IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 23, 2019, accepted September 13, 2019, date of publication September 23, 2019, date of current version October 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942843

Predicting Essential Proteins Based on Second-Order Neighborhood Information and Information Entropy

JIE ZHAO AND XIUJUAN LEI, (Member, IEEE)
School of Computer Science, Shaanxi Normal University, Xi'an 710119, China

Corresponding author: Xiujuan Lei (xjlei@snnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61972451, Grant 61672334, and Grant 61902230, and in part by the Fundamental Research Funds for the Central Universities, Shaanxi Normal University, under Grant GK201901010 and Grant GK201804006.

ABSTRACT Essential proteins are critical components of living organisms and indispensable to cellular life. Identification of essential proteins plays a critical role in the survival and development of life process and understanding the function of cell machinery. The experimental methods are usually costly and time-consuming. In order to overcome these limitations, many computational methods have been proposed to discover essential proteins based on the topological features of PPI networks and other biological information. In this paper, a new method named NIE is proposed to predict essential proteins based on second-order neighborhood information and information entropy of protein complex and subcellular localization. Firstly, a number of studies have shown that the RNA-Seq data is more advantageous than traditional gene expression data in predicting essential proteins. Meanwhile, the protein essentiality is closely related to the subcellular localization information, protein complex information and protein GO terms through data analysis. A weighted PPI network is constructed to reduce the impact of false positives and false negatives data on the identification of essential proteins, which integrates the GO terms information with Pearson correlation coefficient of RNA-Seq data. Secondly, the information entropy of protein complexes and subcellular localization is calculated to represent the biological characteristics of proteins. Furthermore, an information propagation model is constructed, which combines the biological properties of the proteins with the second-order neighborhood information in the PPI network to measure the essentiality of the proteins. In the experiments section, the proposed method is implemented on three common datasets (DIP, Krogan and MIPS) of Saccharomyces cerevisiae. A comparison study with other commonly used algorithms, including LAC, NC, PeC, WDC, UC, LIDC and LBCC is performed to evaluate the performance of NIE. The results show that the new method NIE has a better performance in predicting essential proteins.

INDEX TERMS Essential proteins, information entropy, neighborhood information, protein interaction networks.

I. INTRODUCTION

Essential proteins are those proteins to result in lethality or infertility of a cell if one of them has been deleted [1]. Essential proteins are closely related to the structure, function, and regulation of biological systems, play a very important role in the whole life of the cell. The loss of essential proteins causes the cells to become inactive, causing the cells to lose some functions, leading to pathological changes, affecting the survival and evolution of the living body [2], [3]. Identifying essential proteins and studying the properties and mechanisms of essential proteins have great significance in biology and pathology. Biological assays such as RNA interference [4], single gene knockout [5] and conditional gene knockout [6] methods can determine essential proteins accurately, but the experimental cost is high, and the experimental period is long. With the rapid developments of high-throughput technologies and computer technologies, it is a trend to predict and identify essential proteins through bioinformatics and computational biology methods.

The associate editor coordinating the review of this manuscript and approving it for publication was Trivikram Rao Molugu.

Many computational approaches have been proposed for predicting essential proteins.

Most proteins interact with each other to form a protein-protein interaction (PPI) network to participate in various life processes [7], [8], which makes it possible for us to identify essential proteins from a network level. From this point of view, many essential protein recognition methods based on the characteristics of PPI network are proposed. In the PPI network, nodes represent proteins and edges represent interactions between proteins, therefore, the six main indicators for measuring node centrality Degree Centrality (DC)[9], Betweenness Centrality (BC)[10], Closeness Centrality (CC)[11], Subgraph Centrality (SC)[12], Eigenvector Centrality (EC)[13], and Information Centrality (IC)[14] are used to identify essential proteins in the early stages of research. The node centrality provides a good research idea for finding essential proteins, but the accuracy is not ideal. Min *et al.* developed a new method for weighing protein-protein interactions based on the combination of logistic regression-based model and function similarity [15]. Li *et al.* proposed a method to determine the essentiality of proteins based on the local average connectivity (LAC) of nodes and its neighbors [16]. The above node centrality only considers the importance of the proteins, without considering the importance of the interactions between proteins. Wang *et al.* proposed a new centrality measure method NC [17] for identifying essential proteins based on edge clustering coefficient (ECC). Li *et al.* developed a new topology measure by defining and computing the value of each protein's topology potential [18].

The above methods analyze the essentiality of a protein only based on its network topology properties, without considering biological properties of proteins, and the PPI data have many false positives and false negatives which have a great influence on the accuracy of the methods [19]. In order to solve this shortcoming, many researchers began to integrate the network properties and biological properties of proteins to analyze the essentiality of proteins. Li et al. and Tang *et al.* combined with gene expression profiles to propose new centrality measure PeC [20] and WDC [21] based on edge clustering coefficient and Pearson correlation coefficient (PCC). In addition to the gene expression profiles of proteins, studies have shown that the protein complex information, the protein subcellular localization information, the protein domain information, the protein GO function annotation information, the protein orthologous information, and RNA-Seq information all have an impact on the essentiality of proteins. Peng *et al.* designed an iteration method ION for predicting essential protein by integrating the orthologous information with PPI networks [19]. Li *et al.* developed a united complex centrality (UC) for identification of essential proteins by integrating the protein complexes with the topological features of PPI networks [22]. Luo *et al.* proposed a method LIDC for predicting essential proteins used local interaction density combined with protein complexes based on statistical analyses of essential proteins and protein complexes [23]. Peng *et al.* proposed an algorithm UDoNC by integrating the protein domain data with protein-protein interaction data [24]. Li *et al.* developed an essential protein prediction method by integrating the information of subcellular localization, orthologous proteins and PPI networks, named SON [25]. Qin *et al.* proposed a method LBCC based on the combination of local density, betweenness centrality and in-degree centrality of complex [26]. Lei et al designed a new essential proteins prediction method RSG based on RNA-Seq, subcellular localization and GO annotation datasets [27]. These methods all use the network topology properties score and the biological properties score in PPI network to sort proteins, which can determine the essentiality of the proteins based on the score.

It should be pointed out that the construction of PPI networks also has a certain impact on the essentiality of proteins because the available PPI data contains many false positives and false negatives. Luo and Kuang introduce a new method named CDLC to predict essential proteins by integrating dynamic local average connectivity and in-degree of proteins in complexes [28]. Xiao *et al.* proposed a framework for identifying essential proteins from active PPI networks constructed with dynamic gene expression [29]. Li *et al.* constructed a refined protein interaction network TS-PIN by using time-course gene expression data and subcellular location information [30]. Shang *et al.* constructed integrated dynamic PPI networks use RNA-Seq datasets [31]. The identifying accuracy of these methods is higher than the simple use of PPI networks.

It must also be mentioned that some novel computational methods such as machine learning [32], deep learning and swarm intelligence optimization algorithm have also been applied to essential protein recognition. Zeng *et al.* proposed a deep learning framework to automatically learn biological features without prior knowledge [33]. Lei *et al.* designed essential proteins recognition model based on artificial fish swarm optimization [34] and flower pollination algorithm [35], respectively.

