A deep learning framework for identifying essential proteins based on protein-protein interaction network and gene expression data

Min Zeng, Min Li*, Zhihui Fei
School of Information Science and Engineering, Central South University
Changsha, 410083, P.R. China
E-mail: zengmin@csu.edu.cn, limin@mail.csu.edu.cn, zhihuifei@foxmail.com

Fang-Xiang Wu
Division of Biomedical Engineering and Department of Mechanical Engineering
University of Saskatchewan
Saskatoon, SKS7N5A9, Canada
E-mail: faw341@mail.usask.ca

Yaohang Li
Department of Computer Science
Old Dominion University
Norfolk, USA
E-mail: yaohang@cs.odu.edu

Yi Pan
Department of Computer Science
Georgia State University
Atlanta, GA30302, USA
E-mail: yipan@gsu.edu

***Abstract*—Identifying essential proteins is of vital importance for disease study and drug design. A lot of topology-based and machine learning-based methods have been proposed to identify essential proteins. However, traditional topology-based methods only focus on explicitly described characteristics of network topology and are not expressive enough to capture the complexity of connectivity patterns observed in biological networks. In addition, identification of essential proteins is an imbalanced learning problem due to the fact that there are significantly more non-essential proteins than the essential ones. Few machine learning-based methods take the imbalanced nature into consideration. We propose a new deep learning framework, to tackle the above limitations. In our model, we make use of the node2vec technique to learn topological features from protein-protein interaction (PPI) network without manual feature selection. To overcome the problem of the imbalanced nature of dataset, we use a sampling method, which does not bias to the majority and minority classes in a training step and tend to make full use of all samples during the whole training process. To evaluate the performance of our model, we test it on S. cerevisiae dataset. Our results show that it greatly outperforms topology-based methods including DC, BC, CC, EC, NC, LAC, PeC and WDC. It also outperforms machine learning-based methods including support vector machine (SVM), decision tree, random forest and Adaboost.**

Keywords—deep learning, identifying essential proteins, protein-protein interaction network, gene expression, imbalanced learning

I. INTRODUCTION

Proteins are products of gene expressions, which perform many functions in organisms and play important roles in various biological activities [1]. Essential proteins are indispensable in cellular life because they normally ensure the functions of cellular life [2]. An organism cannot survive or develop if one of the essential proteins has been removed [3]. Identification of essential proteins is one of the focuses in bioinformatics research for the following reasons: 1) Determination of essential proteins helps to understand the minimum requirements of the survival and evolution of a cell; and 2) Essential proteins are potential targets of new antibiotics drug. Essential proteins can be identified by biological experiments such as single gene knockout [4], conditional knockout [5], and RNA interference [6]; however, the experiments are expensive and time-consuming. Considering these experimental constraints, it is urgent to develop an accurate computational approach for identifying essential proteins.

Previous studies have shown that the topological properties of proteins in protein-protein interaction (PPI) network have a strong relationship with gene essentiality [7]. Based on the topological features in PPI networks, various centrality methods have been proposed and used for identifying essential proteins [8]. These centrality methods include Degree Centrality (DC) [9], Betweenness Centrality (BC) [10], Closeness Centrality (CC) [11], Subgragh Centrality (SC) [12], Eigenvector Centrality (EC) [13], Information Centrality (IC) [14] and Edge Clustering Coefficient Centrality (NC) [15]. Additionally, some biological information including gene expression profiles [16-18], subcellular localization [19], and protein domains [20] have been incorporated to predict essential proteins.

With the development of proteomics, more proteins data have been obtained with the information of whether they are essential or not being known. The above mentioned centrality methods do not make use of the label information of proteins although they have achieved reasonable results. Recent years, machine learning methods including support vector machine (SVM) [21] [22], decision tree [23] [24], Naive Bayes [25] [26], ensemble method [27] [28], and genetic algorithms [29] have been widely used for identifying essential proteins.

In the previous studies, however, there are three major limitations. First, for topology-based methods, single topological feature cannot characterize the comprehensive topological information of PPI networks. A PPI network usually has thousands of vertices and tens of thousands edges. A single centrality index of a node is just a real number, which is difficult to characterize the topological features of a complex network. Second, for machine learning-based methods, there is lack of a computational framework to automatically select

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

topological features from various proposed topology-based methods. The commonly used approach for the selection topological feature is to select the most appropriate topological features according to the results of statistical methods. As a result, it is difficult to explain why these features were chosen and what roles they play in the classification. Thirdly, few machine learning-based researches have taken into consideration of the imbalanced nature of data distribution. Data imbalance means an uneven distribution of samples between different classes. The imbalanced nature of dataset usually tends to bias towards the majority class and leads to a poor performance [30].

