PC-SENE: A node embedding based method for protein complex detection

Xiaoxia Liu[1], Zhihao Yang[1,*], Shengtian Sang[1], Lei Wang[2,*], Yin Zhang[2], Hongfei Lin[1], Bo Xu[3], Yijia Zhang[1], Liang Yang[1], Kan Xu[1], Jian Wang[1]
1 College of Computer Science and Technology, Dalian University of Technology, Dalian, China 116024
2 Beijing Institute of Health Administration and Medical Information, Beijing, 100850, China
3 School of Software Technology, Dalian University of Technology, Dalian, China 116024
Email:* yangzh@dlut.edu.cn, wangleibihami@gmail.com

***Abstract*—With the accumulation of protein-protein interaction (PPI) datasets, various computational methods have been developed for identifying protein complexes from PPI networks. However, many exiting computational methods have their own limitations: supervised learning approaches need tedious effort for feature engineering and the quality measures used to guide the mining process of unsupervised methods have some drawbacks in reflecting the properties of a protein complex in PPI networks. In this work, we proposed a novel protein complex detection method, named PC-SENE. For given seeds, it uses alias sampling strategy based on protein node embedding similarities to select potential addable nodes, and makes use of a new conductance measure to decide whether to extend current candidate subgraph in order to find protein complexes. Intuitively, a well trained node embedding vector could preserve both the topological characteristics of the PPI network and the diversity of connectivity patterns of nodes in the network, and thus node embedding similarities can better reflect the relationship between nodes. The experimental results show the robustness and effectiveness of PC-SENE.**

***Index Terms*—protein complex, PPI network, seed-extension method, node embedding**

I. Introduction

Understanding mechanisms underlying protein complexes is one of the keys to deciphering the cellular mechanisms. A protein complex is a group of proteins that physically interact with one another to organize various biological processes in the cell. So identification of protein complexes is important for better understanding the protein complex formations, and the principles of cellular organization and function. Recent developments in high-throughput technology, such as yeast two-hybrid (Y2H) and tandem affinity purification (TAP) with mass spectrometry (MS), have enabled scientists to determine, identify and validate pairwise protein interactions and to generate large-scale protein-protein interaction (PPI) datasets. With the increasing amount of PPI datasets, automatically identifying protein complexes from PPI network has become an efficient way.

In this paper, we propose a new method named PC-SENE (**P**rotein **C**omplexes detection using **S**eed-**E**xtension method based on **N**ode **E**embedding similarities) to detect protein complexes from PPI networks. To our best knowledge, this is the first time that node embedding is combined with seed-extension method to find protein complexes. We compare our PC-SENE with four state-of-the-art protein complex detection methods, which are MCL, RRW, CMC, and ClusterONE, respectively. Experimental results on six different yeast PPI networks from different publicly accessible databases indicate that PC-SENE outperforms all competing algorithms according to different evaluation criteria.

II. Methods

A. Terminologies and definitions

1) Protein complex quality measure: A protein-protein interaction network can be represented as an undirected graph $G = (V, E, W)$ where nodes in V denote proteins and edges in E denote their interactions with different edge weights in W. Here, we use pairwise clustering coefficient C_{uv} score as weight between node u and v, based on the concept that two nodes with more common neighbors have higher possibility to connect with each other.

As motivated by the conductance defined in [1], a new conductance is presented as quality measure to identify well separated subgraph in a given graph in this study. For a set of proteins denoted as S, the new conductance of S in G is defined as:

$$con(S) = f(density) \cdot \frac{|E(S,S)|}{|E(S,S) + E(S,\overline{S})|}, S \cup \overline{S} = V \quad (1)$$

where $\overline{S}$ is the complement set of S, $E(\cdot,\cdot)$ represents the set of edges between two node sets, and $|\cdot|$ denotes the size of the set. $f(density)$ is a exponential function of subgraph density.

2) Node embedding similarities: Node embedding algorithms aim to automatically extract features from graph-structured data and learn the representations for nodes in graph. In our method, node2vec [2] is utilized to obtain the node embeddings for a given PPI network. Therefore, given a protein pair v_i and v_j, we can calculate the similarity between them by using their embedding vectors. If two nodes are "close" to each other or share common subgraph, the node embedding similarity value between them will be higher than others.