Although the above methods have made great progress, there are still many shortcomings and room for advancement in predicting essential proteins fields. In this paper, we developed a novel method to predicting essential proteins based on second-order neighborhood information and information extropy, named NIE. Proteins do not work alone[20]. Considering the co-expression and function-related properties of proteins, we first constructed a weighted PPI network using RNA-Seq data and GO functional annotation data. Then, we calculated the information entropy of the protein complex data and subcellular localization data as the feature space of the protein. Finally, matrix operations are used to obtain the second-order neighborhood information of each protein as the final feature score. To test the performance of the proposed method NIE, we carry out experiments on three PPI network of Saccharomyces cerevisiae and compared with seven other competing methods: LAC, NC, PeC, WDC, UC, LIDC and LBCC. The experimental results show that NIE outperform the seven previously proposed methods.

II. METHODS

A. CONSTRUCTING WEIGHTED PPI NETWORK

The PPI network is usually described as an undirected graph $G(V, E)$, the set of nodes $V = \{v_1, v_2, \ldots, v_n\}$ represent the proteins and the set of edges $E = \{e(v_i, v_j)\}$ represents the interactions between protein v_i and v_j. Studies have shown that the available PPI data is incomplete and contains many false positives and false negatives, which impacts the correctness of discovering essential proteins. However, there are other biological properties that can help us analyze the relationship between proteins. The previous studies have shown that the essential proteins tend to interact with each other and functionally related [27]. With the rapid development of high-throughput sequencing technology, RNA-seq datasets have provided us with a new way to study gene expression profiles [31]. The RNA-Seq data has a larger dynamic range and a better ability to detect and quantify unknown transcripts and subtypes [36], thus better describing the co-expression characteristics between genes. We use RNA-Seq data to measure the gene expression intensity and GO functional annotation data to measure functional similarity between proteins and construct a weighted PPI network.

The Person correlation coefficient (PCC) was calculated to evaluate how strong two interacting proteins are co-expression [37]. The PCC value of a pair of RNA-Seq $x = \{x_1, x_2, \ldots, x_m\}$ and $y = \{y_1, y_2, \ldots, y_m\}$, which encode the corresponding paired protein v_i and v_j interacting in the PPI network, is defined as:

$$PCC(v_i, v_j) = \frac{\sum_{k=1}^{m} (x_k - \mu(x))(y_k - \mu(y))}{\sqrt{\sum_{k=1}^{m} (x_k - \mu(x))^2}\sqrt{\sum_{k=1}^{m} (y_k - \mu(y))^2}} \quad (1)$$

where $\mu(x)$ and $\mu(y)$ are the mean RNA-Seq expression value of proteins v_i and v_j, respectively. The value of PCC ranges from -1 to 1, if $PCC(v_i, v_j)$ is a positive value, there is a positive correlation between protein v_i and v_j, if $PCC(v_i, v_j)$ is a negative value, there is a negative correlation between protein v_i and v_j.

The GO functional annotation data is a biological resource that describes the functional properties of genes, if two interacted proteins v_i and v_j have many common GO terms, their functions are more similar. For protein v_i and v_j, the GO similarity between them was computed as follows [27]:

$$GO_{sim}(v_i, v_j) = \frac{|GO_i \cap GO_j|}{|GO_i \cup GO_j|} \quad (2)$$

where GO_i and GO_j represent the GO terms for proteins v_i and v_j, $|GO_i \cap GO_j|$ and $|GO_i \cup GO_j|$ represent the number of GO terms in the intersection and union of GO_i and GO_j. Based on the PCC and GO_{sim}, we build a weighted PPI network (WPIN), the weight between protein v_i and v_j is defined as follows:

$$Weight(v_i, v_j) = PCC(v_i, v_j) \times GO_{sim}(v_i, v_j) \quad (3)$$

B. INFORMATION ENTROPY

Proteins have many biological information, such as protein complex information and subcellular localization information, which are related to the essentiality of proteins. Information entropy is often used as a quantitative indicator of the information content of a system [38]. We use information extropy to calculate the amount of information about protein biological characteristics. The measure of information entropy associated with each possible data value is the negative logarithm of the probability mass function for the value [39]:

$$H(Z) = -\sum_{z \in Z} p(z) \log(p(z)) \quad (4)$$

where Z represents the event, z represents the possible value of the event, and the corresponding probability is $p(z)$. We separately calculated the information entropy of protein complexes and subcellular localization and used them as eigenvalues of proteins.

1) INFORMATION ENTROPY OF PROTEIN COMPLEX

Protein often bind together to form complexes to carry out their biological functions. Mutation or destruction of essential proteins will cause the organism to lose some function, so the essential protein is closely related to protein complexes. The set $C = \{c_1, c_2, \ldots, c_{nc}\}$ is standard protein complexes, where nc is the number of protein complexes, $c_j = \{v_1, v_2, \ldots, v_p\}, j = 1, 2, \ldots, nc$ is the jth protein complex in C, p is the number of proteins contained in the protein complex c_j, the probability of protein complex c_j is calculated as follows:

$$p(c_j) = \frac{|c_j|}{|C|} \quad (5)$$

where $|c_j|$ is the number of proteins contained in protein complex c_j, $|C|$ is the number of proteins contained in the standard protein complex C.

For each protein $v_i \in V$, we construct an nc-dimensional vector $ComInf_i(1 \times nc)$ to represent the protein associated complex information: $ComInf_i = [0\ 1\ 0\ 1\ \ldots\ 1\ 0\ 0\ 1\ 0\ 1]$

$$ComInf_i(j) = \begin{cases} 1, & v_i \in c_j \\ 0, & v_i \notin c_j \end{cases} \quad (6)$$

the information entropy of protein complex of protein v_i is defined as follows:

$$HC(v_i) = -\sum_{j=1}^{\mathrm{nc}} ComInf_i(j) p(c_j) \log(p(c_j)) \quad (7)$$

2) INFORMATION ENTROPY OF SUBCELLULAR LOCALIZATION

The set $S = \{s_1, s_2, \ldots, s_{ns}\}$ is subcellular localization data, where ns is the number of subcellular localization, $s_k = \{v_1, v_2, \ldots, v_q\}, k = 1, 2, \ldots, ns$ is the kth subcellular localization in S, q is the number of proteins contained in subcellular localization s_k, the probability of subcellular localization

s_k is calculated as follows:

$$p(s_k) = \frac{|s_k|}{|S|} \tag{8}$$

where $|s_k|$ is the number of proteins contained in subcellular localization s_k, $|S|$ is the number of proteins contained in the subcellular localization data S.

