DOI: 10.7641/CTA.2017.60581

针对蛋白质复合体检测的自学习图聚类

朱 佳[1], 武兴成[1], 林雪琴[1], 肖丹阳[1], 肖 菁[1], 黄 晋[1†], 贺超波[2]

(1. 华南师范大学 计算机学院, 广东 广州 510631;
2. 仲恺农业工程学院 信息科学与技术学院, 广东 广州 510225)

摘要: 蛋白质复合体是由两条或多条相关联的多肽链组成, 在生物过程中起着重要作用. 假如用图表示蛋白质–蛋白质相互作用(protein-protein interactions, PPI)网络数据, 那么从中找出紧密耦合的蛋白质复合体是非常困难的, 特别是在近年来PPI网络的容量大大增加的情况下. 在本文中, 通过对称非负矩阵分解, 针对蛋白质复合体检测问题提出了一种图聚类方法, 该方法可以有效地从复杂网络中检测密集的连通子图. 并且将此方法和当前最先进的一些方法在3个PPI数据集中用同一个基准进行比较. 实验结果表明, 本文的方法在3个拥有不同大小和密度的数据集中均显著优于其它方法.

关键词: 图聚类; 蛋白质复合体; 非负矩阵分解

中图分类号: TP273 **文献标识码**: A

A self-learning graph clustering approach for protein complexes detection

ZHU Jia[1], WU Xing-cheng[1], LIN Xue-qin[1],
XIAO Dan-yang[1], XIAO Jing[1], HUANG Jin[1†], HE Chao-bo[2]

(1. School of Computer Science, South China Normal University, Guangzhou Guangdong 510631, China;
2. School of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou Guangdong 510225, China)

Abstract: Protein complex is a group of two or more associated polypeptide chains which plays essential roles in biological process. Given a graph representing protein-protein interactions (PPI) data, it is important but non-trivial to find protein complexes, the subsets of proteins that are closely coupled, from it, particularly in the condition that the PPI network has increased greatly in capacity in the recent years. In this paper, we propose a graph based clustering approach by adopting symmetric non-negative matrix factorization, which can effectively detect densely connected subgraphs from complex networks. We compare the performance of our approach with state-of-the-art approaches in three PPI networks with a well known benchmark complexes. The experimental results show that our approach significantly outperforms other methods in three PPI networks with different data sizes and densities.

Key words: graph clustering; protein complexes; non-negative matrix factorization

1 Introduction

Protein complex is a complex graph structure that is linked by non-covalent protein-protein interactions (PPI)[1–2], which plays an essential role in biological process and discovering drugs in pharmaceutical process. Therefore, correctly identifying protein complexes in PPI network is useful in the field of biomedical. However, with the huge increase of PPI data, only a small amount of protein complexes are identified in vitro because of the bottleneck of experimental approaches. Besides, it requires a large amount of labor resource[3–5].

To overcome the technological limit of experimental approaches for protein complexes detection, computational approaches are used. The PPI network can be represented as a graph where proteins are represented as vertices and their interactions as edges. Each protein complex consists of two or more proteins that are shown as densely connected subgraphs, which indicates graph based clustering methods should be utilized to discover them.

For example, Liu et al[6] used clique finding algorithms to predict protein complexes from PPI network. They devised their own methods to merge overlapping

Received 3 August 2016; accepted 23 March 2017.
†Corresponding author. E-mail: 1936079@qq.com; Tel.: +86 20-26274857.
Recommended by Associate Editor YU Zhu-liang.
Supported by Natural Science Foundation of Guangdong Province, China (2015A030310509), National Science Foundation of China (61370229, 61272067, 61303049) and S&T Planning Key Projects of Guangdong (2014B010117007, 2015B010109003, 2015A030401087, 2016A030303055, 2016B030305004, 2016B010109008).

cliques as protein complexes. Besides, Ref. [7] introduced Markov clustering (MCL) as graph partitioning method by simulating random walks, which used two operators called expansion and inflation to boost strong connections and demotes weak connections. Later, Ref. [8] showed the robustness of MCL with comparison to three other clustering algorithms for protein complexes detection. One of the recent emerging methods is to first identify cores of a protein complex, and then add attachments into these cores to form protein complexes[9–10]. Ref. [11] further evaluated the implementation of this method called COre-AttaCHment based method (COACH) against other methods, and proved that COACH outperforms others in two PPI data sets.

