Protein Complexes Detection Based on Global Network Representation Learning

Bo Xu*[†], Kun Li*, Xiaoxia Liu[‡], Delong Liu[§], Yijia Zhang[‡], Hongfei Lin[‡], Zhihao Yang[‡], Jian Wang[‡]
and Feng Xia*[†]
*School of Software, Dalian University of Technology, China
[†]Key Laboratory for Ubiquitous Network and Service Software of Liaoning, China
[‡]College of Computer Science and Technology, Dalian University of Technology, China
[§]China Institute of Water Resources and Hydropower Research, China

***Abstract*—Detecting protein complexes from protein-protein interaction (PPI) networks allows biologists reveal the principle of cellular organization and functions. Existing computational methods try to incorporate biological evidence to enhance the quality of predicted complexes. However, it is still a challenge to integrate biological information into complexes discovery process under a unified framework. Recently, network embedding methods showed their effectiveness in graph data analysis tasks. It provides a framework for incorporating both network structure and additional node attribute information. This salient feature is particularly desirable in the context of protein complexes identification. However, none of the existing network embedding methods take node attribute proximity and high-order structure proximity into account at the same time. In this paper, we propose a novel global network embedding method, which preserves global network structure and biological information. We utilize this global representation learning method to learn vector representation for proteins. Then, we use a seed-extension clustering method to discover overlapping protein complexes with the embedding results. This novel protein complexes detection method we called GLONE. Evaluated on five real yeast PPI networks, our method outperforms the competing algorithms in terms of different evaluation metrics.**

***Index Terms*—Protein complexes identification, Network embedding, PPI network**

I. Introduction

Protein complexes are significant components of cell and life, which are necessary for us to understand the principles of cellular function and other biological mechanisms within the cell. With the development of high-throughput experimental techniques, a large quantity of PPI networks have been produced to discover protein complexes. A PPI network can be represented as an undirected graph, in which vertices represent proteins and edges represent interactions between proteins. In recent years, several computational algorithms have been proposed to discover protein complexes from PPI networks, which mainly based on the observation that cliques or densely connected subgraphs often represent protein complexes. These methods like ClusterOne [1], MCODE [2], RNSC [3], CMC [4], MCL [5] [6], IPCA [7] solely make use of the topology of the network.

Besides topological information of the network, biological information such as gene expression data, functional information can be used for prediction. The method proposed by Rungsarityotin et al. [8], the method proposed by Jung et al. [9], COACH [10] detects protein complexes with other biological information. However, it is still a challenge to combine biological evidence with network topological information in a unified framework.

Network embedding methods transform the network into a low dimensional space. After network embedding, most of the existing network analytic tasks can be solved more effectively. Network embedding methods can be categorized into two categories [11].

The methods of the first category embed the network by preserving the structural neighboring information. The structures of a network include first-order structure and high-order structure. Several classical network embedding methods like DeepWalk [12], Node2vec [13] and LINE [14] aim to preserve these structures .

The methods of the second category aim to preserve the various information such as node label, node text feature, node attributions, etc. MMDW [15], TADW [16], the method Pan et al. [17] proposed and AANE [18] can capture the structure proximities and attribute proximities.

Utilizing network embedding methods can booost complexes identification performance when incorporating network structure information and additional node information. Therefore, we attempt to utilize the network embedding in protein complexes detection. We propose a new global network embedding method which preserves both high-order structure proximity and biological attribute proximity for identifying protein complexes.

In this study, we propose GLONE to detect protein complexes. Firstly, we propose a novel global network representation learning method to learn vector representation for each protein. Secondly, we use a seed-extension clustering method to discover protein complexes based on embedding results. We compared with six classic protein complexes detection methods to evaluate the performance of our method, which are COACH, CMC, MCODE, ClusterOne, IPCA, MCL on five yeast PPI networks respectively. Experiment results show that our method outperforms the state-of-the-art methods.

This work was supported in part by the Natural Science Foundation of China under Grant 61502071 and Grant 61572094 and in part by the Fundamental Research Funds for the Central Universities under Grant DUT18RC(3)004.
* Corresponding Author: boxu@dlut.edu.cn

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

To summarize, we make the following contributions:

(1) We propose a novel global network representation learning method which preserves both high-order structure proximity and node attributed proximity. It can be used in other network analytic tasks.

(2) Our method provides a framework that integrates many valuable and different biological information into protein complex detection.

(3) The proposed global network representation learning can be used to other biological tasks.

