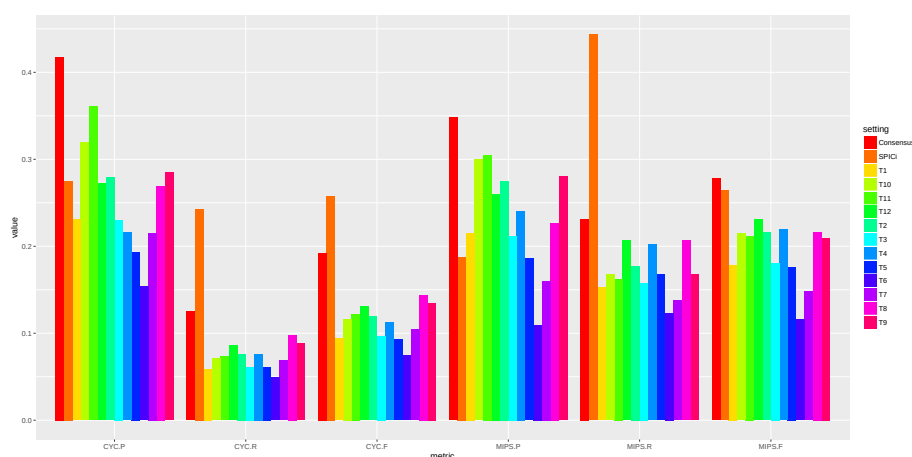


Figure 3: The F-measure of identified complexes from each time point layer, consensus graph and the baseline algorithm. 'CYC-P' means the precision evaluated by CYC2008 [41] and 'MIPS-R' means the recall evaluated by MIPS [42]. 'F' is the F-measure defined as in (12).

In order to validate the biological meaning of the identified complexes, we conduct enrichment analysis with STRING [45], which provides interfaces to access the annotation resources as well as enrichment result. Furthermore, we compare the results of Algorithm 1 and SPICi [43] in a statistical way, and count the number of complexes which are significantly enriched by at least one GO term (biological processes and KEGG pathways) at a given FDR (≤ 0.05) cutoff. From table 2 we can see that Algorithm 2 found relatively larger complexes on average, and the resulted complexes are enriched by more KEGG pathways. The ratio of meaningful complexes is also higher than the result from static PPI network. See the complete enrichment analysis of these modules from two algorithms in supplementary table SS1 and SS2.

Table 2: Overview of the resulted complexes of Algorithm 2 and SPICi [43]. *#Items* means the total GO terms of all complexes and *Ratio* is defined as the proportion of complexes which are significantly enriched by at least one GO term at a given FDR (≤ 0.05).

Method	Modules	Avg. size	Enrichment		
				BP	KEGG
SPICi	298	4.21	#Items	8482	599
			Ratio	0.81	0.72
Algorithm 2	277	5.06	#Items	7114	1088
			Ratio	0.81	0.88

4 Conclusion

This paper presents a continuous optimization approach to extract active modules from PPI networks. The method tries to optimize the objective combining both topological and active properties of a module, which reflects the functional organization mechanism in a complex system. We also generalize the proposed method to the multilayer dynamic PPI networks, using an aggregation of multiple single layers. Empirical studies shows the advantages of the optimization approach as well as the aggregation on multilayer networks.

References

- [1] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell’s functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.
- [2] Victor Spirin and Leonid A Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, 100(21):12123–12128, 2003.
- [3] Jingchun Chen and Bo Yuan. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006.
- [4] Feng Luo, Yunfeng Yang, Chin-Fu Chen, Roger Chang, Jizhong Zhou, and Richard H Scheuermann. Modular organization of protein interaction networks. *Bioinformatics*, 23(2):207–214, 2006.
- [5] Clara Pizzuti and Simona E Rombo. Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, 30(10):1343–1352, 2014.
- [6] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

- [7] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.
- [8] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.
- [9] Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732, 2013.
- [10] Trey Ideker, Owen Ozier, Benno Schwikowski, and Andrew F Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl 1):S233–S240, 2002.
- [11] Dilip Rajagopalan and Pankaj Agarwal. Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, 21(6):788–793, 2005.
- [12] Șerban Nacu, Rebecca Critchley-Thorne, Peter Lee, and Susan Holmes. Gene expression network analysis and applications to immunology. *Bioinformatics*, 23(7):850–858, 2007.
- [13] Igor Ulitsky and Ron Shamir. Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics*, 25(9):1158–1164, 2009.
- [14] Weiqi Chen, Jing Liu, and Shan He. Prior knowledge guided active modules identification: an integrated multi-objective approach. *BMC systems biology*, 11(2):8, 2017.
- [15] Marcus T Dittrich, Gunnar W Klau, Andreas Rosenwald, Thomas Dandekar, and Tobias Müller. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–i231, 2008.
- [16] Xing-Ming Zhao, Rui-Sheng Wang, Luonan Chen, and Kazuyuki Aihara. Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic acids research*, 36(9):e48–e48, 2008.
- [17] Haisu Ma, Eric E Schadt, Lee M Kaplan, and Hongyu Zhao. Cosine: Condition-specific sub-network identification using a global optimization method. *Bioinformatics*, 27(9):1290–1298, 2011.
- [18] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.
- [19] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *arXiv preprint physics/0506133*, 2005.

