A New Scheme for Essential Protein Identification Based on Uncertain Networks

WEI LIU, LIANGYU MA, LING CHEN, (Member, IEEE), AND BYEUNGWOO JEON, (Member, IEEE)

1College of Information Engineering, Yangzhou University, Yangzhou 225127, China
2School of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon 40-746, South Korea

ABSTRACT Identifying essential proteins is important for not only understanding cellular activity but also detecting human disease genes. A series of centrality measures have been proposed to identify essential proteins based on the protein-protein interaction (PPI) network. Although, existing studies have focused on the topological features of the PPI network and the intrinsic characteristics of biological attributes, it is still a big challenge to further improve the prediction accuracy of essential proteins. Moreover, there are substantial amounts of false-positive data in PPI networks; thus, a PPI network should be modelled as an uncertain network. How to identify essential proteins more accurately and conveniently has become a research hotspot. In this paper, we proposed a new essential protein discovery method called ETB-UPPI on uncertain PPI networks. The algorithm detects essential proteins by integrating topological features with biological information. Experimental results on four Saccharomyces cerevisiae datasets have shown that ETB-UPPI can not only improve the prediction accuracy but also outperform other prediction methods, including the most commonly-used centrality measures (DC, SC, BC, IC, EC, and NC), topology-based methods (LAC) and biological-data-integrating methods (PeC, WDC, UDONC, LBCC, TEGS, and RSG).

INDEX TERMS Essential proteins, simrank algorithm, uncertain PPI network, biological information.

I. INTRODUCTION Proteins are biological macromolecular compounds formed by the aggregation of many amino acids. They are one of the most basic substances in life and are widely found in various biological tissue cells. Among all, essential proteins are indispensable in cellular organisms. The deletion of these essential proteins will cause the loss of protein complex function and lead to organisms being unable to function [1]. Therefore, the discovery of essential proteins plays a significant role in disease detection and the manufacture of new drugs.

Traditional experimental techniques, including single-gene knockouts [2], RNA interference [3], and conditional knockouts [4], have been used to discover essential proteins. However, these experimental techniques are expensive. Therefore, it is almost impossible to use experimental techniques to detect essential proteins for a whole organism. Now, due to the development of high-throughput experimental technologies, many high-quality and large-scale PPI datasets have been accumulated[5], which provide relatively reliable and sufficient data for computational methods to detect essential proteins, functional modules and protein complexes. These computational methods in the PPI network also can be applied to many other fields [6]–[8].

Jeong et al. [9] have proposed a centrality-lethality rule that demonstrated that proteins that are closely related to each other have greater potential to be the indispensable proteins in protein-protein interaction networks. After that, a series of centrality methods were proposed to identify essential proteins based on the topological features of PPI networks such as the degree centrality (DC) [10], eigenvector centrality (EC) [10], betweenness centrality (BC) [11], closeness centrality (CC) [12], subgraph centrality (SC) [13], information centrality (IC) [14], edge clustering coefficient centrality (NC) [15], and local average connectivity centrality (LAC) [16]. The experimental results have verified that...
these centrality methods are more efficient in and convenient for finding essential proteins than traditional experimental methods.

However, there are still shortcomings to these methods. On the one hand, PPI data collected from high-throughput experimental techniques are always noisy and include false positives, which have a strong influence on the prediction accuracy of essential proteins. On the other hand, the centrality-based methods ignore the biological information of proteins, which leads to low prediction accuracy. To overcome the shortcomings of these algorithms, substantial amounts of biological information have been applied to the identification of essential proteins in a more accurate manner. For example, Acencio and Lemke [17] introduced a supervised machine learning method by combining network topology and biological characteristics; among them, biological characteristics include subcellular localization information and biological process information. Li et al. [18] proposed PeC and Tang et al. introduced a modified prediction method called WDC [19], which integrates the gene expression data and topological features. Considering the domain information, Peng et al. [20] proposed a method named UDoNC. Zhiwei et al. [21] proposed a method named GOS to predict essential proteins based on gene expression data, orthology, and subcellular localization information. Zhang et al. [22] proposed a method named OGN by integrating orthology, gene expression data and PPI networks. Meanwhile, clustering-based methods have also been used to predict protein complexes [23]. Li et al. [24] proposed a method called SPP to detect essential proteins. It firstly partitioned the PPI network into sub-networks by subcellular location information and then assigned the priority of sub-networks. Finally, it employed the priority and the connections between two nodes to efficiently identify essential proteins. In [25], the authors addressed a novel method called TEGS to detect essential proteins based on gene expression data, GO annotation information and protein subcellular localization data. Lei et al. [26] constructed weighted PPI networks and efficiently discovered essential proteins based on RNA-Seq, GO annotation and subcellular localization information. Li et al. [27] constructed a refined PPI network called TS-PIN by integrating gene expression information and subcellular location information to efficiently identify essential proteins. Afterwards, they adopted a deep learning method [28] to detect essential proteins without any prior knowledge. In addition, some scholars have addressed the discovery of essential proteins from the perspective of building new models, such as the RWEN[29] and E-POC [30] models, in recent years and have achieved good results. Compared with network-based centrality methods, these methods of integrating biological information have greatly improved the prediction accuracy.

Other studies have shown that proteins that frequently appear in different protein complexes are more likely to be essential proteins [31]. Therefore, in recent years, many detection methods for essential proteins incorporating protein complexes have been presented. Li et al. [32] developed a new essential protein discovery method called UC that used protein complexes to increase the prediction accuracy. Chao et al. [33] presented a new algorithm called LBCC for integrating the network topological features and protein complex information. The results showed that LBCC achieved good performance. Lei et al. [34] introduced a new method called PCSD, which integrated the subgraph density and participation degree in the protein complex to predict essential proteins.

Since the interactions between proteins are dynamic, many scholars have focused on the identification of essential proteins on dynamic PPI networks in recent years. A set of time-series temporary sub-networks have been constructed by integrating the time-series gene expression profiles and a deterministic PPI network. For example, Luo et al. [35] proposed a new method CDLC to identify essential proteins by integrating dynamic network topologies with protein complexes. Xiao et al. [36] presented a new essential protein detection method that constructed dynamic networks using gene expression data. In addition, they proved that most of the essential proteins were active. Shang et al. [37] built a new dynamic PPI network using RNA-Seq datasets to identify essential proteins. The experimental results showed that the proposed scheme could not only significantly improve the prediction accuracy but also be extended to many existing methods.