To tackle the above limitations, we employ a network representation learning technique, which is a newly developed network feature extraction technique. Network representation learning aims to encode network topology into a low-dimensional space. It can automatically learn a low-dimensional dense vector for each vertex to represent the topological information of a network without manual topology-based features selection. In addition, the learned representations encode semantic and topological roles of vertices in the network, which can be used to measure topological similarity between vertices. Through network representation learning, the learned dense vector has a richer representation of PPI network than a single centrality index. Then we apply a sampling method to overcome the challenge of the imbalanced dataset. The proposed sampling method utilizes a balanced subset from the raw dataset for training in each training step. After many training steps, in high probability, it can make full use of all samples of the raw dataset to train the model.

In addition to overcoming the above limitations, our deep learning framework also offers other attractive advantages. We transform gene expression profiles into an image to better extract its patterns. In such a way, the effective machine learning techniques for image classification can be used to identify essential proteins. As we know, multi-scale convolutional neural network is a powerful deep learning architecture which has been widely used in image classification. Inspired by their success in image classification, we use multi-scale convolutional neural network to extract the patterns of gene expression profiles [31].

We carry out our experiments on S. cerevisiae data. Accuracy, precision, recall, F-measure and AUC (Area Under receiver operating characteristic Curve) obtained by our model are 0.823, 0.582, 0.518, 0.548 and 0.807, respectively. Our experimental results show that our method yields better performance than topology-based methods including DC, BC, CC, EC, NC, LAC [32], PeC [16], and WDC [17]. It also outperforms the commonly used machine learning methods of SVM, decision tree, random forest (RF) and Adaboost.

II. MATERIALS AND METHODS

A. *Overview of our model*

Our deep learning framework for identifying essential protein is illustrated in Fig.1, which consists of feature extraction and classification part. The inputs to our deep learning network are two types of biological data, gene expression data and PPI network. The feature extraction part is responsible for extracting features and patterns from different biological data. We treat the gene expression data as an image and use multi-scale convolutional layer and pooling layer to extract features. For the PPI network, we apply a network representation learning technique called node2vec to learn a dense vector for each vertex to capture the topological information of a network. After feature extraction, the output vectors are concatenated together as the input for classification. The classification part consists of a fully connected hidden layer and an output layer. The fully connected layer with softmax activation function is used for preliminary processing. On top of the fully connected layer, an output layer performs the essential protein prediction task. Considering the imbalanced nature of the essential protein dataset, we apply a sampling method to train the parameters of deep learning model.

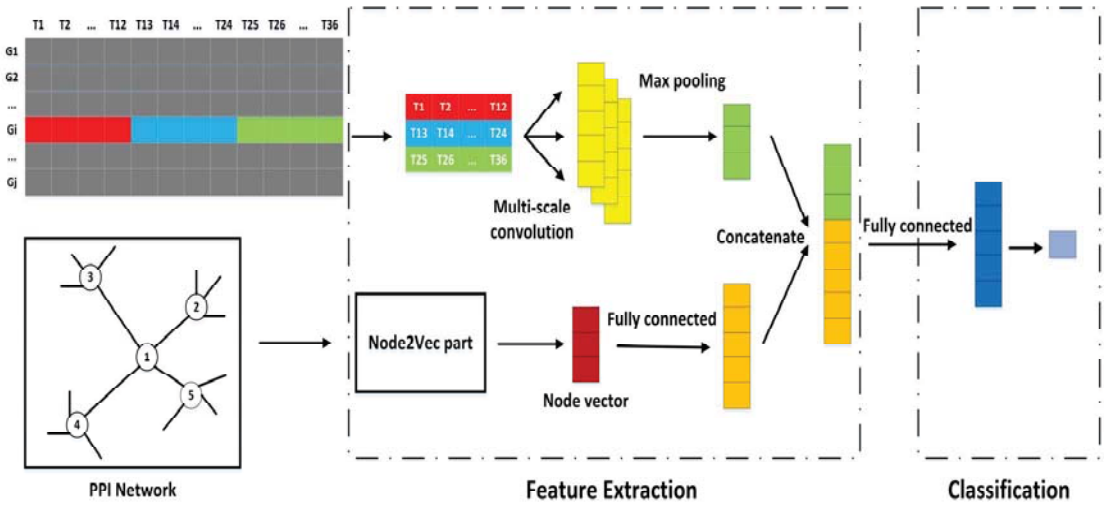

Fig. 1. An overview of our proposed deep learning framework for identifying essential proteins.

