frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 04 October 2022
DOI 10.3389/fmicb.2022.963704

OPEN ACCESS

EDITED BY
Qi Zhao,
University of Science and Technology Liaoning, China

REVIEWED BY
Dingjie Wang,
The Ohio State University, United States
Mingzhi Liao,
Northwest A&F University Apple Research Center, China

*CORRESPONDENCE
Yuan Zhu
zhuyuan@cug.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 07 June 2022
ACCEPTED 16 August 2022
PUBLISHED 04 October 2022

CITATION
Wang C, Zhang H, Ma H, Wang Y, Cai K, Guo T, Yang Y, Li Z and Zhu Y (2022) Inference of pan-cancer related genes by orthologs matching based on enhanced LSTM model. *Front. Microbiol.* 13:963704. doi: 10.3389/fmicb.2022.963704

Inference of pan-cancer related genes by orthologs matching based on enhanced LSTM model

Chao Wang[1†], Houwang Zhang[2†], Haishu Ma[3,4,5], Yawen Wang[6], Ke Cai[3,4,5], Tingrui Guo[3,4,5], Yuanhang Yang[6], Zhen Li[6] and Yuan Zhu[3,4,5,7*]

[1]Department of Surgery, Hepatic Surgery Center, Institute of Hepato-Pancreato-Biliary Surgery, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2]Department of Electrical Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR, China, [3]School of Automation, China University of Geosciences, Wuhan, China, [4]Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan, China, [5]Engineering Research Center of Intelligent Technology for Geo-Exploration, Wuhan, China, [6]School of Mathematics and Physics, China University of Geosciences, Wuhan, China, [7]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Shanghai, China

Many disease-related genes have been found to be associated with cancer diagnosis, which is useful for understanding the pathophysiology of cancer, generating targeted drugs, and developing new diagnostic and treatment techniques. With the development of the pan-cancer project and the ongoing expansion of sequencing technology, many scientists are focusing on mining common genes from The Cancer Genome Atlas (TCGA) across various cancer types. In this study, we attempted to infer pan-cancer associated genes by examining the microbial model organism Saccharomyces Cerevisiae (Yeast) by homology matching, which was motivated by the benefits of reverse genetics. First, a background network of protein-protein interactions and a pathogenic gene set involving several cancer types in humans and yeast were created. The homology between the human gene and yeast gene was then discovered by homology matching, and its interaction sub-network was obtained. This was undertaken following the principle that the homologous genes of the common ancestor may have similarities in expression. Then, using bidirectional long short-term memory (BiLSTM) in combination with adaptive integration of heterogeneous information, we further explored the topological characteristics of the yeast protein interaction network and presented a node representation score to evaluate the node ability in graphs. Finally, homologous mapping for human genes matched the important genes identified by ensemble classifiers for yeast, which may be thought of as genes connected to all types of cancer. One way to assess the performance of the BiLSTM model is through experiments on the database. On the other hand, enrichment analysis, survival analysis, and other outcomes can be used to confirm the biological importance of the prediction results. You may access the whole experimental protocols and programs at https://github.com/zhuyuan-cug/AI-BiLSTM/tree/master.

KEYWORDS

microbe-disease, orthologs, essential proteins, deep learning, BiLSTM model

1. Introduction

Cancer is a malignant and complex kind of disease that seriously endangers human existence. Because of its rapid spread, early onset, and high death rate, cancer is a disease that is challenging to cure. According to the American Cancer Society, there will be 608,570 cancer-related deaths and 1,898,160 new cases of cancer in the nation in 2021 (Xia et al., 2022). The prevention and treatment of cancer have evolved into a public health issue that requires collective human effort. A growing number of scholars are dedicating themselves to pan-cancer research as it is a hot topic globally. The therapeutic treatment of viral diseases, genetic diseases, and other diseases may be improved by the use of gene therapy (Ma et al., 2020). Therefore, accurate detection of pan-cancer genes is essential for understanding cancer and provides better benefits for its prevention, treatment, and development of anti-cancer drugs, which is relevant from a social and economic perspective (Aromolaran et al., 2021).

Currently, the identification of essential genes is the main source of the issue with pan-cancer associated genes prediction. In previous decades, biological experiments including single gene knockout, conditional knockout, and RNA interference were used as the typical methods for identifying essential proteins. These experimental techniques require lengthy and expensive procedures, and the experimental settings frequently affect the outcomes. The same organism may respond differently to different experimental settings (Zhong et al., 2021). An enormous number of protein-protein interactions (PPI) enriched with gene expression data have been available in recent years benefiting from the advancement of high-throughput technology (Li et al., 2017).

According to the two sides, studies on cancer-related genes can be roughly split into two categories. It is intended to investigate the tissue-specific driver genes, on the one hand. The ideas pertaining to complex network analysis were transferred and utilized to biological network analysis by merging cancer sample data onto biological networks. Each node in the network structure had its level of importance evaluated, and the genes with the highest value were found to be the cancer driver genes. Since genes only selectively express proteins, essential proteins can be used to discover essential genes. Numerous effective network-based techniques have been put forth over years to identify crucial proteins from PIN. The most well-known and straightforward one is degree centrality (DC) (Jeong et al., 2001). According to a molecular theory known as the centrality-lethality rule, the highly linked nodes within the PIN serve as its fundamental structural components and are generally more significant than other nodes (Jeong et al., 2001; Zotenko et al., 2008). Other node topological feature-based methods, such as subgraph centrality (SC) (Estrada and Rodriguez-Velazquez, 2005), eigenvector centrality (EC) (Bonacich, 1987), betweenness centrality (BC) (Joy et al., 2005), closeness centrality (CC) (Wuchty and Stadler, 2003), information centrality (IC) (Stephenson and Zelen, 1989) and others, are also used to identify proteins in addition to DC. These techniques assess each node according to its topological structure. In general, network-based approaches are extensively employed in the early stages since they can predict important proteins directly without the need for further information. However, these techniques feature low recall rates and identification precision due to the abundance of false positive and false negative data in PPI networks (Li et al., 2016). The intrinsic biological importance of necessary proteins is also disregarded by these techniques, which ignores essential proteins with low connectivity (Li et al., 2016). Recent research has attempted to incorporate biological knowledge into network-based techniques, which not only reduce the impact of false positives in PPI data but also significantly increase the prediction accuracy of essential proteins (Li et al., 2012; Zhang et al., 2019; Wang et al., 2021). Ess-NEXG (Wang et al., 2020) and DeepEP (Zeng et al., 2019; Liu et al., 2022b) are two related algorithms for finding essential proteins that have been developed as a result of the rapid growth of deep learning. Other algorithms have also been presented to predict other associations (Zhang et al., 2021; Liu et al., 2022a,b).

On the other side, it seeks to identify potential disease-related genes across a variety of malignancies. Several computational methods have been proposed to uncover pan-cancer related genes or driver module types by integrating multi-omics data across various malignancies (Cao and Zhang, 2016; Zhang and Zhang, 2016, 2017; Yang et al., 2017; Li et al., 2020), which is motivated by the objectives of the cancer genome program named The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013). By combining existing information on cancer from various types of tumors, potential patterns and biological processes are investigated. For example, Park et al. (2016) proposed an algorithm called NTriPath based on matrix decomposition to identify and complement pathogenic gene pathways, which overcomes the limitation of studying a single cancer and can complement the existing set of pathogenic pathway genes in multiple cancers. In order to identify possible pan-cancer related genes, Zhu et al. (2022) combined the network representation method with differential expression analysis.

Geneticists have long noted that functional relationships frequently exist between mutations that result in the same biological manifestation. Utilizing these predictions to connect particular genes to phenotypes opens the door to using similar techniques to directly find new disease genes in the study of human genes. In reverse genetics, it is feasible to infer linked phenotypes based on linkages in functional gene networks (Sommer, 2008). Homologous genes are genes found in several species that descended vertically from a single gene found in the last common ancestor, which is how organisms evolved from a common ancestor. Highly identical DNA sequences

between two homologous genes, which may also have the same function, are extremely likely to be found in two animals with very close affinity (Müller, 2003). The concept of homology allows us to more easily study human genes with gene sequences from other species. Similar structures and functionalities are shared by genes that are crucial for life's functions in model organisms. Furthermore, there is mounting evidence that model species are essential for addressing issues connected to the gene variations that underlie human disease. Using model organisms for homology mapping can help us understand human pathogenic genes (Bleackley and MacGillivray, 2011).

Due to its genetic flexibility, small genome size, and manipulability, yeast is one of the model organisms with the highest genetic adaptability. Yeast is a single-cell eukaryote that helps to uncover many fundamental concepts in biology and reveals the activity of human cells. Consequently, yeast is essential for identifying genetic variations in human genes related to illnesses and encoding genetic variations in proteins engaged in multiple pathways. The study revealed a link between the microbiota and associated diseases, and it is crucial to understand the molecular mechanisms of these diseases in order to develop new microbiome-based therapies. Microbiota is the microbial population colonizing multiple organ systems in humans and impacting the outcomes of microbiota-related diseases (Belkaid and Hand, 2014; Sun et al., 2022). Among them, gut microbiota, a dense microbial community in human intestines, has been found closely associated to acute kidney injury (Lei et al., 2022), atherosclerosis (Anto and Blesso, 2022), reduced bone mineral density (Wan et al., 2022), age-related neuroinflammation and cognitive decline (Alsegiani and Shah, 2022), carcinogenesis and cancer immunotherapy resistance (Hersi et al., 2022), and metabolic disorders such as hyperlipidemia, hyperglycemia, hypertension, obesity and diabetes (Beg et al., 2022). Manipulation of the gut microbiota has broad application prospects on diseases. Fecal microbiota transplant (FMT) is one of the microbiome-based therapeutics with clinical application potential in clostridioides difficile colitis, graft-vs.-host disease, and inflammatory bowel disease (Sorbara and Pamer, 2022). In addition, engineered bacteria, postbiotics, and phages are also used as precision microbiome-centered therapies (Bajaj et al., 2022).