For each protein $v_i \in V$, we construct an ns-dimensional vector $SubInf_i$ $(1 \times ns)$ to represent the protein associated subcellular localization information:

$$SubInf_i = [1\,0\,0\,1\,\ldots\,1\,0\,1]$$

$$SubInf_i(k) = \begin{cases} 1, & v_i \in s_k \\ 0, & v_i \notin s_k \end{cases} \tag{9}$$

the information entropy of subcellular localization of protein v_i is defined as follows:

$$HS(v_i) = -\sum_{k=1}^{ns} SubInf_i(k)p(s_k)\log(p(s_k)) \tag{10}$$

For each protein $v_i \in V$, after the information entropy of protein complexes $HC(v_i)$ and the information entropy of subcellular localization $HS(v_i)$ is calculated, we build the feature matrix F for the set of nodes $V = \{v_1, v_2, \ldots, v_n\}$:

$$\begin{cases} f(v_i) = [HC(v_i)\; HS(v_i)] \\ F = [f(v_1), f(v_2), \ldots, f(v_n)] \end{cases} \tag{11}$$

C. PREDICTING ESSENTIAL PROTEINS BASED ON NIE

According to the established weighted PPI network (WPIN), we established a second-order neighborhood information-based propagation model to calculate the score matrix $Score$ of the proteins. A is the adjacency matrix of the WPIN, I_n is identity matrix:

$$\begin{cases} M = A + I_n \\ F_1 = M \times F \\ Score = M \times F \quad {}_1 \end{cases} \tag{12}$$

For each protein $v_i \in V$, the final score is calculated as follows:

$$score(v_i) = \sum Score(i,\ :) \tag{13}$$

In summary, the pseudocode of algorithm NIE is shown in Table 1.

TABLE 1. The pseudocode of algorithm NIE.

```
Algorithm 1: NIE algorithm
Input: The PPI network $G(V, E)$;
       The RNA-Seq data: RS;
       The GO annotation data: $GO$;
       The protein complex data: $C$;
       The subcellular localization data: $S$;
Output: Top $K$ of proteins sorted by $score$ in descent order
Begin
    For each $e(v_i, v_j) \in E$ do
        Calculate $PCC(v_i,v_j)$ by Equation (1) using RNA-Seq data $RS$ ;
        Calculate $GO_{sim}(v_i,v_j)$ by Equation (2) using GO annotation data $GO$;
        Calculate $Weight(v_i,v_j)$ by Equation (3);
    End for
    Construct weighted PPI network WPIN
    Calculate the adjacency matrix $A$ of WPIN
    For each $v_i \in V$ do
        Calculate $HC(v_i)$ by Equation (7) using protein complex data $C$;
        Calculate $HS(v_i)$ by Equation (10) using subcellular localization data $S$;
    End for
    Calculate feature matrix $F$ by Equation (11)
    Calculate score matrix $Score$ by Equation (12)
    For each $v_i \in V$ do
        Calculate final score $score(v_i)$ by Equation (13)
    End for
    Sort $V$ by the value of $score(v_i)$;
    output top $K$ proteins ($K$=100, 200, 300, 400, 500, 600, respectively)
End
```

III. RESULTS AND DISCUSSION

In order to evaluate the accuracy and the efficiency of the proposed NIE algorithm, we implemented it on three common datasets DIP, Krogan and MIPS in Matlab R2015b and executed on a quad-core processor 3.30GHz PC with 8G RAM. For comparison, we implemented seven classic methods LAC, NC, PeC, WDC, UC, LIDC and LBCC. Among these seven methods, LAC and NC rely on the network characteristics of the PPI network, PeC and WDC integrate gene expression data, and UC, LIDC and LBCC integrate protein biometric data. There are also six classic network centrality methods DC, IC, EC, SC, BC, CC and the literature has proven that these methods are not very accurate. Therefore, in this article, we do not compare these six methods. First, the results are sorted in descending order. Then, we select the top ranked proteins as candidate essential proteins respectively. Finally, comparing the candidate essential proteins with the standard essential proteins to judge how many candidates are true essential proteins.

We adopted five types of popular comparison methodologies to evaluate the algorithm performance: 1) Histogram comparison methodology. 2) Jackknife curves methodology. 3) Precision-recall curves methodology. 4) ROC curves methodology. 5) Statistical measures.

A. EXPERIMENTAL DATA

All the data used were Saccharomyces cerevisiae. The experimental data used in this paper are as following:

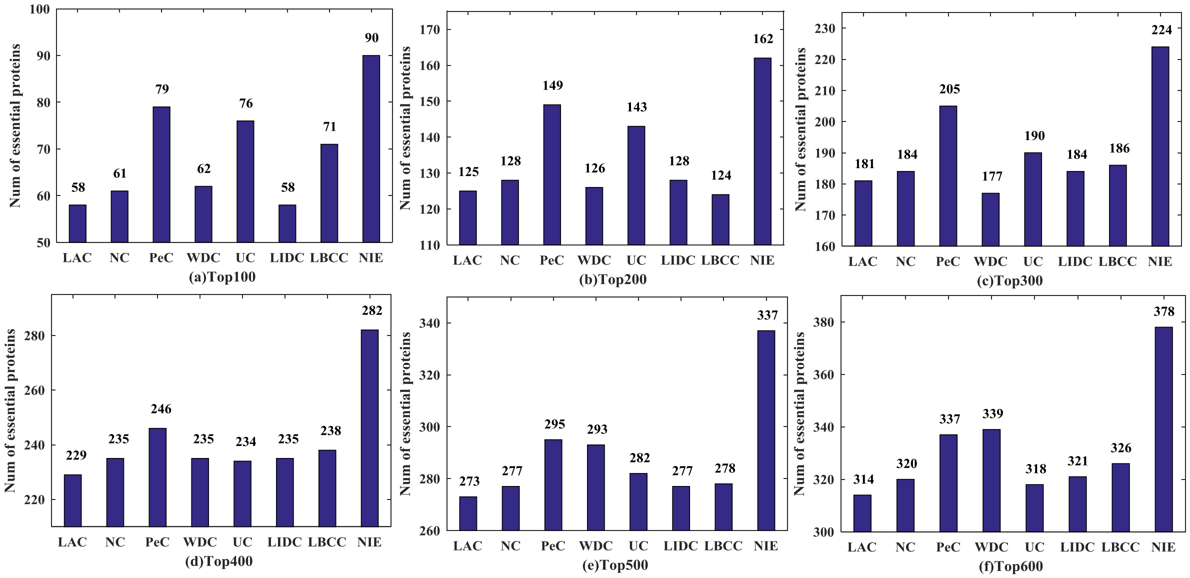

FIGURE 1. Comparison of the number of essential proteins detected by NIE and other seven methods in DIP dataset (a) Top 100 (b) Top 200 (c) Top 300 (d) Top 400 (e) Top 500 (f) Top 600.

1) PPI network data. Three commonly used protein interaction databases DIP [40](version of 20160114), Krogan [41] and MIPS [42] were employed. After preprocessing, the DIP database contains 5028 proteins and 22302 interactions, the network density is 0.0018, the Krogan database contains 2674 proteins and 7075 interactions, the network density is 0.0020, the MIPS database contains 4546 proteins and 12319 interactions, the network density is 0.0012.

