Identifying of protein complexes and functional modules in *E.coli* PPI networks

Ping Kong

Shanghai University of Medicine and Health Sciences

Wei Liu (✉ liuyiwei1030@hotmail.com)

Nantong University <https://orcid.org/0000-0002-8016-0295>

Research article

Keywords: protein complexes, PPI networks, functional modules, clustering algorithm, E. Coli binary network

Posted Date: January 10th, 2020

DOI: <https://doi.org/10.21203/rs.2.20590/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: *Escherichia coli* has been at the center of microbial research for decades, making it a standard microorganism for studying molecular mechanism. Molecular complexes, operons and functional modules are important molecular functional domains of *Escherichia coli*. Most previous studies focused on the detection of *E. coli* protein complexes based on the experimental methods. While the research of prediction of protein complexes in *E. coli* based on large-scale proteomic data, especially the functional modules of *E. coli* are relatively few. Identifying protein complexes and functional modules of *E. coli* is crucial to reveal principles of cellular organizations, processes and functions.

Results: In this study, the protein complexes and functional modules of two high-quality binary interaction datasets of *E. coli* are predicted by an efficient edge clustering algorithm (ELPA) for complex biological network, respectively. According to the gold standard protein complexes and function annotations provided by EcoCyc dataset, the experimental results show that most topological modules predicted in the two datasets match very well with the real protein complexes, cellular processes and biological functions. By analyzing the corresponding complexes and functional modules shows that all predicted protein complexes are fully covered by one or more functional modules. Furthermore, we compared the results of ELPA with a famous node clustering algorithm (MCL) on the same PPI network of *E. coli*, and found that ELPA outperforms MCL in terms of matching with gold standard complexes.

Conclusions: As a consequence, we surmise that topological modules of PPI network detected by ELPA fits well with real protein complexes and functional units. In most predicted topological modules, the protein complexes and corresponding functional modules are highly overlapping. ELPA is an effective tool to predict protein complexes and functional modules in PPI networks of *E. coli*.

1. Introduction

Escherichia coli is a primary model organism for microbe, and perhaps it is the most intensively studied species of bacteria [1–4]. Even so, only two-third of the protein–coding gene products of *E. coli* K-12 currently have experimental evidence indicative of a biological role, and others remain functionally unannotated (orphans) [5]. The growing large-scale genomic data make *E. coli* particularly well-suited to systematic investigations of microbial protein components and functional relationships from a global perspective. Experiments and data analysis (PPI network model) are two effective methods to identify protein complexes and functional modules of *E. coli*. As we all know, the research of *E. coli* based on experimental method is absolutely dominant [5–7] for decades. The experimental method has the advantages of direct verification, but it also has the limitations of high false positive and false negative results. With the development of genomic techniques, some high-quality, binary protein-protein interactions (PPIs) maps of *E. coli* have been released. Although the construction and understanding of these networks are far from complete, the protein complexes and functional modules predicted based on them are useful complements to the experimental methods.

Protein complex is formed by the interaction of more than two functional related peptide chains through disulfide bonds or other proteins, which performs some given biological functions. Functional module is basic functional unit of proteins, which implying the complex relationships involving multiple biological interaction types [8]. It is an important topic for us to understand the basic biological functions of protein by unveiling protein complexes and functional modules from PPI network of *E. coli*. Some studies have been performed to predict protein complexes based on PPI networks of *E. coli* [5, 8, 9], still some works focused on the functional relations about transcriptional regulatory [10–13] and metabolic pathways [13–15] of *E. coli*. However, few jobs involved in predicting functional modules of *E. coli*, especially comparing and analyzing protein complexes and functional modules simultaneously.