However, there are still some tough challenges in detecting essential proteins in PPI networks. On the one hand, interactions within PPI networks change continuously over time. On the other hand, PPI networks are inaccurate and incomplete due to the limitations of experimental methods. Therefore, deterministic PPI networks cannot truly reveal the connections between proteins in living organisms. From this point of view, a PPI network should be modelled as an uncertainty graph. In this paper, we proposed a new method called ETB-UPPI based on network topological features and biological information on uncertain protein-protein interaction networks. First, we convert the Simrank problem in an uncertain network into a Simrank calculation of a deterministic network. Then, we employ the Simrank method to calculate the protein similarity in a PPI network. Finally, we performed feature selection[38] and integrated multivariate biological data to measure the essentiality of the protein. To evaluate the performance of our proposed algorithm, ETB-UPPI, we conduct experiments over four datasets from different sources and compare with other methods, i.e., DC, EC, BC, SC, IC, NC, LAC, PeC, WDC, UDoNC, LBCC, TEGS and RSG. The experimental results have shown that ETB-UPPI can improve the prediction accuracy and is superior to other prediction methods.

II. METHODS
In this section, we first convert the Simrank problem in an uncertain network into a Simrank calculation of a deterministic network. Then, we employ the Simrank
method to calculate the protein similarity in a PPI network. Afterwards, we integrate multivariate biological data such as gene ontology (GO) similarity, the Pearson correlation coefficient, and subcellular localization information. Finally, we give a scoring function to measure the essentiality of the protein.

The basic idea of our method is mainly based on the following premise: (1) A protein with a high similarity score is more likely to be an essential protein. (2) Essential proteins in the same cluster are more likely to be co-expressed. (3) The more often a protein appears in the nucleus, the more likely it is to become an essential protein. To describe the proposed method more clearly, we give the following definitions.

A. PROBLEM DEFINITIONS AND CONCEPTS

Definition 1 (Uncertain PPI Network): An uncertain PPI network can be represented as an undirected graph $G = (V, E, P, A)$, where $V$ denotes a set of protein node, $E$ denotes a set of interactions between pairs of proteins, $P$ represents the probability matrix of $G$, and $A$ represents the adjacent matrix of the undirected graph $G$.

1) SIMRANK SIMILARITY SCORE

In this study, we detect future interactions between pairs of proteins in an uncertain PPI network based on a similarity method. According to the Simrank similarity score, we obtain the similarity measure.

Simrank was proposed by Jeh and Widom [39]. It is an intuitive graph-theoretic model and uses an iterative approach to calculate the similarity between nodes in a network. The algorithm measures the similarity between nodes completely based on the network topology. Simrank assumes that if two objects are referenced by similar objects, the two objects are also similar. In addition, Simrank can be used to reflect direct and indirect connections between vertices in the network.

Due to uncertain interactions in a PPI network, we employed the Simrank method to calculate the similarity between proteins. The method converts the Simrank problem in the uncertain PPI network[40] into a Simrank calculation in a deterministic PPI network. To avoid enumerating all possible worlds of an uncertain network, we constructed an adjacency subgraph of each node, which allowed us to perform transformation matrix calculations within the subgraph.

Definition 2 (Simrank Similarity): Given an uncertain graph $G = (V, E, P, A)$ for each node $v \in V$, let $I(v)$ be the set of neighbours of vertex $v$ in the protein-protein interaction network, and let $|I(v)|$ be the number of vertices in $I(v)$ with $I_i(v)$ the $i$-th neighbour of $v$, $1 \leq i \leq |I(v)|$. The Simrank similarity between nodes, which is denoted as $s(u, v)$, can be defined as follows:

$$s(u, v) = \begin{cases} \frac{c}{|I(u)||I(v)|} \sum_{i=1}^{|I(u)|} \sum_{j=1}^{|I(v)|} s(I_i(u), I_j(v)) & u \neq v \\ 1 & u = v. \end{cases}$$

where $c \in (0, 1)$ is the damping factor. There is a special case to be noted, that is, when $|I(u)| = 0$ or $|I(v)| = 0$, $s(u, v) = 0$.

Initially, $s(u, v)$ is set as $s(u, v) = A(u, v)$, where $A$ is an $n \times n$ matrix that represents the adjacent matrix of $G$, i.e.,

$$s(a, b) = \begin{cases} 1 & a = b \\ 0 & a \neq b. \end{cases}$$

Thus the matrix form equivalent to equation (1) is

$$S = cW^T SW + (1 - c)I$$

where $W = [w_{ij}]$ is the transformation matrix, and its elements are defined as $w_{ij} = \frac{g_{ij}}{\sum_{k=1}^{n} g_{ik}}$. In our method, $c$ is set to 0.8.

Based on the Simrank similarity to detect the interactions between proteins in an uncertain PPI network $G = (V, E, P, A)$, we first convert the uncertain protein network into a deterministic network $\bar{G} = (\bar{V}, \bar{E}, \bar{W})$ by calculating the transformation matrix $\bar{W}$.

Definition 3 (Possible World): Let $G = (V, E, P, A)$ denote an uncertain graph, where $P$ represents the probability matrix, $G'$ denotes a subgraph of $G$ and the probability is $P$. To satisfy such conditions, $G'$ is considered the possible world of $G$.

The probability of a possible world $G'$ is expressed as $P_r[G']$:

$$P_r[G'] = \prod_{e \in E'} p(e) \prod_{e \in E \setminus E'} (1 - p(e)).$$

where $p(e)$ is the probability that each edge $e$ is included in $G'$.

Definition 4 (The Transformation Matrix): Given an uncertain PPI network $G$ and the deterministic network $\bar{G} = (\bar{V}, \bar{E}, \bar{W})$, where $\bar{W}$ denotes the transfer matrix of $\bar{G}$, $\bar{E} = E \cup S$, and $S = \{(u, u) | u \in V\}$, the transfer matrix $\bar{W} = [\bar{w}(u, v)]$ is formulated as

$$\bar{w}(u, u) = \prod_{(u, q) \in E} (1 - p(u, q))$$

$$\bar{w}(u, v) = \sum_{G \in \Omega(u, v)} \frac{A(u, v)}{\sum_{(u, q) \in G'} A(u, q)} P_r[G'](u \neq v)$$

where $(u, v)$ denotes an edge between proteins $u$ and $v$. $A$ represents the adjacent matrix, $\Omega(u, v)$ denotes the set of all possible worlds in $G$, and these possible worlds contain edge$(u, v)$.

