Essential Proteins Identification by Improved Node and Edge Clustering Coefficient Integrated with Biological Information

Zhang Houwang Zhu Yuan Wu Chong

Abstract Computational methods are often applied to identify essential proteins from protein interaction networks. In this paper, we propose an improved node and edge clustering coefficient which has a better combination of dual characteristics of node and edge compared to node and edge clustering coefficient (NEC). It is called new node and edge clustering coefficient (NNEC) and is applied for essential protein identification. In order to further improve the performance of NNEC, we propose a version of NNEC which considers biological information to reduce the effect of false positives of PPI data. Here, we integrate NNEC with gene expression information, functional similarity, and protein-protein sequence similarity, we call the new method INNEC. We apply INNEC method to PPI data of *Saccharomyces cerevisiae* (Yeast) and *Escherichia coli* (E. coli), then we compare it with some state-of-the-art methods(DC, NC, PeC and LAC). The experimental results show that our proposed methods achieve better results in terms of prediction accuracy, AUC of PR-curve, and Jackknife methodology.

Key words Protein-Protein Interaction, Essential Protein, PPI Network, Biological Information

1 Introduction

In cellular life activities, proteins are indispensable components. They perform varied functions like DNA replication, catalyzing metabolic reactions, and transporting molecules in living organisms [1]. And among them, there is a kind of proteins called essential proteins. They are the most important. Living organisms will die or be infertile without them [2, 3]. Some studies have found that some essential proteins are related to human disease genes [4]. Therefore, the study of identification of essential proteins is very necessary. With the development of high-throughput technologies, a lot of protein-protein interactions (PPIs) have been accumulated [5]. And it makes using computational methods to study the identification of essential proteins become possible. Generally, protein-protein interactions are constructed to an undirected network which is called protein interaction network (PIN). Most of computational methods of studying identification of essential proteins are focusing on using the topological characteristics of PIN to identify essential proteins and they are called network-based methods.

Until now, a lot of studies have proposed some efficient network-based methods to identify essential proteins from PIN. Among them, the most well-known and simplest network-based essential protein search method is degree centrality (DC) which is known for centrality-lethality rule [6]. Some studies have confirmed that proteins which have large degree tend to be essential proteins [6, 7]. Like DC, other centrality measures have been also used to identify essential proteins, such as subgraph centrality (SC) [8], eigenvector centrality (EC) [9], betweenness centrality (BC) [10], closeness centrality (CC) [11], information centrality (IC) [12], and *etc.* Additionally, some studies have proposed a few of edge-aided methods [13–16] to identify essential proteins from PIN, such as LAC [14] and edge clustering coefficient [15]. Li *et al.*[16] improved its performance and proposed an improved edge clustering coefficient (NC). And recently, some studies have found that integrating some biological information into edge-aided methods will reduce the effect of false positives of PPI data and highly improve the prediction accuracy of essential proteins [17, 18]. Li *et al.*[18] integrated gene expression data into NC and enhanced its performance (PeC). However, above methods like DC, CC, and *etc.*, they just focus on the importance of nodes in the network but ignore the importance of edges which are functioned as bridges between the connecting nodes [19]. And edge-aided methods like ECC, NC, PeC, and *etc.*, they have not combined the dual characteristics of node and edge effectively and do not consider the advantages of each algorithm [19]. Hence, Jiang *et al.*[19] proposed an idea of combining clustering coefficient with edge clustering coefficient (NEC). However, NEC gives the same weight to topological characteristics of node and edge which causes NEC get a not good combination of dual characteristics of node and edge and give low importance to some vital nodes of high degree. Hence, the performance of NEC is much affected by this point and in some cases, it is inferior to NC. What is more, NEC also does not consider the effect of false positives of PPI data.

