Performance Analysis of Graph Laplacian Matrices in Detecting Protein Complexes

Dong Yun-yuan[1], Keith C.C. Chan[2], Liu Qi-jun[3] and Wang Zheng-hua[1]

[1]*College of Computer, National University of Defense Technology, Changsha 410073, China*
[2]*Department of computing, Hong Kong Polytechnic University, Hong Kong, China*
[3]*College of Science, National University of Defense Technology, Changsha 410073, China*
happydongyy@gmail.com, cskcchan@inet.polyu.edu.hk, ivanliuqj@nudt.edu.cn, zhhwang188@sina.com

Abstract

Detecting protein complexes is an important way to discover the relationship between network topological structure and its functional features in protein-protein interaction (PPI) network. The spectral clustering method is a popular approach. However, how to select its optimal Laplacian matrix is still an open problem. Here, we analyzed the performances of three graph Laplacian matrices (unnormalized symmetric graph Laplacians,, normalized symmetric graph Laplacians and normalized random walk graph Laplacians, respectively) in yeast PPI network. The comparison shows that the performances of unnormalized and normalized symmetric graph Laplacian matrices are similar, and they are better than that of normalized random walk graph Laplacian matrix. It is helpful to choose proper graph Laplacian matrix for PPI networks' analysis.

Keywords: *Protein-protein interaction network, Protein complex, Spectral clustering method, Graph Laplacian matrix*

1. Introduction

Protein complexes, also known as dense sub-networks in a PPI network, are groups of proteins that could carry out many vital functions (e.g. replication, transcription and gene expression, etc.). Protein complexes are important for understanding principles of cellular organization and function. Recently, many high-throughput techniques have been used to produce PPI data. As one of the most important biological networks, PPI network is fascinating for us to understand the whole image of biological processes and cellular systems [1].

However, because experimental methods are both time-consuming and expensive, the number of discovered protein complexes is far from complete. There is a crucial need to develop effective computational methods to accurately identify protein complexes from large-scale PPI networks.

Many computational methods can be used to discover protein complexes. Because a PPI network can be represented as a graph, identifying protein complexes is modeling as network partition problem or graph clustering problem. In recent years, spectral clustering method has become one of the most popular clustering algorithms. It is widely used in network partition and graph clustering, and it can also be used for detecting protein complexes. Spectral clustering method very often outperforms traditional clustering algorithms, such as k-means

or single linkage; and it is simple to implement. Bu and Lu et al applied the spectral clustering method to budding yeast PPI network and identified the hidden topological structure [2,3]. Sen et al used eigenvalue/ eigenvector decomposition in the connectivity matrix and found that proteins in a same eigenvector tended to interact with each other, although they had various degrees [4]. Qin et al focused on constructing similarity graphs and determining the number of clusters, and studied spectral clustering for detecting protein complexes in PPI network [5].

Note that, the most important part of spectral clustering method is the definition of graph Laplacian matrix. Ulrike von Luxburg defined three basic graph Laplacian matrices [6]. A fundamental problem with spectral clustering method is which of the three graph Laplacians matrices should be used. However, as far as we know, there is still lack of references on how to choose a proper graph Laplacian matrix for detecting protein complexes.

In the following sections, we'll first compare the performances of three graph Laplacian matrices under the spectral clustering method in yeast PPI network; and then make a detail analysis to explain the reason from the perspective of mathematics. Finally, we conclude that how to choose a proper graph Laplacian matrix for different PPI network. If the PPI network is connected, the normalized random walk graph Laplacian matrix is the best choice. While when the PPI network is disconnected, the normalized symmetric graph Laplacian matrix is the proper one.

2. Materials and Methods

2.1. Experimental Data

The dataset 'DIP_core' (Database of Interacting Protein core), obtained from the DIP website (http://dip.doe-mbi.ucla.edu/dip), which contains the most reliable interactions, as judged both manually by Deane et al. and computational approaches[7]. By removing self-interacting interactions and redundancy interactions, the resultant network contains 2164 protein and 4303 interaction. The dataset of protein complexes is collected from MIPS database[8], containing 1138 protein complexes.