II. Methods

GLONE is a two-step procedure. Firstly, it learns the vector representation for each protein from the GO attributed PPI network by global network embedding method. Secondly, it utilizes a seed-extension method to identify protein complexes based on embedding results. The major steps of GLONE are presented in Fig.1.

A. Learning vector representations for proteins

Motivated by TADW [16] and AANE [18], we propose a new global network embedding method which preserves both high-order structure proximity and biological attribute proximity simultaneously. The corresponding representation learning method is described below.

1) Modeling attribute information: Given a PPI network $G = (V, E)$, V and E represent proteins and edges in the network separately. In order to integrate the protein attribute into the embedding result, the dot product of protein embedding vector is used to approach their attribute proximity. The loss function is defined as:

$$J_A = ||S - HH^{\mathrm{T}}||_F^2 = \sum_{i=1}^{|V|}\sum_{j=1}^{|V|}(s_{ij} - h_i h_j^{\mathrm{T}})^2 \tag{1}$$

where $H \in R^{|V| \times d}$ is the vector representation matrix and d is the dimension. h_i, $h_j$represent the vector representation for protein i and j. $S \in R^{|V| \times |V|}$ is the attribute affinity matrix, where s_{ij} represents the attribute similarity between protein i and j. The matrix S is calculated in Equation (2):

$$s_{ij} = \frac{\sum_{k=1}^{m} a_{ik} \times a_{jk}}{\sqrt{\sum_{k=1}^{m} a_{ik}^2} \times \sqrt{\sum_{k=1}^{m} a_{jk}^2}} \tag{2}$$

where $A \in R^{|V| \times m}$ is an attribute matrix for proteins and m represents the number of protein attribute information. Here, we select Gene Ontology (GO) slims as attribute information for proteins. a_{ik} represents whether protein i has a corresponding GO slim k with $a_{ik} = 1$ or 0. Since GO slims for cellular component (Cc) inlcude some protein complexes information, we only select GO slims for biological process (Bp) and modecular function (Mf) as protein attributes.

2) Modeling structure information: We hope that our algorithm can well grasp the globality of the structure. TADW mines the correlation of more unconnected nodes in the network by the multi-step transfer. We will adopt it in our structural model.

Motivated by natural language processing, DeepWalk [12] uses local information to learn latent representations by treating walks as the equivalent of sentences. TADW proves that DeepWalk is equivalent to factorize a matrix M.

$$M = HH^{\mathrm{T}} \tag{3}$$

where DeepWalk takes the matrix H as vertex representation. $M \in R^{|V| \times |V|}$ is logarithm of the average probability that node v_i randomly walks to v_j in a fixed step.

$$M = log(\frac{T^1 + T^2 + \cdots + T^t}{t}) \tag{4}$$

where T denotes the transition matrix and T_{ij}^t represents the probability that node v_i randomly walks to node v_j at t steps. W denotes the adjacency matrix of this network, d_i denotes the degree of node i and $T_{ij}^1 = \frac{W_{ij}}{d_i}$. $T^t$is the multiplication of matrix T by t times.

In order to preserve high-order structural proximity in the PPI network, a loss function is defined as :

$$J_G = ||M - HH^{\mathrm{T}}||_F^2 = \sum_{i=1}^{|V|}\sum_{j=1}^{|V|}(m_{ij} - h_i h_j^{\mathrm{T}})^2 \tag{5}$$

where H is low-dimensional representations of proteins.

3) Jointing and optimizing model for representation learning: In order to preserve both of the above aspects, we jointly model the two types of information in the following function:

$$\min_H J = J_A + J_G = ||S - HH^{\mathrm{T}}||_F^2 + \lambda||M - HH^{\mathrm{T}}||_F^2 \tag{6}$$

where λ is a parameter that controls the trade-off between structural and attributed model.

For the calculation of the final loss function, we use the ADMM [19] method, The procedure of optimizing objective function is as following.