By the use of the transformation matrix, we can calculate the Simrank similarity in an uncertain PPI network. This can be written as the matrix form as follows:

$$S = c\bar{W}^T S\bar{W} + (1 - c)I$$

Since the calculation of equation (3) needs to enumerate all possible worlds, the complexity of the computation is $O(2^{|E|} - 1)$. Therefore, we construct a neighbour sub-graph for each node to solve the problem.

Definition 5 (Neighbour Subgraph of Node): For each node $u$ in $\bar{G} = (\bar{V}, \bar{E}, \bar{W})$, let $\Gamma(u) = \{|v|(u, v) \in \bar{E}\}$ represent the set of neighbours of node $u$, and the neighbour subgraph
of node \( u \) is described as \( G(u) = (V(u), E(u), \bar{W}_u) \), where \( V(u) = \{ u \} \cup \Gamma(u), E(u) = \{(w, v) | (w, v) \in \bar{E}, v, w \in V(u) \} \) and \( \bar{W}_u \) is the transfer matrix of \( G(u) \).

The set of edges outside \( G(u) \) is defined as \( \bar{E}(u) \). Assume that the number of possible worlds in \( G(u) \) is \( m \), i.e. \( G_1(u), G_2(u), \ldots, G_m(u) \). For each possible world \( G_i(u) \), let \( E_i(u) \) be the set of edges in \( G_i(u) \) and \( E'_i(u) \) be the set of edges not in \( G_i(u) \).

According to \( G_i(u) \), a possible world of \( \bar{G} \) is defined as \( \bar{G}_{ik}(u) = E_i(u) \cup \bar{E}_{ik}(u) \), where \( \bar{E}_{ik}(u) \in \bar{E}(u) \) denotes the set of edges, and these edges are not in \( E(u) \). \( \bar{G}_{ik}(u) \) is an extension of \( G_i(u) \).

For the possible world \( \bar{G}_{ik}(u) \), the set of edges can be represented as \( \bar{E}_{ik}(u) = \{ e | e \in \bar{E}(u), e \notin \bar{E}_{ik}(u) \} \), and the following can be obtained:

\[
\bar{w}(u, v) = \sum_{G' \in \bar{G}(u,v)} \sum_{(u,q) \in G} \frac{A(u,q)}{P_r[G']}
\]

\[
= \sum_{i=1}^{m} \sum_{k=1}^{m_i} \sum_{(u,q) \in G_{ik}} A(u,q) P_r[G_{ik}]
\]

Since

\[
P_r[G_{ik}] = \prod_{e \in \bar{G}_{ik}} \prod_{e \in \bar{G}_{ik}} [1 - p_r(e)]
\]

\[
= \prod_{e \in \bar{E}_{ik}(u)} \prod_{e \in \bar{E}_{ik}(u)} [1 - p_r(e)]
\]

Then,

\[
\bar{w}(u, v) = \prod_{i=1}^{m} \frac{A(u,v)}{P_r[G_i(u)]} \prod_{e \in \bar{E}_{ik}(u)} [1 - p_r(e)]
\]

\[
\prod_{e \in \bar{E}_{ik}(u)} [1 - p_r(e)]
\]

Because \( \prod_{e \in \bar{E}_{ik}(u)} [1 - p_r(e)] \) is the sum of the probabilities of all possible worlds in \( \bar{E}(u) = \{ e | e \notin G(u), e \in \bar{E} \} \), its value is 1.

Thus, we can obtain

\[
\bar{w}(u, v) = \sum_{i=1}^{m} \frac{A(u,v)}{P_r[G_i(u)]} \prod_{e \in \bar{E}_{ik}(u)} [1 - p_r(e)]
\]

\[
= \sum_{i=1}^{m} \frac{A(u,v)}{P_r[G_i(u)]} P_r(G_i(u)) = \bar{w}_u(u, v)
\]

where \( \bar{w}_u \) represents the transfer matrix of \( G(u) \). Therefore, the transfer weight from node \( u \) to its neighbour \( v \) in \( G \) can be calculated by

\[
\bar{w}(u, v) = \begin{cases} 
\bar{w}_u(u, v) & \text{if } v \in G(u) \\
0 & \text{otherwise.}
\end{cases}
\]  

From formula (6), we can find that \( \bar{W}(u, v) \) can be computed in the subgraph \( G(u) \), which will greatly reduce the time complexity.

Through the above-mentioned method, we can calculate the Simrank similarity score between nodes \( u \) and \( v \), where \( u \) and \( v \) interact in the neighbour subgraph \( G(u) \) of node \( u \).

**Definition 6 (Simrank Similarity Score):** For a protein \( u \), its Simrank similarity score can be defined as the sum of similarity scores between protein \( u \) and all its neighbours.

The calculation of the Simrank similarity score is shown as follows:

\[
S\_score(u) = \sum_{v \in I(u)} S(u, v)
\]

where \( S \) is a protein similarity score matrix and \( I(u) \) is the set of neighbours of protein \( u \).

2) BIOLOGICAL INFORMATION SOURCES

Our proposed algorithm, ETB-UPPI, integrates three biological information sources: gene ontology data, gene expression data, and subcellular localization data. Gene ontology (GO) is a description of the function of a particular gene and provides a comprehensive source of information about the gene function. It is the basis for the computational analysis of large-scale molecular biology and genetics experiments in biomedical research. Usually, we believe that two essential proteins linked by an edge have a greater probability of participating in the same biological process. It is well known that gene expression refers to the process of protein synthesis under gene guidance. Therefore, we employ the Pearson correlation coefficient (PCC) to measure the intensity of co-expression of two interacting proteins. Moreover, subcellular localization refers to the specific site of a certain protein or expression product in a cell; for example, in the nucleus, in the cytoplasm or on the cell membrane. Previous studies have shown that in a certain compartment, such as the cytoplasm, the wider the distribution of proteins is, the greater the possibility of becoming an essential protein[41].