Multiple biological data are currently available due to the advancement of sequencing technologies, enabling it to integrate multi-omics data from various tumors to uncover genes associated to pan-cancer. In this study, we use the yeast network to predict human disease genes. We gathered a pathogenic gene set from multiple cancers. Homologous mapping is then utilized to locate the homologs integrating all of the pathogenic genes of ten tumors. We propose a parameter adaptive model for characterizing node representation ability by merging Subcellular localization information, Gene expression data, and Protein Complexes data with the specifically designed topological properties of the PPI network, which is called PSGN score for short. Additionally, the BiLSTM, a LSTM model with adjacency constraint and multiple features, is proposed for the prediction of essential proteins. The yeast genes that are similar to the seed genes are screened as candidate genes using the BiLSTM algorithm. In order to identify the final predicted human pan-cancer associated genes, homolog mapping of these candidate genes was performed.

Comparative experiments were conducted on the publicly accessible PPI data of Yeast, in order to validate the effectiveness of the proposed evaluation PSGN score and the classification results of BiLSTM. We verified the efficacy of the new proposed score by contrasting the performance of PSGN with classic unsupervised approaches including DC, BC, CC, EC NC, LAC, PeC, and WDC. Further, we compared our BiLSTM model to established machine learning techniques like SVM, decision tree, ensemble learning-based methods, and the most recent deep learning-based approach put forth by Zeng et al. (2019). According to the experimental findings, BiLSTM may identify essential proteins with superior overall outcomes than other cutting-edge techniques. Besides, some biological significance experiments were conducted on real datasets, the results validated the effectiveness of the new proposed algorithm from the reverse genetics perspective. The remaining parts are organized as follows. Section 2 presents the material and methods of the new proposed method. Experimental results and discussions are illustrated in Section 3. Finally, Section 4 concludes the work.

2. Materials and methods

2.1. Datasets

PPI networks: among other species, the PPI network dataset of yeast is the most reliable and complete, making it popular for use in evaluating and identifying essential proteins. Therefore, in this study, we also selected the yeast PPI network dataset. The DIP database is used to gather the PPI data of yeast (Xenarios et al., 2002). There are 5,093 proteins and 24,743 interactions in total after subtracting self-interactions and repetitive interactions.

Essential protein datasets: A list of essential proteins of yeast were collected from the following databases: MIPS (Mewes et al., 2006), SGD (Cherry et al., 2012), and DEG (Zhang and Lin, 2009). A protein in the yeast protein interaction network is considered as an essential protein if it is marked as essential at least in one database. This data has 1,285 essential proteins, 1,167 of which are included in the PPI network constructed from the DIP database. Hence, we take the 1,167 proteins as essential proteins and the rest 3,926 proteins as non-essential ones.

Subcellular localization dataset: the dataset is available in the knowledge channel of COMPARTMENTS database (Binder et al., 2014), which combines the UniProtKB (Magrane, 2011),

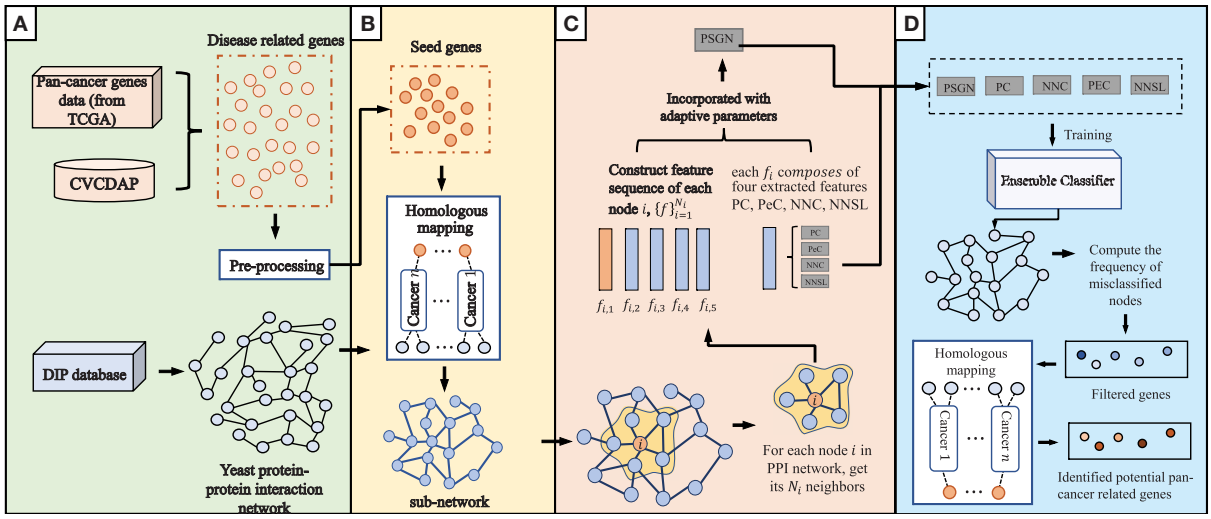

FIGURE 1
An overview of our proposed method to predict pan-cancer related genes *via* orthologs matching. The main algorithm consists of **(A–D)** four parts. **(A)** data integration and pre-processing. **(B)** Homologous mapping. **(C)** Evaluation of node representation ability in graph. **(D)** Prediction of pan-cancer related genes.

MGD (Eppig et al., 2012), SGD (Cherry et al., 2012), FlyBase (Mcquilton et al., 2012), and WormBase datasets (Harris et al., 2010). There are 206,831 subcellular localization records in this dataset, which can be further subdivided into 830 categories.

Protein complex datasets (Luo and Qi, 2015): it is comprised of four real protein complex sets (CM270, CM425, CYC408, and CYC428). Seven hundred and forty-five protein complexes are included in the consolidated dataset.

Gene Expression Omnibus (GEO) dataset: GSE3431 derives from GEO and samples 12 time points during each of three yeast successive metabolic cycles (the interval between two time points is 25 min). The dataset contains 36 samples with 6,777 genes.

Online Mendelian Inheritance in Man (OMIM) dataset: we retained only disease-related variants linked to a genetic disorder listed in the OMIM database. Cross-references were used to directly access annotations for each OMIM disease by downloading the DO (Human Disease Ontology) OBO (Open Biological and Biomedical Ontology) file release. Each retrieved leaf DO term connected to a single OMIM was expanded to include all ancestors and the ontological root term. Term expansion was calculated by parsing the OBO file with an impromptu script.

The Cancer Genome Atlas (TCGA) Database: the Human Genome Research Institute (HGRI) and National Cancer Institute (NCI) launched the Cancer Genome Mapping Project in 2006. The database contains more than 20,000 samples from 33 cancer types, including transcriptome expression data, genome variation data, methylation data, clinical data, and others, which can be accessed *via* https://portal.gdc.cancer.gov/exploration.

2.2. Overview of the new proposed method

The current proposed method, which consists of four main steps of data integration and pre-processing, homologous mapping, evaluation of node representation ability, and prediction of pan-cancer related genes, is shown in detail in Figure 1.

2.2.1. Data integration and pre-processing

The gene expression data of 10 cancers were obtained from TCGA database, including esophageal carcinoma, pancreatic cancer, lung cancer (lung adenocarcinoma, lung squamous cell carcinoma), breast invasive carcinoma, colon adenocarcinoma, rectum adenocarcinoma, cholangiocarcinoma, gastric cancer and ovarian cancer. Due to the duplications and deletions in the pathogenic genes of each cancer, they are used as experimental data after sorting and deletion. We uploaded the TCGA data of 10 cancers selected in the CVCDAP database (https://omics.bjcancer.org/cvcdap/home.do), successfully generated the pan-cancer related pathogenic gene set, and completed the analysis of the pan-cancer network driving genes with the help of the analysis tool of CVCDAP database. We obtained the data of Yeast protein interaction network on DIP database (https://dip.doe-mbi.ucla.edu/dip/Main.cgi) and

downloaded the connection information between Yeast protein nodes directly.

2.2.2. Homologous mapping

We believe that the homologous genes of co-ancestors express themselves similarly. The NCBI Homologene database (https://www.ncbi.nlm.nih.gov/homologene/) compiles homologous gene data for species with complete genome sequencing. In this section, we used Homologene package in the R language, the imported human pathogenic genes were annotated by homology mapping, and the homologous genes of human and yeast genes were taken as seed genes. After identifying the proteins expressed by the seed genes, the interaction network among these yeast proteins can be determined by the STRING database (https://cn.string-db.org/).

2.2.3. Evaluation of node representation ability

Yeast is one of the most genetically model organism. In this study, we firstly explore the essential proteins in the yeast PPI network to further find potential disease related genes. Thus, a new score is defined to evaluate the node representation ability. The PPI network is denoted as graph $G = (V, E)$, where $V = \{v_1, \cdots, v_m\}$ and $E = \{e_{i,j}, 1 \leq i, j \leq m\}$ represent the node set and edge set of the graph, respectively. Specifically, v_i denotes the i-th protein while e_{ij} denotes protein-protein interaction linkage between protein v_i and v_j. $|V| = m$ represents the number of total proteins within G.

