BMC Bioinformatics

RESEARCH Open Access

Constructing multilayer PPI networks based on homologous proteins and integrating multiple PageRank to identify essential proteins

He Zhao[1,2], Huan Xu[1,2], Tao Wang[1,2] and Guixia Liu[1,2*]

*Correspondence:
liugx@jlu.edu.cn

[1] College of Computer Science and Technology, Jilin University, Changchun, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China

Abstract

Background: Predicting and studying essential proteins not only helps to understand the fundamental requirements for cell survival and growth regulation mechanisms but also deepens our understanding of disease mechanisms and drives drug development. Existing methods for identifying essential proteins primarily focus on PPI networks within a single species, without fully exploiting interspecies homologous relationships. These homologous relationships connect proteins from different species, forming multilayer PPI networks. Some methods only construct interlayer edges based on homologous relationships between two species, without incorporating appropriate biological attributes to assess the biological significance of these edges. Furthermore, homologous proteins are often highly conserved across multiple species, and expanding homologous relationships to more species allows for a more accurate assessment of interlayer edge importance.

Results: To address these issues, we propose a novel model, MLPR, which constructs a multilayer PPI network based on homologous proteins and integrates multiple PageRank algorithms to identify essential proteins. This study combines homologous protein data from three species to construct interlayer transition matrices and assigns weights to interlayer edges by integrating the biological attributes of homologous proteins and cross-species GO annotations. The MLPR model uses multiple PageRank methods to comprehensively consider homologous relationships across species and designs three key parameters to find the optimal combination that balances random walks within layers, global jumps, interlayer biases, and interspecies homologous relationships.

Conclusions: Experimental results show that MLPR outperforms other state-of-the-art methods in terms of performance. Ablation experiments further validate that integrating homologous relationships across three species effectively enhances the overall performance of MLPR and demonstrates the advantages of the multiple PageRank model in identifying essential proteins.

Keywords: Essential proteins, Homologous proteins, Multilayer PPI networks, Multiple PageRank

Background

Essential proteins are crucial for cellular viability, their absence can lead to impaired or lost cellular function [1]. The prediction and study of essential proteins not only help uncover the fundamental requirements for cell survival and growth regulation but also play a significant role in understanding disease mechanisms and advancing drug development [2]. Although biological experiments offer high accuracy in identifying essential proteins, they are often expensive, time-consuming, and inefficient. Moreover, such methods are typically constrained by species-specific factors. With the rapid advancement of high-throughput technologies, large-scale protein-protein interaction (PPI) data can now be obtained more efficiently [3], enabling researchers to utilize computational approaches for identifying essential proteins, thus providing a more scalable and efficient solution [4].

Existing computational methods for essential protein identification primarily focus on the PPI networks of single species, utilizing network topological properties or biological attributes. Common topology-based methods include Local Average Connectivity (LAC) [5], Neighborhood Centrality (NC) [6], and Subgraph Centrality (SC) [7], all of which have been proven effective in predicting essential proteins. Tools such as CytoNCA [8] integrate these topological features, while SIGEP [9] calculates p-values based on multiple network topological features (e.g., degree and local clustering coefficient), outperforming methods that rely solely on single topological properties.

Recent studies reveal that certain biological attributes are closely associated with protein essentiality. Some methods combine biological features with network topology for identifying essential proteins. For instance, TS-PIN [10] uses gene expression data and subcellular localization to construct networks for essential protein identification. RWEP [11] applies a random walk algorithm to balance topological and biological features for prediction. During the random walk, walkers move among neighboring nodes with a probability of λ, reflecting the influence of local network topology, and jump to any node in the network with a probability of $1 - \lambda$, capturing global network features. Similarly, SESN [12] employs a seed expansion approach that integrates PPI subnetworks and various biological data for predictions. Additionally, researchers have extracted biological and network topological features relevant to essential proteins, using them as inputs for machine learning or deep learning models. For example, EPNBC [13] combines biological information, a naive Bayes classifier, and the PageRank algorithm for essential protein prediction. DeepEP [14] extracts biological features from gene expression data and captures PPI network topology using node2vec [15], leveraging both feature sets for prediction. ACDMBI [16] extracts features from PPI networks, gene expression data, and subcellular localization data, integrating these into a deep neural network for prediction. MBIEP [17] is a deep learning-based model for predicting essential proteins, by integrating multi-dimensional features from the topological structure of PPI networks, subcellular localization information, and gene expression data, the model significantly enhances prediction performance. However, since PPI datasets are typically imbalanced, these machine learning or deep learning methods tend to suffer from bias when dealing with imbalanced data, which adversely affects prediction accuracy.