Even though the clustering methods that we mentioned above proved their competency in small size of PPI data, they showed poor performance in large size of tightly interconnected PPI network according to our study[12]. In other words, the existing clustering methods are not suitable to detect protein complexes from the tightly interconnected network because these methods usually result in an eigenvalue decomposition problem[13]. With the fast growing of PPI data, the current network will become even more tightly interconnected. To overcome the limitation of existing algorithms, we propose a graph based clustering approach by adopting symmetric non-negative matrix factorization (SNMF), which can effectively detect densely connected subgraphs from complex networks considering the PPI network is a undirected network and protein complexes are densely connected subgraphs. SNMF is a non-negative matrix factorization (NMF) based algorithm for undirected network, where NMF originally is designed for the purpose of finding matrix factors with sound performance[14].

Assume that we have a PPI network represented as an undirected graph, it is difficult for spectral clustering methods to know which protein complex a protein should belong to because it has equal number of links to both protein complexes. To avoid this kind of issue, SNMF can achieve better results by treating the graph as an adjacency matrix and minimize the general loss using non-negative matrix factorization, which factorizes the graph into a cluster membership matrix and a matrix contains the linking information within each cluster. In this study, we used the Euclidean loss to calculate the general loss, which is a one of most common calculation methods. The technical details will be given in the Section 3. To prove that our approach is robust and feasible, we evaluated our approach on three PPI datasets with a well known benchmark complexes.

The rest of the paper is organized as follows: Section 2 presents the latest works related to this study; Section 3 introduces basic concepts and explain how we adopted SNMF for protein complexes detection in detail; Section 4 reports the experimental results; Section 5 summarizes and concludes this paper with future improvement suggestion for protein complexes detection.

2 Related works

This section will discuss recent works that are related to protein complexes detection using different data mining methods including supervised and unsupervised learning. The following review of some of previous important works is presented in theme base.

Ref. [7] introduced Markov clustering (MCL) as graph partitioning method by simulating random walks. It used two operators called expansion and inflation, which boost strong connections and demotes weak connections. Iterative expansion and inflation separate the graphs into many subgraphs.

Ref. [8] showed the robustness of MCL[7] with comparison to the restricted neighborhood search clustering algorithm (RNSC)[15] and molecular complex detection (MCODE)[16] for protein complexes detecting. Each clustering algorithm was applied to binary PPI data in order to test the ability to extract complexes from the networks and the clusters were compared with the annotated the munich information center for protein sequences (MIPS) complexes. However, Ref. [17] proved that if the interaction networks are accurate and complete, then maximal clique finding algorithm can be ideal for detecting protein complexes from the PPI network.

One of the recent emerging techniques is to use protein core attachments method called COACH proposed by Ref. [18] which first detected protein-complex cores as the 'hearts' of protein complexes and then included attachments into these cores to form biologically meaningful structures. Later, Ref. [9] proposed the other core attachments method called CORE, which can identify protein-complex cores and add attachments into these cores to form protein complexes.

Ref. [11] showed that COACH performed better than seven other clustering algorithms in various datasets.

Therefore, we chose COACH as one of typical algorithms to compare with our approach. Ref. [12] utilized the neural network with the semi-supervised learning mechanism to detect the protein complexes. By retraining the neural network model recursively, they could find the optimized parameters for the model to detect the protein complexes. Their comparison results showed that the algorithm can identify protein complexes that are missed by other methods[19].

Ref. [20] proposed a method called ClusterOne for detecting potentially overlapping protein complexes. The method uses a greedy approach to calculate a score called cohesiveness and detecting groups of proteins.

Ref. [12] introduced new algorithm called B3Clustering which finds clusters by adjusting the density of subgraphs to be flexible according to its size, their experimental result supported the efficiency and robustness of B3Clustering for protein complex prediction in PPI networks compared to existing approaches.

Ref. [21] proposed an algorithm to detect which proteins share closely located bottleneck proteins. The proposed algorithm has two steps, the first step is to calculate the shortest distances between all node pairs, the second step is to search dense protein subnetworks(protein complexes) of which proteins share closely located bottleneck proteins according to the results from previous step. Though their experiments showed better performance than some of existing methods but not as good as our approach. The experimental results can be found in Section 4.