In recent years, the development of high-throughput sequencing technologies for pairwise protein interactions (such as two hybrid systems and mass spectrometry) makes it possible to construct PPI networks of *E. coli* at the genomic level. In this study, the protein complexes and functional modules of two high-quality PPI networks of *E. coli* are predicted by ELPA method [16]. According to the gold standard protein complexes and function annotations provided by EcoCyc dataset, the experimental results show that most topological modules predicted by ELPA match well with real protein complexes, cellular processes or biological functions. For example, in the PPI network provided by Hu et al., 75.8% of predicted topological modules match with one or more real protein complexes of *E. coli*, 88.1% of real protein complexes match with one or more topological modules, and 88.3% of predicted topological modules match with at least one functional unit of *E. coli*. On the same PPI network of *E. coli*, we compared the results of ELPA with Markov Clustering algorithm (MCL, the most popular protein complexes detecting method) [17], and found that ELPA outperforms MCL in terms of matching with gold standard complexes. At last, we focused on comparative analysis protein complex with functional module in the same topological module. The results show that the protein complexes and corresponding functional modules in most topological modules are highly overlapping, and the function of protein complexes are basically consistent with the corresponding functional modules. As a consequence, we surmise that ELPA may serve as an effective method to predict protein complexes and functional modules in PPI networks of *E. coli*.

2. Materials And Methods

2.1. PPI Datasets of *E. coli*

In order to prove the validity of this method better, two different large-scale PPIs datasets of *Escherichia coli* are assembled in this study. One is combined interaction dataset (AKA 'Core-experimental') of *E. coli* released by Bacteriome.org [18], which integrates the classic PPI dataset of Hu et al. [5] and functional datasets [18]. It collected 7613 binary interactions between 2283 *E. coli* proteins, which were derived by large-scale tandem affinity purification followed by mass spectrometry (AP/MS) experiments. The other is a PPI dataset provided by Rajagopala et al. [7], which were performed by yeast-two-hybrid method (Y2H). It collected 3929 binary-Y2H interactions between 2039 *E. coli* proteins. In order to predict the protein complex more effectively, only those binary interactions, which associated with the proteins of

known complexes are considered. In the end, 3280 interactions between 1298 proteins were collected from the first PPI dataset, and 1602 interactions between 1144 proteins were collected from the second PPI dataset.

2.2. Protein complexes and functional classes

The last literature-curated protein complexes of E. Coli released by EcoCyc database [19] is taken as the “gold standard” protein complexes. As we all know, many protein complexes of E. Coli contain only two proteins, so those protein complexes that contain at least two proteins showed in E. Coli K-12 dataset are retained. Finally, 297 “gold standard” protein complexes were preserved as benchmark, 732 proteins included. Figure 1 shows the overlap between proteins involved in the E. Coli datasets of Hu et al., Rajagopala et al., and EcoCyc. In addition, the last protein functional annotations of EcoCyc Gene Ontology (GO) database are taken as the benchmark functional classes of E. Coli. From the functional annotations, 32355 protein annotations are singled out as the benchmark functional classes at last.

2.3. Clustering method

Many clustering algorithms for complex network have been developed in the past decade and most of these methods are based on node clustering. However, few clustering algorithms can be used in complex biological networks directly. In this study, a novel algorithm based on edge clustering of complex network, which named edge label propagation algorithm (ELPA) [16] was used to identify topological modules of E. Coli in the above two PPI networks. Relative to node clustering, edge clustering has the natural advantages of compatible node attributes and link attributes of complex network, and can reflect the network topology structure better. In the following, it will be compared with an excellent node clustering algorithm for further explanation. These protein topological modules detected by ELPA will be matched with “gold standard” protein complexes and functional annotations of EcoCyc to identify corresponding protein complexes and functional modules, respectively.

2.4 Evaluation metrics

Precision, recall and F-measure are three types of commonly used evaluation metrics, which are used to measure the quality of predicted complexes [20], and to define how well a predicted complex matches a real complex. Let N_p denotes the number of complexes predicted by a clustering method, and N_B denotes the number of real ones in the gold standard protein complexes. Let N_{PC} be the number of predicted complexes which match at least one real complex, and N_{BC} be the number of real ones that match at least one predicted complex. The three metrics are then defined as follows:

$$Precision = N_{PC} / N_p \quad 1$$

$$Recall = N_{BC} / N_B \quad 2$$

$$F = 2 \times (Precision \times Recall) / (Precision + Recall) \quad 3$$

The matching score (MS_{PB}) [20] between a predicted complex and the real one in EcoCyc dataset is defined as:

$$MS_{PB} = 2 \times Precision \times Recall$$

4

The set of true-positive predictions can be obtained by selecting predictions with matching score larger than a threshold.

