International Journal of Modern Physics B
Vol. 34, No. 10 (2020) 2050090 (14 pages)

DOI: 10.1142/S0217979220500903

A mixed clustering coefficient centrality for identifying essential proteins

Pengli Lu* and JingJuan Yu†

School of Computer and Communication,
Lanzhou University of Technology,
Lanzhou 730050, Gansu, P. R. China
**lupengli88@163.com*
†*yujingjuanmercy@163.com*

Received 5 December 2019
Accepted 11 February 2020
Published 22 April 2020

Essential protein plays a crucial role in the process of cell life. The identification of essential proteins not only promotes the development of drug target technology, but also contributes to the mechanism of biological evolution. There are plenty of scholars who pay attention to discover essential proteins according to the topological structure of protein network and biological information. The accuracy of protein recognition still demands to be improved. In this paper, we propose a method which integrates the clustering coefficient in protein complexes and topological properties to determine the essentiality of proteins. First, we give the definition of In-clustering coefficient (IC) to describe the properties of protein complexes. Then we propose a new method, complex edge and node clustering (CENC) coefficient, to identify essential proteins. Different Protein–Protein Interaction (PPI) networks of Saccharomyces cerevisiae, MIPS and DIP are used as experimental materials. Through some experiments of logistic regression model, the results show that the method of CENC can promote the ability of recognizing essential proteins by comparing with the existing methods DC, BC, EC, SC, LAC, NC and the recent UC method.

Keywords: Protein interaction network; essential protein; protein complex; assessment method.

PACS number: 87.14.Ee

1. Introduction

Protein is a crucial component of all cells and organizations. It is considered as essential proteins as the proteins are necessary to maintain the life of the organism. Not only can essential proteins promote the development of drug target technology, but also help in the study of biological evolution mechanism.[1] Removing the essential proteins can lead to cell death or its inability to replicate and reproduce.[2]

*Corresponding author.

The recognition and protection of essential proteins are the basis of drug development, which provide valuable theories and methods for the diagnosis of diseases, drug design, etc.[3]

In biology, the identification methods of essential proteins mainly rely on biological experiments, such as conditional knockouts,[4] RNA interference[5] and single gene knockouts,[6] coupled with the survival ability of infected organisms being tested. These biological experimental results are clear and effective, but consume large amounts of time, costs and resources. With the improvement of prior technology, several protein–protein interaction (PPI) networks are generated.[7,8] Nowadays, it has been a crucial research direction in the field of bioinformatics for predicting essential proteins from a large number of biological experiments by using the theory of technology from PPI networks. The methods for identifying essential proteins can be divided into several categories.

Based on the centrality–lethality rule which was put forward by Jeong *et al.*, the essentiality of proteins is associated with the topological structure in PPI networks.[9] Thus, a large number of scholars have proposed many indicators based on topological centrality.[12,16] Some of them considered the topology of nodes in networks, such as degree centrality (DC) which considers the connection nodes,[12,17,18] betweenness centrality (BC) which considers the global characteristic,[13,45] subgraph centrality (SC),[19] local average centrality (LAC),[20] eigenvector centrality (EC),[21] information centrality (IC)[27] and closeness centrality (CC),[15] and topology potential-based method (TP).[42] Some of them considered the topology of edges in networks, including edge clustering coefficient (ECC),[23] improved node and edge clustering (INEC) coefficient,[43] integrated edge weights (IECs)[24] and network centrality (NC).[25] CytoNCA, an app of Cytoscape for analyzing the centrality methods, have been a valuable tool to identify the essentiality of proteins.[26]

With the increase of high-throughput biological data, scholars have tried to combine with biology information to improve the accuracy of identifying essential proteins. Considering the functional annotations of genes, a weighted protein–protein network is constructed, by combining ECCs with gene expression data correlation coefficients, a method of PeC is proposed.[32] There is an esPOS method that uses the information of gene expression and subcellular localization.[29] SPP method is based on sub-network division and sequencing by integrating subcellular positioning.[14] Extended pareto optimality consensus (EPOC) model mixes neighborhood CC and orthology information together.[28] Go terms information is also used to predict essential proteins such as RSG method.[33]

There are some studies which recognize essential proteins from the perspective of complexes and functional modules. Hart *et al.* found that the essentiality is an attribute of the protein complex and the protein complexes often determine the essential proteins.[30] Li *et al.* proved that the frequency of the essential protein that occurs in the complex is higher than in the whole network.[29,47] Luo *et al.* proposed a method of (LIDC) combining the local interaction density and protein complexes for predicting essential proteins.[44] Li *et al.* proposed united complex centrality (UC)

which takes into account the frequency of protein appeared in the complex and edge properties.[31]

In this paper, considering the protein complexes information and topological properties, a new method of complex edge and node clustering (CENC) coefficient is proposed to identify essential proteins. To assess the quality of CENC method, different datasets of Saccharomyes cerevisiae, MIPS and DIP are applied. By the comparison of seven existing methods, containing DC, BC, EC, SC, LAC, NC and UC, the experimental results show that our method can be more effective in determining the essentiality of proteins than the existing measures.