In this paper, we revise the defect of NEC and propose a new centrality method called new node and edge clustering coefficient (NNEC) which has a better combination of dual characteristics of node and edge. And in order to reduce the effect of false positives of PPI data and further improve the performance of NNEC, we propose a version of NNEC called INNEC which integrates biological information(gene expression information, functional similarity, and protein-protein sequence similarity). Compared to other state-of-the-art methods, INNEC both considers the topological characteristics of node and edge. What is more, it also considers the effect of false positives of PPI data. We apply it on the publicly-available PPI data of *Saccharomyces cerevisiae* (Yeast) and *Escherichia coli* (E. coli). We compare the performance of INNEC method with DC, NC, PeC, and LAC. The experimental results show that our proposed INNEC achieve better results than other state-of-the-art methods used in this paper.

2 Material and Methods

2.1 Data Source

2.1.1 The Protein-Protein Interaction Network

The PPI data of *Saccharomyces cerevisiae* (Yeast) used in this paper are downloaded from the DIP [20] database (release of Oct.18th, 2012). The PPI dataset contains 4,979 proteins and 22,061 interactions.

Equally, we downloaded the PPI data of *Escherichia coli* (E. coli) from the DIP [20] database. But as Wang *et al.*[21] said, over 1,000 protein-protein interactions are either single-pair interactions or part of small and unconnected networks with fewer than five nodes, and Wang *et al.* removed them from analyses, so we finally took their

data. The E. coli PPI network consists of 2,528 proteins and 11,496 interactions.

2.1.2 Essential Proteins List

The essential genes list of *Saccharomyces cerevisiae* (Yeast) used in this paper is collected from the OGEE [22]. It classifies all genes into three categories: essential, nonessential, and conditional. In this paper, we refer to paper [13] and consider conditional genes as essential. The Yeast network consists of 1,209 essential proteins, 3,322 nonessential proteins, and 448 unknown proteins. The unknown proteins are collected from DIP.

The essential gennes list of *Escherichia coli* (E. coli) used in this paper is also downloaded from the OGEE. In the E. coli network, 444 are essential, 1,403 are nonessential, and 671 unknown ones.

2.1.3 Gene Expression Data

The gene expression dataset of *Saccharomyces cerevisiae* (Yeast): GSE3431 [23] is downloaded from GEO [24], which samples 12 time points during each of three Yeast successive metabolic cycles (the interval between two time points is 25 minutes). The dataset contains 36 samples with 6,777 genes. Among them, 4,858 are involved in the previous PPI network of Yeast.

E. coli gene expression data GSE6425 [25] was also downloaded from GEO, which has expression data of two E. coli strains, MG1655 and UTI89, harvested at multiple time points during aerobic or anaerobic growth in Luria-Bertani medium. Here, we used the MG1655 data which contains 22 samples with 4,345 genes.

2.1.4 Gene Expression Similarity

We use the Pearson correlation coefficient (PCC) to evaluate the gene expression similarity of two interacting proteins [18]. The PCC value of two interacting proteins x and y is calculated as follow,

$$PCC(x,y) = \frac{1}{a-1}\sum_{i=1}^{a}(\frac{X_i - \hat{X}}{\sigma X})(\frac{Y_i - \hat{Y}}{\sigma Y}). \quad (1)$$

where X and Y are the gene expression levels of proteins x and y, respectively, a is the number of samples in the gene expression data, X_i and Y_i are the expression levels of genes X and Y in the sample i under a specific condition, respectively, σX and σY represent the standard deviations, $\hat{X}$ and $\hat{Y}$ represent the mean values.

2.1.5 Functional Similarity

Gene ontology (GO) [26] is widely used to represent the relationships between genes and biological terms. GO semantic similarity is based on the biological characteristics of genes. It is used to represent genes functional similarity [27]. There is an assumption that two essential proteins which interact with each other are more likely to perform functions in the same biological processes [28], in other words, two proteins of high functional similarity are more likely to interact with each other. In this paper, using biological process category of GO, genes functional similarity (GFS) between two proteins is calculated using the algorithm proposed in paper [29], the formula is as follows,

$$\begin{aligned} GFS(x,y) &= max\frac{2*\ln P_{ms}(t_x,t_y)}{\ln P(t_x)+\ln P(t_y)},\\ P_{ms}(t_x,t_y) &= \min_{t\in F(t_x,t_y)} P(t). \end{aligned} \quad (2)$$

where, t_x and t_y are the terms of proteins x and y respectively, $F(t_x,t_y)$ is the set of common ancestors of terms t_x and t_y, and $P(t)$ is the probability of encountering an instance of term t.