2) RNA-Seq data. The RNA-Seq data is collected from the NCBI SPA database (SRX362640)[43]. The data contains 7108 genes at 12 time points after removing the unavailable ones, with 4957 genes involved in the DIP dataset, 2673 genes involved in the Krogan dataset and 4531 genes involved in the MIPS dataset.

3) GO annotation data. The GO database is currently one of most comprehensive ontology databases in bioinformatics. GO-slims are cut-down versions of the GO ontologies containing a subset of the terms in the whole GO [44]. They give a broad overview of the ontology content without the detail of the specific fine-grained terms. Among the 7014 proteins in the GO-slims data, 4939 proteins present in the DIP dataset,2671 proteins present in the Krogan dataset and 4508 proteins present in the MIPS dataset.

4) Protein complex data. CYC2008[45] is a comprehensive catalogue of manually curated 408 heteromeric protein complexes in S. Cerevisiae reliably backed by small-scale experiments from the literature. There are1472, 1152 and 1369 proteins in the DIP, Krogan and MIPS datasets appear in the CYC2008 dataset, respectively

5) Subcellular localization data. The subcellular localization information of proteins was retrieved from knowledge channel of COMPARTMENTS database (April 6, 2017)[46]. After preprocessing, there are 4908 proteins, which could be classed into 11 categories: 1) Cytoskeleton, 2) Cytosol, 3) Endoplasmic, 4) Endosome, 5) Extracellular, 6) Golgi, 7) Mitochondrion, 8) Nucleus, 9) Peroxisome, 10) Plasma and 11) Vacuole. There are 4054, 2350 and 3654 proteins in the DIP, Krogan and MIPS datasets have subcellular localization data, respectively

6) Standard essential protein data. The essential proteins data was selected form SGD [47], DEG [48]. The essential proteins data includes 1285 essential proteins [25]. Among the 1285 essential proteins, 1152, 784 and 1016 essential proteins present in DIP, Krogan and MIPS dataset, respectively.

7) Gene expression data. Gene expression data was retrieved from GEO (Gene Expression Omnibus) with accession number GSE3431[21]. The data contained 7074 genes at 36 time points in the 3 cell life cycles after preprocessing, with 4876, 2644 and 4445 genes involved in the DIP, Krogan and MIPS datasets, respectively. The GEO data is used to implement PeC algorithm.

Detailed data intersection information of experimental data is shown in Table 2.

B. PERFORMANCE COMPARISON WITH HISTOGRAM METHODOLOGY

The performance of NIE is compared with seven existing methods LAC, NC, PeC, WDC, UC, LIDC and LBCC. We select top 100, 200, 300, 400, 500, 600 ranked proteins predicted by LAC, NC, PeC, WDC, UC, LIDC, LBCC and NIE as candidate essential proteins respectively. Fig. 1, Fig.2 and Fig. 3 show the results of the comparison in DIP, Krogan and MIPS dataset respectively. From Fig. 1, Fig. 2 and Fig. 3, it is easy to see that the NIE outperforms LAC, NC, PeC, WDC, UC, LIDC and LBCC obviously.

TABLE 2. The data intersection information of the experimental data.

PPI network	RNA-Seq data	GO annotation data	Protein complex data	Subcellular localization data	Essential proteins data	Gene expression data
DIP (5028 proteins)	4957	4939	1472	4054	1152	4876
Krogan (2674 proteins)	2673	2671	1152	2350	784	2644
MIPS (4546 proteins)	4531	4508	1369	3654	1016	4445

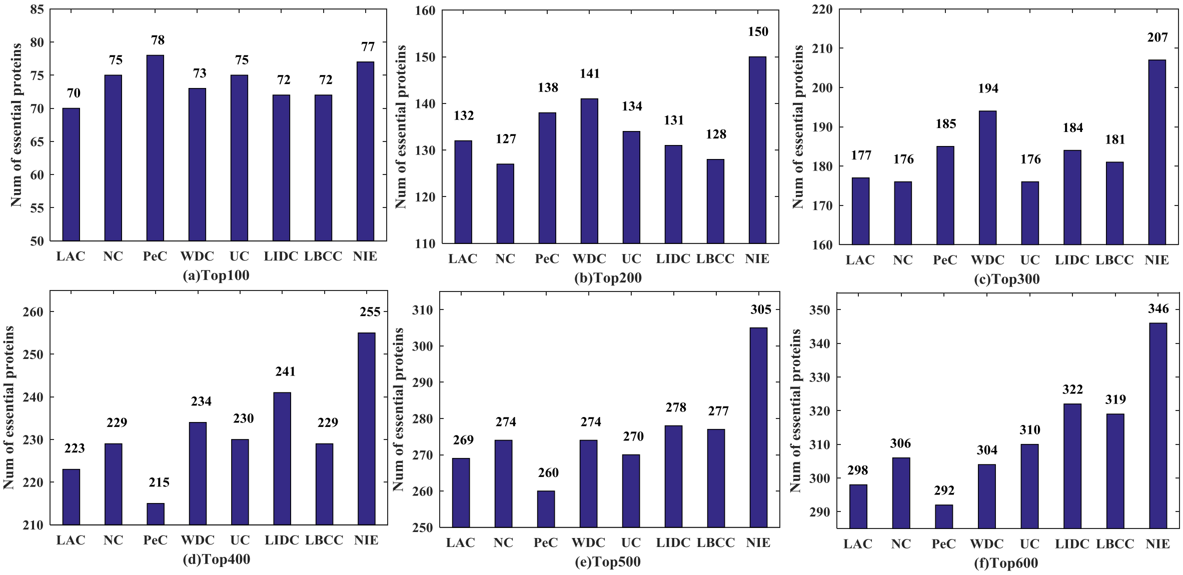

FIGURE 2. Comparison of the number of essential proteins detected by NIE and other seven methods in Krogan dataset (a) Top 100 (b) Top 200 (c) Top 300 (d) Top 400 (e) Top 500 (f) Top 600.