Most recently, Ref. [22] presented an approach of integrating PPI datasets with the PPI data from biomedical literature for protein complexes detection. The approach applied a natural language processing system called PPI extractor, to extract PPI data from biomedical literature. These data were then integrated into the PPI datasets for complex detection. However, though their approach can get additional information from biomedical literature, it comes with large noise data. Their experimental results also showed that the approach did not have huge improvement compared with other methods.

3 Proposed approach

PPI data come in the form of connections between proteins, which is easily described as a graph model. Proteins are represented as vertices and their interactions are represented as edges in the graph. SNMF can achieve better results by treating the graph as an adjacency matrix, and minimize the general loss using nonnegative matrix factorization. It factorizes the graph into a cluster membership matrix and a matrix contains the linking information within each cluster. However, SNMF normally requires a good initialization to obtain high accuracy[23]. For our research, the initialization parameter k of SNMF means the size of original matrix to store the data and the number of clusters being produced. This requirement is unlikely satisfied by protein complexes detection since we do not know how many protein complexes there will be in a PPI network. In this section, we will introduce the technical details how we adopted SNMF using a algorithm that can obtain the best initialized parameter for SNMF during protein complexes detection process.

3.1 Problem formulation

Assume that we have an undirected graph G consists of k clusters with size $l_1, \cdots, l_k$. Each cluster contains one or more vertices. In an ideal situation, without loss of generality, the vertices in different clusters have no connectivity with each other, and the adjacency matrix of G can be represent as:

$$G = \begin{bmatrix} C_1 & 0 & \dots & 0 \\ 0 & C_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & C_k \end{bmatrix},$$

where C_i is a $l_i \times l_i$ matrix, then the G can be factorized as $G = XCX^{\mathrm{T}}$, where

$$X = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}, \; C = \begin{bmatrix} c_1 & 0 & \dots & 0 \\ 0 & c_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & c_k \end{bmatrix}.$$

The matrix X represents the cluster membership where as the matrix C represents the connectivity of vertices within each cluster. Thus, if we use the Euclidean loss to calculate the general loss of the factorization, we have

$$l(G, XCX^{\mathrm{T}}) = ||G - XCX^{\mathrm{T}}||_{\mathrm{F}}^2. \tag{1}$$

We then absorb C into X using $\hat{X} = XC^{1/2}$ because G is symmetric (undirected graph), then the Eq. (1) can be rewritten as $l(G, \hat{X}\hat{X}^{\mathrm{T}}) = ||G - \hat{X}\hat{X}^{\mathrm{T}}||_{\mathrm{F}}^2$. Clearly, the problem is to minimize the $l(G, \hat{X}\hat{X}^{\mathrm{T}})$ because the smaller the loss is, the better clustering results we can obtain.

3.2 Solution

In this section, we will discuss how we solve the preceding problem in SNMF. In addition, we also introduce an algorithm to learn the best value to initialize SNMF for protein complexes detection.

As we discussed earlier, the problem is to minimize the $l(G, \hat{X}\hat{X}^{\mathrm{T}})$. According to [24], we can use gradient decent method to converge the local minimum. The multiplicative updating rule is:

$$\hat{X}_{ik} \leftarrow \frac{1}{2}[\hat{X}_{ik}(1 + \frac{(G\hat{X})_{ik}}{(\hat{X}\hat{X}^{\mathrm{T}}\hat{X})_{ik}})]. \tag{2}$$

From the Eq.(2), we learn that there are two factors to consider. The first factor is the number of times to update $\hat{X}$ until convergence. According to our observations, the effect for clustering performance is little when the general loss $l(G, \hat{X}\hat{X}^{\mathrm{T}})$ is less than 1. Thus, considering the computation cost, we stop the update process once the general loss is less than 1 rather than give a fixed iterative number.

The other factor is the value of k, which is used to initialize SNMF. It is difficult to estimate a value to initialize because PPI network usually contains thousands of proteins and the number of interactions among proteins can be millions. To obtain the best performance,

we proposed an algorithm to learn the value of k rather than random estimation. The learning and clustering steps are presented below:

Step 1 We first set $k = N$ as the initialization of SNMF, where N is the number of vertices in G because it is the maximum number of clusters theoretically. After the first iteration, we have M clusters that contain proteins, $M < N$, as some proteins has been clustered, which left some clusters null.