2.2 Methods

2.2.1 Node and Edge Clustering Coefficient

Jiang *et al.*[19] have proposed a centrality method called node and edge clustering coefficient (NEC) which combines the clustering coefficient and the edge clustering coefficient. The formula of NEC is as follows,

$$\begin{aligned} NEC(x) &= \sum_{y\in N_x} IECC(x,y),\\ IECC(x,y) &= \frac{NTE(x,y)^2 * \prod_{i=\{x,y\}} C(i)}{\prod_{i=\{x,y\}}(d(i)-1)},\\ C(i) &= \frac{2E_i}{d(i)(d(i)-1)}. \end{aligned} \quad (3)$$

where, $NTE(x,y)$ is the number of triangles consist of edge (x,y), x and y represent proteins x and y respectively, N_x is the set of nearest neighbors of protein x, $C(i)$ is the clustering coefficient of protein i, E_i is the number of non-repetitive edges consist of all nearest neighbors of protein i, and $d(i)$ is the degree of protein i. When NEC is applied to a network which has some one degree nodes, the $d(i)-1$ in above equations should be replaced by $d(i)$.

However, from above equations, we can see that NEC gives equal weight to topological characteristics of node and edge which causes the combination of dual characteristics of node and edge is not so good that NEC gives low importance to some vital nodes of high degree. What is more, other researchers' previous studies have shown that the topological characteristics of edge in the network is more useful than node to help us search the essential proteins [13, 15, 16]. Hence, the weight of topological characteristics of edge should be bigger than the weight of topological characteristics of node.

2.2.2 New Node and Edge Clustering Coefficient

In this paper, we propose a new centrality measure called new node and edge clustering coefficient (NNEC) which has a better combination of dual characteristics of node and edge compared to NEC. And in our generalization test, it is more efficient in searching vital nodes in a random network than NEC. Considering the length of article constraints, we do not show this experiment in this paper. The formula of NNEC is as follows,

$$\begin{aligned} NNEC(x) &= \sum_{y\in N_x} NECC(x,y),\\ NECC(x,y) &= \frac{NTE(x,y)^3 * C(y)}{\prod_{i=\{x,y\}}(d(i)-1)}. \end{aligned} \quad (4)$$

2.2.3 Improved New Node and Edge Clustering Coefficient

In order to further improve the performance of NNEC and inspected by paper [18], in this paper, we integrate biological information into NNEC to reduce the effect of false positives of PPI data and propose an improved version of new node and edge clustering coefficient (INNEC). And the formula of INNEC is as follow,

$$INNEC(x) = NNEC(x)\prod_{j=1}^{n} Bio(j). \quad (5)$$

where, $Bio(j)$ is the biological information measure j, and n is the number of biological information measure. When

$n = 0$, INNEC is changed to NNEC. Hence NNEC is a special case of INNEC. A protein has a high score of INNEC value is more likely to be an essential protein.

As we can see, INNEC is an open framework which can integrate different biological information using different n. Here, we select gene expression similarity, functional similarity and protein-protein sequence similarity in this paper. Inspected by paper [18], they integrated gene expression similarity to NC and improved its performance, so we want to verify whether we can improve the performance of NNEC through integrating more kinds of biological information. And through our experiments, the results show INNEC can get the highest prediction accuracy in the high ranking score interval compared with some state-of-the-art methods(DC, NC, PeC and LAC). Hence, INNEC is proposed and implemented.