2. New Centrality: CENC

An undirected simple graph $G(V, E)$ can be used to express a network of protein interaction. Proteins can be regarded as node set V of a network and the connections between two proteins can be regarded as edge set E. In this study, we present a new method of CENC coefficient to judge the essentiality of proteins by combining the features of protein complex and topology of nodes and edges. The basic considerations of CENC are as follows: (1) the essential proteins that appear in complexes can have more frequency and (2) both the topology of node and edge are important factors to affect the essentiality of proteins.

First, we present a classical method of clustering coefficient (C).[22]

$$C(v) = \frac{2E_v}{k_v(k_v - 1)}, \tag{2.1}$$

where E_v is the actual number of edges shared with local neighbors of node v, k_v is the degree of the node v.

Then a clustering coefficient of a node to an edge was generalized by Radicchi *et al.*[23] The ECC is defined as[23]

$$\mathrm{ECC}_{v,u} = \frac{z_{v,u}}{\min(k_v - 1, k_u - 1)}, \tag{2.2}$$

where $z_{v,u}$ is the number of triangles that includes the edge $e(v, u)$ in network. k_v and k_u are the degrees of node u and node v, respectively.

Based on the numbers of connection edges for a node and the clustering coefficient of each edge, the sum of ECCs NC is proposed[25] as

$$\mathrm{NC}(v) = \sum_{u \in N_v} \mathrm{ECC}(v, u), \tag{2.3}$$

where N_v denotes the set of all neighbors of node v.

Furthermore, we propose a new definition In-clustering coefficient (IC) which combines the feature in complexes.

$$\mathrm{IC}(v) = \sum_{i \in \mathrm{ComplexSet}(v)} C(v)_i. \tag{2.4}$$

A subset of protein complexes that contains protein v can be represented as ComplexSet(v), the value of $C(v)$ for the ith protein complex which belongs to ComplexSet(v) can be represented as $C(v)_i$.

Now, based on the above-described definition, we propose our new method CENC coefficient for estimating the essentiality of proteins as follows:

$$\mathrm{CENC}(v) = a * \frac{\mathrm{IC}(v)}{\mathrm{IC}_{\max}} + b * \frac{\mathrm{NC}(v)}{\mathrm{NC}_{\max}} + c * \frac{C(v)}{C_{\max}}, \tag{2.5}$$

where a, b, c are random factors ranging from 1 to 10. Under the amounts of experiments, we can get the best result of the method CENC when a, b and c are 10, 1 and 1, respectively.

3. Experimental Data and Assessment Methods

3.1. *Experimental data*

The experiment data are conducted from Saccharomyes cerevisiae, whose proteins are more complete. Two sets of PPI network data, namely MIPS[35] and DIP[34] are used. In the protein network, all self-interactions and repetitive interactions are deleted as a data preprocessing of these PPIs. Specific properties for these two networks are presented in Table 1. The MIPS network includes 4546 proteins and 12,319 interactions, whose clustering coefficient is about 0.0879. In the DIP network, there are 5093 proteins and 24,743 interactions, whose clustering coefficient is about 0.0973. The known essential proteins are derived from four databases: MIPS,[46] Saccharomyces Genome Database (SGD),[41] Saccharomyces Genome Deletion Project SGDP[4] and Database of Essential Genes (DEG).[35] The protein complex set is from CM270,[46] CM425,[37] CYC408 and CYC428 datasets[38,39] which can gain from Ref. 29, containing 745 protein complexes (including 2167 proteins).

Table 1. Data details of the two protein networks: DIP, MIPS.