a: GENE ONTOLOGY (GO) SIMILARITY

Gene ontology (GO) is an ontology widely used in the field of biological information. It mainly consists of three branches: biological processes, molecular functions, and cellular components. GO not only provides valuable information and convenient methods to study the functional similarity of genes but also improves the accuracy of constructing networks[42]. We introduce the functional similarity method proposed by Wang et al. [43] to calculate the functional similarity between genes.

\[
Sim_{match}(G_1, G_2) = \frac{|\text{Term}(G_1) \cap \text{Term}(G_2)|}{|\text{Term}(G_1) + |\text{Term}(G_2)|}
\]

where \( \text{Term}(G_1) \) and \( \text{Term}(G_2) \) represent the set of GO terms that annotate genes \( G_1 \) and \( G_2 \), respectively.
which is calculated as follows:

\[
PCC(X, Y) = \frac{1}{m} \sum_{i=1}^{m} \left( \frac{g(X, i) - \bar{g}(X)}{\sigma(X)} \right) \cdot \left( \frac{g(Y, i) - \bar{g}(Y)}{\sigma(Y)} \right)
\]

where \(m\) is the number of samples in the gene expression data. \(g(X, i)\) and \(g(Y, i)\) represent the expression level of genes \(X\) and \(Y\) in sample \(i\), respectively. \(\bar{g}(X)\) and \(\bar{g}(Y)\) represent the mean expression level of genes \(X\) and \(Y\), respectively. \(\sigma(X)\) and \(\sigma(Y)\) represent the standard deviations of the expression level of gene \(X\) and gene \(Y\), respectively.

\[c: \text{SUBCELLULAR LOCALIZATION SCORE}\]

Subcellular localization refers to the specific site of a certain protein or expression product in a cell, which is divided into different compartments[45]. As shown in Figure 1-2, by calculating the number of proteins and essential proteins in seven different compartments, we found that the numbers of proteins and essential proteins in the nucleus were the highest. This fully demonstrates that protein localization plays a crucial role in the essentiality of proteins.

\[d: \text{THE IMPORTANCE SCORE}\]

Although the Simrank similarity considers the uncertainty of the protein-protein interaction network, it does not fully determine the essentiality of the protein. Biological information plays an important role in the identification process of essential proteins. Therefore, we use a method that integrates the Simrank similarity and biological properties to identify essential proteins. In ETB-UPPI, the essentiality of a protein is measured by the importance score. The importance score of a protein consists of the Simrank similarity score, the gene ontology (GO) similarity, the Pearson correlation coefficient (PCC) and the subcellular localization score (NSL). For a protein \(u\), the importance score \((E_{\text{score}})\) of protein \(u\) is defined as follows:

\[
E_{\text{score}}(u) = \alpha S_{\text{score}}(u) + (1 - \alpha)\left( \sum_{v \in I(u)} \text{Sim}_{\text{match}}(u, v) \right) + \sum_{v \in I(u)} PCC(u, v) + \text{NSL}(u)
\]

Here, \(I(u)\) is the set of all neighbour nodes of protein \(u\), \(\alpha \in (0, 1)\) is a parameter used to adjust the contribution of the Simrank similarity and other biological property similarities.

\[B. \text{ALGORITHM ETB-UPPI}\]

1) MAIN STEPS OF THE ALGORITHM

Fig. 3 shows the workflow of our ETB-UPPI method. The method is composed of three main steps:

1) Converting the Simrank problem in an uncertain network into a Simrank calculation of a deterministic network.

2) Employing the Simrank method to calculate the protein similarity in a PPI network.

3) Integrating multivariate biological data, such as the gene ontology (GO) similarity, the Pearson correlation coefficient, and subcellular localization information.

4) Giving a scoring function to measure the essentiality of the protein based on the Simrank similarity, gene ontology (GO) similarity, Pearson correlation coefficient and subcellular localization information.

The framework of the proposed ETB-UPPI algorithm (essential protein identification based on topological features and biological information on uncertain protein-protein interaction networks) is shown in Algorithm 1.
Suppose that \( n \) is the total number of protein nodes in \( G \), and the computational time of step 1 is \( O(n^3) \). The computational time of steps 2, 3 and 4 are \( O(n^2) \). Thus, the time of calculating \( \text{E_score} \) is \( O(n^3) \). The time complexity of sorting protein nodes is \( O(n \log n) \). Therefore, the total time complexity of the algorithm is \( O(n^3) \).

### III. EXPERIMENTAL RESULTS AND ANALYSIS

#### A. DATASET

In this section, in order to evaluate the performance of our proposed ETB-UPPI algorithm, we conduct experiments on Saccharomyces cerevisiae data which is the most reliable and completely species.

The PPI datasets were collected from four different sources, including the DIP [46], MIPS [47], Gavin [45], and Krogan [48]. All self-interactions and repeated interactions were filtered. The detailed information of the four different datasets (DIP, MIPS, Gavin, and Krogan) are introduced in Table 1. The essential proteins are collected from the following four different databases: MIPS [47], SGD [49], DEG [50], and SGDP (http://www.sequence.stanford.edu/group/). Gene expression data is downloaded from the Gene Expression Omnibus (GEO) database [51] (http://www.ncbi.nlm.nih.gov/geo/) with accession number GSE3431. It includes three metabolism cycles with a total of 36-time points. Gene expression data set includes 9336 genes. The GO data applied in our method is extracted from the GO Consortium [52]. And the subcellular localization information is obtained from the COMPARTMENTS database [53].

#### B. THE EFFECT OF PARAMETER ON PERFORMANCE

In our method, the importance score of the proteins consists of two parts: the Simrank similarity score and the biological information. We adjusted them using a parameter \( \alpha \), which ranges from 0 to 1. Tables 2-5 list the identification results of ETB-UPPI on four different networks obtained with different values of \( \alpha \). When \( \alpha = 1 \), the \( \text{E_score} \) only considers the topology information, and when \( \alpha = 0 \), the \( \text{E_score} \) only considers the biological information. As shown in Tables 2-4, we can find that when \( \alpha = 0.6 \), the ETB-UPPI algorithm can achieve better performance.

To further reflect the impact of \( \alpha \) on our algorithm, we use Figure 4 to show how the F-measure of the ETB-UPPI method fluctuates under different values of \( \alpha \). From Figure 4(a), (b), and (d), we can clearly observe that the F-measure is maximized when \( \alpha = 0.6 \). It can be noticed from Figure 4(c) that the F-measure obtains the best result when \( \alpha = 0.5 \). Thus, we set \( \alpha = 0.6 \) in our experiments.