The features of our new proposed score considers node-aided biological information, edge-aided biological information and network topological features. We'll go through how to use and integrate this data to create the attributes needed to determine a protein's essentiality in the subsections that follow. The establishment process requires three specific steps.

Step 1: Construction of node represented features

1) Protein complexes score: previous studies indicated that intracellular proteins always tend to connect with their neighbors to form densely connected modules, which are called protein complexes and by this way proteins could take part in more complex and diverse biological activities and functions (Luo and Qi, 2015). Given that essential proteins are crucial in maintaining the main structure and functions in protein complexes (Zotenko et al., 2008), protein complexes data could be used for the identification of essential proteins (Lei et al., 2018).

For the protein v_i, the essentiality tends to be higher if it is found in more protein complexes. In order to calculate the protein complexes (PC) score, we do the following:

$$PC(i) = |Complex(i)| \tag{1}$$

where $Complex(i)$ denotes the sets of protein complexes including v_i, and $|Complex(i)|$ is the number of protein complexes including v_i.

2) Subcellular localization score: it has been proved that proteins must be localized at their appropriate subcellular compartments to perform their desired functions and thus the subcellular localization information is beneficial for the identification of essential proteins (Peng et al., 2015). To ensure the relationships of subcellular localization with the topological features of PPI network, refer to Li et al. (2016), we firstly use the previous feature NNC to sort the proteins within the PPI network, and then calculate the numbers of subcellular location l where the top k% proteins appear and where the bottom k% proteins appear, respectively.

Given the data's false positives, counting proteins at higher rates may result in more errors; as a result, we use $k = 5$ in this work as Li et al. (2016) sets, i.e., that the top/bottom 5% proteins are selected. Besides, we define T_l as the frequency of the localization l where the top k% proteins appear and B_l as the frequency of the localization l where the bottom k% proteins appear. Subcellular localization correlation coefficient SLCC(l) can be calculated by Equation (2)

$$SLCC(l) = \begin{cases} 1 - \dfrac{B_l}{T_l}, & T_l < B_l; \\ \dfrac{T_l}{B_l} - 1, & otherwise, \end{cases} \tag{2}$$

when $T_l < B_l$, it means that more proteins with low NNC values tend to appear in the location l and it is assumed that the relationship between the location l and the essentiality of proteins is negative. On the other hand, when $T_l \geq B_l$, there should also be a positive correlation between the location l and the essentiality of proteins. When $T_l = 0$, we set $SLCC(l)$ as the maximum of $1 - \dfrac{B_l}{T_l}$ with $T_l \neq 0$. And when $B_l = 0$, we set $SLCC(l)$ as the maximum of $\dfrac{T_l}{B_l} - 1$ with $B_l \neq 0$.

Besides, considering that a protein may appear in multiple subcellular locations, take protein v_i for instance, its subcellular localization score $SL(i)$ could be calculated as the sum of $SLCC(l)$ of all the subcellular locations where it appears. Moreover, the normalized value $NSL(i)$ of SL for each protein v_i is used by Equation (3)

$$NSL(i) = \frac{SL(i) + max\_SL}{Max(SL(i) + max\_SL)}, \tag{3}$$

where max_SL represents the maximum value of $SL(i)$ for all the proteins within the PPI network. Maxin the denominator takes for all the nodes within the PPI network.

In order to strengthen the identification precision of subcellular localization, we combine the NSL score with a network topological feature NNEC that is proposed in Zhu and Wu (2018) and has a good compatibility with biological

information. The combined feature is called NNSL for short, for each protein v_i, its $NNSL(i)$ score can be calculated by Equation (4)

$$NNSL(i) = NSL(i) \times NNEC(i), \tag{4}$$

where $NNEC(i) = \sum_{j \in \mathbb{N}(i)} NECC(i,j)$ and $NECC$ can be obtained by Equation (5)

$$NECC(i,j) = \frac{T(i,j)^3 \times C(j)}{\prod_{t=\{i,j\}}(d(t)-1)}, \tag{5}$$

where $T(i,j)$ denotes the number of triangles made up of proteins v_i and v_j, $C(j) = \frac{2\mathbb{E}_j}{d(j)(d(j)-1)}$ is the clustering coefficient of protein v_j, $\mathbb{E}_j$ is the number of non-repetitive edges consisting of all nearest neighbors of v_j. $d(t)$ denotes the degree for protein t, for $t = i$ or j.

Step 2: Construction of edge represented features

Gene expression data is a type of biological information that has been utilized for a long time to compute edge correlations and identify essential proteins. *PeC* is a method that combines gene expression data with edge clustering coefficient *ECC* in order to reduce the impact of false positives on the PPI network. As a result, we apply *PeC* in this study to extract pertinent information from gene expression data. For a protein v_i, its PeC score $PeC(i)$ can be computed by Equation (6)

$$PeC(i) = \sum_{j \in \mathbb{N}(i)} ECC(i,j) \times PCC(i,j), \tag{6}$$

where $ECC(i,j)$ is the edge coefficient between edge $e_{i,j}$, $PCC(i,j)$ is the Pearson's correlation coefficient of a pair of proteins (v_i and v_j). s denotes the length of the gene expression data, which can be calculated by Equation (7)

$$PCC(i,j) = \frac{1}{s-1}\sum_{t=1}^{s}[\frac{g(i,t)-\bar{g}(i)}{\sigma(i)}] \times [\frac{g(j,t)-\bar{g}(j)}{\sigma(j)}], \tag{7}$$

where $g(i,t)$ and $g(j,t)$ are the expression level of v_i and v_j in the sample time t under a specific condition, $\bar{g}(i)$ and $\bar{g}(j)$ represent the mean expression level of v_i and v_j, and $\sigma(i)$ and $\sigma(j)$ represent the standard deviation of expression level of v_i and v_j, respectively.

To extract the topological information of proteins within the PPI network, it is necessary to construct an effective feature representing the network structures of the nodes and connections with neighbors. Network centrality (NC) is a representative topology based method widely used for predicting essential proteins (Wang et al., 2012). Hence, we choose it for network topological feature construction. For the protein v_i, its network centrality $NC(i)$ can be calculated as the sum of edge clustering coefficients $ECC(i,j)$ of each edge $e_{i,j}$ connected with v_i by Equation (8)

$$\begin{aligned} NC(i) &= \sum_{j \in \mathbb{N}(i)} ECC(i,j) \\ &= \sum_{j \in \mathbb{N}(i)} \frac{T(i,j)}{min(d_i - 1, d_j - 1)}, \end{aligned} \tag{8}$$

where $\mathbb{N}(i)$ is the set of nodes which directly connect with protein v_i.

In order to match other features based on biological information, here we use the normalized NC value (denoted as NNC) for each protein. Then for v_i, its normalized value $NNC(i)$ is defined by Equation (9)

$$NNC(i) = \frac{NC(i)}{max(NC(i))}, \tag{9}$$

where $max(NC(i))$ denotes the maximum NC value of all the proteins in the graph G, and the value of $NNC(i)$ will be normalized between 0 and 1.

Step3 : Feature integration by liniear model with adaptive parameters

The structure of hetergeous feature integration involves Protein complex $PC(i)$, Subcellular localization $NNSL(i)$, Gene expression $PeC(i)$ and Network topology $NNC(i)$ multiple information (PSGN). Here, we reconcile these features using a linear model in order to fully integrate this information. Take protein v_i within the PPI network for instance, its evaluation score could be calculated by PSGN score presented in Equation (10)

$$\begin{aligned} PSGN(i) = &((PC(i) + NNSL(i) \times a + PeC(i) \times (1-a)) \\ &\times b + NNC(i) \times (1-b) \end{aligned} \tag{10}$$

where a and b are two weights to balance these heterogeneous features. And a is utilized for combining the node based and edge based biological features, b is set to integrate the topological features and biological features.

When integrating numerous pieces of information, several methods for identifying essential proteins require for the adjustment of parameters and the setting of an optimal one for feature combinations. In contrast, our approach proposes an adaptable parameter strategy to deal with various information based on the unique number of essential proteins that must be identified. These are the concepts: depending on the number of essential proteins we need to identify, the adaptive domain of each piece of information varies.

For example, we use PC and NNC two features to identify essential proteins of Yeast PPI dataset respectively. Through

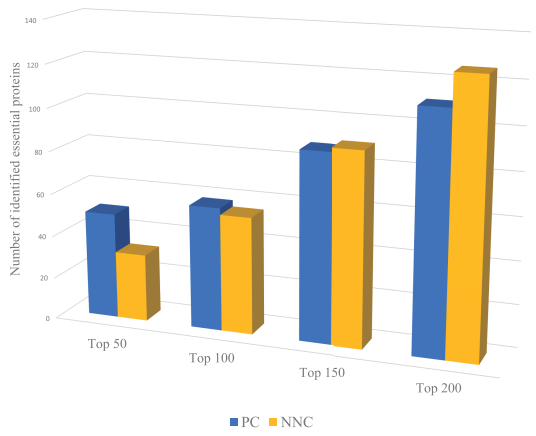

FIGURE 2
The number of essential proteins predicted by PC and NNC in different scales.

Figure 2, PC can capture more essential proteins compared with NNC when dealing with proteins with higher ranking positions. And for proteins with lower positions in rank, the effect of PC is not so significant as NNC. That means when we need to identify the proteins with higher ranking positions (like Top 50, Top 100), we need to assign larger weights on PC. On the contrary, to predict essential proteins with lower ranking positions (like Top 150, Top 200), NNC should be assigned with larger weights. However, most methods will give constant parameters which ignore the variation of functions of different biological information for identification when the number of essential proteins needed to be predicted changes.