Dataset	Proteins	Interactions	Average degree	Essential proteins	Clustering coefficient
MIPS	4546	12,319	5.42	1016	0.0879
DIP	5093	24,743	9.72	1167	0.0973

3.2. *Assessment methods*

According to the values of CENC, proteins are sorted in descending orders. First, some numbers of top proteins in sequence are selected as predictive essential proteins, then compared with the real essential proteins. This allows us to know the quantity of true essential proteins. Therefore, the sensitivity (SN), specificity (SP), F-measure (F) and accuracy (ACC), positive predictive value (PPV), negative predictive value (NPV) can be calculated.[36,37]

The following are the formulas for calculating these six statistical indicators:
Sensitivity:

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

Specificity:

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}.$$

Positive predictive value:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}.$$

Negative predictive value:

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}.$$

F-measure:

$$F = \frac{2 * \mathrm{SN} * \mathrm{PPV}}{\mathrm{SN} + \mathrm{PPV}}.$$

Accuracy:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N},$$

where TP stands for the quantity of true essential proteins which is correctly selected as essential proteins. FP is the quantity of nonessential proteins which is incorrectly selected as essential. TN is the quantity of nonessential proteins which is correctly selected as nonessential. FN is the quantity of essential proteins which is incorrectly selected as nonessential. P and N stand for the sum number of essential and nonessential proteins, respectively.

4. Results

4.1. *Comparison with other centrality measures*

We follow the principle of "sorting-screening" to evaluate the performance of CENC. Comparisons of CENC method with other seven previous measures — DC,[12] BC,[13,45] EC,[21] SC,[19] LAC,[20] NC,[25] UC[31] — are carried out in the MIPS and DIP datasets. To be specific, proteins are sorted in descending order on the basis of their values of CENC and other seven previous measures. Then predictive essential proteins are chosen according to the top 100, 200, 300, 400, 500 and 600 proteins. Finally, by comparing with the known essential proteins, the quantity of true essential proteins among these predictive essential proteins can be obtained. The experimental results of these measures are shown in Figs. 1 and 2.

From Fig. 1, the quantity of true essential proteins judged by CENC are 67, 124, 171, 209, 243 and 260 from the top 100 to the top 600, respectively, being the best among the eight methods in MIPS network. Although the UC method

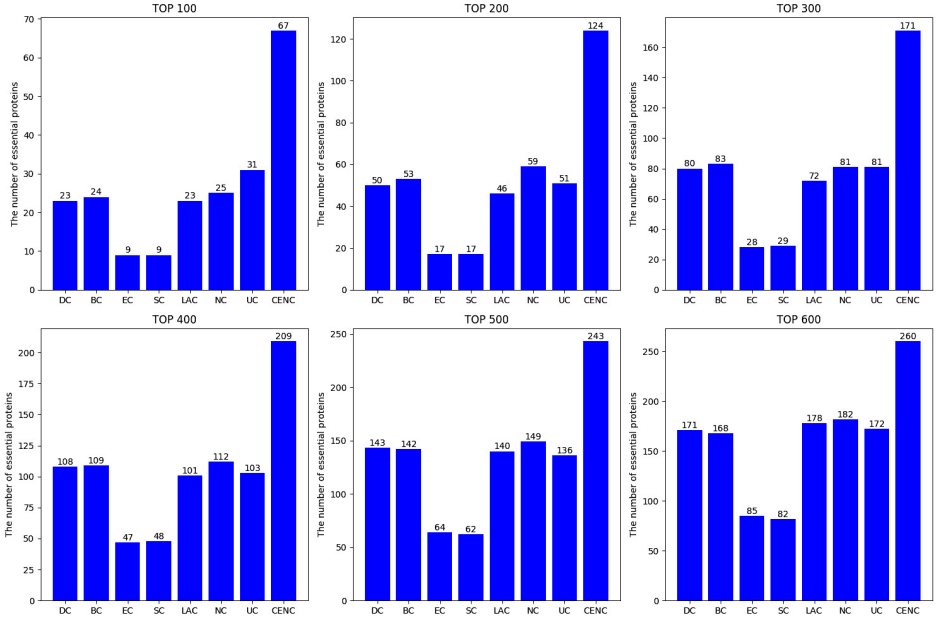

Fig. 1. The quantity of true essential proteins determined by CENC and other seven previous methods from the MIPS network.