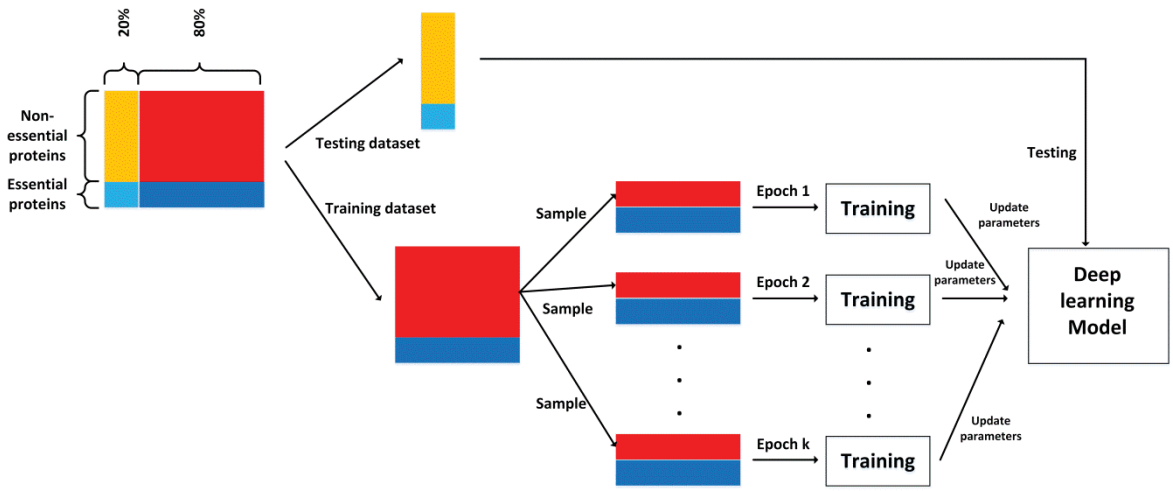

Fig. 2. An illustration of sampling method. In the sampling process, we used 80% samples for training and 20% samples for testing.

B. Network representation learning

Network topological feature extraction plays an important role in the study of identifying essential protein. Various network representation learning techniques have been proposed in recent years. Node2vec [33], a deep learning method, learns vector representations of vertexes based on local network information. It utilizes random walk algorithm to obtain each vertex's sequence. Then the Skip-Gram model [34] is employed to predict surrounding context words given a center vertex by maximizing the co-occurrence likelihood between a target vertex and its context vertices. During learning iterations, the learned vectors are successively updated using the Skip-Gram model. After completing the training step, the outputs of node2vec are dense vectors for all vertices in the network. These dense vectors are considered to be semantic and topological representation of the network.

C. Sampling method

Data imbalance, an uneven distribution of instances between different classes, is a very common phenomenon in real-world data sets. There have been a lot of studies on how to solve the imbalanced data learning problem including sampling methods, cost-sensitive learning methods, kernel-based learning methods, and active learning methods [30]. Sampling method is widely used and very effective among these methods. However, traditional sampling methods including random undersampling, random oversampling, and SMOTE [35] are not suitable for direct use in our deep learning framework.

To improve the prediction performance, we apply a sampling method to our model. We denote M as the number of minority class instances (essential proteins) and N as the number of majority class instances (non-essential proteins) in the training dataset where M<<N. M instances were sampled from the majority class at each epoch, then we combine the M instances in the majority class and all instances in the minority class together to train our deep learning model. This process is carried out k times to train our model. Such a sampling method ensures that each instance in the majority class can be picked and trained with equal number of two class instances to avoid overfitting. Fig. 2 illustrates the sampling method.

D. Multi-scale convolutional neural network

Convolutional neural network (CNN) is a class of deep neural networks that have been successfully applied to computer vision [36]. CNNs utilize layers with convolving filters that are applied to local features [37], which allows CNN layers to automatically learn low-level features from input data. Multi-scale CNN uses different sizes of kernels to extract local features, which has been showed to be an efficient method to combine different features for classification [38].

Inspired by the success of multi-scale CNN [39], we treat a gene expression profile as an image in order to better extract gene expression profiles features. A gene expression profile has three successive metabolic cycles. There are 12 time points in a cycle, and the time interval between two time points is 25 minutes. Specifically, we transform a one-dimensional vector with 36 real values into an image with 1 channel * 3 rows * 12 columns (as illustrated in Figure 1). Then we use multi-scale CNNs to extract local information and to explore the relationship between cycles.