In general, the effect of biological information is more reliable than topological features of network when dealing with proteins with higher ranking positions. Therefore, the weight should be adjusted adaptively according to the number of essential proteins needed to be identified. The parameter adaptive model is proposed by Equation (11)

$$P = \alpha_i + \beta_i \times input, \tag{11}$$

where input is the expected number of essential proteins needed to be identified. In this research, $i = 1$ or 2, when $i = 1$, $P = a$, by test, we take $\alpha_1 = 0.49$, $\beta_1 = -0.0005$, when $i = 2$, $P = b$, by test, we take $\alpha_2 = 1$, $\beta_2 = -0.0003$. This parameter model means that, the weights of biological information are greater when calculating the top ranked essential proteins, especially the node-aided biological information (PC and NNSL). With the increase of input, the weight of network-based topological feature (NNC) gradually increases, and the weight of edge-aided biological information (PeC) also increases gradually.

2.3. Prediction of potential pan-cancer related genes

As we discussed above, for proteins in the PPI network, the proteins' feature can be represented by NNC, NNSL, PC, PeC and PSGN. As it is shown in Figure 1B, the seed proteins are labeled as 1 and other proteins in yeast PPI network are labeled as 0. The final prediction results *via* enhanced BiLSTM model *via* repeated experiments as shown in Figure 3. Then the representation can be divided into training dataset and testing dataset, we sample the data from the embedding vector of pan-cancer network based on cross-validation. As shown in Figure 1, this process is trained by multiple classifiers on the sampled data. After obtaining the trained classifiers, we use them to pre- dict pan-cancer-related genes. For each predicted node, the frequency of the node can be considered as the decision metric in the training processes. Finally, the final node representation ability can be calculated by counting the frequency. We take nodes with proper frequencies as potential candidate pan-cancer genes. The whole procedures of our proposed approach AI-BiLSTM are presented in Algorithm 1.

3. Results and discussion

In this research, we investigate the interaction network of the model microbial Yeast, and find potential pan-cancer related genes by homologous mapping. Firstly, the LSTM model was used to categorize the essential genes in the Yeast interaction network, and then homology matching was used to further mine the disease genes. Therefore, the experimental analysis was carried out from two aspects. On the PPI datasets for yeast, we compared the performance of the novel proposed BiLSTM model with several conventional approaches. Secondly, we validated biological significance of the predicted genes through GO enrichment analysis, pathway analysis, survival analysis, clustering analysis and so forth. All of the approaches that are compared in this study adopt their default parameters. All the experiments are run on a personal computer with Windows 10 OS, Intel Core i7 2.3GHz CPU, and 16GB memory.

3.1. Effectiveness of the new proposed BiLSTM model

3.1.1. Evaluation of PSGN

For PSGN, similar to most of validation methods for the identification of essential proteins, we also ranked all proteins by using each essential protein identification method in a descending order. And then we selected a certain number of top

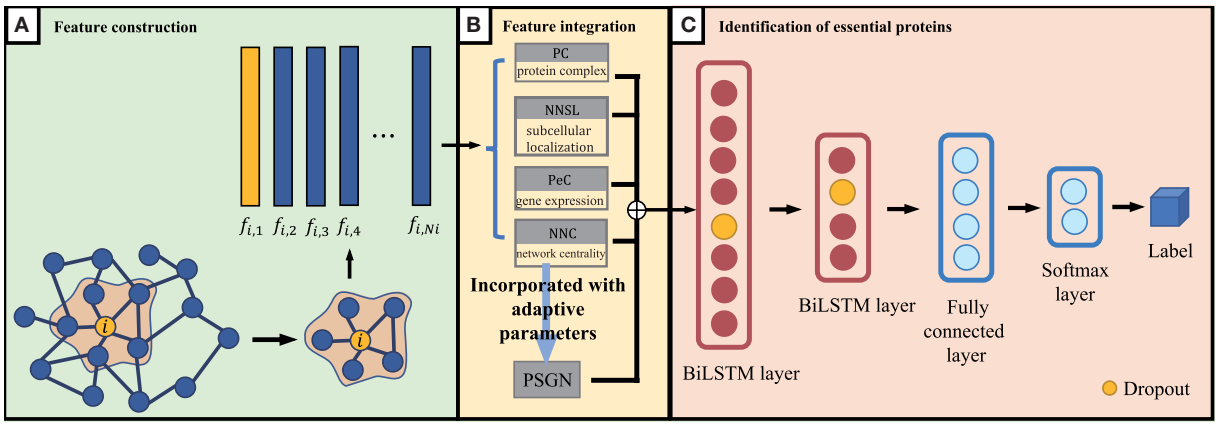

FIGURE 3
An overview of our proposed BiLSTM model. **(A)** Feature construction, **(B)** Feature integration, and **(C)** Identification of essential proteins.

Input: The PPI network $G = (V, E)$, protein complex, subcellular localization, gene expression data, threshold
Output: The classification label for proteins;
Calculate PC for each protein by using Equation (1);
Calculate NNSL for each protein by using Equation (4);
Calculate PeC for each protein by using Equation (6);
Calculate NNC for each protein by using Equation (9);
Incorporate PC, NNSL, PeC and NNC by using the adaptive parameters through Equation 10 to obtain PSGN score;
Integrate protein feature representation enhanced by [NNC, NNSL, PC, PeC, PSGN];
for i $(1 \rightarrow n)$ do
Random select (1/n)% data as training dataset, others as test dataset;
Classification and fix the protein label by BiLSTM;
end
Count the frequency of the predicted essential gene (labeled 1)
return The genes with frequencies greater than the threshold.

Algorithm 1. BiLSTM for prediction of potential pan-cancer related genes.

ranked proteins as the essential protein candidates (like top 100), after that the accuracy of identification could be computed by counting the number of true essential proteins.

Figure 4 gives a specific comparison of the results of identification of essential proteins. As shown in the figure, PSGN can identify more essential proteins compared with the other eight methods. The number of true essential proteins identified by PSGN is higher than other methods in the top 100, top 200, top 300, top 400, top 500, and top 600 proteins. In addition, by observing the results of the top 100 proteins, we find that PSGN can obtain a prediction precision of 90%, which is much higher than other methods.

For better comparison, the precision-recall (PR) curve, a common methodology for evaluating the performance of essential proteins identification methods, is used in this paper. The comparison of our method with the other methods for predicting essential proteins on the Yeast PPI network by using the PR curve is shown in Figure 5. The PR curve of PSGN obtains the better result compared to the PR curves of other methods. Our method significantly exceeds other methods with the largest AUC value, illustrating the effectiveness of our method.

To further evaluate its effectiveness, we take the jackknife curve to compare the prediction results of our proposed method PSGN with other methods. The results are shown in Figure 6. The x-axis denotes the number of proteins ranked by each essential protein identification method and the y-axis is the number of truly identified essential proteins of each method. The areas under the jackknife curves can measure the performances of the method for identifying essential proteins. As shown in Figure 6, the jackknife curve of our proposed method PSGN can identify more essential proteins from the Yeast PPI network compared with other methods, demonstrating that PSGN is more effective and can get better results than other state-of-art methods.

For interpreting the advantages of our method in deeper levels, we also choose 5 widely used metrics (sensitive, specificity,

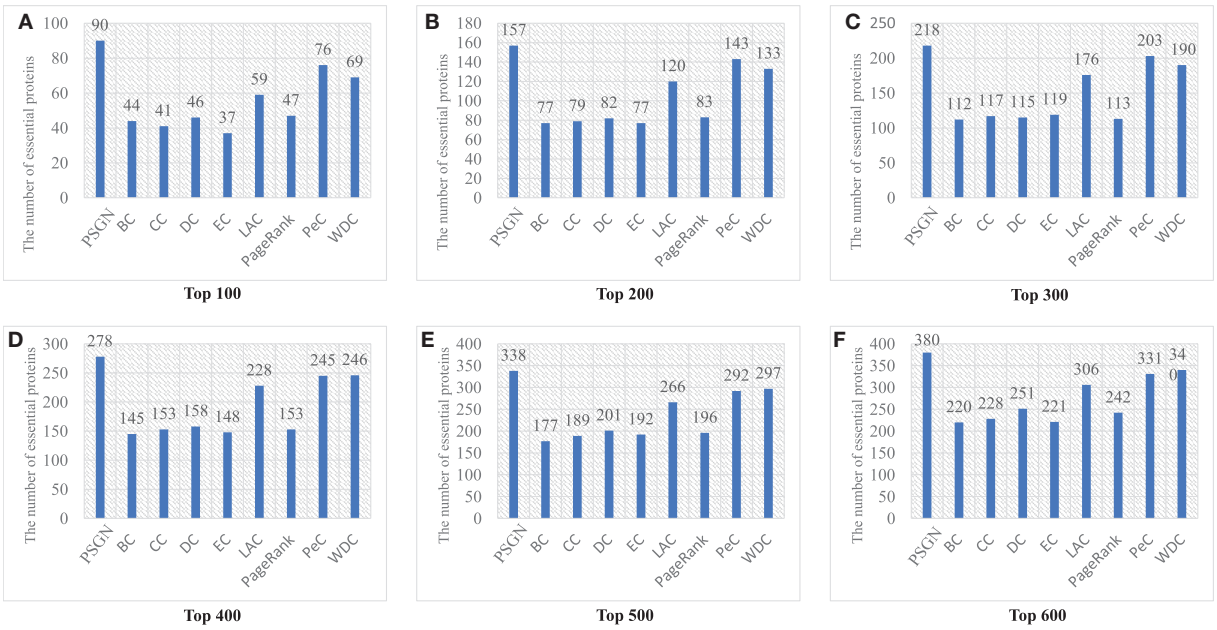

FIGURE 4
The number of essential proteins predicted by PSGN, BC, CC, DC, EC, LAC, PageRank, PeC, and WDC. **(A–F)** show the results of these methods when selecting top 100 to 600 ranked proteins as candidates of essential proteins.