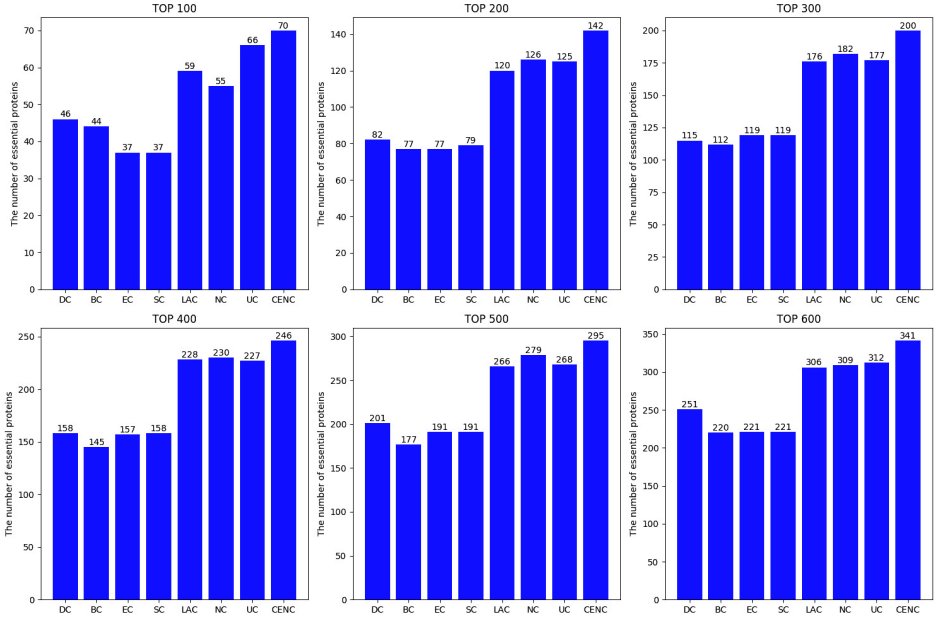

Fig. 2. The quantity of true essential proteins determined by CENC and other seven previous methods from the DIP network.

has good performance in the yeast PPI network, it is still poor in MIPS network. Among these seven proposed methods, SC is the lowest indicator of recognition of essential proteins. Compared to the SC method, our CENC method improves the rate of 86.56%, 86.29%, 83.04%, 77.03%, 74.49%, 68.46% in the top 100 to top 600, respectively. When we choose the best performance for each top, the CENC method can still obtain 53.73%, 52.42%, 51.46%, 46.41%, 38.68% and 30% improvements in predicting essential proteins.

From Fig. 2, it can be seen that the CENC method performs better than the existing methods of DC, BC, EC, SC, LAC, NC and UC in DIP network. Compared with the best result among these seven methods, the true essential proteins determined by CENC method are increased by 4, 16, 18, 16 and 29 from the top 100 to the top 600, respectively. Moreover, the quantity of essential proteins is much more than the previous methods including DC, BC, SC and EC.

4.2. *Evaluation of six statistical methods and the precision–recall curves*

The six statistical methods are used to evaluate the indicator of CENC as well as other seven identification measures, mentioned in Sec. 3.2. Proteins are sorted from high to low order on the basis of their values of these methods. Then the top proteins of 20% are taken into account as predictive essential proteins, the remaining 80% can be considered as candidates for nonessential proteins. On the two different networks, the comparisons among the values of CENC and other seven measures are executed as shown in Table 2. For DIP network, these six statistic values for CENC are higher than other previous measures, which show that CENC

Table 2. Comparing the results of sensitivity (SN), specificity (SP), F-measure (F), positive predictive value (PPV), negative predictive value (NPV) and accuracy (ACC) of CENC and other seven previous algorithms.

Dataset	Methods	SN	SP	PPV	NPV	F-measure	ACC
MIPS	DC	0.254	0.803	0.291	0.772	0.271	0.671
	BC	0.197	0.796	0.278	0.716	0.231	0.629
	EC	0.139	0.773	0.163	0.738	0.150	0.620
	SC	0.138	0.773	0.162	0.739	0.149	0.620
	LAC	0.271	0.812	0.314	0.779	0.291	0.682
	NC	0.281	0.814	0.325	0.781	0.302	0.686
	UC	0.271	0.812	0.314	0.778	0.291	0.682
	CENC	**0.317**	**0.827**	**0.368**	**0.792**	**0.341**	**0.704**
DIP	DC	0.353	0.834	0.409	0.80	0.379	0.716
	BC	0.308	0.823	0.361	0.785	0.333	0.70
	EC	0.323	0.824	0.374	0.789	0.347	0.701
	SC	0.316	0.822	0.366	0.787	0.339	0.698
	LAC	0.405	0.852	0.472	0.815	0.436	0.743
	NC	0.40	0.850	0.463	0.813	0.428	0.739
	UC	0.391	0.850	0.458	0.811	0.422	0.737
	CENC	**0.422**	**0.858**	**0.491**	**0.820**	**0.454**	**0.751**

has a better prediction accuracy. For MIPS network, these values of SN, SP, PPV, NPV, F-measure and ACC determined by CENC are 0.317, 0.827, 0.368, 0.792, 0.341 and 0.704, respectively, being higher than the previously proposed methods of DC, BC, EC, SC, LAC, NC and UC. These results indicate that CENC method has a better performance than the existing seven methods.