B. PC-SENE algorithm

Given a graph, in primary, embedding vectors of each node are generated, and the similarities between two nodes based on embedding vectors and pairwise clustering coefficient are

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

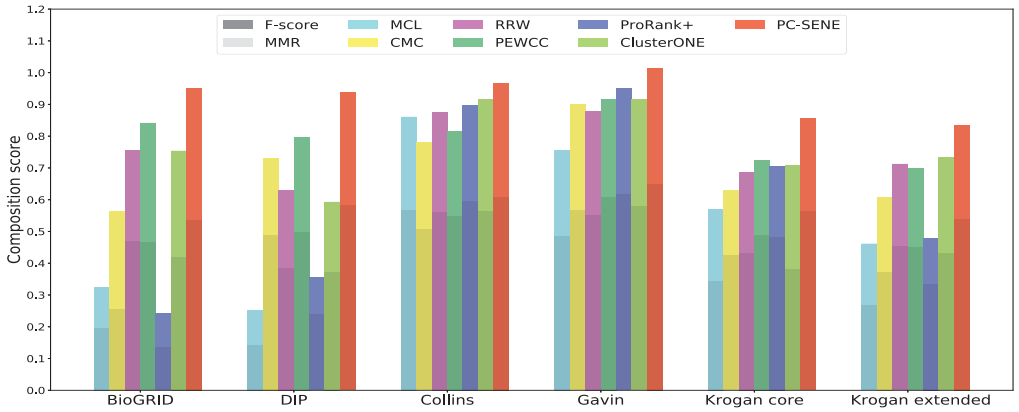

Fig. 1. Comparison of all competing algorithms in terms of the composite scores of F-score and MMR. Shades of the same color indicate different evaluating scores. Each bar height reflects the value of the composite score.

calculated. Secondly, the nodes whose degree is above the average degree of the nodes in the graph are selected as initial seeds and the PC-SENE algorithm uses alias sampling strategy based on node embedding similarities to select addable nodes. Then the algorithm decides whether to add the new node or remove node from current candidate subgraph based on the new conductance value. Thirdly, highly overlapping subgraph pairs are merged. Finally, subgraphs that contain less than three nodes are discarded and the remaining subgraphs are outputted as final protein complexes..

III. RESULTS

A. Datasets

We carried out all the experiments on six real world yeast PPI networks: Gavin, Collins, Krogan core and Krogan extended, BioGRID and DIP. To compare the results with reference complexes, we have constructed a golden standard complexes set by selecting all the protein complexes that had at least three proteins from MIPS, CYC2008, SGD, Aloy and TAP06. Consequently, there was a total 789 protein complexes in the reference set.

B. Evaluation Metrics

To formally evaluate the performance of our method, we used four statistic measures which are widely used in the previous literature for complex detection tasks: precision, recall, F-score and MMR (Maximum matching ratio). Note that a predicted complex is defined to be matched with a known complex if the neighborhood affinity score between them is not less than 0.25 as suggested by previous studies [3].

C. Performance comparison

Figure 1 shows the comparison results on six PPI networks. The overall results, which are represented by the composite scores, demonstrate that PC-SENE obviously outperforms other algorithms on all six networks. With respect to the MMR, ClusterONE is the best for krogan core and krogan extend, but PC-SENE performs the best for the rest. Furthermore, the performance of PC-SENE remains stable in F-score for all the six networks. Specifically, PC-SENE attains the best F-score on all the networks. PC-SENE consistently achieves the best performance among all the unweighted networks, suggesting PC-SENE is capable for detecting protein complexes from large-scale unweighted networks.

IV. CONCLUSION

We propose a new approach, named PC-SENE, for identifying protein complexes in protein-protein interaction networks. We found that PC-SENE outperforms other six state-of-the-art algorithms in identifying protein complexes. The experimental results show that our method can better preserve the diversity of connectivity patterns of nodes and the structures of the networks by utilizing alias sampling strategy based on node embedding similarities to select potential addable nodes, and can better capture the topological structure of a protein complex by using a new conductance as complex quality measure. We hope our work may help the bioinformatics researcher to explore more undiscovered protein complexes.

ACKNOWLEDGEMENTS

This work was supported by the grants from the National Key Research and Development Program of China (No. 2016YFC0901902), Natural Science Foundation of China (No. 61272373, 61572102, 61572098 and 61502071), and Trans-Century Training Program Foundation for the Talents by the Ministry of Education of China (NCET-13-0084).

REFERENCES

[1] R. Andersen, F. Chung, and K. Lang, "Local graph partitioning using pagerank vectors," in *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*. IEEE, 2006, pp. 475–486.

[2] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 855–864.

[3] X. Liu, Z. Yang, Z. Zhou, Y. Sun, H. Lin, J. Wang, and B. Xu, "The impact of protein interaction networks characteristics on computational complex detection methods," *Journal of theoretical biology*, vol. 439, pp. 141–151, 2018.