Ess-NEXG: Predict Essential Proteins by Constructing a Weighted Protein Interaction Network Based on Node Embedding and XGBoost

Nian Wang, Min Zeng, Jiashuai Zhang, Yiming Li, and Min Li(✉)

School of Computer Science and Engineering, Central South University, Changsha 410083, People's Republic of China
{wang_nian,zengmin,jiashuaizhang,lynncsu}@csu.edu.cn, limin@mail.csu.edu.cn

Abstract. Essential proteins are indispensable in the development of organisms and cells. Identification of essential proteins lays the foundation for the discovery of drug targets and understanding of protein functions. Traditional biological experiments are expensive and time-consuming. Considering the limitations of biological experiments, many computational methods have been proposed to identify essential proteins. However, lots of noises in the protein-protein interaction (PPI) networks hamper the task of essential protein prediction. To reduce the effects of these noises, constructing a reliable PPI network by introducing other useful biological information to improve the performance of the prediction task is necessary. In this paper, we propose a model called Ess-NEXG which integrates RNA-Seq data, subcellular localization information, and orthologous information, for the prediction of essential proteins. In Ess-NEXG, we construct a reliable weighted network by using these data. Then we use the node2vec technique to capture the topological features of proteins in the constructed weighted PPI network. Last, the extracted features of proteins are put into a machine learning classifier to perform the prediction task. The experimental results show that Ess-NEXG outperforms other computational methods.

Keywords: Essential proteins · RNA-Seq data · Subcellular localization · Weighted protein-protein interaction network · Node embedding · XGBoost

1 Introduction

Essential proteins are very important in organisms and play a crucial role in the life process [1]. If the absence of a certain protein would lead to organisms to become disability or death, it can be said that this protein is essential [2]. Identification of essential proteins not only helps us to deepen the understanding of the life activities of cells

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Z. Cai et al. (Eds.): ISBRA 2020, LNBI 12304, pp. 95–104, 2020.
https://doi.org/10.1007/978-3-030-57821-3_9

but also provides a theoretical basis for the study of the pathogenesis of complex diseases and the discovery of drug targets [3, 4]. Thus, it is important for biologists to identify essential proteins. Conventional methods for the identification of essential proteins are biological experiments including RNA interference [5], conditional knockout [6], and single-gene knockout [7]. However, these experimental methods are expensive and time-consuming. Therefore, it is necessary to identify essential proteins by using computational approaches.

The rule of Centrality-Lethality, which indicates that nodes with high connectivity in the networks tend to be essential proteins, has been proposed in 2001 [8]. After that, several computational methods have been developed to identify essential proteins. These computational methods can be roughly divided into two classes: topology-based and machine learning-based methods. There are many topology-based methods, such as Degree Centrality (DC) [9], Betweenness Centrality (BC) [10], Closeness Centrality (CC) [11], Subgraph Centrality (SC) [12], Eigenvector Centrality (EC) [13], Information Centrality (IC) [14], and Local Average Connectivity (LAC) [15]. These methods focus on node centrality and provide a decent performance to identify essential proteins.

With the development of high-throughput sequencing technology, an increasing number of protein data are available to obtain. These protein data lay the foundation for the identification of the essential proteins. Many researchers integrated PPI network and biological information to improve the performance of the essential protein identification. The representative methods are PeC [16], UDoNC [17], ION [18], and CoTB [19]. Besides, many traditional machine learning algorithms are applied to this task. These machine learning algorithms include support vector machine (SVM) [20], Naïve Bayes [21], genetic algorithm [22], and decision tree [23]. Recently, deep learning techniques also have been applied to essential protein prediction and achieve good performance. Zeng *et al*. [24] proposed a novel computational framework to predict essential proteins based on deep learning techniques which can automatically learn features from three kinds of biological data. Zeng *et al*. also proposed a method named DeepEP [25, 26] which integrates PPI network and gene expression profiles.

Both in topology-based and machine learning-based methods, PPI networks play an important role. Studies showed that there are many false positive and false negative edges in PPI networks [27, 28], which can influence the performance of essential protein prediction [29]. Thus, to reduce the effects of these noises, it is imperative to construct a reliable weighted network to improve the performance of essential protein prediction by using other biological information. In this study, we used three kinds of biological data: RNA-Seq data, the subcellular localization information, and the orthologous information.