3. Results

3.1. Identification of topological modules

Based on above two complexes-related PPIs, two corresponding PPI networks are constructed, which named as NetH (Network provided by Hu et al.) and NetR (Network provided by Rajagopala et al.), respectively. In this study, protein complexes and functional modules of E.coli will be analyzed based on the two PPI networks. 120 and 171 topological modules are predicted by ELPA in NetH and NetR, respectively. In complex biological networks, small and sparse modules have been proved to perform important biological functions, so the identification of small and sparse topological modules is equally important with the larger and dense ones. Besides large and dense modules, ELPA can uncover those small and sparse modules as well. Since there are many gold-standard protein complexes contain only two proteins, those topological modules which consist of only two proteins are retained. The size of predicted topological modules ranges from two to hundreds of proteins. Furthermore, we found that many topological modules detected by ELPA are overlapping with each other. This phenomenon fits well with real protein complexes and functional modules, which means some proteins involved in multiple complexes or functional modules. The study of overlapping proteins across different complexes or functional modules is also an important research topic.

3.2. Identification of protein complexes

The protein modules of NetH and NetR predicted by ELPA are matched with 297 gold standard benchmark complexes of EcoCyc dataset, respectively. In the NetH, 222 benchmark complexes (88.1%) matched 91 predicted protein modules (75.8%), which mean that each protein module matches one or more real complexes. Most of the benchmark complexes consist of no more than ten proteins, so the larger protein modules should contain multiple complexes. While in the NetR, 134 benchmark complexes (53.2%) matched 70 predicted protein modules (40.9%). Through comparison, it is found that the scale of PPI network has a great influence on matching quality. In order to evaluate the predicted complexes, the set of effective matching complexes were obtained by selecting whose matching score larger than a threshold. Then Precision, Recall and F-measure were used to evaluate the quality of each module based on these effective matching complexes, respectively. The results show that most protein complexes predicted by ELPA are matched well with corresponding real complexes in both NetH and NetR. For example, In the NetH, 50th protein module consists of 8 proteins, among them potF, potH and potI are three proteins in putrescine ABC transporter complex; potA, potB, potC and potD cover all the four proteins of putrescine/spermidine ABC transporter complex (showed in Fig. 2a). 54th protein module consists of 5

proteins, it fully covered by two complexes: ferrichrome transport system and ferric coprogen transport system. fhuA, fhuB, fhuC and fhuD are ferrichrome transport system proteins, while fhuB, fhuC, fhuD and fhuE are ferric coprogen transport system proteins. Figure 2b shows that fhuB, fhuC and fhuD are the common proteins of the two complexes. As Fig. 3 shows, in the NetR, among the 8 proteins of 76th protein module, ccmA, ccmC, ccmD and ccmE are Protoheme IX ABC transporter proteins, while ccmE, ccmF and ccmH are CcmEFGH holocytochrome synthetase proteins, and ccmE links the two complexes. The above analysis shows that ELPA can effectively predict protein complex implied in E. Coli PPI network.