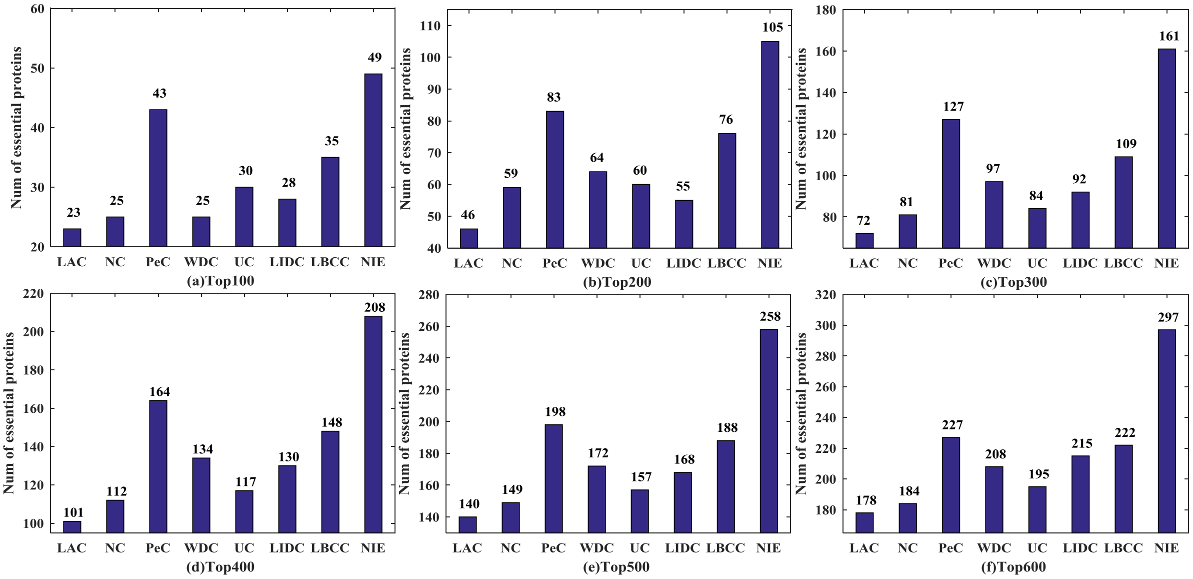

FIGURE 3. Comparison of the number of essential proteins detected by NIE and other seven methods in MIPS dataset (a) Top 100 (b) Top 200 (c) Top 300 (d) Top 400 (e) Top 500 (f) Top 600.

In the DIP dataset, among the top 100 predicted essential proteins, the NIE method correctly predicted 90 essential proteins, LAC and LIDC only correctly predicted 58. The more the predicted correct number of top ranked proteins, the more prominent the advantage of NIE. The accuracy of the NIE method is much higher than the other seven methods.

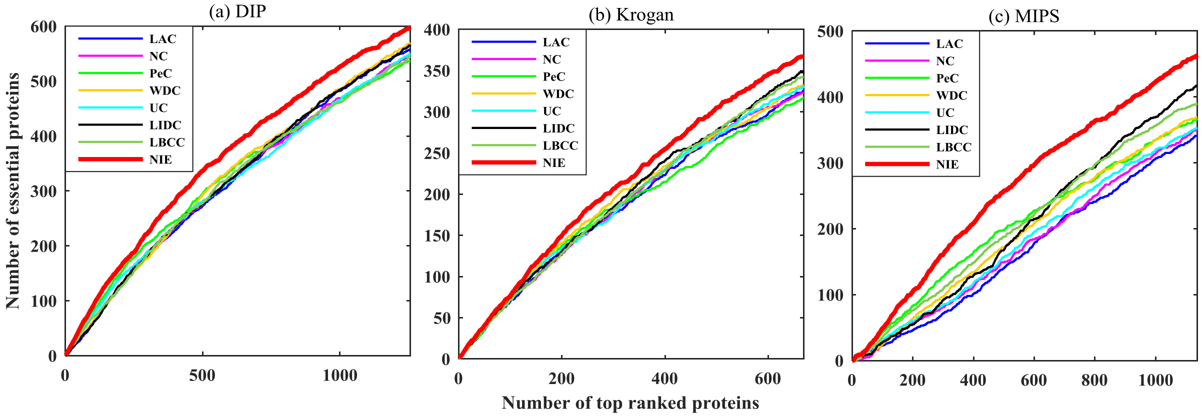

FIGURE 4. Jackknife curves of NIE and other seven methods in (a) DIP dataset (b) Krogan dataset and (c) MIPS dataset.

It can also be seen from the Fig. 1, Fig. 2 and Fig. 3 that the accuracy of the method based on network characteristics such as LAC and NC are not as good as other methods. Owing to the edge aggregation coefficient and characteristics of gene expression are fully combined, the prediction performance of PeC and WDC algorithms is highly improved. However, in the DIP and MIPS datasets, the PeC method is better than WDC, but in the Krogan dataset, WDC performance is better than PeC. As we all know, the density of Krogan is greater than DIP and MIPS. Therefore, the performance of the PeC and WDC methods have a certain relationship with the structure of the PPI network. The NIE method performs consistently on three datasets.

C. VALIDATED BY JACKKNIFE METHODOLOGY

In order to further analyze the stability of NIE and other seven algorithms, we adopt Jackknife methodology developed by Holman et al. [49]. Jackknife methodology is an evaluation method for continuously displaying the prediction results of algorithms. The experimental results validated by Jackknife methodology in DIP, Krogan and MIPS are shown in Fig. 4. The horizontal axis represents the number of candidate essential proteins, and the vertical axis represents the predicted correct number of candidate essential proteins, which can reflect the correctness of the methods. In the Histogram comparison methodology, we only analyzed the top 100, 200, 300, 400, 500 and 600 prediction results, but it does not explain the stability of the algorithm well. In the comparison of Jackknife, we compared the top 25% of the algorithm results. Fig. 4(a) shows the Jackknife curve in DIP dataset, the horizontal axis represents top ranked proteins range from 0 to 1257. Fig. 4(b) shows the Jackknife curve in Krogan dataset, the horizontal axis represents top ranked proteins range from 0 to 669. Fig. 4(c) shows the Jackknife curve in MIPS dataset, the horizontal axis represents top ranked proteins range from 0 to 1137.

As can be seen from Fig. 4, on DIP, Krogan and MIPS datasets, the correctness of the NIE algorithm is much greater than other algorithms. This result is consistent with Fig. 1, Fig. 2 and Fig. 3. The density of DIP is 0.0018, the density of Krogan is 0.0020, the density of MIPS is 0.0012. In these three datasets, the correctness of the methods PeC and WDC are better than LIDC and LBCC before about top 15% of the results, but the correctness between top 15% and top 25% is lower than LIDC and LBCC. The density of the three datasets are different, but the method NIE has higher correctness throughout the whole 25% of the results, which fully demonstrates that the NIE method has good stability.

D. VALIDATED BY PRECISION-RECALL CURVES METHODOLOGY

The histogram comparison methodology and the Jackknife curves methodology can reflect the accuracy and stability of the methods in a local range, in order to measure the overall performance of the methods, we have drawn the *Precision-Recall* curve (PR-curve) as shown in Fig, 5. The horizontal and vertical axis represents *recall* and *precision* respectively, defined as follows:

$$precision = \frac{TP}{TP + FP} \tag{14}$$

$$recall = \frac{TP}{TP + FN} \tag{15}$$

where *TP* is true positives, refers to essential proteins correctly predicted as essential, *FP* is false positives, refers to nonessential proteins incorrectly predicted as essential, *FN* is false negatives, refers to essential proteins incorrectly predicted as nonessential. *Precision* indicates the predicted correct ratio of the predicted essential proteins and *recall* indicates the proportion of the standard essential proteins that are predicted correctly, the greater the *precision* and *recall*, the better the classification performance of the method.