In addition, the precision–recall curve, a statistical method for evaluating stability, can be used for CENC method and other previous seven measures which are defined as follows:

$$\text{Precision } (n) = \frac{\text{TP}(n)}{\text{TP}(n) + \text{FP}(n)},$$

$$\text{Recall } (n) = \frac{\text{TP}(n)}{\text{TP}(n) + \text{FN}(n)},$$

where the definitions of TP, FP, FN are depicted in Sec. 3.2. The results are revealed in Figs. 3 and 4. In DIP network, our method of CENC has a better performance than the other methods. The same results are shown in MIPS network.

4.3. *Validation by the receiver operating characteristic curve and AUC*

The Receiver Operating Characteristic (ROC) is a valuable tool to measure the imbalance in classification.[48] It is used to evaluate the pros and cons of a binary classifier. Predicting essential proteins can be regarded as a two-classification case. Their definitions are as follows:

$$\text{TPR}(n) = \frac{\text{TP}(n)}{\text{TP}(n) + \text{FN}(n)},$$

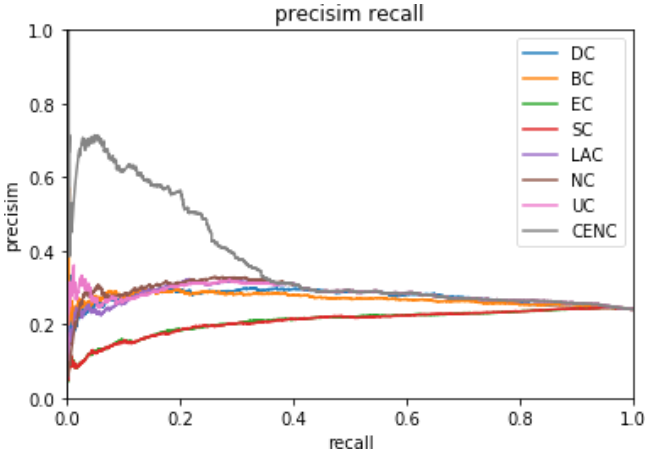

Fig. 3. (Color online) Precision and recall curves of CENC and other seven methods for MIPS network.

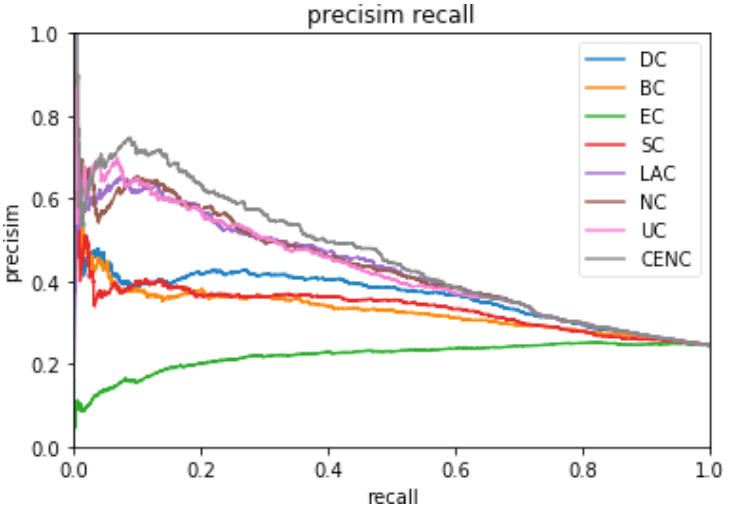

Fig. 4. (Color online) Precision and recall curves of CENC and other seven methods for DIP network.

$$\mathrm{FPR}(n) = \frac{\mathrm{FP}(n)}{\mathrm{FP}(n) + \mathrm{TN}(n)},$$

where the meanings of TP, FP, FN and TN are described in Sec. 3.2. As shown in Figs. 5 and 6, the ROC curve of CENC is slightly higher than that of the other seven methods, indicating that the method of CENC is more effective.

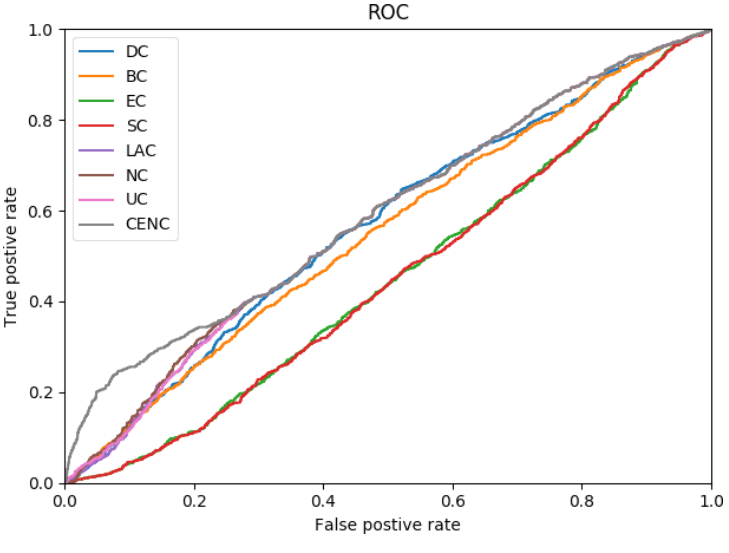

Fig. 5. (Color online) ROC curves of the CENC and the other seven methods for the MIPS network.