In this paper, we propose a novel computational framework named Ess-NEXG to identify essential proteins. First, to eliminate the noises in the PPI network, the PPI network is weighted by integrating RNA-Seq data, subcellular localization information, and orthologous information. Different from using score function in traditional computational methods, the weights of edges are calculated by dimension reduction from these data. Second, the network representation learning technique is used to learn the topological features of each protein in the weighted PPI network. Finally, the extracted features are used as the input of XGBoost model to identify potential essential proteins.

The effectiveness of Ess-NEXG is validated on the PPI network of saccharomyces cerevisiae (S. cerevisiae) [30]. Compared with the current topology-based methods including BC, CC, EC, IC, LAC, NC, SC, PeC, SPP [31], WDC [32], RSG [33] and NIE [34], Ess-NEXG achieves a better performance. Besides, Ess-NEXG also outperforms other machine learning-based methods.

2 Materials and Methods

2.1 Data Source and Preprocessing

In this study, we used multiple biological data to identify essential proteins: PPI network, RNA-Seq data, subcellular localization information, and orthologous information. These biological data are widely used in the prediction of essential proteins. The PPI network dataset is downloaded from BioGRID database. After the removal of self-cycle interactions and discrete nodes, there are 5,501 proteins and 52,271 interactions in the dataset. Proteins and interactions represent nodes and edges in the PPI network, respectively.

The essential proteins are downloaded from Four databases: MIPS [35], SGD [36], DEG [37], OGEE [38]. After integrating information of essential proteins in the four databases, the dataset contains 1285 essential proteins. The RNA-Seq data is collected from the NCBI SRA database by Lei *et al.* [39]. This dataset contains gene expression data of 7108 proteins. The subcellular localization information is downloaded from the knowledge channel of COMPARTMENTS database [40]. The orthologous information is gathered from InParanoid database [41].

2.2 Constructing Weighted PPI Network

Formally, the PPI network is described as an undirected graph G (V, E) consisting of a set of nodes $V = \{v_1, v_2, \ldots, v_n\}$ and edges $E = \{e(v_i, v_j)\}$. A node $v_i \in V$ represents a protein and an edge $e(v_i, v_j) \in E$ represents the interaction between protein v_i and v_j. As mentioned above, the PPI network plays an indispensable role in essential protein prediction. However, recent studies showed that there are some noises in the current PPI network, which can affect the identification performance. In order to improve the performance of the essential protein prediction, it is necessary to construct a reliable PPI network.

In this paper, we used RNA-Seq data, subcellular localization information, and orthologous information to weigh the PPI network to reduce the effects of noises. The three types of biological data represent the co-expression of two interacting proteins, the spatiality of proteins, and the conservatism of proteins, respectively. Thus, they can be used to filter the noises and calculate the weight of interacting proteins.

2.2.1 Obtain Better Representation with Principal Component Analysis

We have three different types of biological data. If we combine them directly, each protein has a 24-dimensional feature vector. However, the three kinds of biological data are from different sources and have the following properties:

1. The ranges of values in the three types of biological data vary a lot. The range of values in RNA-Seq data is from zero to tens of thousands; the values in subcellular localization information are binary (0 and 1); the range of values in orthologous information is from 0 to 99.
2. The subcellular localization information is very sparse; the RNA-Seq data and orthologous information are dense.
3. The dimensionality of three kinds of data is different. To order to extract useful features, we used principal component analysis (PCA) to reduce the dimensionality and obtain better representations of proteins.

It is important to find a good way that combines three kinds of biological data for calculating how strong two proteins interact. In consideration of the differences in those three kinds of biological data, we use PCA to reduce the dimension of the 24-dimensional vector to get a better protein representation vector. PCA is a useful tool for feature extraction. The samples are projected from high-dimensional space into low-dimensional space by PCA through linear transformation, which can obtain a dense protein vector and be more suitable for calculating the weight of edges. After the steps of PCA, we can obtain a dense vector which is a better representation.

2.2.2 Calculate the Strength of Interacting Proteins by Pearson's Correlation Coefficient

Pearson's correlation coefficient (PCC) is used to calculate how strong two proteins interact in the raw PPI network. After PCA, each protein has a dense representation vector $W_i = (\omega_1, \omega_2, \ldots \omega_{n'})$. So the strength of two interacting proteins $v_i = (x_1, x_2, \ldots x_{n'})$ and protein $v_j = (y_1, y_2, \ldots y_{n'})$ is calculated by PCC. The value of PCC ranges from -1 to 1, if PCCv_i, v_j is a positive value, it means that the relationship between protein v_i and v_j is positive. On the contrary, if PCCv_i, v_j is a negative value, it means that the relationship between protein and v_j is negative.