2.3 Performance Evaluation

We compare INNEC method with DC, NC, PeC, and LAC. Some studies have found that edge-aided methods like NC and PeC perform better than the centrality-based methods previously published (BC, CC, DC, EC, IC, SC, and *etc.*) [13]. Hence, we only choose the DC as the control method to represent previously published centrality-based methods. To verify the performance of INNEC, we select two performance evaluation measures: Jackknife curve and precision-recall curve.

2.3.1 Jackknife Curve

The first performance evaluation measure we select is Jackknife curve [30]. Jackknife curve is used to represent the number of samples that are correctly predicted among a top ranked prediction list. In a 2D Jackknife curve, the x-axis represents the number of essential protein candidates sorted in a descending order. And the y-axis represents the number of true essential proteins in the essential protein candidates.

2.3.2 Precision-recall Curve (PR-curve)

The second performance evaluation measure used in this paper is precision-recall curve (PR-curve). It can be calculated as follows,

$$\begin{aligned} Pre &= \frac{TP}{TP+FP}, \\ Re &= \frac{TP}{TP+FN}. \end{aligned} \tag{6}$$

where, TP is the number of true positive, FP is the number of false positive, and FN is the number of false negative. The larger area under the curve (AUC) means better performance of the method.

3 Experiments and Results

We implemented INNEC and all other comparetive methods. Then, we applied them on the PPI network of Yeast and E.coli. As most of validation methods used in the experiments of identification of essential proteins, we also ranked all proteins by using each essential protein search method and selected a certain number of top ranked proteins as essential protein candidates (top 24%). Fig.1 shows the PR-curves got on the Yeast dataset of all methods. Fig.2 shows the PR-curves got on the E.coli dataset of all methods. From Fig.1 and Fig.2 we can see that, compared to other methods, INNEC achieves the largest area under the curve (AUC). What is more, its AUC value is much higher than PeC which has confirmed the efficiency of INNEC. From Fig.1, we can see an interesting phenomenon.

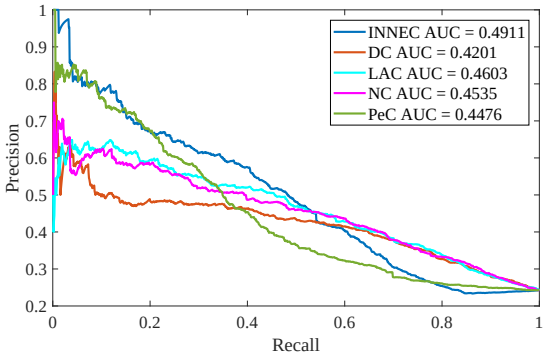

Figure 1 The PR-curves obtained by INNEC, DC, LAC, NC, PeC on the Yeast dataset.

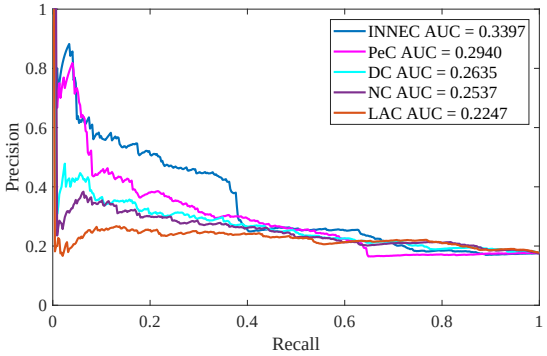

Figure 2 The PR-curves obtained by INNEC, DC, LAC, NC, PeC on the E.coli dataset.

Compared to the original NC, the AUC value of PeC has a little decline. However, PeC achieves much higher AUC value in the interval of $0 < Recall < 0.4$. And this means that it can identify more true essential proteins in the interval of high ranking score. Fig.3 and Fig.4 shows the Jackknife curves obtained by all methods. It can be seen clearly that INNEC method achieve much better results than rest methods. It has further proved the efficiency of INNEC . From Fig.5, we can see that INNEC identifies more true essential proteins in top 200, 400, 500, and 600 essential protein candidates than other methods. Although PeC identifies more true essential proteins in top 100 and 300 essential protein candidates than INNEC, there is no big difference between PeC and INNEC. From Fig.6, we can see that INNEC identifies more true essential proteins in top 100, 200, 300, 400, 500, and 600 essential protein candidates than all other methods. Considering the difference of the two datasets, the dataset of E.coli has higher false positives than the dataset of Yeast. INNEC gets better results than PeC in the dataset of E.coli also illustrates that INNEC is more rubust when the datasset has false positives.