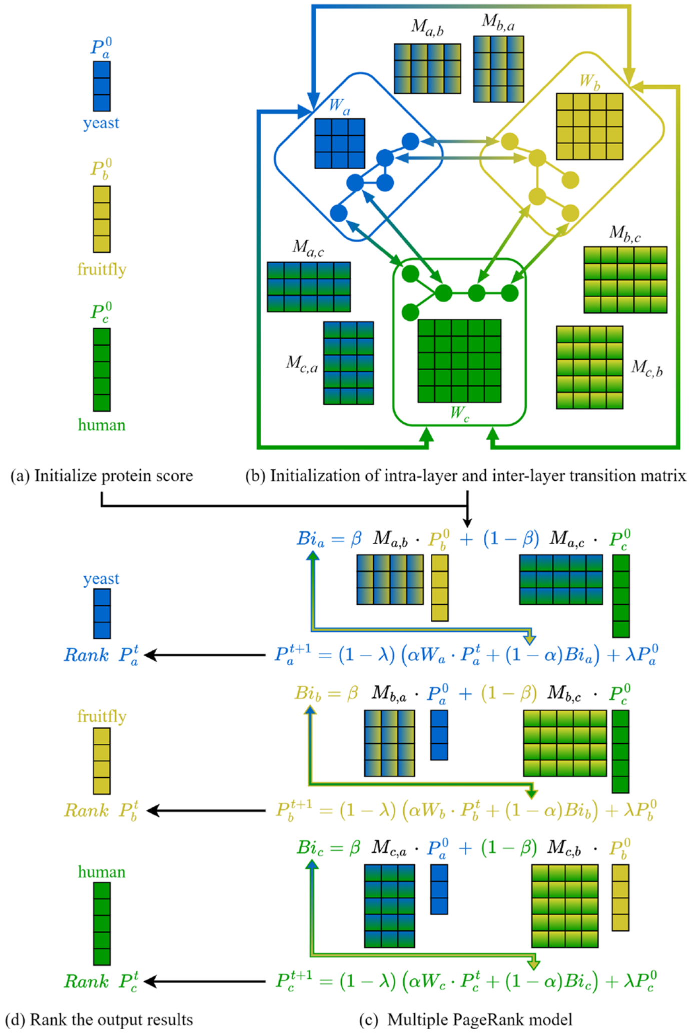
The above methods mainly focus on the PPI network of a single species and fail to fully exploit interspecies homology relationships. Homologous proteins are often highly conserved across species, and predicting essential proteins in one species can help identify homologous proteins with similar critical functions in other species. These homologous relationships connect proteins from different species, forming multilayer PPI networks. In complex network research, significant advances have been made in multilayer network analysis [18–20]. In the field of essential protein prediction, the RWO method [21] constructs multilayer PPI networks using orthologous relationships between yeast and human PPI networks. It employs a random walk algorithm to iteratively update protein scores, controlling interlayer transition probabilities and the probability of intralayer random walks that lead to proteins with interlayer connections. However, the RWO method relies solely on pairwise homology relationships to construct interlayer edges, without integrating biological attributes (e.g., GO annotations) to evaluate their importance. Moreover, as homologous proteins are often highly conserved across multiple species, extending homology relationships to more species can facilitate a more comprehensive evaluation of interlayer edges and their importance.

Despite significant progress in the field of essential protein identification, several limitations still exist in current methods. First, most approaches are limited to the analysis of PPI networks within a single species and fail to fully exploit homologous relationships across species, resulting in incomplete assessments of protein importance. Second, certain existing methods have shortcomings in constructing and evaluating the weights of inter-layer edges. These methods generally connect proteins from different species based on homologous relationships between two species, but they lack proper strategies for assigning weights to inter-layer edges. Lastly, existing methods for parameter tuning in PageRank models are limited, as they primarily focus on balancing the probabilities of random walks and global jumps. In models involving only two species, tuning strategies mainly adjust inter-layer transition probabilities and the probabilities of intra-layer jumps to inter-layer nodes. However, these approaches struggle to handle multiple species effectively. They fail to optimize parameter balance across complex multi-species PPI networks and necessitate a fundamental redesign of the PageRank model to better integrate homology relationships across multiple species.