We set $Z = H$ as constraints and rewrite the final optimization function:

$$\min_H J = ||S - HZ^{\mathrm{T}}||_F^2 + \lambda||M - HZ^{\mathrm{T}}||_F^2 \quad s.t. \quad Z = H \tag{7}$$

We can use the extended Lagrangian multiplier method [20] and the optimization function can be rewritten as:

$$\begin{aligned} L =&||S - HZ^{\mathrm{T}}||_F^2 + \lambda||M - HZ^{\mathrm{T}}||_F^2 \\ &+ \frac{\rho}{2}(||H - Z + U||_F^2 - ||U||_F^2) \end{aligned} \tag{8}$$

where $U \in R^{|V| \times d}$ is the extended matrix and ρ is a penalty parameter. To minimize the value of the function L, we can find a saddle value for L by iteratively updating H, Z, and U in Equation (9)(10)(11).

$$\begin{aligned} H^{k+1} =& \arg\min_{H^k}(||S - H^k Z^{k^{\mathrm{T}}}||_F^2 + \lambda||M - H^k Z^{k^{\mathrm{T}}}||_F^2 \\ &+ \frac{\rho}{2}||H^k - Z^k + U^k||_F^2) \end{aligned} \tag{9}$$

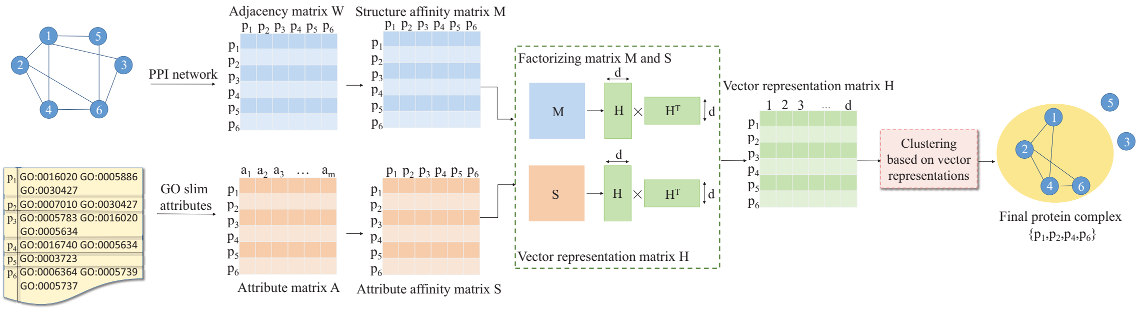
1
5
2
3
4
6
PPI network
Adjacency matrix W
p1 p2 p3 p4 p5 p6
Structure affinity matrix M
p1 GO:0016020 GO:0005886 GO:0030427
p2 GO:0007010 GO:0030427
p3 GO:0005783 GO:0016020 GO:0005634
p4 GO:0016740 GO:0005634
p5 GO:0003723
p6 GO:0006364 GO:0005739 GO:0005737
GO slim attributes
a1 a2 a3 … am
Attribute matrix A
Attribute affinity matrix S
Factorizing matrix M and S
M
H
×
H^T
d
S
Vector representation matrix H
Vector representation matrix H
1 2 3 … d
Clustering based on vector representations
Final protein complex
{p1,p2,p4,p6}

Fig. 1. The basic idea of GLONE to predict protein complexes.

$$Z^{k+1} = \arg\min_{H^k}(||S - H^{k+1}Z^{k^{\mathrm{T}}}||_F^2 + \lambda||M - H^{k+1}Z^{k^{\mathrm{T}}}||_F^2 + \frac{\rho}{2}||H^{k+1} - Z^k + U^k||_F^2) \tag{10}$$

$$U^{k+1} = U^k + (H^{k+1} - Z^{k+1}) \tag{11}$$

Since all of them are convex functions, we can get the update rules by deriving it equal to 0.

$$H^{k+1} = \frac{2SZ^k + 2\lambda MZ^k + \rho(Z^k - U^k)}{4Z^{k^{\mathrm{T}}}Z^k + \rho I} \tag{12}$$

$$Z^{k+1} = \frac{2SH^{k+1} + 2\lambda MH^{k+1} + \rho(H^{k+1} - U^{k+1})}{4Z^{k^{\mathrm{T}}}Z^k + \rho I} \tag{13}$$

Since each row of H and Z are independent of each other, we can calculate each row i of H and Z in parallel.

$$h_i^{k+1} = \frac{2s_iZ^k + 2\lambda m_iZ^k + \rho(z_i^k - u_i^k)}{4Z^{k^{\mathrm{T}}}Z^k + \rho I} \tag{14}$$

$$z_i^{k+1} = \frac{2s_iH^{k+1} + 2\lambda m_iH^{k+1} + \rho(h_i^{k+1} - u_i^k)}{4Z^{k^{\mathrm{T}}}Z^k + \rho I} \tag{15}$$

In final, each protein is represented as a vector.

B. Clustering based on the seed-extension method

We propose a seed-extension method which is motivated by IPCA [7] to detect protein complexes. It can be divided into three major parts: weighting proteins, selecting seeds, and extending clusters.