#### C. COMPARING ETB-UPPI WITH OTHER METHODS

To evaluate the performance of ETB-UPPI, we compare our ETB-UPPI method with other prediction methods (DC [10], EC [10], BC [11], SC [13], IC [14], NC [15], LAC [16], PeC [18], WDC [19], UDONC [20], LBCC [33], TEGS [25], and RSG [26]) on four different datasets. For convenience,
TABLE 2. The number of essential proteins correctly predicted by ETB-UPPI with different $\alpha$ on DIP network.

| $\alpha$ | top1\% | top5\% | top10\% | top15\% | top20\% | top25\% |
|----------|--------|--------|---------|---------|---------|---------|
| 0        | 14     | 79     | 171     | 238     | 316     | 381     |
| 0.1      | 15     | 83     | 177     | 242     | 321     | 388     |
| 0.2      | 17     | 85     | 181     | 247     | 325     | 394     |
| 0.3      | 17     | 88     | 181     | 252     | 328     | 398     |
| 0.4      | 18     | 88     | 186     | 252     | 331     | 401     |
| 0.5      | 20     | 90     | 190     | 255     | 336     | 405     |
| 0.6      | 19     | 92     | 191     | 257     | 338     | 405     |
| 0.7      | 19     | 91     | 190     | 256     | 337     | 404     |
| 0.8      | 17     | 89     | 188     | 251     | 335     | 401     |
| 0.9      | 17     | 86     | 184     | 251     | 331     | 397     |
| 1.0      | 16     | 84     | 181     | 249     | 328     | 393     |

TABLE 3. The number of essential proteins correctly predicted by ETB-UPPI with different $\alpha$ on MIPS network.

| $\alpha$ | top1\% | top5\% | top10\% | top15\% | top20\% | top25\% |
|----------|--------|--------|---------|---------|---------|---------|
| 0        | 10     | 52     | 106     | 151     | 201     | 236     |
| 0.1      | 10     | 53     | 109     | 158     | 206     | 242     |
| 0.2      | 10     | 55     | 110     | 162     | 209     | 247     |
| 0.3      | 12     | 55     | 114     | 165     | 209     | 249     |
| 0.4      | 13     | 57     | 117     | 168     | 212     | 253     |
| 0.5      | 13     | 59     | 119     | 170     | 212     | 254     |
| 0.6      | 13     | 59     | 117     | 170     | 213     | 254     |
| 0.7      | 13     | 58     | 117     | 168     | 211     | 253     |
| 0.8      | 12     | 58     | 115     | 168     | 211     | 253     |
| 0.9      | 12     | 56     | 113     | 165     | 207     | 251     |
| 1.0      | 16     | 84     | 181     | 249     | 328     | 393     |

TABLE 4. The number of essential proteins correctly predicted by ETB-UPPI with different $\alpha$ on Gavin network.

| $\alpha$ | top1\% | top5\% | top10\% | top15\% | top20\% | top25\% |
|----------|--------|--------|---------|---------|---------|---------|
| 0        | 10     | 52     | 106     | 151     | 201     | 236     |
| 0.1      | 10     | 53     | 109     | 158     | 206     | 242     |
| 0.2      | 10     | 55     | 110     | 162     | 209     | 247     |
| 0.3      | 12     | 55     | 114     | 165     | 209     | 249     |
| 0.4      | 13     | 57     | 117     | 168     | 212     | 253     |
| 0.5      | 13     | 59     | 119     | 170     | 212     | 254     |
| 0.6      | 13     | 59     | 117     | 170     | 213     | 254     |
| 0.7      | 13     | 58     | 117     | 168     | 211     | 253     |
| 0.8      | 12     | 58     | 115     | 168     | 211     | 253     |
| 0.9      | 12     | 56     | 113     | 165     | 207     | 251     |
| 1.0      | 11     | 55     | 112     | 163     | 202     | 248     |

TABLE 5. The number of essential proteins correctly predicted by ETB-UPPI with different $\alpha$ on Krogan network.

| $\alpha$ | top1\% | top5\% | top10\% | top15\% | top20\% | top25\% |
|----------|--------|--------|---------|---------|---------|---------|
| 0        | 17     | 92     | 161     | 225     | 273     | 319     |
| 0.1      | 19     | 95     | 166     | 229     | 279     | 324     |
| 0.2      | 19     | 97     | 169     | 233     | 283     | 328     |
| 0.3      | 20     | 99     | 172     | 237     | 288     | 332     |
| 0.4      | 21     | 99     | 174     | 240     | 292     | 336     |
| 0.5      | 22     | 102    | 177     | 240     | 293     | 338     |
| 0.6      | 22     | 102    | 179     | 242     | 292     | 339     |
| 0.7      | 22     | 102    | 178     | 240     | 292     | 337     |
| 0.8      | 22     | 101    | 176     | 237     | 290     | 334     |
| 0.9      | 20     | 101    | 176     | 237     | 287     | 331     |
| 1.0      | 20     | 99     | 173     | 232     | 286     | 327     |

FIGURE 4. The effect of $\alpha$.

UDONC and LBCC are only applied to the DIP dataset. We calculated the importance scores of the proteins by different prediction methods and ranked them in descending order according to their importance score of each protein. Then, we chose six levels (the top 1, top 5, top 10, top 15, top 20, and top 25 percent of all proteins) as candidate essential proteins. Finally, we compared the known essential proteins with candidate essential proteins to determine the number of truly essential proteins in the candidate essential proteins.

The prediction results at six levels on the DIP dataset are shown in Figure 5. It can be shown that our ETB-UPPI method is more superior to all the other methods (DC, EC, BC, LAC, PeC, WDC, UDONC, LBCC) at six levels from the top 1% to the top 25% of ranked proteins. From Figure 5, we can see that the total numbers of essential proteins correctly predicted by our method ETB-UPPI were 42, 206, 332, 436, 548, and 621 at six levels. For the top 1% of identified essential proteins, it can be noticed that the ETB-UPPI algorithm performs much better than DC, EC, BC, LAC, PeC, WDC, UDONC, LBCC, and TEGS (90, 75, 75, 44, 10, 31, 13, 13 and 2 percent, respectively). On the top 10% and top 15% of identified essential proteins, our algorithm is slightly inferior to TEGS and RSG. However, in most cases, our algorithm still outperforms TEGS and RSG.
Therefore, it can be concluded that the ETB-UPPI algorithm outperforms the ten other methods.