Step 2 We set $k = M$ and rerun SNMF, and we will obtain P clusters that contain proteins, $P \leqslant M$.

Step 3 If $P = M$, the process is stopped and outputs the current clustering results. Otherwise, we take M as the high bound High, and P as the low bound Low, and set $k = \text{ceil}((\text{High} + \text{Low})/2)$ to rerun SNMF.

Step 4 If the number of clusters that contain proteins is equal to the k we set in Step 3, the process is stopped and outputs the current clustering results. Otherwise, we repeat the Step 3 but give the k value of current iteration to the high bound High to calculate the k value for next iteration.

Step 2 seems duplicate to Step 1 but it is necessary to narrow down the range of k. The Steps 3 and 4 adjust the value of k for SNMF via learning the number of clusters contain proteins. In other words, the process is stopped when there is no empty clusters. The pseudocode is followed:

INPUT: A graph $G = (V, E)$(G is a PPI network).
OUTPUT: S (a set of clusters, each cluster contains one or more proteins).
Initialize $k = N$, where $N = \text{NumOfVertices}(G)$;
$S = \text{SNMF}(G, k)$;
$M = \text{NumOfClustersContainElements}(S)$;
$S = \text{SNMF}(G, M)$;
$P = \text{NumOfClustersContainElements}(S)$;
if $P == M$ **then**
  return S
else
  $\text{High} = M$;
  $\text{Low} = P$;
  $k = \text{ceil}(\dfrac{\text{High} + \text{Low}}{2})$;
  $S = \text{SNMF}(G, k)$;
  while $k! = \text{NumOfClustersContainElements}(S)$ **do**
    $\text{High} = k$;
    $k = \text{ceil}(\dfrac{\text{High} + \text{Low}}{2})$;
    $S = \text{SNMF}(G, k)$;
  end while
  return S
end if

Lastly, because one protein may belong to multiple protein complexes [25], which indicates 'soft clustering' is required in SNMF. Thus, we assign $G_{ij} = 1$ if there is an edge between vertex i and j; and $G_{ij} = 0$ otherwise when we construct the adjacency matrix G. We then pick top H proteins from each column in G for clustering based on the possibility assigned by SNMF. Because most of other existing approaches are 'hard clustering', which means each protein only belongs to one cluster, we set $H = 1$ for this reason in our experiments. More details can be found in Section 4.

4 Evaluation

In this section, we show our experiments on three PPI data sets to demonstrate the performance of our approach by comparing it with state-of-the-art methods. The experiments were performed on a desktop with Pentium(R) CPU dual core 2.60 GHz and 4 GB memory. Our algorithm is slower than others due to the learning step to find the best initialization for SNMF. However, the calculation of the whole process still can be completed in less than one hour on all three data sets, which is quite acceptable. In addition, since PPI data clustering usually is one-off process in the real world, we do not focus on running time improvement and time complexity analysis in this research as clustering quality is much important.

4.1 Data corpus and evaluation metrics

We used the latest three popular PPI data sets for Saccharomyces cerevisiae, namely, Krogan[26], Dip[27] and Biogrid[28]. The Krogan and dip data sets were used by Li et al.[11] to evaluate the performance of several clustering algorithms. As shown in Table 1, Krogan and Dip data sets have similar number of average degree and density, but Biogrid has much higher average degree and density than them. Because PPI data can be represented as a undirected graph $G = (V, E)$, thus, the average degree is calculated as $\dfrac{2 \times |E|}{|V|}$, and the density is calculated as

$$\frac{2 \times |E|}{|V| \times (|V| - 1)}.$$

Table 1 Features of PPI datasets

Data set	Vertices	Edges	Average degree	Density
Krogan	5364	61289	22.85	0.0043
Dip	4972	17836	7.17	0.0014
Biogrid	6242	255510	81.87	0.013

PPI data have a high rate of false positives, which has been estimated to be about 50%[29]. The noise of the data disturbs clustering methods to detect protein complexes from PPI data. Thus, we used CYC2008 complexes as a reference data set, which was published by Pu et al.[30]. CYC2008 provides a comprehensive catalogue of manually curated 408 protein complexes in Saccharoyces cerevisiae, and has 90% more complexes

than the other popular data set MIPS[31].