1) Weighting proteins: For an input network $G = (V, E)$, we assign a weight for each protein in the PPI network.

$$WP_v = \sum_{i,v\in V\bigcap(i,v)\in E} cosine(h_i, h_v) \tag{16}$$

where protein i is the neighbor of protein v in PPI network and h_i, h_v are their vector represenations. We define cosine similarity of vector representations as the weight of each pair of proteins and assign the sum of weights of incident edges as protein's weight.

2) Selecting seeds: We sort proteins in a non-increasing order by their weight WP in a queue $SeedSet$. We pick the first protein in $SeedSet$ as a seed to grow. Once the cluster is completed, proteins in this cluster are removed from the queue $SeedSet$. Then we pick the first protein remaining in the queue $SeedSet$ as the seed. This process repeats until $SeedSet$ is empty.

3) Extending clusters: After selecting a seed, this cluster S is extended by adding proteins from its neighbors. We utilize two parameters to determine whether a neighbor can be added.

- We define a priority weight to measure how strongly a protein v is connected to a cluster S:

$$CM_{vS} = \frac{\sum_{i,v\in V\bigcap i\in S\bigcap(i,v)\in E} cosine(h_i, h_v)}{\sum_{i,j\in V\bigcap i,j\in S\bigcap(i,j)\in E} cosine(h_i, h_j)} \tag{17}$$

  where i, j are proteins in cluster S and v is a neighbor of S. h_i, $h_j$and h_v are vector representation of protein i, j and v. We calculate cosine similarity of embedding results as the weight of each pair of proteins.
- We define the diameter dm of a subgraph as the average length between each pairs of proteins in this subgraph.

The cluster S is extended by adding proteins recursively from its neighbors according to the above two parameters. The specific process is as follows:

- Neighbors of S are sorted in descending order by CM_{vS}.
- We select the candidate protein v with the highest priority. If its CM_{vS} is not less than a threshold value θ and the diameter of $S + v$ is bounded by dm ($dm = 2$ is used according to previous analysis [7]), this protein can be added to this cluster. When either of these two conditions is not satisfied, this candidate protein cannot be added to this cluster. Then the next highest priority neighbor of this cluster is taken into consideration.
- Once any neighboring protein is added into S, the cluster and its neighbors are updated.

This process repeats until all the neighbors of the cluster cannot satisfy these two conditions. All the proteins formed the complete complex will be removed from the $SeedSet$. Once the $SeedSet$ is empty, the clustering is finished. The obtained clusters are considered as protein complexes.

III. Experiment results

A. Datasets

We implemented experiments on five networks: DIP [21], Krogancore [22], Biogrid [23], Gavin [24], Collins [25]. The GO slim information was downloaded from https://downloads.yeastgenome.org/curation/literature/go_slim_mapping.tab. The benchmark complex dataset was downloaded from http://wodaklab.org/cyc2008/.

B. Evaluation metrics

The main measures to evaluate the performance of methods are $Precision$, $Recall$, $F - score$, Sn, PPV and Acc.

C. Performance comparison

We compared our approach GLONE with six protein complex identification methods: COACH, CMC, MCODE, ClusterONE, IPCA and MCL on five PPI datasets.

Fig.2 shows the comparison results of $F - score$ and Acc in five PPI networks. The overall results which are represented by the composite scores demonstrate that GLONE clearly outperforms other algorithms for all five networks.

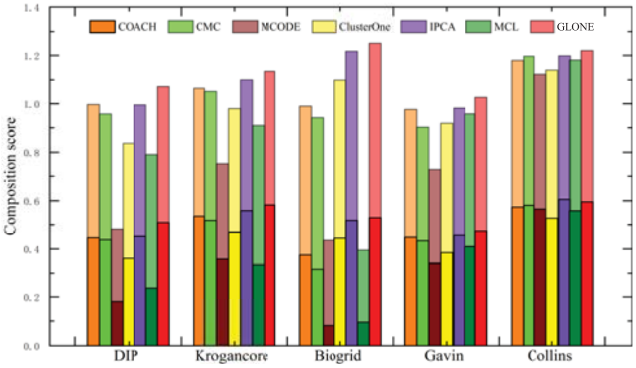

Fig. 2. Comparison of other six algorithms in terms of the composite scores of F-score and Acc.