In order to plot the PR curve, firstly, the method results are sorted in descending order. Secondly, the former *k* proteins are treated as essential proteins, and the rest are treated as non-essential proteins, and the corresponding *precision* and *recall* values are calculated. The value of *k* ranges from

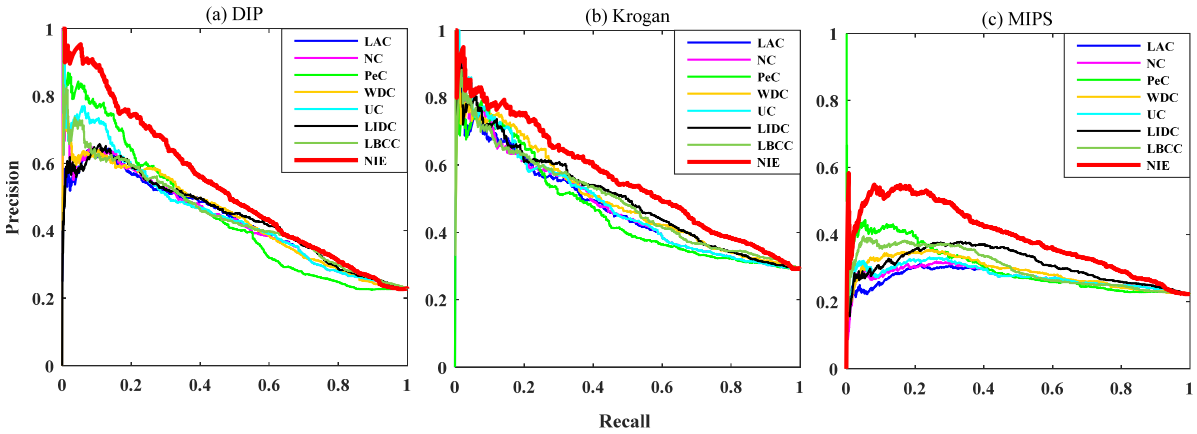

FIGURE 5. Precision-recall curves of NIE and other seven methods in (a) DIP dataset (b) Krogan dataset and (c) MIPS dataset.

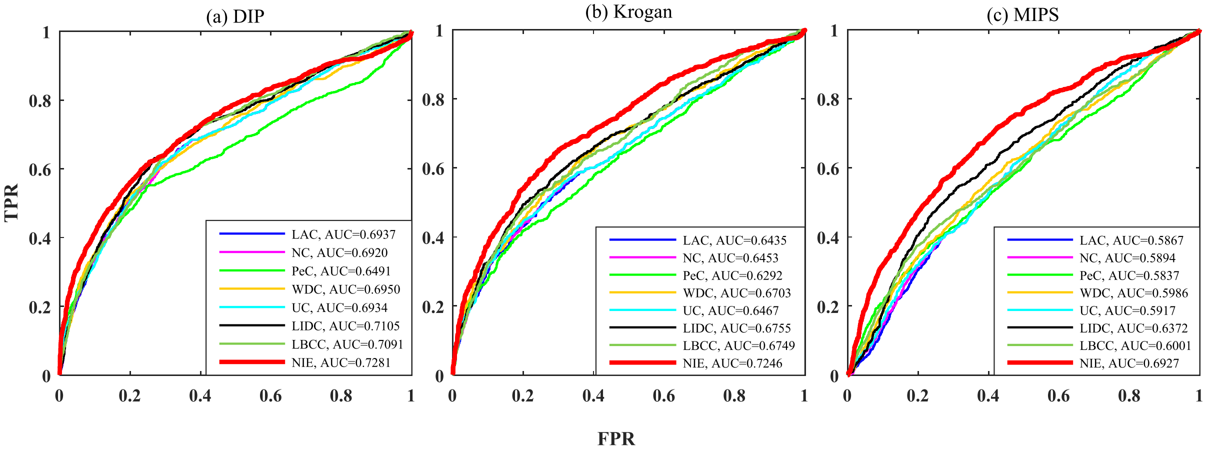

FIGURE 6. ROC curves of NIE and other seven methods in (a) DIP dataset (b) Krogan dataset and (c) MIPS dataset.

1 to the number of the proteins. Thirdly, draw the PR curve according to the *precision* and *recall*.

Fig. 5(a), Fig. 5(b) and Fig. 5(c) show the PR curve of NIE and other methods in DIP, Krogan and MIPS dataset respectively. In the DIP and MIPS datasets, the precision value of PeC is better than LIDC method, and then lower than LIDC method, which is consistent with the trend of the Jackknife curve. Similarly, the performance of the WDC, UC and LBCC methods differ in different datasets. As can be seen from Fig. 5, the PR curve of NIE achieves the best performance among all the methods and in all the three datasets.

E. VALIDATED BY ROC CURVES METHODOLOGY

To further measure the performance of the methods, the Receiver Operating Characteristic (ROC) curve has been used. ROC curve is a good way of visualizing a classifier's performance in order to select a suitable operating point, or decision threshold [50]. The ROC curves demonstrate the tradeoff between the false positive rate (FPR) and the true positive rate (TPR).

$$FPR = \frac{FP}{FP + TN} \tag{16}$$

$$TPR = \frac{TP}{TP + FN} \tag{17}$$

where TN is true negatives, refers to nonessential proteins correctly predicted as nonessential. FP, TP and FN have described in the previous section. The area under the ROC curves (AUC) is used to measure the performance of corresponding algorithms, the bigger the area is, the better prediction performance the algorithm has.

Fig. 6(a), Fig. 6(b) and Fig. 6(c) show the comparison of ROC curves and AUC values of different methods in DIP, Krogan and MIPS datasets, respectively. In these three datasets, the AUC value of the NIE method are all the largest 0.7281, 0.7246, 0.6927, and the AUC value of the PeC method are all the smallest 0.6491, 0.6292, 0.5837. It demonstrated that the NIE method is excellent and robust.

F. EVALUATION USING STATISTICAL MEASURES

As can be seen from the histogram and Jackknife curve, the accuracy of the PeC method is higher than that of LBCC and LIDC method, but the ROC value is the smallest. Therefore, we must comprehensively consider the performance of the methods. We compare the following statistical measures

TABLE 3. Comparison of the values of *SN*, *SP*, *PPV*, *NPV*, *F-measure* and *ACC* of NIE and other seven methods.