To overcome the limitations of existing methods, this study proposes a novel approach for cross-species essential protein identification by integrating homologous relationships across multiple species, refining the evaluation of inter-layer edge weights, designing a multiple PageRank model, and introducing an efficient parameter-tuning strategy. (1) This study incorporates homologous protein data from three species (yeast and fruit fly, yeast and human, fruit fly and human) to construct inter-layer transition matrices. Since homologous proteins are typically highly conserved across species, extending homologous relationships to additional species provides a more accurate representation of inter-layer connectivity and protein importance. (2) By leveraging the biological attributes of homologous proteins and cross-species Gene Ontology (GO) annotation data, biological weights are assigned to inter-layer edges, allowing for a more precise evaluation of inter-species protein interactions. This improves the reliability of essentiality assessments and enhances the biological relevance of inter-layer connections. (3) A novel multiple PageRank model is developed based on multi-layer PPI networks, where essentiality scores are iteratively updated by comprehensively considering intra-layer interactions and inter-layer transitions. Three key parameters are introduced to regulate intra-layer random walks and global jumps, inter-layer transition biases, and the influence of homologous relationships. To mitigate the high computational cost of traditional parameter tuning, a two-step optimization strategy is proposed, significantly reducing complexity and time while ensuring model performance.

Methods

The overall process of MLPR is illustrated in Fig. 1. The MLPR algorithm involves three species, represented in blue, yellow, and green, denoted as a, b, and c, respectively. The algorithm comprises four main parts. First, as shown in Fig. 1a, various biological data, including homologous proteins, gene expression, subcellular localization, and protein complexes, are integrated to initialize the initial scores of proteins. Second, as depicted in Fig. 1b, the intra-layer transition matrices and inter-layer transition matrices for the multilayer PPI network are constructed. The intra-layer transition matrices for species a, b, and c are denoted as W_a, W_b, and W_c, respectively, while the inter-layer transition matrices between species include $M_{a,b}$, $M_{a,c}$, $M_{b,a}$, $M_{b,c}$, $M_{c,a}$, and $M_{c,b}$. Next, as


P_a^0
yeast
P_b^0
fruitfly
P_c^0
human
$M_{a,b}$
$M_{b,a}$
W_a
W_b
$M_{a,c}$
$M_{b,c}$
$M_{c,a}$
$M_{c,b}$
W_c
$Bi_a = \beta \ M_{a,b} \cdot P_b^0 + (1-\beta) \ M_{a,c} \cdot P_c^0$
yeast
$Rank \ P_a^t$
$P_a^{t+1} = (1-\lambda)\left(\alpha W_a \cdot P_a^t + (1-\alpha) Bi_a\right) + \lambda P_a^0$
$Bi_b = \beta \ M_{b,a} \cdot P_a^0 + (1-\beta) \ M_{b,c} \cdot P_c^0$
fruitfly
$Rank \ P_b^t$
$P_b^{t+1} = (1-\lambda)\left(\alpha W_b \cdot P_b^t + (1-\alpha) Bi_b\right) + \lambda P_b^0$
$Bi_c = \beta \ M_{c,a} \cdot P_a^0 + (1-\beta) \ M_{c,b} \cdot P_b^0$
human
$Rank \ P_c^t$
$P_c^{t+1} = (1-\lambda)\left(\alpha W_c \cdot P_c^t + (1-\alpha) Bi_c\right) + \lambda P_c^0$

(a) Initialize protein score (b) Initialization of intra-layer and inter-layer transition matrix

(d) Rank the output results (c) Multiple PageRank model

Fig. 1 The overall workflow of the MLPR

shown in Fig. 1c, the multiple PageRank model is implemented. During each iteration, the importance scores of proteins are updated by combining intra-layer interactions and inter-layer biases, thereby incorporating homologous relationships across different species to provide a more comprehensive evaluation of protein importance. Finally, as illustrated in Fig. 1d, once the model meets the convergence criteria, the protein scores are ranked in descending order, and the top-ranked proteins are identified as the predicted essential proteins.