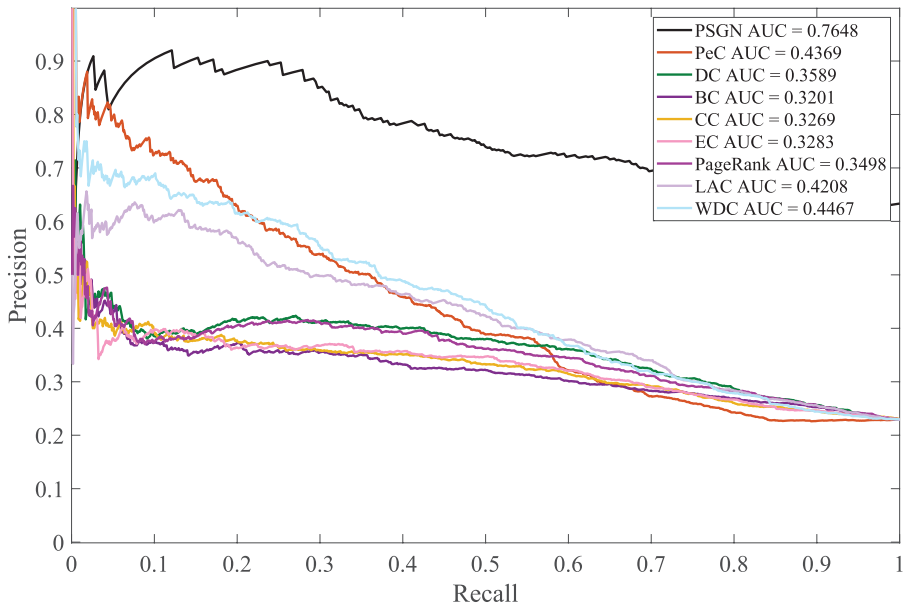

FIGURE 5
Comparison of PSGN, BC, CC, DC, EC, LAC, PageRank, PeC, and WDC using precision-recall (PR) curve method.

precision, F-measure, and accuracy) to evaluate all the methods. Figure 7 shows the results of 5 evaluation metrics obtained by all identification methods on the PPI network of Yeast. As shown in the figure, it is obvious that our proposed PSGN can outperform other methods significantly in terms of all 5 evaluation metrics.

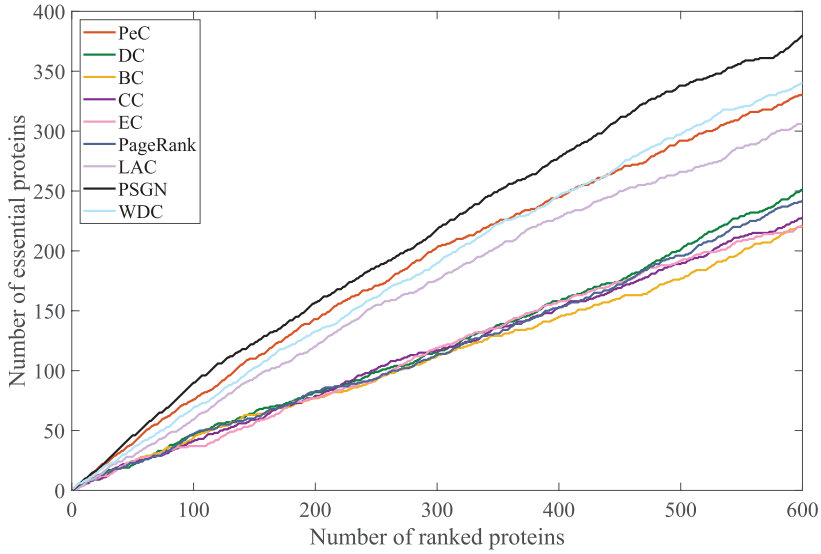

FIGURE 6
Comparison of PSGN, BC, CC, DC, EC, LAC, PageRank, PeC, and WDC using Jackknife method.

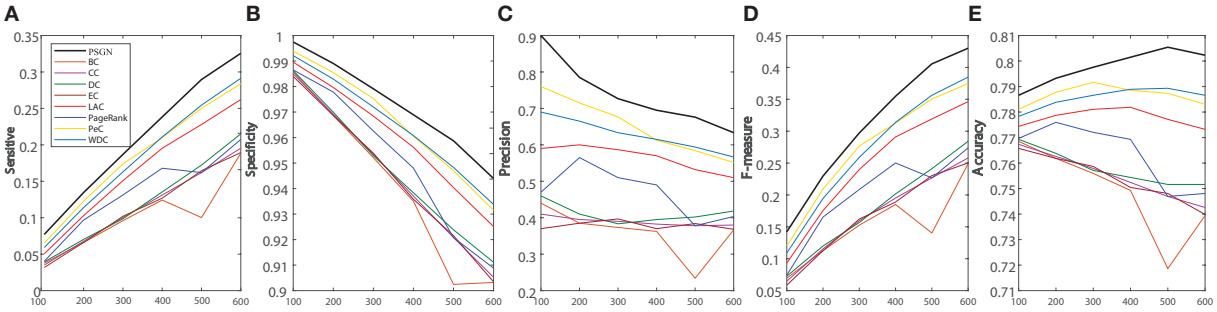

FIGURE 7
Comparative experiments on the Yeast PPI networks in terms of sensitive **(A)**, specificity **(B)**, precision **(C)**, F-measure **(D)**, and accuracy **(E)** obtained by PSGN, BC, CC, DC, EC, LAC, PageRank, PeC, and WDC.

3.1.2. Evaluate of the classified performance of BiLSTM

Machine learning algorithms like SVM, decision tree (DT), random forest (RF) and adaboost are widely used in the tasks of bioinformatics. For fair comparison with these machine learning methods, as the setting in the work of Zeng et al. (2019), we use the sequences composed of integrated biological features PC, PeC, NNSL, the topological feature NNC and the integrate feature PSGN as the input of these machine learning algorithms for training and testing. Besides, we also compared with the algorithm proposed by Zeng et al. (2019).

AI-BiLSTM proposed in this research achieved improved performance compared with other state-of-the-art algorithms with the highest value marked in bold in Table 1. Our model obtains recall, F-measure and AUC with values of 0.5674, 0.5134, and 0.6781, respectively, which are better than SVM, decision tree, random forest, Adaboost, and Zeng et al. (2019). Although

TABLE 1 Comparison of performance between our model and other machine learning algorithms.

Classifier	Accuracy	Precision	Recall	F-measure	AUC
SVM	**0.7654**	**0.4931**	0.3037	0.3759	0.6045
RF	0.7252	0.4295	0.5527	0.4833	0.6651
DT	0.7134	0.3809	0.3713	0.3760	0.5942
Adaboost	0.7409	0.4347	0.3797	0.4054	0.6150
Zeng et al. (2019)	0.7055	0.3802	0.4219	0.3999	0.6067
BiLSTM	0.7369	0.4803	**0.5742**	**0.5231**	**0.6829**

our model does not show the highest values in terms of accuracy and precision and the performance is slightly weaker than SVM in these two assessments, our model owns much better recall, F-measure and AUC. In general, BiLSTM is superior to all other methods.

Besides, for verifying the significance of each feature, we make an ablation test on the features including PC, PeC, NNC, and NNSL. In the ablation experiments, we remove a feature to observe its effect on the identification of essential proteins.

Table 2 shows that NNSL takes the most crucial role in prediction of essential proteins (lowest value marked in bold). The score of accuracy, F-measure, and AUC will drop dramatically without NNSL.

In this section, we compared our BiLSTM with the traditional methods like DC, CC, BC, EC, NC, LAC, PeC, WDC, and PSGN. For fair comparision, 20% of top ranked proteins scored by classical methods are treated as the essential proteins, the rest are regarded as non-essential proteins. Comparing with the list of essential proteins, we can calculate the scores of accuracy, precision, recall, F-measure and AUC of each method.

As the experimental results shown in Table 3, we can find that the scores of our BiLSTM in terms of precision, recall, F-measure, and AUC are significantly higher than the results of DC, BC, CC, EC, NC, LAC, PeC, WDC, and PSGN, which also illustrates the remarkable performance of our method for identifying essential proteins.

TABLE 2 Experimental results for ablation test.

Features	Accuracy	Precision	Recall	F-measure	AUC
Without PC	0.7409	0.4615	**0.4918**	0.476	0.6556
Without NNSL	**0.7222**	0.4111	0.5086	**0.4547**	**0.6469**
Without PeC	0.7242	**0.3833**	0.5769	0.4606	0.6694
Without NNC	0.7311	0.4491	0.5637	0.5000	0.6736
Without PSGN	0.7340	0.4688	0.5674	0.5134	0.6781
BiLSTM	0.7369	0.4803	0.5742	0.5231	0.6829

TABLE 3 Comparison of performance between our proposed non-local GNN and other classical methods.