2.2. Spectral Clustering Method

The intuitive goal of clustering is to partition similar nodes into the same group and dissimilar ones into different groups. So, the problem of clustering can now be reformulated using the similarity graph: we want to find a partition of the graph, which satisfy that the edges between different groups have very low weight (i.e. nodes in different clusters are dissimilar from each other) and the edges within a group have high weight (i.e. nodes within the same cluster are similar to each other).

In the framework of classical spectral clustering to detect protein complexes, a similarity graph is firstly built from the original dataset, and then its Laplacian matrix, which is calculated by the similarity graph, is used for the clustering. Based on the clustering results, the PPI networks are decomposed into multi-group structure by the mapping graph nodes into proteins and clusters into protein complexes.

2.2.1. Graph notations: The PPI network can be modeled as an undirected graph $G = (V, E)$, in which V is the set of nodes(proteins) and E is the set of edges(protein interactions). The adjacency matrix of graph G is the matrix $A= a_{ij}$, i, j=1,2, …, n. If there is an edge between node i and node j, a_{ij}=1; otherwise a_{ij}=0. The degree matrix D is a diagonal matrix and the weighted adjacency matrix of graph G is the matrix $W=\{w_{ij}\}$.

The main tools for spectral clustering are graph Laplacian matrices. Spectral graph theory is the study of those matrices. There are three forms of graph Laplacian matrices, defined as follows[6].

The unnormalized graph Laplacians matrix: $L=D-W$

The normalized symmetric graph Laplacians matrix: $L_{sys}=D^{-1/2}L\ D^{-1/2}=I-\ D^{-1/2}WD^{-1/2}$

The normalized random walk graph Laplacians matrix: $L_{rw}=D^{-1}L=I-\ D^{-1}W$

2.2.2. Construction of weighted PPI network: There are many ways to reconstruct a PPI network by different weighted adjacency matrices *W*. Here, according to the topological structure of the PPI network, we get a weighted PPI network. If a pair of interacted proteins had more common neighbors, they would have stronger functional associations; and it was not random[9]. The statistical signification of common neighbors for each pair of interacting proteins can be used to weight the PPI network. We count the number of distinct ways in which two proteins *i* and *j* with n_1 and n_2 neighbors respectively and *m* in common. The total number of proteins in the PPI network is represented as *n*. *P*-value is used to weight edges.

$$P(n,n_1,n_2,m)=\frac{\binom{n}{m}\binom{n-m}{n_1-m}\binom{n-n_1}{n_2-m}}{\binom{n}{n_1}\binom{n}{n_2}}$$

2.2.3. Spectral clustering method: We apply a spectral clustering algorithm based on the three matrices. It only has one parameter gap, and no need to give the exact number of protein complexes at first.

Input: weighted matrix *W*, degree matrix *D* and parameter *gap*

Procedure:

1. Compute the second smallest eigenvalue λ_2 and its corresponding normalized eigenvector x_2.

2. Sort the elements in x_2 ascendingly and find the max interval between two adjacent elements. The elements whose values are more than the max interval are in one cluster, while the less ones are in another cluster.

3. Repeat step 2, until the max interval in one cluster is less than gap.

Output: Complexes C_1, C_2 …

2.3. Evaluation Criteria

To evaluate the effectiveness of graph Laplacian matrices, we use overlapping score[10] to determine how accurate a predicted complex is matched with a known protein complex.
$O(C_i, C_j)=\frac{s}{n_{C_i}*n_{C_j}}$, where C_i is the predicted complex, C_j the known protein complex is, *s* is the number of common proteins shared by C_i and C_j, n_{C_i} is the number of proteins in C_i, and $n_{C_j}$is the number of proteins in C_j. The value of overlapping score is ranging from 0 to 1. If *O* is equal to 0, it means no protein in C_i is found in C_j (i.e. zero matching); if *O* is equal to 1, it means all proteins in C_i are found in C_j (i.e. perfect matching).

3. Results and Discussion

3.1. Parameter *gap*

Parameter *gap* is the threshold of interval between two elements in the eigenvector. As shown in Figure 1, with the increase of *gap*, the number of complexes decreases. As *gap* is increased, the similar requirement of two adjacent elements declines, which leads to the reducing of the number of protein complexes; and vice versa. Meanwhile, when *gap* is too small, the proteins in a complex are partitioned into several complexes; and when *gap* is too large, the proteins originally in different complexes are assigned into one complex. Because we use the dataset of protein complexes from MIPS as benchmark, the number of protein complexes is labeled in Figure 1.