Experimental datasets

The experiments are based on three species: yeast, fruitfly, and human. The biological data involved include PPI datasets, essential proteins, protein complexes, Gene Ontology (GO) and subcellular localization data, homologous protein data between species, and gene expression data. To standardize protein IDs across different datasets, we use the UniProt platform (https://www.uniprot.org/). The sources and processing methods for each dataset are detailed below:

PPI datasets: The yeast PPI dataset is obtained from DIP [22], while the PPI datasets for fruitfly and human are sourced from BioGRID [23]. After acquiring the data, basic cleaning steps are performed, including removing duplicate entries and self-loop interactions. The basic information for the processed PPI data is shown in Table 1.

Essential proteins: The benchmark essential protein datasets are collected from multiple databases: yeast data is derived from MIPS [24], SGD [25], DEG [26], and OGEE [27]; fruitfly data is obtained from DEG and OGEE; human data is sourced from DEG. After standardizing the IDs with the PPI datasets and removing duplicates, the number of essential proteins for each species is listed in Table 1.

Protein complexes: The yeast protein complex data is collected from MIPS, SGD, ALOY [28], and CYC2008 [29, 30]. Only complexes containing two or more proteins are retained, resulting in 745 protein complexes. Fruitfly protein complex data is sourced from AP-MS [31], and after standardizing IDs with the PPI dataset, 1637 protein complexes are obtained. Human protein complex data is collected from CORUM [32], and after ID unification, 2351 protein complexes are retained.

GO and subcellular localization: GO annotation data for yeast is obtained from the SGD database (https://downloads.yeastgenome.org/curation/literature/go_slim_mapping.tab). GO annotation data for fruitfly and human is sourced from the COMPARTMENTS database [33]. Subcellular localization data is also extracted from the COMPARTMENTS database using knowledge-based channels for each species.

Homologous proteins: Homologous protein data is sourced from the InParanoid database [34]. Version 7.0 is used to obtain homologous protein information between

Table 1 Information of PPI datasets

Species	Number of proteins	Number of interations	Number of essential proteins
Yeast	5093	24,743	1167
Fruitfly	7783	35,015	493
Human	26,410	822,859	8707

yeast and fruitfly, yeast and human, as well as fruitfly and human. Homologous protein pairs with a confidence score of 100% are retained. After unifying IDs with the PPI datasets, 1868, 1868, and 4268 homologous protein pairs are obtained for the respective species pairs.

Gene expression data: Gene expression data is downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/browse/) with dataset IDs GSE3431, GSE7763, and GSE45878 for yeast, fruitfly, and human, respectively. To map the data with the PPI datasets, SOFT format family files are downloaded from GEO. If multiple probe data corresponds to the same ID in the PPI datasets, the average value of the probes is taken. After preprocessing, 4981, 7378, and 15,413 gene expression data entries are obtained for yeast, fruitfly, and human, respectively.

Initializing protein score vectors based on multi-biological data

The MLPR algorithm integrates the following biological data: homologous proteins, gene expression data, subcellular localization, and protein complexes to generate the initial score vector for proteins, accurately reflecting their essentiality.

Homologous Proteins: Homologous proteins originate from a common ancestral gene and typically maintain high structural and functional similarity. Although genes may undergo variations during evolution, the fundamental structure and functions of homologous proteins are preserved across species. Previous studies have shown that proteins with homologous relationships exhibit significant conservation in terms of function and structure [35, 36]. For instance, certain homologous proteins between yeast and humans are not only highly consistent in function but also retain interaction patterns and many ancestral subcellular localization features [37]. Due to the high conservation of homologous proteins across species, they often play indispensable roles in vital processes. Therefore, homology is an important factor in the study of essential proteins. Research suggests that if a protein exhibits high homology across multiple species and is functionally important, it is likely critical for the survival of these species [21]. By studying homologous proteins, it is possible to identify potential essential proteins in one species and infer their critical functions in other species. Based on this premise, this study utilizes known homologous protein data in yeast, fruit fly, and human to identify essential proteins with similar functions across these species. Since the high conservation of homologous proteins is often closely associated with their essentiality, we analyze and evaluate the homology relationships among the three species (yeast and fruit fly, yeast and human, fruit fly and human).