Conclusions

In this study, we proposed a novel protein complexes detection method. Firstly, we proposed a novel global network embedding method to learn vector representation for each protein from GO attributed PPI networks. Secondly, we used a seed-extension clustering method to discover overlapping protein complexes based on our embedding results. Experiments showed that our method GLONE outperforms five protein complexes detection methods on five different datasets. We concluded that GLONE can effectively enhance the quality of predicted complexes. In the future, we will apply our novel global network embedding method to other biological networks.

References

[1] T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks." *Nature Methods*, vol. 9, no. 5, pp. 471–472, 2012.

[2] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *Bmc bioinformatics*, vol. 4, no. 1, p. 2, 2003.

[3] A. D. King, N. Przulj, and I. Jurisica, "Protein complex prediction via cost-based clustering." *Bioinformatics*, vol. 20, no. 17, pp. 3013–3020, 2004.

[4] G. Liu, L. Wong, and H. N. Chua, "Complex discovery from weighted ppi networks." *Bioinformatics*, vol. 25, no. 15, pp. 1891–7, 2009.

[5] S. Van Dongen, "Graph clustering by flow simulation," *Phd Thesis University of Utrecht*, 2000.

[6] J. B. Pereiraleal, A. J. Enright, and C. A. Ouzounis, "Detection of functional modules from protein interaction networks." *Proteinsstructure Function & Bioinformatics*, vol. 54, no. 1, pp. 49–57, 2004.

[7] M. Li, J. E. Chen, J. X. Wang, B. Hu, and G. Chen, "Modifying the dpclus algorithm for identifying protein complexes based on new topological structures," *Bmc Bioinformatics*, vol. 9, no. 1, pp. 1–16, 2008.

[8] W. Rungsarityotin, R. Krause, A. Schodl, and A. Schliep, "Identifying protein complexes directly from high-throughput tap data with markov random fields," *BMC Bioinformatics*, vol. 8, no. 1, pp. 482–482, 2007.

[9] S. Jung, W. H. Jang, H. Hur, B. Hyun, and D. Han, "Protein complex prediction based on mutually exclusive interactions in protein interaction network," *Genome Informatics*, vol. 21, pp. 77–88, 2008.

[10] M. Wu, X. Li, C. K. Kwoh, and S. K. Ng, "A core-attachment based method to detect protein complexes in ppi networks," *Bmc Bioinformatics*, vol. 10, no. 1, p. 169, 2009.

[11] P. Cui, X. Wang, J. Pei, and W. Zhu, "A survey on network embedding," *IEEE Transactions on Knowledge & Data Engineering*, vol. PP, no. 99, pp. 1–1, 2017.

[12] B. Perozzi, R. AlRfou, and S. Skiena, "Deepwalk: online learning of social representations," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 701–710.

[13] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 855–864.

[14] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," *international world wide web conferences*, pp. 1067–1077, 2015.

[15] C. Tu, W. Zhang, Z. Liu, and M. Sun, "Maxmargin deepwalk: discriminative learning of network representation," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 3889–3895.

[16] C. Yang, D. Zhao, D. Zhao, E. Y. Chang, and E. Y. Chang, "Network representation learning with rich text information," in *International Conference on Artificial Intelligence*, 2015, pp. 2111–2117.

[17] S. Pan, J. Wu, X. Zhu, C. Zhang, and Y. Wang, "Tri-party deep network representation," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 1895–1901.

[18] X. Huang, J. Li, and X. Hu, "Accelerated attributed network embedding," in *Proceedings of the 2017 SIAM International Conference on Data Mining.*, 2017, pp. 633–641.

[19] S. Boyd, N. Parikh, E. Chu, and B. Peleato, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations & Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[20] D. Hallac, J. Leskovec, and S. Boyd, "Network lasso: Clustering and optimization in large graphs," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 387–396.

[21] I. Xenarios, ukasz Salwnski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg, "Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, no. 1, p. 303, 2002.

[22] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, and A. P. Tikuisis, "Global landscape of protein complexes in the yeast saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637–43, 2006.

[23] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: a general repository for interaction datasets," *Nucleic Acids Research*, vol. 34, no. Database issue, pp. 535–9, 2006.

[24] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, and B. Dmpelfeld, "Proteome survey reveals modularity of the yeast cell machinery," *Nature*, vol. 440, no. 7084, pp. 631–6, 2006.

[25] S. R. Collins, P. Kemmeren, X. Zhao, J. Greenblatt, F. Spencer, F. C. P. Holstege, J. S. Weissman, and N. J. Krogan, "Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae," *Molecular & Cellular Proteomics*, vol. 6, no. 3, pp. 439–450, 2007.