We used neighbourhood affinity score to see whether a complex detected by an algorithm is matched with protein complexes in the CYC2008, which was used by Li et al.[11]. We then used it to calculate the precision, recall, and F-measure to evaluate the performance of an algorithm. The neighbourhood affinity score $NA(p,b)$ is defined as follows:

$$NA(p,b)=\frac{|V_p\cap V_b|^2}{|V_p|\cdot|V_b|},$$

where $P=(V_p,E_p)$ is a predicted complex and $B=(V_b,E_b)$ is a benchmark complex. We then have the precision calculated as follows: $\text{Precision}=N_{\text{cp}}/|P|$, where $N_{\text{cp}}=|\{p\mid p\in P, NA(p,b)\geqslant\omega,\ \text{for}\ \exists b\in B\}|$.

The recall is calculated as follows: $\text{Recall}=N_{\text{cb}}/|B|$, where $N_{\text{cb}}=|\{b\mid b\in B, NA(p,b)\geqslant\omega, \text{for}\ \exists p\in P\}|$.

The F-measure is the precision, recall, and F-measure harmonic mean of precision and recall as follows:

$$\text{F-measure}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}.$$

The ω is a threshold, which indicates if a protein complex is identified for any protein complex in the benchmark data set. According to our experiments and the recommendation by [11], we set the neighbourhood affinity score threshold as 0.25, which made the difference of performance among various algorithms.

In addition, we also used three indicators to measure the quality of clustered protein complexes, fraction (Frac), maximum matching ratio (MMR)[20] and geometry accuracy (Acc)[8]. Frac is an indicator that measures the fraction of pairs between two protein complexes with an overlap score θ larger than 0.25, where Frac(θ) is calculated as below:

$$\theta(A,B)=\frac{|A\cap B|}{|A||B|},\tag{3}$$

where A and B are two protein complexes.

Acc is the geometric mean of two other measures: the clustering-wise sensitivity (Sn) and the clustering-wise positive predictive value (PPV) as follows:

$$\text{Sn}=\frac{\sum\limits_{i=1}^{n}\max\limits_{j=1}^{m}t_{ij}}{\sum\limits_{i=1}^{n}n_i},\ \text{PPV}=\frac{\sum\limits_{j=1}^{m}\max\limits_{i=1}^{n}t_{ij}}{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}t_{ij}},$$

where n are the number of proteins of reference protein complexes and m are the number of proteins of clustered protein complexes. The element t_{ij} refers to the number of proteins that are found in both complexes. Because Sn can be inflated by putting every protein in the same complex while the PPV can be maximized by putting every protein in its own complex, we then have these two measures to compute the geometric mean of Sn and PPV: $\text{Acc}=\sqrt{\text{Sn}\times\text{PPV}}$.

MMR represents the two sets of clustered protein complexes as a bipartite graph where the two sets of nodes represent the reference and predicted complexes, respectively, and an edge connecting a reference complex with a predicted one is weighted by the overlap score. The overlap score between two protein complexes is computed by Eq.(3). The value of the MMR is given by the total weight of particular subset of edges that have maximum weight, divided by the number of reference protein complexes. This measure expresses how well the clustered protein complexes represent the reference ones.

4.2 Evaluation results

To evaluate our approach, we compared the performance of our approach with five state-of-the-art approaches, MCL[7], COACH[18], B3Clustering[12], the algorithm proposed by Ahn et al.[21], and ClusterOne[20]. MCL, COACH and ClusterOne are representative algorithms and cited by many other researchers. B3-Clustering is our previous work, and the algorithm proposed by Ahn et al. is the latest work for protein complexes detection to the best of our knowledge.

4.2.1 Comparison test

As a fair comparison, we set $H=1$ to compare with other algorithms because most of the existing approaches are 'hard clustering', which means each protein only belongs to one cluster. Table 2 shows the number of protein complexes detected by different algorithms. The results on different data sets are presented in Figs.1–3, respectively.

From the results, we learn that our approach outperforms others on all the three data sets in terms of precision and F-measure. Particularly on the Biogrid data set that has high density, our approach achieves 0.68 precision rate and 0.58 F-measure, both of which are more than 50% higher than the algorithm in the second place. On the dip data set, our approach achieves the highest 0.78 precision rate, which is nearly double than others. Similar outcomes are also found on Krogan data set.