- [20] Mark EJ Newman. Spectral methods for community detection and graph partitioning. *Physical Review E*, 88(4):042822, 2013.
- [21] Disease module identification dream challenge. <https://www.synapse.org/#!/Synapse:syn6156761>. Accessed: 2017-07-15.
- [22] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.
- [23] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
- [24] Yen-Chuen Wei and Chung-Kuan Cheng. Towards efficient hierarchical designs by ratio cut partitioning. In *Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers., 1989 IEEE International Conference on*, pages 298–301. IEEE, 1989.
- [25] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326, 2011.
- [26] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.
- [27] Chuan-bi Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.
- [28] Paul H Calamai and Jorge J Moré. Projected gradient methods for linearly constrained problems. *Mathematical programming*, 39(1):93–116, 1987.
- [29] Dong Li, Zhisong Pan, Guyu Hu, Zexuan Zhu, and Shan He. Active module identification in intracellular networks using a memetic algorithm with a new binary decoding scheme. *BMC Genomics*, 18(2):209, 2017.
- [30] P ERDdS and A R&WI. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.
- [31] Mikko Kivelä, Alex Arenas, Marc Barthélemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.
- [32] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328(5980):876–878, 2010.
- [33] Helmut Lutkepohl. Handbook of matrices. *Computational Statistics and Data Analysis*, 2(25):243, 1997.

- [34] Andreas Rosenwald, George Wright, Wing C Chan, Joseph M Connors, Elias Campo, Richard I Fisher, Randy D Gascoyne, H Konrad Muller-Hermelink, Erlend B Smeland, Jena M Giltneane, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947, 2002.
- [35] Daniela Beisser, Gunnar W Klau, Thomas Dandekar, Tobias Müller, and Marcus T Dittrich. Bionet: an r-package for the functional analysis of biological networks. *Bioinformatics*, 26(8):1129–1130, 2010.
- [36] Le Ou-Yang, Dao-Qing Dai, Xiao-Li Li, Min Wu, Xiao-Fei Zhang, and Peng Yang. Detecting temporal protein complexes from dynamic protein-protein interaction networks. *BMC bioinformatics*, 15(1):335, 2014.
- [37] Xiwei Tang, Jianxin Wang, Binbin Liu, Min Li, Gang Chen, and Yi Pan. A comparison of the functional modules identified from time course and static ppi network data. *BMC bioinformatics*, 12(1):339, 2011.
- [38] Jianxin Wang, Xiaoqing Peng, Min Li, and Yi Pan. Construction and application of dynamic protein interaction network based on time course gene expression data. *Proteomics*, 13(2):301–312, 2013.
- [39] Bolin Chen, Weiwei Fan, Juan Liu, and Fang-Xiang Wu. Identifying protein complexes and functional modules from static ppi networks to dynamic ppi networks. *Briefings in bioinformatics*, 15(2):177–194, 2014.
- [40] Xiaoli Li, Min Wu, Chee-Keong Kwoh, and See-Kiong Ng. Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC genomics*, 11(1):S3, 2010.
- [41] Shuye Pu, Jessica Wong, Brian Turner, Emerson Cho, and Shoshana J Wodak. Up-to-date catalogues of yeast protein complexes. *Nucleic acids research*, 37(3):825–831, 2008.
- [42] Hans-Werner Mewes, C Amid, Roland Arnold, Dmitriy Frishman, Ulrich Güldener, Gertrud Mannhaupt, Martin Münsterkötter, Philipp Pagel, Normann Strack, Volker Stümpflen, et al. Mips: analysis and annotation of proteins from whole genomes. *Nucleic acids research*, 32(suppl.1):D41–D44, 2004.
- [43] Peng Jiang and Mona Singh. Spici: a fast clustering algorithm for large biological networks. *Bioinformatics*, 26(8):1105–1111, 2010.
- [44] Tamás Nepusz, Haiyuan Yu, and Alberto Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471–472, 2012.
- [45] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein

interaction networks, integrated over the tree of life. *Nucleic acids research*, page gku1003, 2014.