4 Conclusion

In this paper, we propose an improved node and edge clustering coefficient called new node and edge clustering coefficient (NNEC). Compared to NEC, NNEC has a better combination of dual characteristics of node and edge. In order to further improve the performance of NNEC, we in-

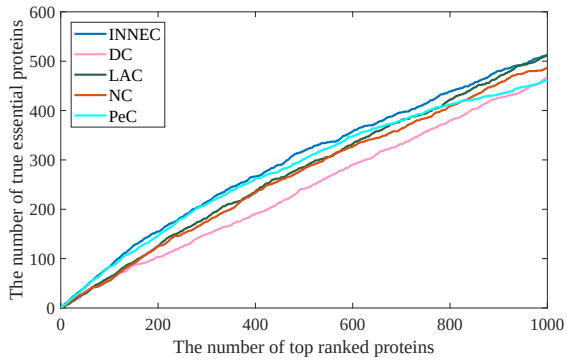

Figure 3 The Jackknife curves obtained by by INNEC, DC, LAC, NC, PeC on the Yeast dataset.

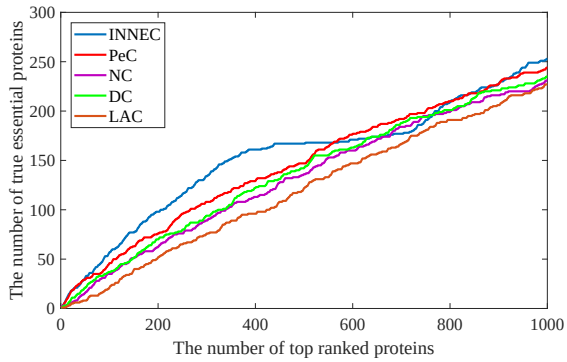

Figure 4 The Jackknife curves obtained by by INNEC, DC, LAC, NC, PeC on the E.coli dataset.

tegrate biological information(gene expression information, functional similarity, and protein-protein sequence similarity) into NNEC to reduce the effect of false positives of PPI data and propose an improved version of NNEC called INNEC. Through the comparative experiments on the PPI dataset of Yeast and E.coli, INNEC proposed in this paper achieve some interesting results compared to DC, NC, PeC and LAC.

References

1 B. Xu, J. Guan, Y. Wang, and Z. Wang, "Essential protein detection by random walk on weighted protein-protein interaction networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2017.

2 R. S. Kamath, A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann *et al.*, "Systematic functional analysis of the caenorhabditis elegans genome using RNAi," *Nature*, vol. 421, no. 6920, p. 231, 2003.

3 M. L. Acencio and N. Lemke, "Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information," *BMC bioinformatics*, vol. 10, no. 1, p. 290, 2009.

4 S. J. Furney, M. M. Albà, and N. López-Bigas, "Differences in the evolutionary history of disease genes affected by dominant or recessive mutations," *BMC genomics*, vol. 7, no. 1, p. 165, 2006.

5 M. Li, P. Ni, X. Chen, J. Wang, F. Wu, and Y. Pan, "Construction of refined protein interaction network for predicting essential proteins," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2017.

6 H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, p. 41, 2001.

7 X. He and J. Zhang, "Why do hubs tend to be essential in protein networks?" *PLoS genetics*, vol. 2, no. 6, p. e88, 2006.

8 E. Estrada and J. A. Rodriguez-Velazquez, "Subgraph centrality in complex networks," *Physical Review E*, vol. 71, no. 5, p. 056103, 2005.