Table 2 Number of protein complexes detected by different algorithms

Data set	Our Approach	B3Clustering	COACH	MCL	Ahn et al.	ClusterOne
Krogan	401	646	570	626	652	342
Dip	388	786	748	840	646	366
Biogrid	550	477	3158	55	510	380

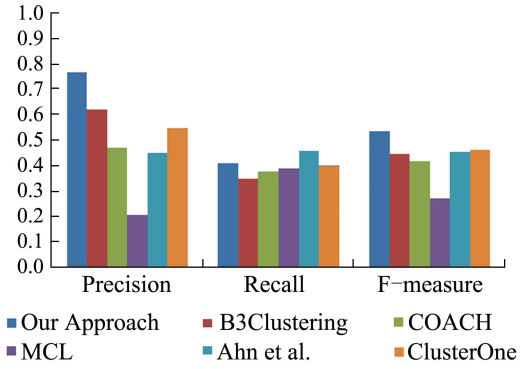

Fig. 1 Comparison results on Krogan data set

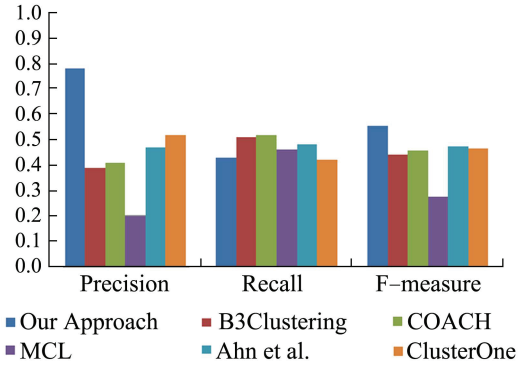

Fig. 2 Comparison results on dip data set

Finally, we note that our approach has lower recall rate on Krogan and dip data sets compared to the algorithm proposed by Ahn et al. because our approach detects fewer number of protein complexes. However, we can improve the recall rate by increasing the value of H if necessary.

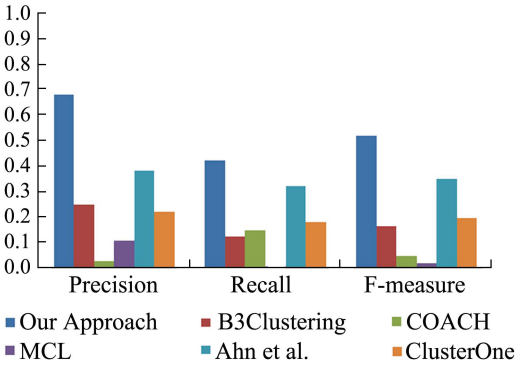

Fig. 3 Comparison results on Biogrid data set

4.2.2 Quality measurement

In this section, we compare the clustering quality of each algorithm. Tables 3–5 show the quality comparisons on Krogan, Dip and Biogrid data set respectively. On Krogan data set, we can see that our approach's overall quality is around 5% higher than ClusterOne, which is the second best algorithm. Similar situation happened on the other two data sets. Our approach outperforms other algorithms, particularly on Biogrid data set obviously.

Table 3 Quality of the clustered protein complexes from Krogan data set

Data set	Our Approach	B3Clustering	COACH	MCL	Ahn et al.	ClusterOne
Frac	0.71	0.6	0.35	0.64	0.65	0.67
Acc	0.68	0.61	0.46	0.64	0.64	0.66
MMR	0.52	0.43	0.19	0.35	0.4	0.42

Table 4 Quality of the clustered protein complexes from dip data set

Data set	Our Approach	B3Clustering	COACH	MCL	Ahn et al.	ClusterOne
Frac	0.81	0.72	0.61	0.44	0.52	0.79
Acc	0.73	0.69	0.58	0.69	0.69	0.71
MMR	0.5	0.45	0.36	0.43	0.43	0.48

Table 5 Quality of the clustered protein complexes from Biogrid data set

Data set	Our Approach	B3Clustering	COACH	MCL	Ahn et al.	ClusterOne
Frac	0.58	0.52	0.14	0.3	0.51	0.55
Acc	0.68	0.35	0.39	0.46	0.61	0.66
MMR	0.3	0.24	0.05	0.15	0.25	0.27