E. Assessment metrics

Essential protein dataset is an imbalanced dataset. To properly evaluate the performance of our model and the other algorithms in identifying essential proteins, we use some assessment metrics for imbalanced learning.

In the following, we first explain four frequently used terms TP, TN, FP, and FN. TP and TN represent the number of samples of the minority and majority class which are classified correctly, respectively, and FP and FN represent the number of samples of the minority and majority class which are misclassified, respectively. Accuracy is defined as:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

Precision, recall, and F-measure are defined as:

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

$$\text{recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (3)$$

$$\mathrm{F-measure} = \frac{(1+\beta^2)\cdot \mathrm{recall}\cdot \mathrm{precision}}{\beta^2\cdot \mathrm{precision}+\mathrm{recall}} \quad (4)$$

where β is a coefficient to adjust the relationship between precision and recall. In this study, we adopt β=1.

AUC and average precision (AP) score are used in our study for evaluation.

III. DATA SOURCES

In this study we use multi-source datasets, including PPI network dataset, essential protein dataset, gene expression dataset. PPI network dataset of S. cerevisiae is the most widely used dataset in the study of identifying essential proteins. This dataset is downloaded from BioGRID database. After removing repeated interaction and self-interactions, the processed dataset contains 5616 proteins and 52833 interactions.

Essential protein dataset is selected from the following databases: MIPS [40], SGD [41], DEG [2], and SGDP, which contains 1199 essential proteins.

Gene expression dataset is retrieved from Tu et al., 2005 [42], which contain 6776 gene products (proteins) and 36 samples in total. This dataset has three successive metabolic cycles with 12 time points in a cycle and, the time interval between two time consecutive points is 25 minutes.

IV. EXPERIMENTAL RESULTS

A. Comparison with results of topology-based methods

To evaluate the performance of our deep learning framework, we compare our model with existing topology-based methods, DC, BC, CC, EC, NC and LAC, which have been widely used for comparison in essential protein prediction. Additionally, we also compare our model with PeC and WDC which are based on the integration of PPI and gene expression data. In this work, we select the top 1185 proteins (our processed PPI network has 1185 essential proteins) ranked by DC, BC, CC, EC, NC, LAC, PeC, and WDC as their predicted essential proteins. The rest of proteins are regarded as non-essential proteins. According to known labels of essential proteins and non-essential proteins, we obtain a confusion matrix which is used to calculate precision, recall, F-measure and accuracy of each method. In Table 1 we compare our predicted accuracy, precision, recall, and F-measure with those of DC, BC, CC, EC, NC, LAC, PeC, and WDC. By inspecting Table 1, one can find that all assessment metrics obtained by our deep learning method significantly outperform DC, BC, CC, EC, NC, LAC, PeC, and WDC. Our model obtains the values of accuracy, F-measure, precision and recall being 0.823, 0.548, 0.582 and 0.518, respectively, which are better than other topology-based methods including DC (0.740, 0.436, 0.430 and 0.433), BC (0.722, 0.398, 0.393 and 0.395), CC (0.665, 0.262, 0.260, and 0.261), EC (0.727, 0.408, 0.401, and 0.404), NC (0.752, 0.468, 0.464 and 0.466), LAC (0.745, 0.467, 0.409 and 0.436), PeC (0.747, 0.438, 0.430 and 0.434), and WDC (0.742, 0.455, 0.459, and 0.457). The experimental results show that our model is not only superior to those simple topology-based methods including DC, BC, CC, EC, NC, and LAC, but also outperforms those methods based on the integration of PPI and gene expression data.

TABLE Ⅰ. COMPARISON OF PERFORMANCES BETWEEN METHODS OF OUR MODEL, DC, BC, CC, EC, NC, LAC, PEC AND WDC.

Models	Accuracy	Precision	Recall	F-measure
DC	0.740	0.436	0.430	0.433
BC	0.722	0.398	0.393	0.395
CC	0.665	0.262	0.260	0.261
EC	0.727	0.408	0.401	0.404
NC	0.752	0.468	0.464	0.466
LAC	0.745	0.467	0.409	0.436
PeC	0.747	0.438	0.430	0.434
WDC	0.742	0.455	0.459	0.457
Our model	0.823	0.582	0.518	0.548

B. Comparison with results of other machine learning algorithms

A lot of machine learning algorithms have been employed for identifying essential proteins. The most commonly used algorithms are SVM, decision tree and ensemble learning-based methods. Here, we compare our deep learning framework with these algorithms. All of these machine learning algorithms are implemented by using scikit-learn python package. To ensure a fair comparison, we use gene expression data and node vectors which are generated by node2vec and then concatenate them into a vector as the input of these machine learning algorithms.