9 P. Bonacich, "Power and centrality: A family of measures," *American journal of sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

10 M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, "High-betweenness proteins in the yeast protein interaction network," *BioMed Research International*, vol. 2005, no. 2, pp. 96–103, 2005.

11 S. Wuchty and P. F. Stadler, "Centers of complex networks," *Journal of Theoretical Biology*, vol. 223, no. 1, pp. 45–53, 2003.

12 K. Stephenson and M. Zelen, "Rethinking centrality: Methods and examples," *Social networks*, vol. 11, no. 1, pp. 1–37, 1989.

13 Y. Wang, H. Sun, W. Du, E. Blanzieri, G. Viero, Y. Xu, and Y. Liang, "Identification of essential proteins based on ranking edge-weights in protein-protein interaction networks," *PloS one*, vol. 9, no. 9, p. e108716, 2014.

14 M. Li, J. Wang, X. Chen, H. Wang, and Y. Pan, "A local average connectivity-based method for identifying essential proteins from the network level." *Computational Biology & Chemistry*, vol. 35, no. 3, pp. 143–150, 2011.

15 F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004.

16 J. Wang, M. Li, H. Wang, and Y. Pan, "Identification of essential proteins based on edge clustering coefficient," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 1070–1080, 2012.

17 M. Li, J. Wang, H. Wang, and Y. Pan, "Essential proteins discovery from weighted protein interaction networks," in *International Symposium on Bioinformatics Research and Applications*. Springer, 2010, pp. 89–100.

18 M. Li, H. Zhang, J.-x. Wang, and Y. Pan, "A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data," *BMC systems biology*, vol. 6, no. 1, p. 15, 2012.

19 Y. Q. Jiang, Y. Wang, G. S. Wang, G. Ou, C. Su, and L. Huang, "Essential protein identification by a bootstrap k-nearest neighbor method based on improved edge clustering coefficient," *Computational Intelligence in Industrial Application*, 2015.

20 I. Xenarios, L. Salwinski, 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, pp. 303–305, 2002.

21 W. Yan, H. Sun, D. Wei, B. Enrico, V. Gabriella, X. Ying, and Y. Liang, "Identification of essential proteins based on ranking edge-weights in protein-protein interaction networks," *Plos One*, vol. 9, no. 9, p. e108716, 2014.

22 W.-H. Chen, P. Minguez, M. J. Lercher, and P. Bork, "OGEE: an online gene essentiality database," *Nucleic acids research*, vol. 40, no. D1, pp. D901–D906, 2011.

23 B. P. Tu, A. Kudlicki, M. Rowicka, and S. L. McKnight, "Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes," *Science*, vol. 310, no. 5751, pp. 1152–1158, 2005.

24 R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 2002.

25 C. S. Reigstad, S. J. Hultgren, and J. I. Gordon, "Functional genomic studies of uropathogenic escherichia coli and host urothelial cells when intracellular bacterial communities are assembled," *Journal of Biological Chemistry*, vol. 282, no. 29, pp. 21 259–67, 2007.

26 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene Ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, p. 25, 2000.

27 Y. Jiang, Y. Wang, W. Pang, L. Chen, H. Sun, Y. Liang, and E. Blanzieri, "Essential protein identification based on essential protein–protein interaction prediction by integrated edge weights," *Methods*, vol. 83, pp. 51–62, 2015.

28 N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis *et al.*, "Global landscape of protein complexes in the yeast saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, p. 637, 2006.

29 D. Lin *et al.*, "An information-theoretic definition of similarity." in *Icml*, vol. 98, no. 1998. Citeseer, 1998, pp. 296–304.

30 A. G. Holman, P. J. Davis, J. M. Foster, C. K. Carlow, and S. Kumar, "Computational prediction of essential genes in an unculturable endosymbiotic bacterium, wolbachia of brugia malayi," *BMC microbiology*, vol. 9, no. 1, p. 243, 2009.