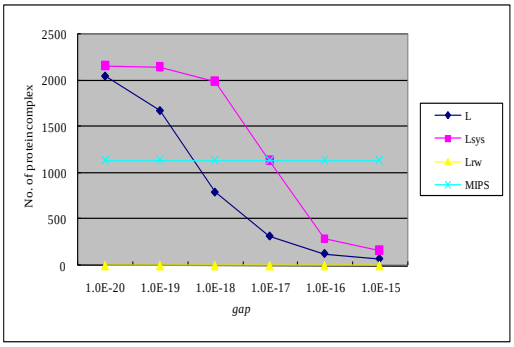

Figure 1. Number of Protein Complex with Different *gap*

3.2. Analysis of the Performance of Three Matrices

The overlapping score of three Laplacian matrices are calculated. The performances of L and L_{sys} are similar, while L_{rw} performs worst, with no matched protein complex.

3.2.1. L and L_{sys}: As shown in Figure 2, the performances of L and L_{sys} are similar. Because of their definition, L and L_{sys} are similar matrices.

As shown in Figure 1 and 2(a), (b), L_{sys} performs a little better than L. There are some explanations, such as graph partitioning[6], consistency[11,12] and so on.

3.2.2. L_{rw}: L_{rw} is a normalized random walk graph Laplacian. A random walk on a graph is a stochastic process which randomly jumps from one vertex to another. It works well in connected networks.

A graph G is connected if and only if $\lambda_2(G) > 0$[13]. For DIP_core, $\lambda_2 = -0.9801 < 0$, which means DIP_core is disconnected and L_{rw} is not suitable for DIP_core. If the PPI network is a connected network, according to reference[6], L_{rw} is best choice.

4. Conclusions

In this study, we analyze the performances of three graph Laplacian matrices when the spectral clustering method is applied for detecting protein complexes in a network context. The results show that the performances of unnormalized and normalized symmetric graph Laplacian matrices are similar, both of which are better than the performance of normalized random walk graph Laplacian matrix.

So we can choose proper graph Laplacian matrix according to the types of PPI networks. If the PPI network is connected, the normalized random walk graph Laplacian matrix (L_{rw}) is

the best choice. While when the PPI network is disconnected, the normalized symmetric graph Laplacian matrix (L_{sys}) is advocated for using.