Finally, the weight of edges in the network is $\text{Weight}(v_i, v_j) = \text{PCC}(v_i, v_j)$. So far, the raw PPI network has been weighted by integrating three types of biological data. Figure 1 plots the workflow of the weighting process.

2.3 Identification of Essential Proteins Based on Network Representation Learning and XGboost

In order to identify essential proteins more correctly, it is necessary to learn better topological features for proteins. In this study, we use node2vec [42] to learn the topological features. Node2vec technique was developed in 2016, it is inspired by word2vec [43] and DeepWalk [44]. It projects every node in the network to a low-dimensional space vector based on unsupervised learning. Node2vec defines two parameters p and q that are used to balance the depth-first search (DFS) and the breadth-first search (BFS), which can preserve the local neighbor node relations and global structure information.

After getting topological features of proteins, the next step is choosing a suitable classifier for essential protein prediction. XGBoost (eXtreme Gradient Boosting) [45] is one of the best available machine learning methods. XGBoost algorithm uses a simple

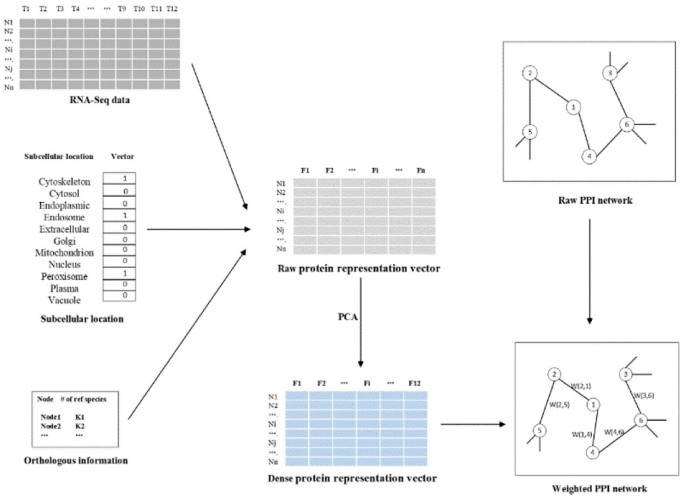

Fig. 1. A diagram of the weighted PPI network construction. The raw protein representation vector is constituted by using RNA-Seq data, subcellular location, and orthologous information. In order to obtain a better representation, we use the PCA technique to reduce the dimension of the raw protein representation vector. Finally, we weight the PPI network by PCC based on the dense representation vector and the raw PPI network.

model to fit the data that can get a general performance. Then, simple models are added to the whole XGBoost model constantly. Until the whole model approaches the complexity of the sample data, the performance of this model is best to identify essential proteins. Figure 2 plots the whole workflow.

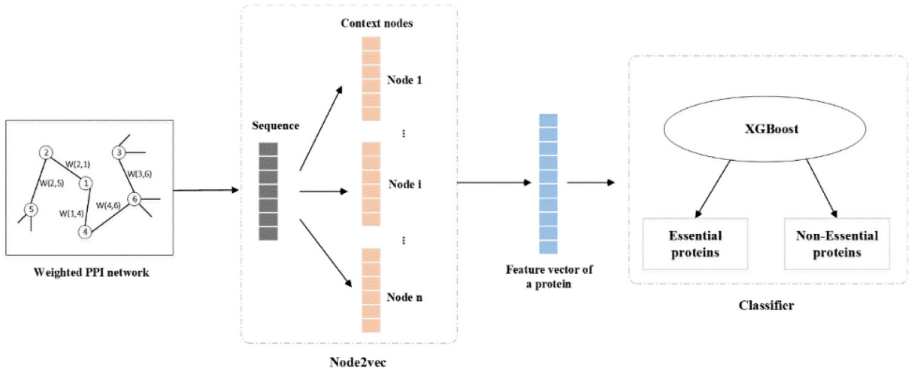

Fig. 2. A diagram of the essential proteins identifying. We use node2vec to extract the protein features, and then use the features we extract as the input of XGBoost to classify proteins.