The prediction results at six levels on the MIPS dataset are shown in Figure 6. Clearly, our ETB-UPPI method outperforms all the other methods (DC, EC, SC, IC, NC, LAC, PeC, WDC, TEGS, and RSG) at five levels. From Figure 6, we can see that the total numbers of essential proteins correctly predicted by ETB-UPPI were 19, 92, 191, 257, 338, and 405 at six levels. ETB-UPPI achieves the best performance for the top 1%, top 10%, top 15%, top 20%, and top 25%, and PeC obtains the best performance at the top 5%. For the top 5% of identified essential proteins, PeC detected 93 truly essential proteins, while ETB-UPPI correctly identified 92 essential proteins. When the number of identified essential protein is large, our ETB-UPPI method slightly outperforms PeC. For example, when predicting the top 25% of essential proteins, ETB-UPPI achieves an approximately 25% increase over PeC.

The prediction results at six levels on the Gavin dataset are shown in Figure 7. Our ETB-UPPI method performs better than all the other methods at six levels from the top 1% to top 25% of sorted proteins. Figure 7 shows that the numbers of essential proteins correctly identified by our ETB-UPPI method are 13, 59, 117, 170, 213, and 254 at six levels. ETB-UPPI achieves the best performance at the top 5%, top 10%, top 15%, top 20%, and top 25%. LAC and PeC obtain the best performance at the top 1%. For the top 25% of identified essential proteins, LAC and ETB-UPPI achieve the same performance. However, compared with DC, EC, SC, IC, NC, PeC, WDC, TEGS, and RSG, the performance of our ETB-UPPI algorithm is increased by 14%, 23%, 35%, 17%, 1%, 9%, 3%, 0.4% and 2%, respectively.

The prediction results at six levels on the Krogan dataset are shown in Figure 8. From Figure 8, we can see that the
total numbers of essential proteins predicted correctly by our ETB-UPPI method are 22, 102, 179, 242, 292, and 339 at six levels. Note that the performance of ETB-UPPI algorithm is only inferior to the RSG algorithm and superior to the other nine algorithms at six levels from the top 1% to top 10% of ranked proteins. When predicting the top 1% of essential proteins, the numbers of essential proteins correctly predicted by NC and ETB-UPPI are the same. Moreover, the ETB-UPPI algorithm performs slightly worse than RSG and TEGS, whereas it outperforms the other eight algorithms at six levels from the top 15% to top 25% of sorted proteins. For the top 25% of identified essential proteins, compared with five centrality methods (DC, EC, SC, IC, NC, and LAC), PeC, and WDC, the performance of our ETB-UPPI algorithm is increased by 6%, 33%, 24%, 7%, 4%, 4%, 6% and 6%, respectively.

Based on the comprehensive analysis of Figures 5-8, it can be seen that our ETB-UPPI method performs best on the DIP, MIPS and Gavin datasets. On the Krogan dataset, ETB-UPPI identified slightly fewer essential proteins than the RSG and TEGS methods. It is obvious that the ETB-UPPI method has a better identification efficiency and strong stability. The experimental results show the correctness as well as the effectiveness of applying ETB-UPPI to essential protein discovery on PPI networks.

D. VALIDATION USING SIX STATISTICAL MEASURES
To further demonstrate the performance of our method, we apply six statistical measures, sensitivity (SN), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), F-measure (F), and accuracy (ACC), to compare the ETB-UPPI method with other methods. The statistical measures are defined as follows.

True positive (TP) denotes the total number of true positives (correctly identified essential proteins). False negative (FN) denotes the total number of truly essential proteins minus TP. True negative (TN) denotes the number of true negatives (correctly identified non-essential proteins). False positive (FP) denotes the total number of non-essential proteins incorrectly predicted as essential proteins. We rank the proteins in descending order based on importance scores obtained by different algorithms and select the top 25% as essential proteins. These statistical measures of the prediction result are calculated as follows:

\[ SN = \frac{TP}{TP + FN} \]  
\[ SP = \frac{TN}{TN + FP} \]  
\[ PPV = \frac{TP}{TP + FP} \]  
\[ NPV = \frac{TN}{TN + FN} \]  
\[ F = \frac{2 \times SN \times PPV}{SN + PPV} \]  
\[ ACC = \frac{TP + TN}{TP + TN + FP + FN} \]

A high value of the six above statistical measures suggests good performance overall. The results of the statistical measures are shown in Tables 6-9. From Table 9, it can be seen that ETB-UPPI performs slightly worse than TEGS and RSG on the Krogan dataset. However, from Tables 6-8, it can be clearly observed that the six statistical measures of ETB-UPPI are higher than those of any other method on the three datasets, which further demonstrates that our ETB-UPPI method can predict essential proteins more accurately.

E. FOLD ENRICHMENT OF ETB-UPPI
Fold enrichment [37] is used to describe how those gold standard essential proteins are enriched in the set of identified essential proteins predicted by ETB-UPPI method and the other methods at six levels on the Krogan dataset.
TABLE 7. The six statistical measures on the MIPS dataset.

| Methods | SN    | SP    | PPV   | NPV   | F     | ACC  |
|---------|-------|-------|-------|-------|-------|------|
| DC      | 0.324 | 0.771 | 0.289 | 0.799 | 0.306 | 0.671|
| EC      | 0.198 | 0.735 | 0.177 | 0.761 | 0.187 | 0.615|
| SC      | 0.198 | 0.735 | 0.177 | 0.761 | 0.187 | 0.615|
| IC      | 0.287 | 0.760 | 0.256 | 0.787 | 0.270 | 0.654|
| NC      | 0.346 | 0.777 | 0.309 | 0.805 | 0.326 | 0.681|
| LAC     | 0.336 | 0.774 | 0.300 | 0.802 | 0.317 | 0.676|
| PeC     | 0.361 | 0.782 | 0.323 | 0.810 | 0.341 | 0.688|
| WDC     | 0.368 | 0.784 | 0.329 | 0.812 | 0.347 | 0.691|
| TEGS    | 0.397 | 0.792 | 0.354 | 0.820 | 0.374 | 0.703|
| RSG     | 0.385 | 0.789 | 0.344 | 0.818 | 0.364 | 0.698|
| ETB-UPPI| 0.399 | 0.793 | 0.356 | 0.821 | 0.376 | 0.704|