TABLE Ⅱ. COMPARISON OF PERFORMANCE BETWEEN OUR MODEL AND OTHER MACHINE LEARNING ALGORITHMS.

Algorithms	Accuracy	Precision	Recall	F-measure	AUC
SVM	0.809	0.71	0.12	0.21	0.72
DT	0.698	0.31	0.39	0.35	0.58
RF	0.809	0.63	0.17	0.27	0.70
Adaboost	0.805	0.54	0.34	0.42	0.73
Our model	0.823	0.58	0.52	0.55	0.81

In Table 2 we compare the performance results of our model with other machine learning algorithms. From the results presented in Table 2, we conclude that our model achieves the state-of-the-art results among these methods. Our model obtains F-measure and AUC with values of 0.55 and 0.81, respectively, which are better than SVM (0.21 and 0.72), decision tree (0.35 and 0.58), random forest (0.27 and 0.70),

Adaboost (0.42 and 0.73). Nevertheless, for individual metrics including accuracy, precision, or recall, our model does not show the highest values. For example, SVM has the highest precision value (0.71), but this is achieved by significantly sacrificing recall (0.12). Anyway, compared to the other machine learning methods, our model exhibits the best overall performance.

V. CONCLUSIONS

We propose a deep learning framework for identifying essential proteins. Our model employs node2vec technique to automatically learn semantic and topological features from PPI network without manual features selection. The technique of node2vec maps vertices to a low-dimensional space and obtains dense vectors, which have richer representation of PPI network than traditional centrality indexes. Hence our method captures comprehensive topological features of PPI network. We further apply a sampling method to solve the problem of imbalanced nature of data distribution. It utilizes a balanced subset from raw training dataset for training at each time and thus the classifier does not bias to any class in a training step. By training enough times, it makes use of all non-essential protein samples in raw training dataset. We use S. cerevisiae dataset to evaluate the performance of our model. Comparison with widely used methods of DC, BC, CC, EC, NC, LAC, PeC, and WDC demonstrates that our model greatly outperforms existing topology-based methods. Our model also outperforms machine learning methods such as SVM, decision tree, random forest, and Adaboost.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grants (No. 61832019, No. 61622213 and No. 61728211), the 111 Project (No.B18059) and the Fundamental Research Funds for the Central Universities of Central South University (No.2018zzts563).

REFERENCES

[1] J. I. Glass, C. A. Hutchison, H. O. Smith, and J. C. Venter, "A systems biology tour de force for a near‐minimal bacterium," *Molecular systems biology,* vol. 5, p. 330, 2009.

[2] R. Zhang and Y. Lin, "DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes," *Nucleic acids research,* vol. 37, pp. D455-D458, 2008.

[3] A. E. Clatworthy, E. Pierson, and D. T. Hung, "Targeting virulence: a new paradigm for antimicrobial therapy," *Nature chemical biology,* vol. 3, p. 541, 2007.

[4] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, *et al.*, "Functional profiling of the Saccharomyces cerevisiae genome," *nature,* vol. 418, p. 387, 2002.

[5] L. M. Cullen and G. M. Arndt, "Genome‐wide screening for gene function using RNAi in mammalian cells," *Immunology & Cell Biology,* vol. 83, pp. 217-223, 2005.

[6] T. Roemer, B. Jiang, J. Davison, T. Ketela, K. Veillette, A. Breton, *et al.*, "Large‐scale essential gene identification in Candida albicans and applications to antifungal drug discovery," *Molecular microbiology,* vol. 50, pp. 167-181, 2003.

[7] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature,* vol. 411, p. 41, 2001.

[8] X. Peng, J. Wang, W. Peng, F.-X. Wu, and Y. Pan, "Protein–protein interactions: detection, reliability assessment and applications," *Briefings in bioinformatics,* vol. 18, pp. 798-819, 2016.

[9] M. W. Hahn and A. D. Kern, "Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks," *Molecular biology and evolution,* vol. 22, pp. 803-806, 2004.

[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, pp. 96-103, 2005.

[11] S. Wuchty and P. F. Stadler, "Centers of complex networks," *Journal of Theoretical Biology,* vol. 223, pp. 45-53, 2003.