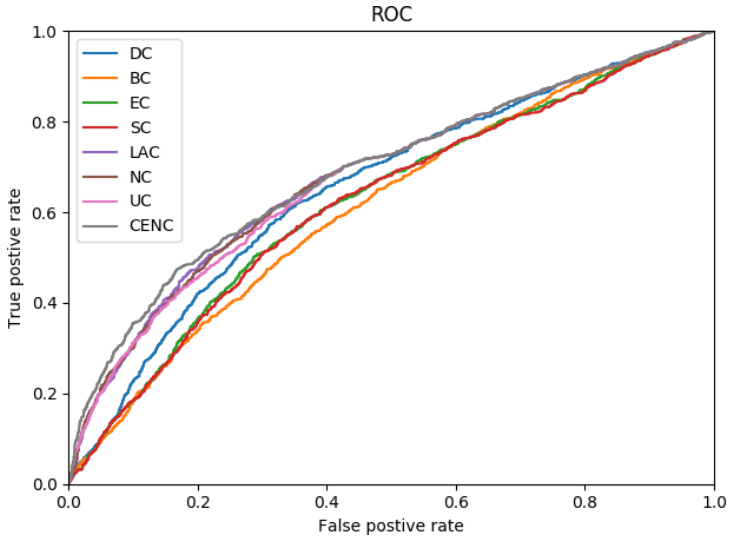

Fig. 6. (Color online) ROC curves of the CENC and the other seven methods for the DIP network.

Table 3. AUC values of CENC and other seven methods in MIPS and DIP networks.

Methods	DC	BC	EC	SC	LAC	NC	UC	CENC
MIPS	0.289	0.277	0.225	0.277	0.289	0.289	0.289	0.300
DIP	0.327	0.307	0.312	0.340	0.340	0.339	0.331	0.340

To further reveal the experimental results of the ROC curves, the area under the ROC curves is used to quantitatively analyze the results, generally called AUC. The AUC results are shown in Table 3. The values of CENC method are much more than the previous existing methods.

4.4. *Evaluation of jackknife methodology*

The jackknife methodology was developed by Holman *et al.*, being an effective universal prediction method.[38] The X-axis represents the quantity of selected predictive essential proteins after sequencing and the Y-axis represents the quantity of true essential proteins in the selected proteins. First, according to the predicted value, proteins are sorted in descending order. Then we choose predictive essential proteins from top 0 to top 800 in each dataset. Last, the jackknife curve is drawn based on the accumulated quantity of real essential proteins. From Figs. 7 and 8, we can see that the prediction efficiency for CENC method is higher than that of other seven centrality measures on the MIPS and DIP networks. Consequently, the jackknife curves reveal that our CENC method is an effective approach for predicting essential proteins.