TABLE 8. The six statistical measures on the Gavin dataset.

| Methods | SN    | SP    | PPV   | NPV   | F     | ACC  |
|---------|-------|-------|-------|-------|-------|------|
| DC      | 0.358 | 0.831 | 0.617 | 0.631 | 0.453 | 0.627|
| EC      | 0.202 | 0.713 | 0.349 | 0.541 | 0.256 | 0.493|
| SC      | 0.305 | 0.791 | 0.525 | 0.599 | 0.386 | 0.581|
| IC      | 0.350 | 0.825 | 0.603 | 0.626 | 0.443 | 0.630|
| NC      | 0.408 | 0.870 | 0.704 | 0.660 | 0.517 | 0.671|
| LAC     | 0.412 | 0.872 | 0.709 | 0.661 | 0.521 | 0.673|
| PeC     | 0.374 | 0.844 | 0.645 | 0.640 | 0.474 | 0.641|
| WDC     | 0.397 | 0.861 | 0.684 | 0.653 | 0.503 | 0.661|
| TEGS    | 0.410 | 0.871 | 0.707 | 0.660 | 0.519 | 0.672|
| RSG     | 0.404 | 0.866 | 0.696 | 0.657 | 0.511 | 0.666|
| ETB-UPPI| 0.412 | 0.872 | 0.709 | 0.661 | 0.521 | 0.673|

TABLE 9. The six statistical measures on the Krogan dataset.

| Methods | SN    | SP    | PPV   | NPV   | F     | ACC  |
|---------|-------|-------|-------|-------|-------|------|
| DC      | 0.406 | 0.814 | 0.475 | 0.768 | 0.438 | 0.694|
| EC      | 0.322 | 0.780 | 0.378 | 0.735 | 0.348 | 0.646|
| SC      | 0.357 | 0.790 | 0.407 | 0.745 | 0.380 | 0.660|
| IC      | 0.401 | 0.812 | 0.469 | 0.766 | 0.432 | 0.691|
| NC      | 0.415 | 0.818 | 0.486 | 0.771 | 0.447 | 0.699|
| LAC     | 0.416 | 0.817 | 0.485 | 0.771 | 0.448 | 0.699|
| PeC     | 0.404 | 0.814 | 0.474 | 0.767 | 0.436 | 0.694|
| WDC     | 0.424 | 0.822 | 0.496 | 0.775 | 0.457 | 0.705|
| TEGS    | 0.473 | 0.864 | 0.555 | 0.821 | 0.511 | 0.761|
| RSG     | 0.478 | 0.844 | 0.560 | 0.796 | 0.516 | 0.737|
| ETB-UPPI| 0.432 | 0.849 | 0.507 | 0.807 | 0.466 | 0.739|

essential proteins. Let $N$ denote the set of identified essential proteins, and $N_e$ denote the number of essential proteins in $N$. $TPR = \frac{N_e}{N}$ denotes the percentage of true positives. Suppose that $V$ denotes the set of all proteins in the PPI network and that $P$ denotes the set of all gold standard essential proteins in $V$. $EPR = \frac{|P|}{|V|}$ denotes the percentage of all essential proteins in $V$. Therefore, the fold enrichment of the identified proteins is calculated as follows:

$$fold\ enrichment = \frac{TPR}{EPR} \quad (19)$$

We also test and compare the fold enrichment of the ETB-UPPI and WDC algorithms on the four datasets. It can be shown from Figure 9 that the fold enrichment of our ETB-UPPI method on four (DIP, MIPS, Gavin, and Krogan) PPI networks is significantly higher than the WDC method.

F. VALIDATION WITH JACKKNIFE METHODOLOGY

In this section, we apply the jackknife methodology to further assess the performance of our ETB-UPPI method as well as the other prediction methods. The horizontal axis of the jackknife curves denotes the number of proteins in the PPI networks sorted in descending order based on their importance scores calculated by the corresponding methods. We analysed the performance of ETB-UPPI and the other
ETB-UPPI performs best on the three different datasets while only performing slightly worse than TEGS and RSG on the Krogan dataset. This indicates that ETB-UPPI can detect more essential proteins.

IV. CONCLUSION

In the post-genomic era, it is important to predict essential proteins and protein complexes. To overcome the shortcomings of traditional experimental methods and achieve higher prediction accuracy, many identification methods have been developed. In this study, we developed a new essential protein identification method called ETB-UPPI based on network-based topological features and biological information. First, we converted the Simrank problem in an uncertain network into a Simrank calculation on a deterministic network. Then, we used the Simrank method to calculate protein similarity in a PPI network. After that, we combined the topology feature of PPI networks with three sources of biological information about proteins. Finally, the importance score of the proteins was obtained from the sum of the GO similarity, Pearson correlation coefficient, subcellular localization information, edge clustering coefficient, and Simrank similarity score. We compared our ETB-UPPI method with other methods, i.e., DC, EC, SC, IC, BC, NC, LAC, PeC, WDC, UDONC, LBCC, TEGS, and RSG, in terms of the number of essential proteins, six statistical measures, and the jackknife curves. The experiment results on four different datasets have proved that ETB-UPPI achieves the best performance.