Method	Accuracy	Precision	Recall	F-measure	AUC
DC	0.7335	0.4050	0.3470	0.3737	0.5977
CC	0.7150	0.3580	0.3067	0.3304	0.5716
BC	0.7139	0.3550	0.3041	0.3276	0.5699
EC	0.7194	0.3690	0.3161	0.3405	0.5777
LAC	0.7563	0.4630	0.3967	0.4273	0.6299
NC	0.7469	0.4390	0.3761	0.4051	0.6166
PeC	0.7555	0.4610	0.3950	0.4254	0.6288
WDC	**0.7630**	0.4800	0.4113	0.4430	0.6394
PSGN	0.7614	0.4771	0.4301	0.4524	0.6450
LSTM-AM	0.7340	0.4688	0.5674	0.5134	0.6781
BiLSTM	0.7369	**0.4803**	**0.5742**	**0.5231**	**0.6829**

Bold values mean best scores.

3.2. Analyze biology significance of the new proposed method

Human disease phenotypes share corresponding orthologs in Yeast gene sets. The BiLSTM model, which was firstly established based on Yeast gene sets, has been further validated in human disease gene sets. In order to reasonably extrapolating the proposed model in microbiota-diseases, genes known to be associated formed a seed set. For the test of human disease gene prediction, we collected sets of Yeast genes whose human orthologs were linked to the same OMIM disease. Human disease phenotypes from OMIM were collapsed into major categories.

3.2.1. Identification of pan-cancer related genes

In the experiments, we selected 10 kinds of cancers as the research objects, including esophageal carcinoma, pancreatic cancer, lung cancer (lung adenocarcinoma, lung squamous cell carcinoma), breast invasive carcinoma, colon adenocarcinoma, rectum adenocarcinoma, cholangiocarcinoma, gastric cancer and ovarian cancer, which can be obtained from the TCGA dataset. Due to the duplications of pathogenic genes between cancers, a total of 17,126 pathogenic genes were obtained after weight removal.

We believed that the common ancestor genes were similar in expression, so we did homology mapping on the background PPI network to find the homologous genes of human genes and Yeast genes. Then we take these genes as seed genes, a total of 1,166 homologous genes were found. Besides, we collected a total of 1,166 proteins expressed by seed genes and obtained the protein-protein interaction network using the STRING database. As it is shown in Figure 8A, it can be found that the corresponding Yeast proteins have a strong correlation with each other, which lays a foundation for our subsequent experiments. Through inputting the seed genes combined with the constructed PSGN features into the proposed BiLSTM algorithm, potential genes which are similar to seed genes will be predicted with corresponding scores. Predicted genes with score greater than 8 were screened out and regarded as candidate genes. By homologous mapping candidate genes, the homologous genes of these genes in human were found as the final predicted genes, and a total of 365 final predicted genes were obtained which is shown in Figure 8B. To further validate the biological significance of the predicted cancer related genes, we conducted a series of biological analysis like GO enrichment

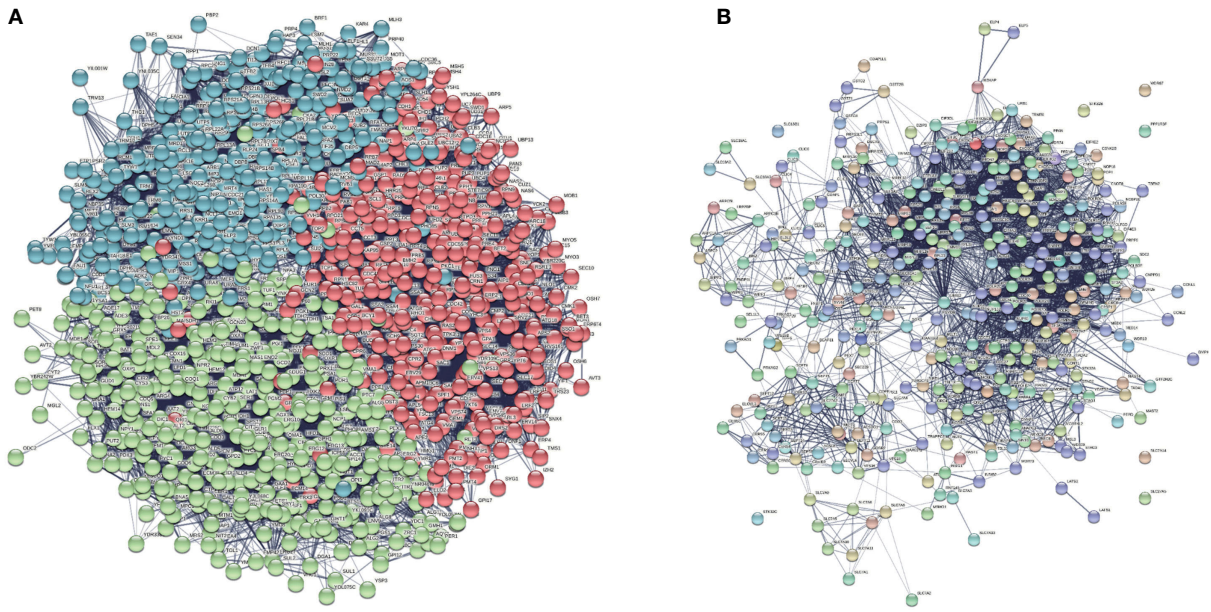

FIGURE 8
We found a total of 1,166 proteins expressed by seed genes and obtained the interaction network between these Yeast proteins screening in the background Yeast PPI network. **(A)** Generated Yeast protein interaction network. **(B)** Clustered protein interaction network with nine communities.

analysis, KEGG pathway analysis, clustering analysis in the following sections.

3.2.2. GO enrichment analysis

For the GO items, we analyzed the relationships of final predicted genes with pan-cancers. According to the ranking of the error rate (FDP), 10 functional annotations with the largest statistical significance were obtained from Biological Process (BP), Cellular Component (CC) and Molecular Function (MF) three branches of GO datasets. As is shown in Table 4, we can find that genes are highly correlated with several important biological processes such as transcription, mRNA splicing, rRNA binding and processing, and cytokinesis, which proved the inner correlations with these predicted genes. What's more, the occurrence sites also involve several cellular sites such as nucleoplasm, ribosome and cytosols, which indicates that these predicted genes are highly related to cell development and possibility with the growth of tumors.

Besides, during clustering analysis, eight modular subneworks M_0 to M_7 enriched in much more CGC genes with higher compactness structures are showed in Figure 9A. Specifically, we find that six of our predicted pan-cancer related genes are enriched in these modulars. Besides, for each module of gene lists, pathway and process enrichment analysis has been carried out with the ontology sources. The results are showed in Figure 9B.

3.2.3. KEGG pathway enrichment analysis

By KEGG pathway enrichment analysis of the predicted genes, we obtained five pathways with the highest correlation with these genes like Proteasome (map03050), Valine, leucine and isoleucine degradation (map00280), Terpenoid backbone biosynthesis (map00900), Mismatch repair (map03430), and Glutathione metabolism (map00480). Among these pathways, the proteasome pathway was the most enriched pathway, which are usually used as an inhibitor in the cancer therapy.

3.2.4. Survival analysis

To verify the biological significance of the experimental results, we conducted further survival analysis. As shown in Figure 10, EIF4A3, NHP2L1, and UBA52 are the three genes with the highest moderate prediction of human genes, which are closely related to RNA metabolic function. Here, we carried out a survival analysis of these three genes, respectively, and it can be seen from the results that all these three genes have a significant impact on the survival time of Bladder urothelial carcinoma (BLCA) patients, which verifies the performance of

TABLE 4 Ten functional annotations with the largest statistical significance for three branches in GO database.

Category	Term	Function	*p*-value	FDR
BP	GO:0006412	Translation	4.59573E-17	3.92797E-20
	GO:0000398	mRNA splicing *via* spliceosome	1.06421E-14	1.81916E-17
	GO:0061640	Cytoskeleton-dependent cytokinesis	2.2982E-11	5.89282E-14
	GO:0006749	Glutathione metabolic process	7.84992E-10	2.68373E-12
	GO:0006364	rRNA processing	1.48788E-09	6.35848E-12
	GO:0002181	Cytoplasmic translation	2.12271E-09	1.25302E-11
	GO:0034613	Cellular protein localization	2.12271E-09	1.27E-11
	GO:1903241	U2-type prespliceosome assembly	4.36481E-08	2.98449E-10
	GO:0006351	Transcription, DNA-templated	2.9991E-07	2.307E-09
	GO:0016575	Histone deacetylation	4.39654E-07	3.75773E-09
CC	GO:0005654	Nucleoplasm	1.91891E-33	5.1171E-36
	GO:0005829	Cytosol	1.29113E-15	8.91902E-18
	GO:0005940	Septin ring	1.29113E-15	1.72151E-17
	GO:0032153	Cell division site	1.29113E-15	1.72151E-17
	GO:0031105	Septin complex	1.29113E-15	1.72151E-17
	GO:0071005	U2-type precatalytic spliceosome	1.28732E-14	2.05972E-16
	GO:0005681	Spliceosomal complex	3.15875E-13	5.89633E-15
	GO:0005840	Ribosome	3.95972E-12	8.44741E-14
	GO:0046540	U4/U6 × U5 tri-snRNP complex	5.04467E-11	1.21072E-12
	GO:0005666	DNA-directed RNA polymerase III complex	4.88524E-10	1.30273E-11
MF	GO:0003735	Structural constituent of ribosome	5.78912E-16	3.91157E-18
	GO:0003899	DNA-directed 5'-3' RNA polymerase activity	7.15936E-13	7.2561E-15
	GO:0005515	Protein binding	1.02098E-11	1.3797E-13
	GO:0060090	Binding, bridging	5.14905E-09	8.69772E-11
	GO:0046933	Proton-transporting ATP synthase activity, Rotational mechanism	1.20677E-06	2.44616E-08
	GO:0003743	Translation initiation factor activity	1.45514E-06	3.44121E-08
	GO:0000340	RNA 7-methylguanosine cap binding	1.51118E-06	4.08426E-08
	GO:0050291	Sphingosine N-acyltransferase activity	1.71826E-06	5.22443E-08
	GO:0015179	L-amino acid transmembrane transporter activity	2.08402E-06	7.04059E-08
	GO:0019843	rRNA binding	2.57108E-05	9.55469E-07

the new proposed prediction method from the respective of homologous matching.