Dataset	Methods	*SN*	*SP*	*PPV*	*NPV*	*F-measure*	*ACC*
DIP	LAC	0.4922	0.8147	0.4412	0.8437	0.4653	0.7408
	NC	0.4844	0.8124	0.4342	0.8413	0.4579	0.7372
	PeC	0.4714	0.8085	0.4226	0.8373	0.4456	0.7313
	WDC	0.4983	0.8165	0.4467	0.8455	0.4711	0.7436
	UC	0.4870	0.8132	0.4366	0.8421	0.4604	0.7384
	LIDC	0.5043	0.8183	0.4521	0.8474	0.4768	0.7464
	LBCC	0.4809	0.8114	0.4311	0.8402	0.4547	0.7356
	NIE	**0.5278**	**0.8253**	**0.4732**	**0.8546**	**0.4990**	**0.7571**
Krogan	LAC	0.6135	0.5744	0.3743	0.7817	0.4650	0.5859
	NC	0.6135	0.5744	0.3743	0.7817	0.4650	0.5859
	PeC	0.5982	0.5680	0.3650	0.7731	0.4534	0.5769
	WDC	0.6607	0.5940	0.4031	0.8084	0.5007	0.6135
	UC	0.6135	0.5744	0.3743	0.7817	0.4650	0.5859
	LIDC	0.6633	0.5950	0.4047	0.8098	0.5027	0.6150
	LBCC	0.6492	0.5892	0.3961	0.8019	0.4920	0.6068
	NIE	**0.7015**	**0.6109**	**0.4280**	**0.8314**	**0.5317**	**0.6375**
MIPS	LAC	0.3750	0.7438	0.2965	0.8052	0.3312	0.6614
	NC	0.3819	0.7458	0.3019	0.8074	0.3372	0.6645
	PeC	0.3917	0.7487	0.3097	0.8104	0.3459	0.6689
	WDC	0.3967	0.7501	0.3136	0.8120	0.3503	0.6711
	UC	0.3799	0.7453	0.3004	0.8067	0.3355	0.6636
	LIDC	0.4577	0.7676	0.3619	0.8310	0.4042	0.6983
	LBCC	0.4203	0.7569	0.3323	0.8193	0.3711	0.6816
	NIE	**0.4980**	**0.7793**	**0.3938**	**0.8436**	**0.4398**	**0.7164**

including sensitivity (*SN*), specificity (*SP*), positive predictive value (*PPV*), negative predictive value (*NPV*), F-measure and accuracy (*ACC*), defined as follows:

$$SN = \frac{TP}{TP + FN} \tag{18}$$

$$SP = \frac{TN}{TN + FP} \tag{19}$$

$$PPV = \frac{TP}{TP + FP} \tag{20}$$

$$NPV = \frac{TN}{TN + FN} \tag{21}$$

$$F - measure = \frac{2 * SN * PPV}{SN + PPV} \tag{22}$$

$$ACC = \frac{TP + TN}{P + N} \tag{23}$$

where *P* and *N* are the total number of essential proteins and nonessential proteins, respectively. *FP*, *TP*, *TN* and *FN* have described in the previous section.

The comparison of *SN*, *SP*, *PPV*, *NPV*, *F-measure* and *ACC* of NIE and other methods are shown in Table 3. Because the essential proteins data includes 1285 essential proteins, we performed a statistical analysis of top 1285 predicted essential proteins predicted by NIE and other seven methods. As shown in Table 3, it is obvious that the *SN*, *SP*, *PPV*, *NPV*, *F-measure* and *ACC* of NIE are higher than that of any other methods on three different datasets, which shows that NIE can identify essential proteins more accurately.

IV. CONCLUSION

Essential proteins help us analyze and understand life activities, meanwhile, discovering essential proteins are useful for disease prediction and drug design. However, PPI data have many false positives and false negatives which have a great influence on the accuracy of predicting essential proteins. So firstly we used RNA-seq data to describe the co-expression properties between proteins, and used GO annotation data to describe the functional correlation between proteins, and then combined with the PPI network to construct weighted PPI network to reduce the impact of false positives and false negatives data on the identification of essential proteins. Furthermore, proteins have different functions at different locations, the same protein has different functions at different locations, and proteins often join together to form a complex to perform a function. Therefore, we calculate the subcellular localization information entropy and protein complex information entropy of proteins, respectively, and use them to construct the feature matrix of proteins. Finally, we established a computational model based on second-order neighborhood information to identify essential proteins named NIE. To verify the accuracy, stability, and effectiveness of the NIE method, we performed experiments on three datasets: DIP, Krogan and MIPS, and compared them with seven classical essential protein prediction methods: LAC, NC, PeC, WDC, UC, LIDC and LBCC. Five types of popular comparison methodologies used to compare algorithm performance. The experimental results clearly indicate that NIE can better performance than seven other methods.

REFERENCES

[1] B. Zhao, J. Wang, X. Li, and F.-X. Wu, "Essential protein discovery based on a combination of modularity and conservatism," *Methods*, vol. 110, pp. 54–63, Nov. 2016.

[2] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, and R. Bangham, "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis," *Science*, vol. 285, no. 5429, pp. 901–906, 1999.

[3] L. M. Steinmetz, C. Scharfe, A. M. Deutschbauer, D. Mokranjac, Z. S. Herman, T. Jones, A. M. Chu, G. Giaever, H. Prokisch, P. J. Oefner, and R. W. Davis, "Systematic screen for human disease genes in yeast," *Nature Genet.*, vol. 31, no. 4, pp. 400–404, 2002.

[4] L. M. Cullen and G. M. Arndt, "Genome-wide screening for gene function using RNAi in mammalian cells," *Immunol. Cell Biol.*, vol. 83, no. 3, pp. 217–223, Jun. 2005.

[5] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, and S. Dow, "Functional profiling of the Saccharomyces cerevisiae genome," *Nature*, vol. 418, no. 6896, pp. 387–391, 2002.

[6] T. Roemer, B. Jiang, J. Davison, T. Ketela, K. Veillette, A. Breton, F. Tandia, A. Linteau, S. Sillaots, C. Marta, N. Martel, S. Veronneau, S. Lemieux, S. Kauffman, J. Becker, R. Storms, C. Boone, H. Bussey, "Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery," *Mol. Microbiol.*, vol. 50, no. 1, pp. 167–181, Oct. 2003.

[7] A. Sikandar, W. Anwar, U. I. Bajwa, X. Wang, M. Sikandar, L. Yao, Z. L. Jiang, and Z. Chunkai, "Decision tree based approaches for detecting protein complex in protein protein interaction network (PPI) via link and sequence analysis," *IEEE Access*, vol. 6, pp. 22108–22120, 2018.

[8] J. Yang, H. Pu, L. Jian, J. Gu, and M. Fan, "Modeling and analysis of protein synthesis and DNA mutation using colored Petri nets," *IEEE Access*, vol. 6, pp. 22386–22400, 2018.

[9] M. W. Hahn and A. D. Kern, "Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks," *Mol. Biol. Evol.*, vol. 22, no. 4, pp. 803–806, 2005.

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

[11] S. Wuchty and P. F. Stadler, "Centers of complex networks," *J. Theor. Biol.*, vol. 223, no. 1, pp. 45–53, 2003.

[12] E. Estrada and J. A. Rodríguez-Velázquez, "Subgraph centrality in complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 71, May 2005, Art. no. 056103.

[13] P. Bonacich, "Power and centrality: A family of measures," *Amer. J. Sociol.*, vol. 92, no. 5, pp. 1170–1182, 1987.

[14] K. Stephenson and M. Zelen, "Rethinking centrality: Methods and examples," *Soc. Netw.*, vol. 11, no. 1, pp. 1–37, 1989.

[15] L. Min, J. Wang, H. Wang, and P. Yi, "Essential proteins discovery from weighted protein interaction networks," in *Proc. Int. Symp. Bioinf. Res. Appl.*, 2010, pp. 89–100.

[16] 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," *Comput. Biol. Chem.*, vol. 35, no. 3, pp. 143–150, Jun. 2011.