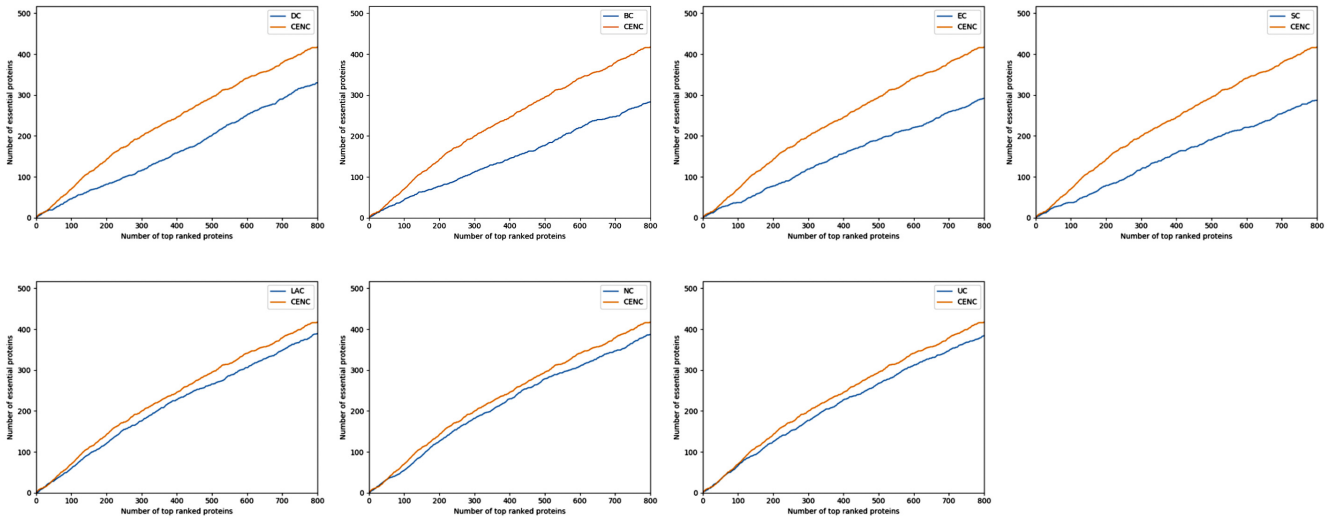

Fig. 7. (Color online) The performances of CENC and other seven centrality measures on the DIP network that are evaluated by a jackknife methodology.

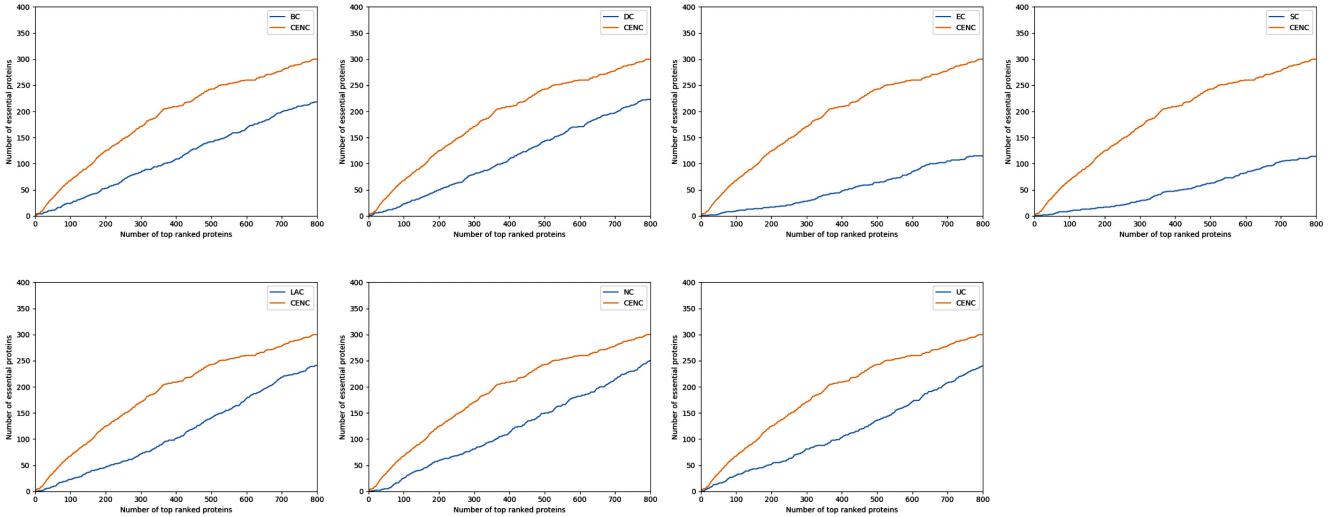

Fig. 8. (Color online) The performances of CENC and other seven centrality measures on the MIPS network are evaluated by a jackknife methodology.

5. Conclusion

Essential proteins are crucial for the survival and normal functioning of all organisms. Improving the recognition accuracy of essential proteins is a challenging task. Plenty of scholars devoted themselves to identify essential proteins in terms of the topological features for the whole network, ignoring the importance of complex and biological information. In this paper, on the basis of the mixed clustering coefficient for complexes and edge topology, a new CENC method is proposed. Then two different datasets of MIPS and DIP are applied. The evaluation methods include "sorting-screening" method, six statistical methods, the precision–recall curves, ROC curve, AUC and jackknife method. Then we compare CENC with other seven proposed methods containing DC, BC, EC, SC, LAC, NC and UC by using these evaluation methods. It is found that our proposed CENC method has the ability to improve the accuracy in predicting essential proteins.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 11361033) and the Natural Science Foundation of Gansu Province (No. 1212RJZA029).