4. Conclusion

High-throughput techniques and machine learning approaches, combined with an increasing understanding of the microbiota and their collective genome from preclinical and large-scale clinical studies, offer exciting opportunities for modernizing microbe-based strategies from untargeted to precision microbiome-centered therapies. Essential proteins have drawn attention for their crucial roles in controlling signal transduction, individual variation in treatment response, and a wide range of other microbiome-related processes. The properties and purposes of biological data used to identify critical proteins are explored in this study. In light of the findings, we suggest a linear adaptive model PSGN, which may adaptively modify the weights for balancing each type of biological or topological property. We have demonstrated that the NNSL feature is significantly more important than other features through experimental validation. Moreover, the new algorithm PSGN improved the ability to represent features discriminatively. In the experiments, we first contrasted the PSGN with established methods including PageRank, DC, BC, CC, EC NC, LAC, PeC, and WDC. The results demonstrated that PSGN outperforms the other approaches in terms of overall performance. Furthermore, we evaluate our BiLSTM with machine learning methods and the most recent deep learning-based methods. The results of experiments

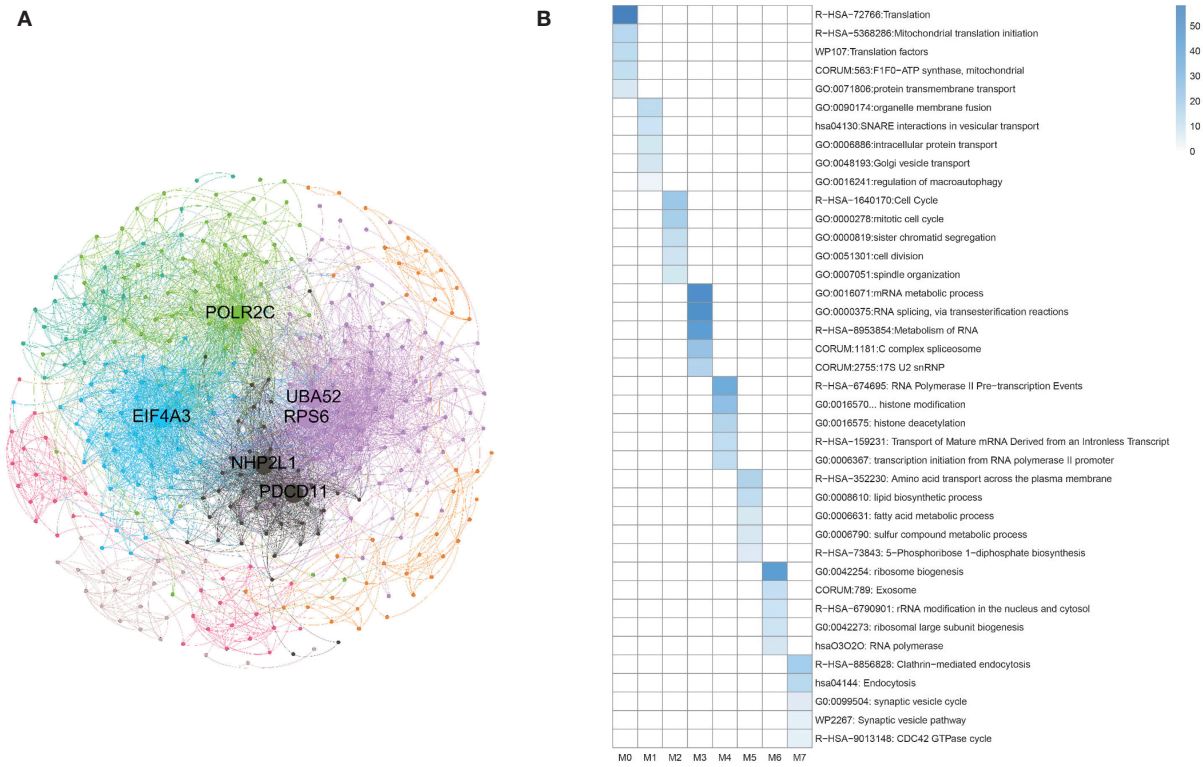

FIGURE 9
Clustering analysis for matched homologous genes in human. **(A)** Topological organization for eight modular sub-networks marked with different colors. **(B)** Enrichment biological functions of pan-cancer sub-networks. Each row represents a GO BP term and each column corresponds to a pan-cancer sub-network for each subnetwork.

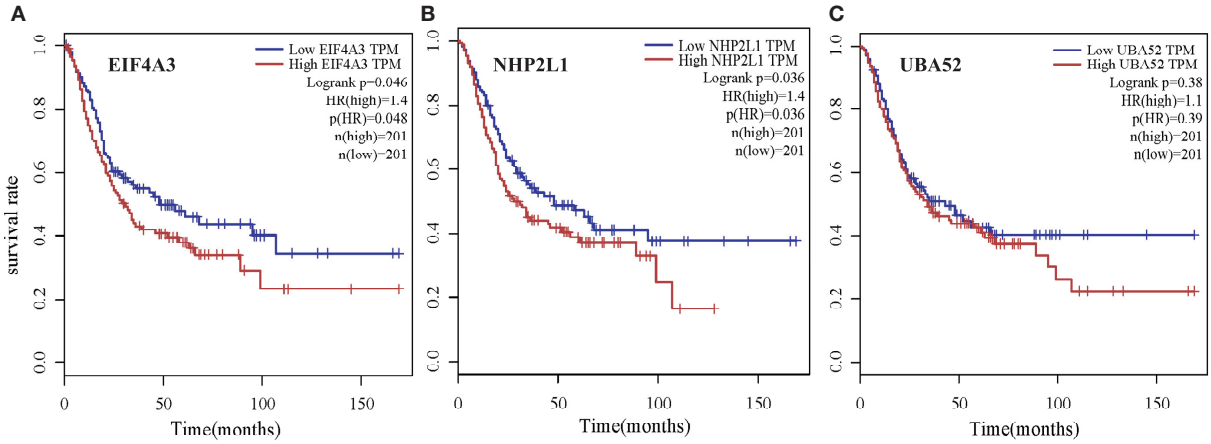

FIGURE 10
Survival analysis of three genes with the highest moderate prediction of human genes, which are closely related to RNA metabolic function for Bladder urothelial carcinoma (BLCA) patients. **(A)** EIF4A3, **(B)** NHP2L1, and **(C)** UBA52.

may potentially establish the capability of the new proposed BiLSTM. Our suggested models for biological information have considerable generality, making them suitable for integrating almost all biological features. In the future, we will continue to test and search for more suitable biological information for identifying essential proteins in more species.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

CW: conceptualization. HZ and HM: methodology. HZ and KC: software. CW, TG, YW, YY, and ZL: validation. HZ and YZ: writing—original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding

This work was partially supported by the National Natural Science Foundation of China (12126367 and 12126305), Chen Xiao-Ping Foundation for the Development of Science and Technology of Hubei Province (CXPJJH12000002-2020058), the Hubei Provincial Natural Science Foundation of China (2015CFA010), Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (CUGGC02), and Shanghai Municipal Science and Technology Major Project (2018SHZDZX01), Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (LCNBI), and ZJLab.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alsegiani, A. S., and Shah, Z. A. (2022). The influence of gut microbiota alteration on age-related neuroinflammation and cognitive decline. *Neural Regenerat. Res.* 17, 2407. doi: 10.4103/1673-5374.335837

Anto, L., and Blesso, C. N. (2022). Interplay between diet, the gut microbiome, and atherosclerosis: role of dysbiosis and microbial metabolites on inflammation and disordered lipid metabolism. *J. Nutr. Biochem.* 105, 108991. doi: 10.1016/j.jnutbio.2022.108991

Aromolaran, O., Aromolaran, D., Isewon, I., and Oyelade, J. (2021). Machine learning approach to gene essentiality prediction: a review. *Brief. Bioinform.* 22, bbab128. doi: 10.1093/bib/bbab128

Bajaj, J. S., Ng, S. C., and Schnabl, B. (2022). Promises of microbiome-based therapies. *J. Hepatol.* 76, 1379–1391. doi: 10.1016/j.jhep.2021.12.003

Beg, R., Gonzalez, K., and Martinez-Guryn, K. (2022). Implications of microbe-mediated crosstalk in the gut: impact on metabolic diseases. *Bioch. Biophys. Acta* 1867, 159180. doi: 10.1016/j.bbalip.2022.159180

Belkaid, Y., and Hand, T. W. (2014). Role of the microbiota in immunity and inflammation. *Cell* 157, 121–141. doi: 10.1016/j.cell.2014.03.011

Binder, J. X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S. I., et al. (2014). Compartments: unification and visualization of protein subcellular localization evidence. *Database* 2014, bau012. doi: 10.1093/database/bau012

Bleackley, M. R., and MacGillivray, R. T. (2011). Transition metal homeostasis: from yeast to human disease. *Biometals* 24, 785–809. doi: 10.1007/s10534-011-9451-4

Bonacich, P. (1987). Power and centrality: a family of measures. *Am. J. Sociol.* 92, 1170–1182. doi: 10.1086/228631