REFERENCES

[1] Y. Chen and D. Xu, “Understanding protein dispensability through machine-learning analysis of high-throughput data,” Bioinformatics, vol. 21, no. 5, pp. 575–581, Oct. 2004.
[2] 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.
[3] G. E. Schulz and R. H. Schirmer, “Principles of protein structure,” Springer Adv. Texts Chem., vol. 118, no. 1, pp. 151–152, 1979.
[4] 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, and 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, Aug. 2003.
[5] K.-B. Kim, “Web-based computational system for protein-protein interaction inference,” J. Inf. Process. Syst., vol. 8, no. 3, pp. 459–470, Sep. 2012.
[6] Y. Chen, J. Xiong, W. Xu, and J. Zuo, “A novel online incremental and decremental learning algorithm based on variable support vector machine,” Cluster Comput., vol. 22, no. S3, pp. 7435–7445, Jan. 2018.
[7] K. Gu, L. Wang, and B. Yin, “Social community detection and message propagation scheme based on personal willingness in social network,” Soft Comput., vol. 23, no. 15, pp. 6267–6285, Jun. 2018.
[8] J. Qiang, P. Chen, W. Ding, T. Wang, F. Xie, and X. Wu, “Heterogeneous-length text topic modeling for reader-aware multi-document summarization,” ACM Trans. Knowl. Discovery from Data, vol. 13, no. 4, pp. 1–21, Aug. 2019.
[9] H. Jeong, Z. N. Oltvai, and A.-L. Barabási, “Prediction of protein essentiality based on genomic data,” Complexus, vol. 1, no. 1, pp. 19–28, Dec. 2002.
[10] R. R. Vallabhajosyula, D. Chakravartli, S. Lutfi, A. Ray, and A. Raval, “Identifying hubs in protein interaction networks,” PLoS ONE, vol. 4, no. 4, Apr. 2009, Art. no. e5344.
X. Tang, J. Wang, J. Zhong, and Y. Pan, “Predicting essential proteins in the yeast protein interaction network,” *J. Biomed. Biotechnol.*, vol. 2005, no. 2, pp. 96–103, 2005.

S. Wuichert and P. F. Stadler, “Centers of complex networks,” *J. Theor. Biol.*, vol. 223, no. 1, pp. 45–53, Jul. 2003.

E. Estrada and J. A. Rodríguez-Velázquez, “Subgraph centrality in complex networks,” *Phys. Rev. E*, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top., vol. 71, no. 5, May 2005, Art. no. 056103.

K. Stephenson and M. Zelen, “Rethinking centrality: Methods and examples,” *Social Netw.*, vol. 11, no. 1, pp. 1–37, Mar. 1989.

J. 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. 2012.

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, Jul. 2011.

M. L. Acencio and N. Lemke, “Towards the prediction of essential genes based on network topology, cellular localization and biological process information,” *BMC Bioinf.*, vol. 10, no. 1, p. 290, Sep. 2009.

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.

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.

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 Trans. Comput. Biol. Bioinf.*, vol. 12, no. 2, pp. 276–288, Mar. 2015.

M. Li, Z. Niu, X. Chen, P. Zhong, F. Wu, and Y. Pan, “A reliable neighbor-based method for identifying essential proteins by integrating gene expressions, orthology, and subcellular localization information,” *Tsinghua Sci. Technol.*, vol. 21, no. 6, pp. 668–677, Dec. 2016.

X. Zhang, W. Xiao, and X. Hu, “Predicting essential proteins by integrating orthology, gene expressions, and PPI networks,” *PLoS ONE*, vol. 13, no. 4, Apr. 2018, Art. no. e0195410.

E. Gultepe and M. Makrehchi, “Improving clustering performance using independent component analysis and unsupervised feature learning,” *Hum.-Centric Comput. Inf. Sci.*, vol. 8, no. 1, Aug. 2018.

M. Li, W. Li, F.-X. Wu, Y. Pan, and J. Wang, “Identifying essential proteins based on sub-network partition and prioritization by integrating subcellular localization information,” *J. Theor. Biol.*, vol. 447, pp. 65–73, Jun. 2018.

W. Zhang, J. Xu, and X. Zou, “Predicting essential proteins by integrating network topology, subcellular localization information, gene expression profile and GO annotation data,” *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.

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.

M. Li, P. Ni, X. Chen, J. Wang, F.-X. 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. 2019.

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.

B. Zhao, Y. Zhao, X. Zhang, Z. Zhang, F. Zhang, and L. Wang, “An iteration method for identifying yeast essential proteins from heterogeneous network,” *BMC Bioinf.*, vol. 20, no. 1, p. 355, Jun. 2019.

G. Li, M. Li, W. Peng, Y. Li, Y. Pan, and J. Wang, “A novel extended Pareto optimality consensus model for predicting essential proteins,” *J. Theor. Biol.*, vol. 480, pp. 141–149, Nov. 2019.

M. Li, Y. Lu, Z. Niu, and F.-X. Wu, “Unified complex centrality for identification of essential proteins from PPI networks,” *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 2, pp. 370–380, Mar. 2017.

M. Li, Y. Lu, Z. Niu, F.-X. Wu, and Y. Pan, “Identification of essential proteins by complexes and interaction network,” *Bioinform. Res. Appl.*, pp. 255–265, 2014.

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, Aug. 2016, Art. no. e0161042.
LIANGYU MA was born in Nantong, Jiangsu, China, in 1994. She is currently pursuing the master's degree with the College of Information Engineering, Yangzhou University, Yangzhou, China. Her research topic is bioinformatics.

LING CHEN (Member, IEEE) was born in 1951. He graduated from the Mathematics Department, Yangzhou Teachers’ College. He is currently a Professor of computer science with the Information Technology College, Yangzhou University. He has published more than 200 articles in journals and conferences. He has also authored/coauthored six books. He has 15 research projects supported by the Chinese Natural Science Foundation and other organizations. His research interests include artificial intelligence, data mining, system optimization, and complex network analysis. He is a member of ACM. He received five Awards of Progress in Science and Technology from the Government of Anhui and Jiangsu Province. He was awarded the Government Special Allowance by the State Council.

BYEUNWOO JEON (Member, IEEE) received the B.S. degree (magna cum laude) and the M.S. degree in electronics engineering from Seoul National University, Seoul, South Korea, in 1985 and 1987, respectively, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, in 1992. From 1993 to 1997, he was with the Signal Processing Laboratory, Samsung Electronics, South Korea, where he conducted research and development on video compression algorithms, the design of digital broadcasting satellite receivers, and other MPEG-related research for multimedia applications. Since September 1997, he has been the Faculty with the School of Electronic and Electrical Engineering, Sungkyunkwan University, South Korea, where he is currently a Full Professor. He served as the Project Manager of digital TV and broadcasting at the Korean Ministry of Information and Communications, from March 2004 to February 2006, where he supervised all digital TV-related R and D, South Korea. He has authored many articles in the areas of video compression, preprocessing/postprocessing, and pattern recognition. His research interests include multimedia signal processing, video compression, statistical pattern recognition, and remote sensing. Dr. Jeon is also a member of Tau Beta Pi and Eta Kappa Nu. He is also a member of SPIE, IEEK, KICS, and KSOBE. He was a recipient of the 2005 IEEK Haedong Paper Award from the Signal Processing Society, South Korea.

* * *