References

1. H. B. Fraser *et al.*, *Science* **296**, 750 (2002).
2. B. Xu *et al.*, *IEEE/ACM Trans. Comput. Biol. Bioinf.* **16**, 377 (2017).
3. E. A. Winzeler *et al.*, *Science* **285**, 901 (1999).
4. Y. Wang *et al.*, *PLoS One* **9**, e108716 (2014).
5. T. Roemer *et al.*, *Mol. Microbiol.* **50**, 167 (2003).
6. L. M. Cullen and G. M. Arndt, *Immunol. Cell Biol.* **83**, 217 (2005).
7. E. Estrada, *Proteomics* **6**, 35 (2006).
8. W. Peng *et al.*, *BMC Syst. Biol.* **6**, 87 (2012).
9. G. Giaever *et al.*, *Nature* **418**, 387 (2002).
10. H. M. Jeong *et al.*, *Nature* **411**, 41 (2001).
11. B. H. Zhao *et al.*, *IEEE Trans. Nanobiosci.* **13**, 415 (2014).
12. M. W. Hahn and A. D. Kern, *Mol. Biol. Evol.* **22**, 803 (2005).
13. L. C. Freeman, *Sociometry* **40**, 35 (1977).
14. M. Li *et al.*, *J. Theor. Biol.* **447**, 65 (2018).
15. S. Wuchty and P. F. Stadler, *J. Theor. Biol.* **223**, 45 (2003).
16. N. N. Batada, L. D. Hurst and M. Tyers, *PLoS Comput. Biol.* **2**, e88 (2006).
17. C. C. Lin *et al.*, *J. Proteome Res.* **8**, 1925 (2009).
18. H. Liang and W. H. Li, *Trends Genet.* **23**, 375 (2007).
19. E. Estrada and A. Juan, *Phys. Rev. E* **71**, 1 (2005).
20. M. Li *et al.*, *Comput. Biol. Chem.* **35**, 143 (2011).
21. P. Bonacich, *Am. J. Soc.* **92**, 1170 (1987).
22. T. Nie *et al.*, *Phys. A Stat. Mech. Appl.* **453**, 290 (2016).
23. F. Radicchi *et al.*, *Proc. Nat. Acad. Sci. USA* **101**, 2658 (2003).
24. Y. Jiang *et al.*, *IEEE Int. Conf. Bioinformatics and Biomedicine* (IEEE, 2014).
25. J. Wang *et al.*, *IEEE/ACM Trans. Comput. Biol. Bioinf.* **9**, 1070 (2012).

26. Y. Tang *et al.*, *Biosystems* **127**, 67 (2015).
27. K. Stephenson and M. Zelen, *Soc Netw.* **11**, 1 (1989).
28. G. Li *et al.*, *IEEE/ACM Trans. Comput. Biol. Bioinf.* (2018).
29. Z. P. Zhang *et al.*, *J. Theor. Biol.* **480**, 274 (2019).
30. G. T. Hart, I. Lee and E. M. Marcotte, *BMC Bioinf.* **8**, 236 (2007).
31. M. Li *et al.*, *EEE/ACM Trans. Comput. Biol. Bioinf.* **14**, 370 (2017).
32. M. Li, H. H. Zhang and Y. P. Fei, *J. Central South Univ.* **44**, 1024 (2013).
33. X. Lei *et al.*, *Knowl. Based Syst.* **151**, 136 (2018).
34. I. Xenarios *et al.*, *Nucleic Acids Res.* **30**, 303 (2002).
35. R. Zhang and Y. Lin, *Nucleic Acids Res.* **37**, D455 (2009).
36. J. Wang *et al.*, *Trans. Comput. Biol. Bioinf.* **9**, 1070 (2012).
37. C. C. Friedel, J. Krumsiek and R. Zimmer, *Int. Conf. Research in Computational Molecular Biology* (Springer-Verlag, 2008).
38. S. Pu *et al.*, *Nucleic Acids Res.* **37**, 825 (2009).
39. S. Pu *et al.*, *Proteomics* **7**, 944 (2010).
40. A. G. Holman *et al.*, *BMC Microbiol.* **9**, 1 (2009).
41. J. M. Cherry *et al.*, *Nucleic Acids Res.* **26**, 73 (1998).
42. M. Li *et al.*, *IEEE/ACM Trans. Comput. Biol. Bioinf.* **12**, 372 (2015).
43. Y. Zhu and C. Wu, *Proc. 37th Chinese Control Conf.* (2018).
44. J. W. Luo and Y. Qi, *PLoS One* **10**, e0131418 (2015).
45. M. P. Joy *et al.*, *J. Biomed. Biotechnol.* **2005**, 96 (2014).
46. H. W. Mewes *et al.*, *Nucleic Acids Res.* **34**, 169 (2004).
47. J. B. Pereira-Leal *et al.*, *Mol. Biol. Evol.* (2015).
48. P. Bradley, *Pattern Recognit.* **30**, 1145 (1996).