[12] E. Estrada and J. A. Rodriguez-Velazquez, "Subgraph centrality in complex networks," *Physical Review E,* vol. 71, p. 056103, 2005.

[13] P. Bonacich, "Power and centrality: A family of measures," *American journal of sociology,* vol. 92, pp. 1170-1182, 1987.

[14] K. Stephenson and M. Zelen, "Rethinking centrality: Methods and examples," *Social networks,* vol. 11, pp. 1-37, 1989.

[15] 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 (TCBB),* vol. 9, pp. 1070-1080, 2012.

[16] 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, p. 15, 2012.

[17] X. Tang, J. Wang, J. Zhong, and Y. Pan, "Predicting essential proteins based on weighted degree centrality," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),* vol. 11, pp. 407-418, 2014.

[18] M. Li, R. Zheng, H. Zhang, J. Wang, and Y. Pan, "Effective identification of essential proteins based on priori knowledge, network topology and gene expressions," *Methods,* vol. 67, pp. 325-333, 2014.

[19] X. Peng, J. Wang, J. Wang, F.-X. Wu, and Y. Pan, "Rechecking the centrality-lethality rule in the scope of protein subcellular localization interaction networks," *PloS one,* vol. 10, p. e0130743, 2015.

[20] W. Peng, J. Wang, Y. Cheng, Y. Lu, F. Wu, and Y. Pan, "UDoNC: an algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),* vol. 12, pp. 276-288, 2015.

[21] Y.-C. Hwang, C.-C. Lin, J.-Y. Chang, H. Mori, H.-F. Juan, and H.-C. Huang, "Predicting essential genes based on network and sequence analysis," *Molecular BioSystems,* vol. 5, pp. 1672-1678, 2009.

[22] K. Plaimas, R. Eils, and R. König, "Identifying essential genes in bacterial metabolic networks with machine learning methods," *BMC systems biology,* vol. 4, p. 56, 2010.

[23] M. Seringhaus, A. Paccanaro, A. Borneman, M. Snyder, and M. Gerstein, "Predicting essential genes in fungal genomes," *Genome research,* vol. 16, pp. 1126-1135, 2006.

[24] 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, p. 290, 2009.

[25] A. M. Gustafson, E. S. Snitkin, S. C. Parker, C. DeLisi, and S. Kasif, "Towards the identification of essential genes using targeted genome sequencing and comparative analysis," *Bmc Genomics,* vol. 7, p. 265, 2006.

[26] J. Cheng, Z. Xu, W. Wu, L. Zhao, X. Li, Y. Liu, *et al.*, "Training set selection for the prediction of essential genes," *PloS one,* vol. 9, p. e86805, 2014.

[27] J. Deng, L. Deng, S. Su, M. Zhang, X. Lin, L. Wei, *et al.*, "Investigating the predictability of essential genes across distantly related organisms using an integrative approach," *Nucleic acids research,* vol. 39, pp. 795-807, 2010.

[28] Y. Lu, J. Deng, J. C. Rhodes, H. Lu, and L. J. Lu, "Predicting essential genes for identifying potential drug targets in Aspergillus fumigatus," *Computational biology and chemistry,* vol. 50, pp. 29-40, 2014.

[29] J. Zhong, J. Wang, W. Peng, Z. Zhang, and Y. Pan, "Prediction of

essential proteins based on gene expression programming," *BMC genomics,* vol. 14, p. S7, 2013.

[30] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge & Data Engineering,* pp. 1263-1284, 2008.

[31] M. Li, Z. Fei, M. Zeng, F. Wu, Y. Li, Y. Pan*, et al.*, "Automated ICD-9 Coding via A Deep Learning Approach," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* pp. 1-1, 2018.

[32] 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 and chemistry,* vol. 35, pp. 143-150, 2011.

[33] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855-864.

[34] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111-3119.

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017, p. 12.

[38] A. Roy and S. Todorovic, "A multi-scale cnn for affordance segmentation in rgb images," in *European Conference on Computer Vision*, 2016, pp. 186-201.

[39] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, and J. Wang, "Automatic ICD-9 coding via deep transfer learning," *Neurocomputing,* 2018.

[40] H.-W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs*, et al.*, "MIPS: a database for genomes and protein sequences," *Nucleic acids research,* vol. 30, pp. 31-34, 2002.

[41] J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester*, et al.*, "SGD: Saccharomyces genome database," *Nucleic acids research,* vol. 26, pp. 73-79, 1998.

[42] 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, pp. 1152-1158, 2005.