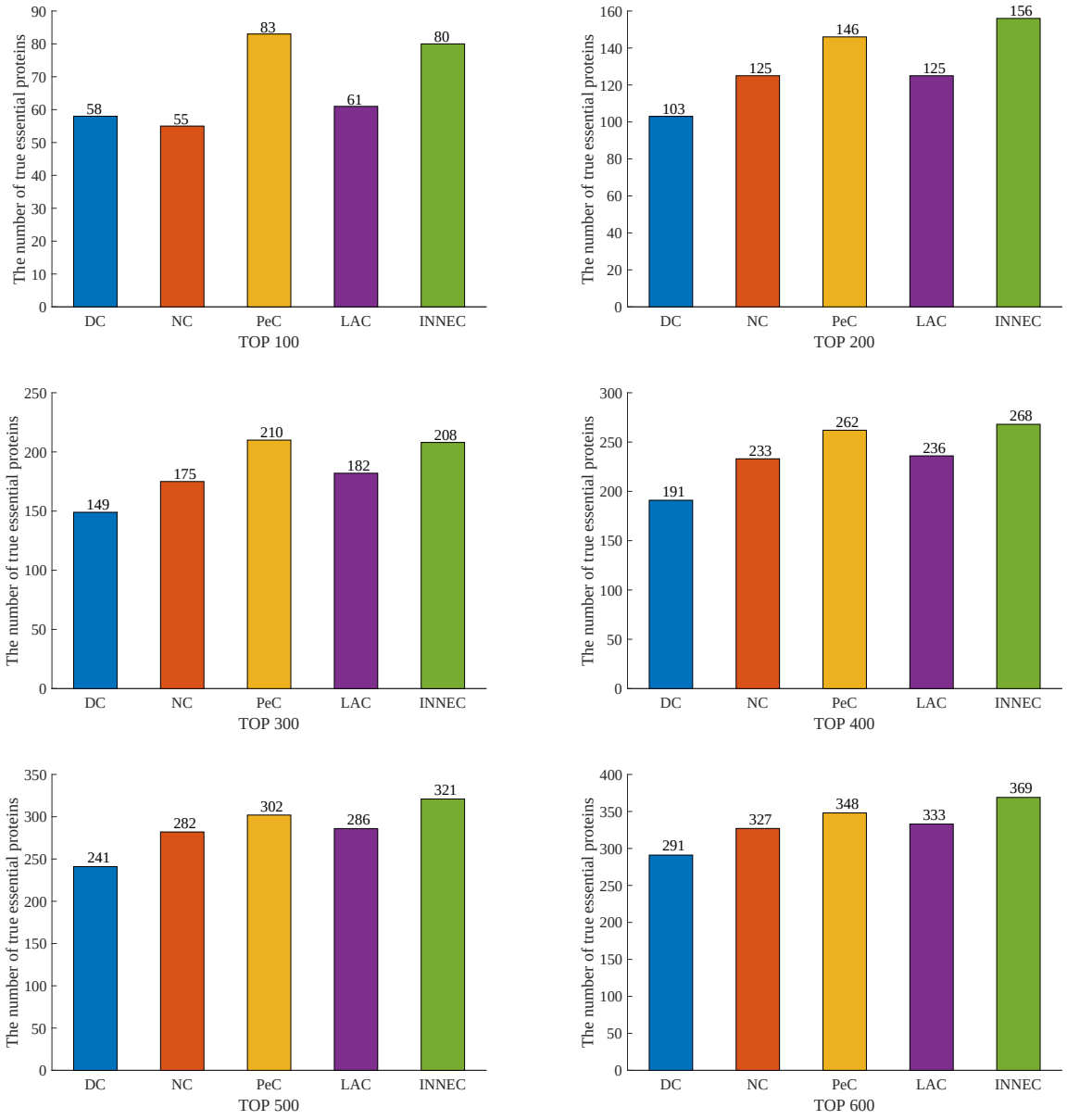

Figure 5 Comparison of the number of essential proteins identified from PIN of Yeast by using DC, NC, PeC, LAC, INNEC.