5 Conclusions and future work

Detecting protein complexes in PPI network is an important task in the field of biomedical. Thus, with advances in technology, PPI network is growing much faster than ever, which makes the task non-trivial. In this study, we proposed a graph based clustering approach by adopting SNMF[32] with good initialization from a learning algorithm, which can effectively detects densely connected subgraphs from complex networks. Compared with other protein complexes de-

tection methods, our approach can support ‘soft clustering’, which means one protein can be assigned to multiple clusters. Thus, the approach we proposed can be adopted in some real applications according to actual requirements. Extensive experiments performed on various PPI data sets show that our approach is robust and outperforms other state-of-the-art approaches. In the future, we plan to integrate information from biomedical literature as features to calculate the weight for each edge in the graph, which shall further improve the performance of protein complexes detection.

References:

[1] LACOUNT D J, VIGNALI M, CHETTIER R, et al. A protein interaction network of the malaria parasite plasmodium falciparum [J]. *Nature*, 2005, 438(7064): 103 – 107.

[2] RUAL J F O, VENKATESAN K, HAO T, et al. Towards a proteome-scale map of the human protein-protein interaction network [J]. *Nature*, 2005, 437(437): 1173 – 1178.

[3] GAVIN A C, BÖSCHE M, KRAUSE R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes [J]. *Nature*, 2002, 415(6868): 141 – 147.

[4] HO Y, GRUHLER A, HEILBUT A, et al. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry [J]. *Nature*, 2002, 415(6868): 180 – 183.

[5] STELZL U, WORM U, LALOWSKI M, et al. A human protein-protein in-teraction network: a resource for annotating the proteome [J]. *Cell*, 2005, 122(6): 830 – 832.

[6] LIU G M, CHUA H N, WONG L, et al. Complex Ddiscovery from weighted PPI networks [J]. *Bioinformatics*, 2009, 25(15): 1891 – 1897.

[7] DONGEN S V. *Graph clustering by flow stimulation* [D]. Dutch: University of Utrecht, 2000.

[8] BROHWE S, VAN H J. Evaluation of cluster-ing algorithms for protein-protein interaction net-works [J]. *BMC Bioinformatics*, 2006, 7(1): 488 – 496.

[9] LEUNG H C, YIU S M, XIANG Q, et al. Predicting protein complexes from ppi data: a core-attachment approach [J]. *Journal of Computational Biology*, 2009, 16(2): 133 – 144.

[10] WU D D, HU X. An efficient approach to detect a protein community from a seed [C] *//IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. La Jolla, CA, USA: IEEE, 2005: 135 – 141.

[11] LI X, WU M, KWOH C K, et al. Computational approaches for detecting protein com-plexes from protein interaction networks: a survey [J]. *BMC Bioinformatic*, 2010, 11(1): 1 – 19.

[12] CHIN E J, ZHU J. B3clustering: identifying protein complexes from protein-protein interac-tion network [C] *//The 15th Asia-Pacific Web Conference*. Berlin Heidelberg: Springer, 2013: 108 – 119.

[13] WANG F, LI T, WANG X, et al. Community discovery using nonnegative matrix factorization [J]. *Data Mining and Knowledge Discovery*, 2009, 22(3): 493 – 521.

[14] LEE D D, SEUNG H S. Learning the parts of objects by nonnegative matrix factorization [J]. *Nature*, 1999, 401(6755): 788 – 791.

[15] KING A, PRZULJ N, JURISICA I. Protein com-plexes prediction via cost-based clustering [J]. *Bioinformatics*, 2004, 20(17): 3013 – 3020.

[16] BADER G D, HOGUE C W. An automated method for finding molecular complexes in large protein interaction networks [J]. *BMC Bioinformatics*, 2003, 4(1): 2.

[17] TOMITA E, TANAKA A, TAKAHASHI H. The worst-case time complexity for generating all maximal cliques and computational experiments [J]. *Theoretical Computer Science*, 2006, 363(1): 28 – 42.

[18] MIN W, LI X L, KWOH C K, et al. A core-attachment based method to detect protein complexes in ppi networks [J]. *BMC Bioinformatics*, 2009, 10(1): 169.

[19] LEI S, LEI X, ZHANG A. Protein complex detection with semi-supervised learning in protein interaction networks [J]. *Proteome Science*, 2011, 9(1): S1 – S5.