3.3. Identification of functional module

To determine whether the predicted protein modules have biological significance, each topological module is analyzed with the gold-standard proteins functional annotations of EcoCyc dataset. The topological modules of NetH and NetR detected by EPLA are matching with benchmark functional annotations, respectively. If the majority of proteins (> 50%) of a predicted topological module covered by a single functional term, then it is defined as a significance functional module. In the NetH, most of the predicted modules (82.5%) are significance functional modules, among them about 30% match perfectly (fully covered by a single functional term). For example, as Fig. 4a shows, 24 in out of 25 proteins of 19th protein module are annotated by GO:0006810, Obviously, the functions of these proteins are similar; all the 7 proteins of 40th protein module are fully covered by GO:0005886, GO:0016020 and GO:0017004, respectively (showed in Fig. 4b). While in the NetR, 60.8% predicted modules are significance functional modules, and 24.7% of them are fully covered by a single functional term. For example, as Fig. 5 shows, malY, malT, fixB, ybdM, recA and aes are annotated by GO:0005515, aes, recA and yhfW are annotated by GO:0005737, and yhfW and pyrC are annotated by GO:0046872. Obviously, some proteins have more than one function. The above analysis shows that ELPA can effectively detect functional modules implied in E. Coli PPI network as well.

3.4. Comparative evaluations

Most of clustering methods of complex network are based on node clustering, among them MCL have been proven that superior to other methods in identifying the functional modules or protein complexes in most cases [21, 22]. ELPA is a novel edge clustering method, it considers both node and link attitude, and can reflect the network structure better [16, 23]. Next we will compare the clustering results of ELPA and MCL in the same PPI network, respectively. ELPA is a parameter free method, and MCL will take the default parameters.

In order to compare the performance of MCL and ELPA, three metrics: Precision, Recall and F-measure are used to evaluate the predicted quality of protein complexes. Figure 6a shows the comparisons of topological modules of NetH and corresponding protein complexes between the two methods. We observe that the accuracy of ELPA slightly superior to MCL. The value of precision, recall and F-measure of ELPA are 72.5%, 61.5% and 66.5%, while those of MCL are 55.1%, 65.5% and 59.9%, respectively.

Figure 6b shows the comparisons of topological modules of NetH and corresponding functional modules between the two methods. Effective and average matching rate of significance functional modules is utilized to evaluate the predicted quality of functional modules. The effective matching rate and average matching rate of ELPA are 82.5% and 70.9%, while that of MCL are 85.9% and 70% respectively. Similar results are got in NetR, as Fig. 7a shows, The value of precision, recall and F-measure of ELPA are 35.7%, 28.6% and 31.8%, while those of MCL are 24.3%, 32.1% and 27.7%, respectively. As Fig. 7b shows, the effective matching rate and average matching rate of ELPA are 60.8% and 74.3%, while that of MCL are 68.2% and 74.4%, respectively. The above results show that ELPA is an effective method to predict protein complexes and functional modules of E. Coli.

3.5. Comparative analysis of protein complexes and functional modules

PPIs can be divided into Permanent interactions and transient interactions. Permanent interactions are strong and stable, which give rise to protein complexes while the transient interactions vary with cellular processes and form functional modules. Therefore, comparative analysis of protein complexes and corresponding functional modules is of great scientific significance. For example, as Fig. 8a shows, the 21th protein module of NetH mainly consists of three real complexes: NADH: ubiquinone oxidoreductase I (10 out of 11 real proteins of this complex match with this module), hydrogenase 4 (5 proteins of this complex fully covered by this module) and formate hydrogenlyase complex (3 out of 5 real proteins of this complex match with this module). Moreover, GO:0055114 fully covers the 18 proteins of this module. Among them, hycE, hycF and hycG are part of formate hydrogenlyase complex, and the function annotations of them are hydrogenase 3, formate hydrogenlyase complex iron-sulfur protein, and hydrogenase 3 and formate hydrogenlyase complex-HycG subunit, respectively. Which means the function of the three proteins agrees with formate hydrogenlyase complex, and it also hints that hydrogenase 3 complex is related with formate hydrogenlyase complex. Hydrogenase 4 consists of hyfB, hyfD, hyfF, hyfG and hyfI. The function annotations of hyfB, hyfD, hyfF, hyfG and hyfI are hydrogenase 4-component B, D, F, and large, small subunit respectively. This is highly consistent with the function of hydrogenase 4 complex. The remaining ten proteins: nuoB, nuoC, nuoE, nuoF, nuoG, nuoH, nuoI, nuoL, nuoM and nuoN are all related with the function of NADH: ubiquinone oxidoreductase complex. Figure 8b shows the similar results in NetR, such as in the 52th protein module, malG, malE, malF and malK are part of maltose ABC transporter complex (4 out of 5 real proteins of this complex match with this module), and all of them matching with GO:0015768, GO:0042956 and GO:0043190. The function annotations of malE, malG/malF, and malK are maltose ABC transporter-periplasmic binding protein, maltose ABC transporter-membrane subunit, and maltose ABC transporter-ATP binding subunit respectively. Above results show that protein complexes should be highly related with corresponding functional modules.