3 Results

3.1 Comparisons with Current Topology-Based Methods

To validate the performance of Ess-NEXG, we compared Ess-NEXG with some current topology-based methods (BC, CC, EC, IC, LAC, NC, SC, PeC, SPP, WDC, RSG, NIE). In these methods, every node has a score according to corresponding score function. Because there are 1285 essential proteins in the PPI network, we select the top 1285 proteins as candidate essential proteins, and the rest 4216 proteins are candidate non-essential proteins. According to the true label, we calculated the accuracy, precision, recall, and F-score of the 12 computation-based methods. The results of Ess-NEXG and other topology-based methods are shown in Table 1.

Table 1. Comparison of the values of accuracy, precision, recall, and F-score of Ess-NEXG and other topology-based methods.

Methods	Accuracy	Precision	Recall	F-score
BC	0.728	0.411	0.383	0.396
CC	0.670	0.278	0.259	0.268
EC	0.732	0.420	0.391	0.405
IC	0.746	0.454	0.423	0.438
LAC	0.763	0.492	0.458	0.475
NC	0.762	0.490	0.457	0.473
SC	0.732	0.420	0.391	0.405
PeC	0.758	0.480	0.447	0.463
SPP	0.706	0.561	0.479	0.516
WDC	0.758	0.481	0.448	0.464
RSG	0.758	0.475	0.518	0.495
NIE	0.757	0.473	0.528	0.499
Ess-NEXG	0.819	0.600	0.580	0.590

From Table 1, we can see that all assessment metrics obtained by Ess-NEXG are higher than other topology-based methods. According to the results of these topology-based methods, we find that the accuracy of LAC, the precision of SPP, the recall of NIE, and the F-score of SPP are the highest values in these four assessment metrics among these topology-based methods. Compare with the four assessment metrics, Ess-NEXG improves the performance by 7.3%, 7.0%, 9.8%, and 14.3% respectively. In summary, the results indicate that Ess-NEXG outperforms other topology-based methods.

3.2 Comparisons with Other Machine Learning Algorithms

In Ess-NEXG, we choose the XGBoost classifier to identify essential proteins. In order to validate the performance of Ess-NEXG, we compared Ess-NEXG with other machine

learning algorithms including support vector machine (SVM), Naïve Bayes, and decision tree, random forest [46], AdaBoost [47]. To ensure equitably, we also use the same input features of proteins and assessment metrics. The results are shown in Table 2. From Table 2, we can see that Ess-NEXG has the best performance. Figure 3 plots the ROC curve of Ess-NEXG and other machine learning algorithms. We can see that the ROC curve of Ess-NEXG is significantly higher than other machine learning algorithms. Table 2 and Fig. 3 show that Ess-NEXG is better than other machine learning algorithms.

Table 2. Comparison of the values of accuracy, precision, recall, F-score, and AUC of Ess-NEXG and other machine learning algorithms.

Model	Accuracy	Precision	Recall	F-score	AUC
SVM	0.70	0.38	0.62	0.47	0.73
Naïve Bayes	0.79	0.50	0.38	0.43	0.72
Decision tree	0.71	0.35	0.40	0.37	0.62
Random forest	0.80	0.58	0.26	0.36	0.71
AdaBoost	0.79	0.51	0.29	0.37	0.71
Ess-NEXG	0.82	0.60	0.58	0.59	0.82

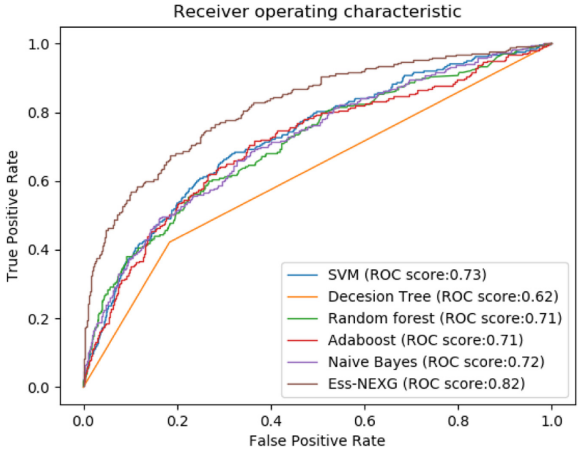

Fig. 3. ROC curves of Ess-NEXG and other machine learning algorithms.