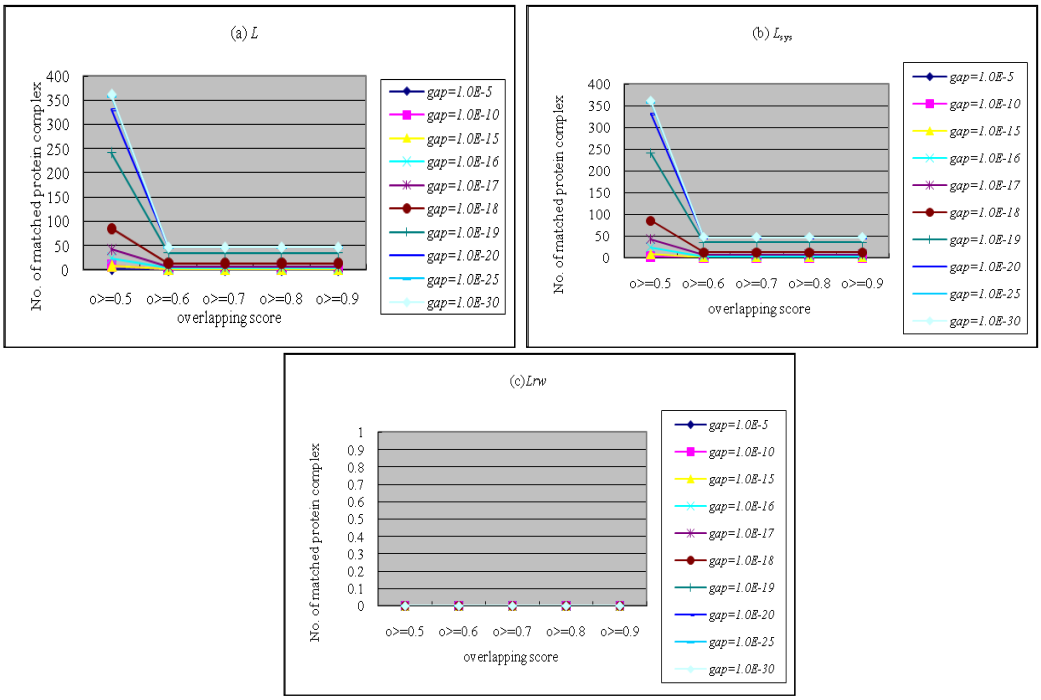

Figure 2. Number of Matched Protein Complex with Three Laplacian Matrices

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No.: 60603054 and 60773021, and Hu'nan Natural Science Foundation Grant No.: 08JJ4021.

References

[1] U. Alon, "Biological networks: the tinkerer as an engineer", Science 301, pp. 1866-1867 **(2003)**.

[2] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li and R. Chen, "Topological structure analysis of the protein-protein interaction network in budding yeast", Nucleic Acids Research 31, pp. 2443-2450 **(2003)**.

[3] H. Lu, X. Zhu, H. Liu, G. Skogerbø, J. Zhang, Y. Zhang, L. Cai, Y. Zhao, S. Sun, J. Xu, D. Bu and R. Chen, "The interactome as a tree-an attempt to visualize the protein-protein interaction network in yeast", Nucleic Acids Research 32, pp. 4804-4811 **(2004)**.

[4] T. Z. Sen, A. Kloczkowski and R. L. Jernigan, "Functional clustering of yeast proteins from the protein-protein interaction network", BMC Bioinformatics 7 **(2006)**.

[5] G. Qin and L. Gao, "Spectral clustering for detecting protein complexes in protein-protein interaction (PPI) networks", Mathematical and Computer Modelling 52, pp. 2066-2074 **(2010)**.

[6] U.v. Luxburg, "A tutorial on spectral clustering", Stat Comput 17, pp. 395-416 **(2007)**.

[7] Deane CM, Salwin´ ski Ł, Xenarios I, E. D, "Protein Interactions: Two Methods for Assessment of the Reliability of High Throughput Observations", Mol Cell Proteomics 1, pp. 349-356 **(2002)**.

[8] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H. W. Mewes, A. Ruepp and D. Frishman, "The MIPS mammalian protein-protein interaction database", Bioinformatics 21, pp. 832-834 **(2005)**.

[9] M. P. Samanta and S. Liang, "Predicting protein functions from redundancies in large-scale protein interaction networks", Proc. Natl. Acad. Sci., pp. 12579-12583, **(2003)** USA.

[10] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks", BMC Bioinformatics 4, 2 **(2003)**.

[11] U. von Luxburg, O. Bousquet and M. Belkin, "On the convergence of spectral clustering on random samples: the normalized case", in: J.S.-T.a.Y. Singer (Ed.), Proceedings of the 17th Annual Conference on Learning Theory (COLT), Springer, pp. 457-471 **(2004)** , New York, USA.

[12] U. von Luxburg, O. Bousquet and M. Belkin, "Limits of spectral clustering", in: Y.W. L. Saul, and L. Bottou (Ed.), Advances in Neural Information Processing Systems (NIPS) MIT Press, Cambridge, MA, pp. 857-864 **(2005)**.

[13] P. Miroslav Fiedler, "Algebraic connectivity of graphs", Czechoslovak Mathematical Journal 23, pp. 298-305 **(1973)**.

Authors

DONG Yun-yuan

She is a PhD candidate in computer science of the National University of Defense Technology. Her current research interestes is bioinformatics.

LIU Qi-jun

PhD and lecturer in college of science of the National University of Defense Technology. His main research interests include bioinformatics and system biology.

WANG Zheng-hua

Professor and PhD supervisor in computer science of the National University of Denfense Technology. His main research interests include bioinformatics, parallel computing and computational fluid dynamics.