4. Conclusion

Most of researches on protein complexes or functional modules are based on experimental methods. For a long time, due to the lack of the support of large-scale PPI dataset, and the existence of too many orphan proteins, making it difficult to predict protein complexes or functional modules based on network model. Some high-throughput PPI datasets of E. Coli are released [7, 18, 19] in recent years, which make it possible to predict protein complexes and functional modules of E. Coli on large-scale based on computational methods. In this study, we focus on identifying and analyzing protein complexes and functional modules from two large-scale PPI networks of E. Coli. Now, we all know that clustering analysis method of PPI network is a useful supplement to the experimental methods.

Node clustering and edge clustering are two different types of methods to uncover network structure from different perspective. Edge clustering has a natural advantage over node clustering in network community detection. Based on the edge clustering method (ELPA), some interesting protein complexes and functional modules are identified. Furthermore, some comparative analysis are performed to investigate corresponding protein complexes and functional modules, which related with the same topological modules. This helps to understand the dynamic relationship between protein complexes and functional modules. In addition, the results of ELPA are compared with that of MCL method, and it performs better than MCL in most cases. Our conclusion is that ELPA may be used as an effective cluster analysis tool for different types of biological networks.

Abbreviations

E. Coli: Escherichia coli

PPI: protein-protein interaction

ELPA: edge label propagation algorithm

MCL: Markov clustering algorithm

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: Not applicable

Funding: This research was funded by the Talent Foundation of Ludong University, grant number LA2016007; Natural Science Foundation of Jiangsu, grant number BK2016245; and Natural Science Foundation of Shandong, grant number ZR2017MF052 and National Natural Science Foundation of China, grant number 81830052&81830053;and Shanghai Key Laboratory of Molecular Imaging, grant number 18DZ2260400

Competing Interest: The authors declare that they have no competing interests.

Author Contributions: methodology, validation, formal analysis and writing—original draft preparation, W.L.; writing, review, visualization and editing, P.K. All authors have read and approved the manuscript.

Acknowledgments: We thank those researchers who have graciously shared the *E. Coli* PPIs with us.

References

1. Arifuzzaman M., Maeda M., Itoh A., *et al.* Largescale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res.*, **2006**, vol 16, pp. 686–691.
2. Butland G., Peregrin-Alvarez J.M., Li J., *et al.* Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, **2005**, vol. 433, pp. 531–537.
3. Joyce A.R., Reed J.L., White A., *et al.* Experimental and computational assessment of conditionally essential genes in Escherichia coli. *Bacteriol.*, **2006**, vol. 188, pp. 8259–8271.
4. Riley M., Abe T., Arnaud M.B., *et al.* Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. *Nucleic Acids Res.*, **2006**, vol. 34, pp. 1–9.
5. Hu P., Janga S.C., Babu M., *et al.* Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. *PLoS Biol.*, **2009**, Vol. 7(4), pp. 929-947.
6. Butland G., Joyce L., Wehong Y., *et al.*, Interaction network containing conserved and essential protein complex in Escherichia coli. *Nature*, **2005**, vol. 433, pp. 531-537.
7. Seesandra V., Rajagopala *et al.*, The binary protein-protein interaction landscape of Echerichia coli. *Biotechnol.*, **2014**, Vol. 32, pp. 285-290.
8. Victor Spirin and Leonid A. Mirny. Protein complexes and functional modules in molecular networks. *Natl. Acad. Sci. USA*, **2003**, vol. 100 (21), 12123-12128.
9. Lei Shi, Xiujuan Lei, Aidong Zhang. Protein complex detection with semi-supervised learning in protein interaction networks. *Proteome science*, **2011**, vol 9 (suppl 1), S5.
10. Osbaldo R.A., Julio A., Ricardo M.M., *et al.* Modular analysis of the transcriptional regulatory network of E. Coli. *TRENDS in Gentetics*, **2005**, vol. 21 (1), pp. 16-20.
11. Faith J.J., Hayete B., Thaden J.T., *et al.* Largescale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **2007**, vol. 5, e8.
12. Barrett C.L., Herring C.D., Reed J.L., Palsson B.O. The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states. *Natl. Acad. Sci. USA*, **2005**, vol. 102, pp. 19103–19108.
13. Gama-Castro S., Jimenez-Jacinto V., Peralta-Gil M., *et al.* RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **2008**, vol. 36, pp. D120–124.
14. Feist A.M., Henry C.S., Reed J.L., *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol.*, **2007**,