Cao, Z., and Zhang, S. (2016). An integrative and comparative study of pan-cancer transcriptomes reveals distinct cancer common and specific signatures. *Sci. Rep.* 6, 1–13. doi: 10.1038/srep33398

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., et al. (2012). Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40, 700–705. doi: 10.1093/nar/gkr1029

Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., and Richardson, J. E. (2012). The mouse genome database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.* 40, 881–886. doi: 10.1093/nar/gkr974

Estrada, E., and Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E* 71, 056103. doi: 10.1103/PhysRevE.71.056103

Harris, T. W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., et al. (2010). WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.* 38, 463–467. doi: 10.1093/nar/gkp952

Hersi, F., Elgendy, S. M., Al Shamma, S. A., Atell, R. T., Sadiek, O., and Omar, H. A. (2022). Cancer immunotherapy resistance: the impact of microbiome-derived short-chain fatty acids and other emerging metabolites. *Life Sci.* 300, 120573. doi: 10.1016/j.lfs.2022.120573

Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41–42. doi: 10.1038/35075138

Joy, M. P., Brock, A., Ingber, D. E., and Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.* 2005, 96. doi: 10.1155/JBB.2005.96

Lei, J., Xie, Y., Sheng, J., and Song, J. (2022). Intestinal microbiota dysbiosis in acute kidney injury: novel insights into mechanisms and promising therapeutic strategies. *Ren. Fail.* 44, 571–580. doi: 10.1080/0886022X.2022.2056054

Lei, X., Fang, M., Wu, F., and Chen, L. (2018). Improved flower pollination algorithm for identifying essential proteins. *BMC Syst. Biol.* 12, 129–140. doi: 10.1186/s12918-018-0573-y

Li, G., Li, M., Wang, J., Wu, J., Wu, F., and Pan, Y. (2016). Predicting essential proteins based on subcellular localization, orthology and PPI networks. *BMC Bioinform.* 17, 571–581. doi: 10.1186/s12859-016-1115-5

Li, M., Ni, P., Chen, X., Wang, J., Wu, F., et al. (2017). Construction of refined protein interaction network for predicting essential proteins. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 1386–1397. doi: 10.1109/TCBB.2017.2665482

Li, M., Zhang, H., Wang, J., and Pan, Y. (2012). A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Syst. Biol.* 6, 1–9. doi: 10.1186/1752-0509-6-15

Li, Y., Jiang, T., Zhou, W., Li, J., Li, X., et al. (2020). Pan-cancer characterization of immune-related lncRNAs identifies potential

oncogenic biomarkers. *Nat. Commun*. 11, 1–13. doi: 10.1038/s41467-020-14802-2

Liu, W., Jiang, Y., Li, P., Sun, X., Gan, W., et al. (2022a). Inferring gene regulatory networks using the improved Markov blanket discovery algorithm. *Interdisc. Sci. Comput. Life Sci*. 14, 168–181. doi: 10.1007/s12539-021-00478-9

Liu, W., Lin, H., Huang, L., Peng, L., Tang, T., Zhao, Q., et al. (2022b). Identification of miRNA-disease associations *via* deep forest ensemble learning based on autoencoder. *Brief. in Bioinform*. 23, bbac104. doi: 10.1093/bib/bbac104

Luo, J., and Qi, Y. (2015). Identification of essential proteins based on a new combination of local interaction density and protein complexes. *PLoS ONE* 10, e0131418. doi: 10.1145/2818302

Ma, C., Wang, Z., Xu, T., He, Z., and Wei, Y. (2020). The approved gene therapy drugs worldwide: from 1998 to 2019. *Biotechnol. Adv*. 40, 107502. doi: 10.1016/j.biotechadv.2019.107502

Magrane, M. (2011). Uniprot knowledgebase: a hub of integrated protein data. *Database* 2011, bar009. doi: 10.1093/database/bar009

Mcquilton, P., Pierre, S. E. S., and Thurmond, J. (2012). Flybase 101-the basics of navigating flybase. *Nucleic Acids Res*. 40, D706-D714. doi: 10.1093/nar/gkr1030

Mewes, H., Frishman, D., Mayer, K. F. X., Munsterkotter, M., Noubibou, O., Pagel, P., et al. (2006). MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res*. 34, 169–172. doi: 10.1093/nar/gkj148

Müller, G. B. (2003). Homology: the evolution of morphological organization. *Origin. Organ. Beyond Gene Dev. Evolut. Biol*. 2, 51. doi: 10.7551/mitpress/5182.001.0001

Park, S., Kim, S.-J., Yu, D., Pena-Llopis, S., Gao, J., Park, J. S., et al. (2016). An integrative somatic mutation analysis to identify pathways linked with survival outcomes across 19 cancer types. *Bioinformatics* 32, 1643–1651. doi: 10.1093/bioinformatics/btv692

Peng, X., Wang, J., Wang, J., Wu, F., and Pan, Y. (2015). Rechecking the centrality-lethality rule in the scope of protein subcellular localization interaction networks. *PLoS ONE* 10, e0130743. doi: 10.1371/journal.pone.0130743

Sommer, R. J. (2008). Homology and the hierarchy of biological systems. *Bioessays* 30, 653–658. doi: 10.1002/bies.20776

Sorbara, M. T., and Pamer, E. G. (2022). Microbiome-based therapeutics. *Nat. Rev. Microbiol*. 20, 365–380. doi: 10.1038/s41579-021-00667-9

Stephenson, K., and Zelen, M. (1989). Rethinking centrality: methods and examples. *Soc. Networks* 11, 1–37. doi: 10.1016/0378-8733(89)90016-6

Sun, F., Sun, J., and Zhao, Q. (2022). A deep learning method for predicting metabolite-disease associations *via* graph neural network. *Brief. Bioinform*. 23, bbac266. doi: 10.1093/bib/bbac266

Wan, X., Eguchi, A., Fujita, Y., Ma, L., Wang, X., Yang, Y., et al. (2022). Effects of (R)-ketamine on reduced bone mineral density in ovariectomized mice: a role of gut microbiota. *Neuropharmacology* 213, 109139. doi: 10.1016/j.neuropharm.2022.109139

Wang, C., Han, C., Zhao, Q., and Chen, X. (2021). Circular RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform*. 22, bbab286. doi: 10.1093/bib/bbab286

Wang, J., Li, M., Wang, H., and Pan, Y. (2012). Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 9, 1070–1080. doi: 10.1109/TCBB.2011.147

Wang, N., Zeng, M., Zhang, J., Li, Y., and Li, M. (2020). Ess-"NEXG: predict essential proteins by constructing a weighted protein interaction network based on node embedding and xgboost," in *International Symposium on Bioinformatics Research and Applications* (Moscow: Springer), 95–104.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet*. 45, 1113–1120. doi: 10.1038/ng.2764

Wuchty, S., and Stadler, P. F. (2003). Centers of complex networks. *J. Theor. Biol*. 223, 45–53. doi: 10.1016/S0022-5193(03)00071-7

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S., and Eisenberg, D. (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*. 30, 303–305. doi: 10.1093/nar/30.1.303

Xia, C., Dong, X., Li, H., Cao, M., Sun, D., et al. (2022). Cancer statistics in china and united states, 2022: profiles, trends, and determinants. *Chin. Med. J*. 135, 584–590. doi: 10.1097/CM9.0000000000002108

Yang, X., Gao, L., and Zhang, S. (2017). Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns. *Brief. Bioinform*. 18, 761–773. doi: 10.1093/bib/bbw063

Zeng, M., Li, M., Wu, F.-X., Li, Y., and Pan, Y. (2019). DeepEP: a deep learning framework for identifying essential proteins. *BMC Bioinform*. 20, 506–510. doi: 10.1186/s12859-019-3076-y

Zhang, J., and Zhang, S. (2016). The discovery of mutated driver pathways in cancer: models and algorithms. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 15, 988–998. doi: 10.1109/TCBB.2016.2640963

Zhang, J., and Zhang, S. (2017). Discovery of cancer common and specific driver gene sets. *Nucleic Acids Res*. 45, e86-e86. doi: 10.1093/nar/gkx089

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021). Using network distance analysis to predict lncRNA-miRNA lnteractions. *Interdisc. Sci. Comput. Life Sci*. 13, 535–545. doi: 10.1007/s12539-021-00458-z

Zhang, R., and Lin, Y. (2009). Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res*. 37, 455–458. doi: 10.1093/nar/gkn858

Zhang, W., Xu, J., and Zou, X. (2019). Predicting essential proteins by integrating network topology, subcellular localization information, gene expression profile and go annotation data. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 17, 2053–2061. doi: 10.1109/TCBB.2019.2916038

Zhong, J., Tang, C., Peng, W., Xie, M., Sun, Y., Tang, Q., et al. (2021). A novel essential protein identification method based on ppi networks and gene expression data. *BMC Bioinform*. 22, 1–21. doi: 10.1186/s12859-021-04175-8

Zhu, Y., and Wu, C. (2018). "Identification of essential proteins using improved node and edge clustering coefficient," in *The 37th Chinese Control Conference (CCC)* (Wuhan), 1543–1547.

Zhu, Y., Zhang, H., Yang, Y., Zhang, C., Ou-Yang, L., et al. (2022). Discovery of pan-cancer related genes *via* integrative network analysis. *Brief. Funct. Genomics* 21, 325–338. doi: 10.1093/bfgp/elac012

Zotenko, E., Mestre, J., O'Leary, D. P., and Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol*. 4, e1000140. doi: 10.1371/journal.pcbi.1000140