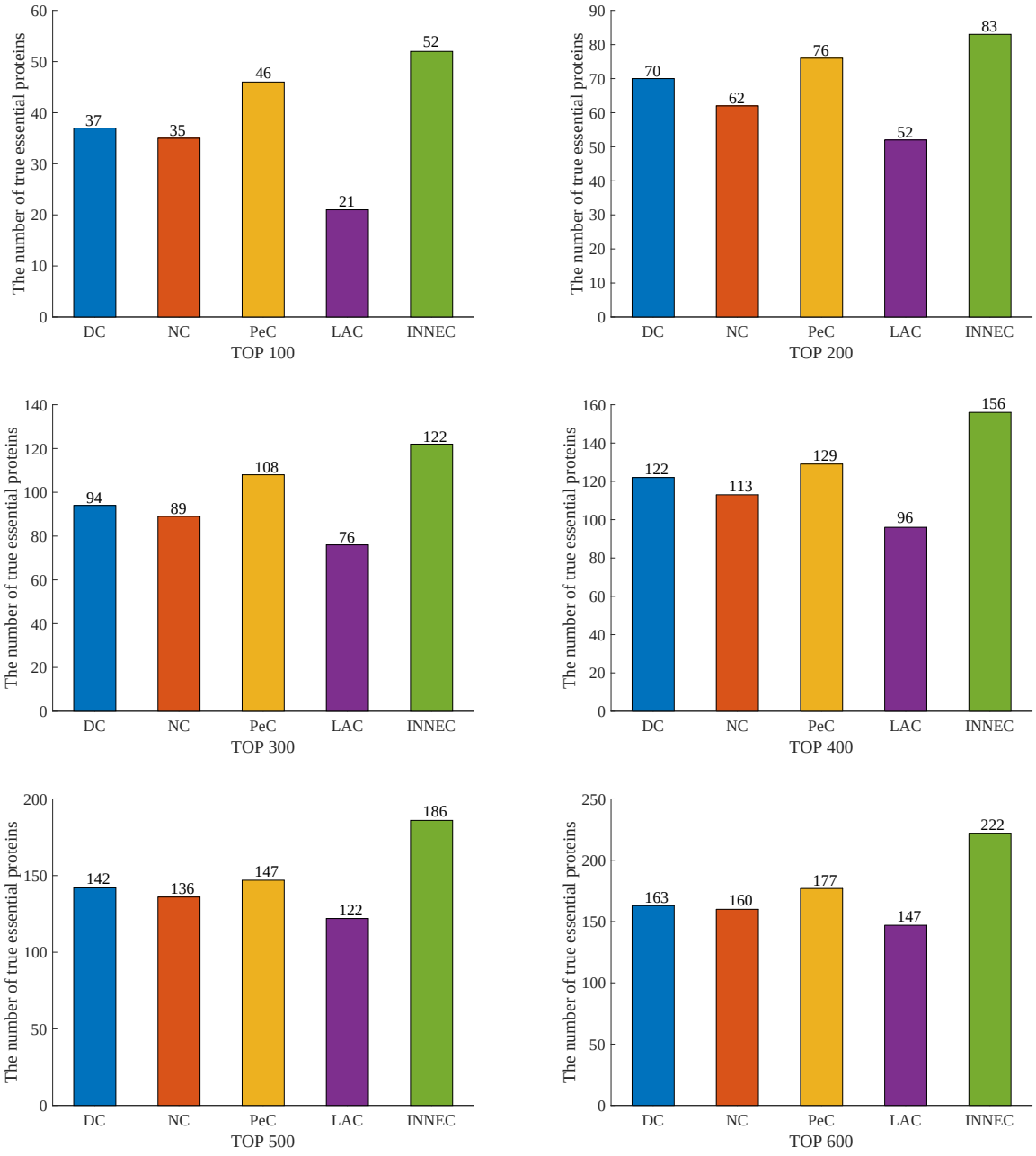

Figure 6 Comparison of the number of essential proteins identified from PIN of E.coli by using DC, NC, PeC, LAC, INNEC.