vol. 3, pp. 121.

15. Geryk J., Sianina F. Modules in the metabolic network of E. Coli with regulatory interactions. *J. Data Min. Bioinform.*, **2013**, vol. 8(2), 188-202.
16. Liu W., Jiang X., Pellegrini M., *et al.* Discovering communities in complex networks by edge label propagation. *Scientific Reports*, **2016**, 6:22470.
17. Enright A.J., Van Dongen S., Ouzounis C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **2002**, vol. 30, pp. 1575–1584.
18. Peregrin-Alvarez J.M., Xiong X., Su C. *et al.* The modular organization of protein interactions in Echerichia coli. *PLoS comput. Biol.*, **2009**, vol. 5 (10), e1000523.
19. Keseler I.M., Mackie A., Peraltagil M., *et al.*, EcoCyc: fusing model organism databases with systems biology. *Nucl. Acid. Res.*, **2013**, vol 41, pp. D605-612.
20. Chen B., Fan W., Liu J., *et al.* Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks. *Briefings in Bioinformatics*, **2014**, vol 15, pp. 177-194.
21. Brohee S., van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, **2006**, vol. 7, pp. 488.
22. Reid A.J., Ranea J.A., Orengo C.A. Comparative evolutionary analysis of protein complexes in E. Coli and yeast. *BMC Genomics*, **2010**, vol 11, pp. 79-0.
23. Liu W., Wu A., Uncover protein complexes in E.coli network. 2015 IEEE International Conference on Bioinformatics and Biomedicine, IEEE, 2015.

Figures

Figure 1

Overlaps among sets of proteins of E. Coli that are involved in Hu et al., Rajagopala et al. and EcoCyc benchmark datasets.

(a)

(b)

Figure 2

Matches between predicted modules and EcoCyc benchmark complexes in the NetH. (a) The 50th predicted module matching with real complexes. (b) The 54th predicted module matching with real complexes.

Figure 3

The 76th predicted module in the NetR matching with EcoCyc benchmark complexes.

Figure 5

The 10th predicted module in the NetR matching with benchmark GO classifications.

(a)

(b)

Figure 6

The comparative analysis between ELPA and MCL in the NetH. (a) Accuracy comparison based on three evaluation metrics. (b) Frequency comparison based on effective matching rate and average matching rate.

Figure 7

The comparative analysis between ELPA and MCL in the NetR. (a) Accuracy comparison based on three evaluation metrics. (b) Frequency comparison based on effective matching rate and average matching rate.

Figure 8

The comparative analysis between protein complexes and functional modules. (a) The 21th predicted module of NetH covered by three real complexes and fully annotated by GO:0055114. Proteins functions in this module are all related with above three real complexes. (b) The 52th predicted module of NetR mainly covered by maltose ABC transporter complex and fully annotated by three functions, and proteins functions in this complex are all related with maltose ABC transporter.