[20] NEPUSZ T, YU H, PACCANARO A. Detecting overlapping protein complexes in protein-protein interaction networks [J]. *Nature Methods*, 2012, 9(5): 471 – 472.

[21] AHN J, LEE D H, YOON Y M, et al. Improved method for protein complex detection using bottleneck proteins [J]. *BMC Medical Informatics and Decision Making*, 2013, 13(1): S1 – S5.

[22] YANG Z H, YU F Y, LIN H F. Integrating ppi datasets with the P-PI data from biomedical literature for protein complex detec-tion [J]. *BMC Medical Genomics*, 2014, 7(2): S1 – S3.

[23] ALBRIGHT R, COX J, DULING D, et al. Algorithms, initializations and convergence for the nonnegative matrix factorization [J]. *Eprint Arxiv*, 2014: 1 – 18.

[24] WANG D D, LI T, ZHU S H, et al. Multi-document summarization via sentence-level se-mantic analysis and symmetric matrix factorization [J]. *International Acm Sigir Conference on Research & Development in Information Retrieval*, 2008, 5(2): 307 – 314.

[25] HARTWELL L H, HOPFIELD J J, LEIBLER S, et al. From molecular to modular cell biology [J]. *Nature*, 1999, 402(6761): 47 – 52.

[26] KROGAN N, CAGNEY G, YU H, et al. Global landscape of protein complex-es in the yeast saccharomyces cerevisiae [J]. *Nature*, 2006, 440(7082): 637 – 643.

[27] XENARIOS I, SALWINSKI L, DUAN X, et al. Dip, the database of interacting proteins: a research tool for studying cellular networks of ptoein interactions [J]. *Nucleic Acids Research*, 2002, 30(1): 303 – 305.

[28] STARK C, BREITKREUTZ B J, REGULY T, et al. Biogrid: a general repository for interaction datasets [J]. *Nucleic Acids Research*, 2006, 34(1): 535 – 539.

[29] SPRINZAK E, SATTAH S, MAGALIT H. How re-liable are experimental protein-protein interaction data [J]. *Journal of Molecular Biology*, 2003, 327(5): 919 – 913.

[30] PU S, WONG J, TURNER B, et al. Up-to-date catalogues of yeast protein complexes [J]. *Nucleic Acids Research*, 2009, 37(3): 825 – 831.

[31] MEWES H W, AMID C, ARNOLD R, et al. Mips: analysis and annotation of proteins from whole genomes [J]. *Nucleic Acids Research*, 2004, 32(1): 41 – 44.

[32] XUN Ning, ZHANG Yun, SUN Haiwei, et al. Structural features of attribute reduction matrix and layer fast algorithm [J]. *Control Theory & Applications*, 2007, 24(5): 766 – 770.
(徐宁, 章云, 孙海卫, 等. 属性约简矩阵特征结构及分层约简快速算法 [J]. 控制理论与应用, 2007, 24(5): 766 – 770.)

作者简介:

朱 佳 (1980−), 男, 博士, 副教授, 目前研究方向为机器学习, 已在DMKD, APWEB, DASFAA, WWW等知名国际期刊和会议上发表论文多篇, E-mail: jzhu@m.scnu.edu.cn;

武兴成 (1990−), 男, 硕士研究生, 目前研究方向为数据挖掘, E-mail: 798111484@qq.com;

林雪琴 (1991−),女, 硕士研究生, 目前研究方向为社交网络与数据挖掘, E-mail: xqlin@m.scnu.edu.cn;

肖丹阳 (1992−), 男, 硕士研究生, 研究方向为大数据处理与数据挖掘, E-mail: 444875571@qq.com;

肖 菁 (1975−), 女, 博士, 目前主要研究方向为计算智能、数据挖掘和信息检索, 在人工智能杂志和会议上发表了计算智能、Web文本检索和中英文信息抽取等方面的论文30多篇, E-mail: xiaojing@scnu.edu.cn;

黄 晋 (1976−), 男, 博士, 副研究员, 主要从事数据库、数据挖掘、个性化推荐方面的研究, 近年来在KAIS、JCST、电子学报等国内外重要期刊和FSKD、ICMLC等相关领域的重要国际会议上发表论文20余篇, E-mail: 1936079@qq.com;

贺超波 (1982−), 男, 博士, 副教授, 目前研究方向为数据挖掘、社会计算、教育信息化, E-mail: 25034469@qq.com.