[17] J. X. Wang, M. Li, H. Wang, and Y. Pan, "Identification of essential proteins based on edge clustering coefficient," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 4, pp. 1070–1080, Jul./Aug. 2012.

[18] M. Li, Y. Lu, J. Wang, F. X. Wu, and Y. Pan, "A topology potential-based method for identifying essential proteins from PPI networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 2, pp. 372–383, Mar./Apr. 2015.

[19] W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, and Y. Pan, "Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks," *BMC Syst. Biol.*, vol. 6, no. 1, p. 87, Jul. 2012.

[20] 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 Syst. Biol.*, vol. 6, no. 1, p. 15, 2012.

[21] X. Tang, J. Wang, J. Zhong, and Y. Pan, "Predicting essential proteins based on weighted degree centrality," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 2, pp. 407–418, Mar. 2014.

[22] M. Li, Y. Lu, Z. Niu, and F.-X. Wu, "United complex centrality for identification of essential proteins from PPI networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 2, pp. 370–380, Mar./Apr. 2017. doi: 10.1109/TCBB.2015.2394487.

[23] J. Luo, Q. Yi, and C. Peter, "Identification of essential proteins based on a new combination of local interaction density and protein complexes," *PLoS ONE*, vol. 10, no. 6, 2015, Art. no. e0131418.

[24] W. Peng, J. Wang, Y. Cheng, "UDoNC: An algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 2, pp. 276–288, Mar./Apr. 2015.

[25] G. Li, M. Li, J. Wang, J. Wu, F.-X. Wu, and Y. Pan, "Predicting essential proteins based on subcellular localization, orthology and PPI networks," *BMC Bioinf.*, vol. 17, no. 8, p. 279, 2016.

[26] C. Qin, Y. Sun, and Y. Dong, "A new method for identifying essential proteins based on network topology properties and protein complexes," *PLoS ONE*, vol. 11, no. 8, 2016, Art. no. e0161042.

[27] X. Lei, J. Zhao, H. Fujita, and A. Zhang, "Predicting essential proteins based on RNA-Seq, subcellular localization and GO annotation datasets," *Knowl.-Based Syst.*, vol. 151, pp. 136–148, Jul. 2018.

[28] J. Luo and L. Kuang, "A new method for predicting essential proteins based on dynamic network topology and complex information," *Comput. Biol. Chem.*, vol. 52, pp. 34–42, Oct. 2014.

[29] Q. Xiao, J. Wang, X. Peng, F.-X. Wu, and Y. Pan, "Identifying essential proteins from active PPI networks constructed with dynamic gene expression," *Bmc Genomics*, vol. 16, no. 3, p. S1, 2015.

[30] 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 Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1386–1397, Jul./Aug. 2019.

[31] X. Shang, Y. Wang, and B. Chen, "Identifying essential proteins based on dynamic protein-protein interaction networks and RNA-Seq datasets," *Sci. China Inf. Sci.*, vol. 59, no. 7, Jul. 2016, Art. no. 070106 .

[32] X. Lei, X. Yang, and H. Fujita, "Random walk based method to identify essential proteins by integrating network topology and biological characteristics," *Knowl.-Based Syst.*, vol. 167, pp. 53–67, Mar. 2019.

[33] M. Zeng, M. Li, Z. Fei, F. Wu, Y. Li, Y. Pan, and J. Wang, "A deep learning framework for identifying essential proteins by integrating multiple types of biological information," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.

[34] X. Lei, Y. Xiaoqin, and W. Fangxiang, "Artificial fish swarm optimization based method to identify essential proteins," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.

[35] X. Lei, M. Fang, F.-X. Wu, and L. Chen, "Improved flower pollination algorithm for identifying essential proteins," *Bmc Syst. Biol.*, vol. 12, no. 4, p. 46, 2018.

[36] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[37] X. Lei, Y. Ding, H. Fujita, and A. Zhang, "Identification of dynamic protein complexes based on fruit fly optimization algorithm," *Knowl.-Based Syst.*, vol. 105, pp. 270–277, Aug. 2016.

[38] G.-Y. Wang, H. Yu, and D.-C. Yang, "Decision table reduction based on conditional information entropy," *Chin. J. Comput.*, vol. 25, no. 7, pp. 759–766, 2002.

[39] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Inf. Process. Manage.*, vol. 42, no. 1, pp. 155–165, 2006.

[40] 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 Res.*, vol. 30, no. 1, pp. 303–305, 2002.

[41] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, and J. Li, "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.

[42] U. Güldener, M. Münsterkötter, M. Oesterheld, P. Pagel, A. Ruepp, H.-W. Mewes, and V. Stümpflen, "MPact: The MIPS protein interaction resource on yeast," *Nucleic Acids Res.*, vol. 34, no.1, pp. D436–D441, 2006.

[43] E. Aslankoohi, B. Zhu, M. N. Rezaei, K. Voordeckers, D. De Maeyer, K. Marchal, E. Dornez, C. M. Courtin, and K. J. Verstrepen, "Dynamics of the Saccharomyces cerevisiae transcriptome during bread dough fermentation," *Appl. Environ. Microbiol.*, vol. 79, no. 23, pp. 7325–7333, Dec. 2013.

[44] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Li, and B. Xu, "Protein complex prediction in large ontology attributed protein-protein interaction networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 3, pp. 729–741, May 2013.

[45] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak, "Up-to-date catalogues of yeast protein complexes," *Nucleic Acids Res.*, vol. 37, no. 3, pp. 825–831, 2009.

[46] J. X. Binder, S. Pletscher-Frankild, K. Tsafou, C. Stolte, and S. I. O'Donoghue, R. Schneider, and L. J. Jensen, "COMPARTMENTS: Unification and visualization of protein subcellular localization evidence," *Database*, vol. 2014, p. bau012, Jan. 2014.

[47] J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, and Y. Jia, "SGD: Saccharomyces genome database," *Nucleic Acids Res.*, vol. 26, no. 1, pp. 73–79, 1998.

[48] R. Zhang and Y. Lin, "DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes," *Nucleic Acids Res.*, vol. 37, no. 1, pp. D455–D458, 2009.

[49] 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 Microbiol.*, vol. 9, no. 1, p. 243, 2009.

[50] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.

JIE ZHAO received the B.S. degree in computer science and technology from Lanzhou University, Lanzhou, China, in 2009, and the M.S. degree in computer software and theory from Shaanxi Normal University, Xi'an, China, in 2015, where he is currently pursuing the Ph.D. degree with the School of Computer Science. His current research interests include bioinformatics, swarm intelligent optimization, and data mining techniques and their applications.

XIUJUAN LEI (M'19) received the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2001 and 2005, respectively. She was a Visiting Scholar with the Department of Computer Science and Engineering, State University of New York at Buffalo, USA, from 2009 to 2010. She is currently a Professor with the School of Computer Science, Shaanxi Normal University, Xi'an. Her research interests include bioinformatics, swarm intelligent optimization, data mining, and big data analysis.