4 Conclusions

Essential proteins are very important proteins in the life process, which can help us understanding life activities in living organisms. The identification of essential proteins is helpful in drug design and disease prediction. In this paper, we propose a novel computational framework to identify essential proteins in the PPI network. Previous studies have shown that the noises in the PPI network affect the performance of essential protein identification. In order to reduce the effects of noises in the PPI network, we propose a weighted method that integrates RNA-Seq data, subcellular localization information, and orthologous information. After obtaining the protein representation vector by using PCA technique, PCC is used to calculate the edge weights in the PPI network. Then node2vec is applied to extract topological features from the weighted PPI network. Finally, the topological features are fed into XGboost model to identify essential proteins. In order to evaluate the performance of Ess-NEXG, we compared it with current topology-based methods. The results show that Ess-NEXG outperforms them. In addition, we also compared Ess-NEXG with machine learning algorithms to show effectiveness. While Ess-NEXG outperforms other computational models, it still has some limitations. The biggest limitation is that we have to collect biological data for each new species, which is expensive and cumbersome. In the future, we would further improve the performance of essential protein prediction by using powerful deep learning techniques [48] and useful biological information [49].

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grants (No. 61832019), the 111 Project (No. B18059), Hunan Provincial Science and Technology Program (2018WK4001).

References

1. Winzeler, E.A., et al.: Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science **285**, 901–906 (1999)
2. Clatworthy, A.E., Pierson, E., Hung, D.T.: Targeting virulence: a new paradigm for antimicrobial therapy. Nat. Chem. Biol. **3**, 541 (2007)
3. Furney, S.J., Albà, M.M., López-Bigas, N.: Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genom. **7**, 165 (2006). https://doi.org/10.1186/1471-2164-7-165
4. Zhao, J., Lei, X.: Detecting overlapping protein complexes in weighted PPI network based on overlay network chain in quotient space. BMC Bioinform. **20**, 1–12 (2019)
5. Roemer, T., et al.: Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. Mol. Microbiol. **50**, 167–181 (2003)
6. Cullen, L.M., Arndt, G.M.: Genome-wide screening for gene function using RNAi in mammalian cells. Immunol. Cell Biol. **83**, 217–223 (2005)
7. Giaever, G., et al.: Functional profiling of the Saccharomyces cerevisiae genome. Nature **418**, 387 (2002)
8. Jeong, H., Mason, S.P., Barabási, A.-L., Oltvai, Z.N.: Lethality and centrality in protein networks. Nature **411**, 41 (2001)
9. Hahn, M.W., Kern, A.D.: Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol. Biol. Evol. **22**, 803–806 (2004)

10. Joy, M.P., Brock, A., Ingber, D.E., Huang, S.: High-betweenness proteins in the yeast protein interaction network. Biomed. Res. Int. **2005**, 96–103 (2005)
11. Wuchty, S., Stadler, P.F.: Centers of complex networks. J. Theor. Biol. **223**, 45–53 (2003)
12. Estrada, E., Rodriguez-Velazquez, J.A.: Subgraph centrality in complex networks. Phys. Rev. E **71**, 056103 (2005)
13. Bonacich, P.: Power and centrality: a family of measures. Am. J. Sociol. **92**, 1170–1182 (1987)
14. Stephenson, K., Zelen, M.: Rethinking centrality: methods and examples. Soc. Netw. **11**, 1–37 (1989)
15. Li, M., Wang, J., Chen, X., Wang, H., Pan, Y.: A local average connectivity-based method for identifying essential proteins from the network level. Comput. Biol. Chem. **35**, 143–150 (2011)
16. Li, M., Zhang, H., Wang, J.-X., Pan, Y.: A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. BMC Syst. Biol. **6**, 15 (2012). https://doi.org/10.1186/1752-0509-6-15
17. Peng, W., Wang, J., Cheng, Y., Lu, Y., Wu, F., Pan, Y.: UDoNC: an algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks. IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) **12**, 276–288 (2015)
18. Peng, W., Wang, J., Wang, W., Liu, Q., Wu, F.-X., Pan, Y.: Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. BMC Syst. Biol. **6**, 87 (2012). https://doi.org/10.1186/1752-0509-6-87
19. Qin, C., Sun, Y., Dong, Y.: A new computational strategy for identifying essential proteins based on network topological properties and biological information. PLoS ONE **12**, e0182031 (2017)
20. Hwang, Y.-C., Lin, C.-C., Chang, J.-Y., Mori, H., Juan, H.-F., Huang, H.-C.: Predicting essential genes based on network and sequence analysis. Mol. BioSyst. **5**, 1672–1678 (2009)
21. Cheng, J., et al.: Training set selection for the prediction of essential genes. PLoS ONE **9**, e86805 (2014)
22. Zhong, J., Wang, J., Peng, W., Zhang, Z., Pan, Y.: Prediction of essential proteins based on gene expression programming. BMC Genom. **14**, S7 (2013). https://doi.org/10.1186/1471-2164-14-S4-S7
23. Acencio, M.L., Lemke, N.: Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinform. **10**, 290 (2009). https://doi.org/10.1186/1471-2105-10-290
24. Zeng, M., et al.: A deep learning framework for identifying essential proteins by integrating multiple types of biological information. IEEE/ACM Trans. Comput. Biol. Bioinform. (2019). https://doi.org/10.1109/TCBB.2019.2897679
25. Zeng, M., Li, M., Wu, F.-X., Li, Y., Pan, Y.: DeepEP: a deep learning framework for identifying essential proteins. BMC Bioinform. **20**, 506 (2019). https://doi.org/10.1186/s12859-019-3076-y
26. Zeng, M., Li, M., Fei, Z., Wu, F.-X., Li, Y., Pan, Y.: A deep learning framework for identifying essential proteins based on protein-protein interaction network and gene expression data. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 583–588. IEEE (2018)
27. Zhang, F., et al.: A deep learning framework for gene ontology annotations with sequence-and network-based information. IEEE/ACM Trans. Comput. Biol. Bioinform. (2020). https://doi.org/10.1109/TCBB.2020.2968882
28. Zhang, F., Song, H., Zeng, M., Li, Y., Kurgan, L., Li, M.: DeepFunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions. Proteomics **19**, 1900019 (2019)
29. Von Mering, C., et al.: Comparative assessment of large-scale data sets of protein–protein interactions. Nature **417**, 399 (2002)

30. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. Nucleic Acids Res. **34**, D535–D539 (2006)
31. Li, M., Li, W., Wu, F.-X., Pan, Y., Wang, J.: Identifying essential proteins based on sub-network partition and prioritization by integrating subcellular localization information. J. Theor. Biol. **447**, 65–73 (2018)
32. Tang, X., Wang, J., Zhong, J., Pan, Y.: Predicting essential proteins based on weighted degree centrality. IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) **11**, 407–418 (2014)
33. Lei, X., Zhao, J., Fujita, H., Zhang, A.: Predicting essential proteins based on RNA-Seq, subcellular localization and GO annotation datasets. Knowl.-Based Syst. **151**, 136–148 (2018)
34. Zhao, J., Lei, X.: Predicting essential proteins based on second-order neighborhood information and information entropy. IEEE Access **7**, 136012–136022 (2019)
35. Mewes, H.-W., et al.: MIPS: a database for genomes and protein sequences. Nucleic Acids Res. **30**, 31–34 (2002)
36. Cherry, J.M., et al.: SGD: saccharomyces genome database. Nucleic Acids Res. **26**, 73–79 (1998)
37. Zhang, R., Lin, Y.: DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res. **37**, D455–D458 (2008)
38. Chen, W.-H., Minguez, P., Lercher, M.J., Bork, P.: OGEE: an online gene essentiality database. Nucleic Acids Res. **40**, D901–D906 (2011)
39. Zhao, J., Lei, X., Wu, F.-X.: Predicting protein complexes in weighted dynamic PPI networks based on ICSC. Complexity **2017**, 1–11 (2017)
40. Binder, J.X., et al.: COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database **2014** (2014)
41. Östlund, G., et al.: InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. **38**, D196–D203 (2009)
42. Grover, A., Leskovec, J.: node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864. ACM (2016)
43. Goldberg, Y., Levy, O.: word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014)
44. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 701–710. ACM (2014)
45. Chen, W., Fu, K., Zuo, J., Zheng, X., Huang, T., Ren, W.: Radar emitter classification for large data set based on weighted-xgboost. IET Radar Sonar Navig. **11**, 1203–1207 (2017)
46. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
47. Freund, Y., Schapire, R., Abe, N.: A short introduction to boosting. J.-Japn. Soc. Artif. Intell. **14**, 1612 (1999)
48. Zeng, M., Li, M., Fei, Z., Yu, Y., Pan, Y., Wang, J.: Automatic ICD-9 coding via deep transfer learning. Neurocomputing **324**, 43–50 (2019)
49. Zeng, M., Zhang, F., Wu, F.-X., Li, Y., Wang, J., Li, M.: Protein–protein interaction site prediction through combining local and global features with deep neural networks. Bioinformatics **36**, 1114–1120 (2020)