Specifically, let yeast, fruit fly, and human be represented by species a, b, and c, respectively. For a protein v in species a, its homology-based essentiality score is defined as Eq. 1:

$$OR_v^a = \frac{\left|orth_v^b \cup orth_v^c\right|}{OR_{max}^a} \tag{1}$$

where $orth_v^b$ and $orth_v^c$ denote the sets of homologous proteins of v in species b and species c, respectively, and $OR_{max}^a = \max\left(OR_v^a\right), (v \in V_a)$. The value of OR_v^a ranges from [0, 1]. This scoring method reflects the potential of a protein to be essential across multiple species by measuring the extent of its homologous relationships.

Gene expression data: Gene expression data is presented in the form of an expression matrix, where each row represents the expression levels of a protein across different sample points, and each column corresponds to the expression value of a sample point. Due to differences in the number of sample points among species, the data processing methods vary. Specifically, yeast contains 12 sample points, fruit fly has 34 sample points, and human includes 837 sample points. To ensure the representativeness of sample point data, the average value of the gene expression at each sample point is used in this study. $expr_i(g)$ represents the gene expression value from the expression matrix, and i denotes the sample point number.

For yeast, the gene expression value at each sample point is obtained by averaging the expression values of 12 interval points, and the calculation formula is as follows:

$$Ge_i(g) = \frac{expr_i(g) + expr_{i+12}(g) + expr_{i+24}(g)}{3}, \quad i \in [0, 11] \tag{2}$$

For fruit fly, the gene expression value at each sample point is obtained by averaging the expression values of 4 consecutive points, and the formula is as follows:

$$Ge_i(g) = \frac{expr_{4\times i}(g) + expr_{4\times i+1}(g) + expr_{4\times i+2}(g) + expr_{4\times i+3}(g)}{4}, \quad i \in [0, 33] \tag{3}$$

For human, the gene expression value at each sample point directly uses the corresponding expression value:

$$Ge_i(g) = expr_i(g), \quad i \in [0, 836] \tag{4}$$

The interaction strength between proteins is measured by the co-expression relationship of their gene expression, and the Pearson correlation coefficient (PCC) is used to calculate the co-expression strength between two proteins [38, 39]. The formula for PCC is as follows:

$$PCC(X, Y) = \frac{\sum_{k=1}^{n}(X_k - \bar{X})(Y_k - \bar{Y})}{\sqrt{\sum_{k=1}^{n}(X_k - \bar{X})^2} \cdot \sqrt{\sum_{k=1}^{n}(Y_k - \bar{Y})^2}} \tag{5}$$

where X and Y represent the gene expression data of protein v and protein u at different sample points, defined as: $X = \{X_1, X_2, \ldots, X_k, \ldots, X_n\}, Y = \{Y_1, Y_2, \ldots, Y_k, \ldots, Y_n\}$, where n denotes the number of sample points, which depends on the number of sample points for each species, as defined in Eqs. 2, 3, and 4. To further normalize the co-expression strength between proteins, PCC is standardized into a co-expression weight GW_{vu}, and the formula is as follows:

$$GW_{vu} = \frac{PCC(X, Y) + 1}{2} \tag{6}$$

After normalization, the value of GW_{vu} ranges from [0, 1], which facilitates subsequent scoring.

The gene expression score GE_v of protein v is the sum of co-expression weights with its neighboring proteins, normalized as follows:

$$GE_v = \frac{\sum_{u \in N_v} GW_{vu}}{GE_{max}} \tag{7}$$

where N_v represents the set of neighbors connected to protein v, and GE_{max} denotes the maximum gene expression score among all proteins, which is used to normalize the scores into the range [0, 1]. Through this process, the gene expression score of each protein effectively reflects its co-expression importance in the network, thereby providing a more precise basis for initializing the essentiality score of proteins.

Subcellular Localization: In the field of essential protein identification, research commonly focuses on 11 subcellular localizations associated with protein essentiality [33]. To identify key subcellular localizations highly associated with essential proteins from these 11 subcellular compartments, the proportion of essential proteins in each subcellular localization *subi* is calculated and defined as $EPI_{subi} = EP_{subi}/P_{subi}$, where EP_{subi} represents the number of essential proteins in *subi*, and P_{subi} represents the total number of proteins in *subi*. Then, a threshold $EPthre = ep/p$ is set to filter out important subcellular localizations, where *ep* is the number of essential proteins in the PPI dataset, and *p* is the total number of proteins in the PPI dataset. If the *EPI* value of a subcellular localization exceeds this threshold, it is selected into the set *SC*. For example, in the yeast PPI network, the selected set of subcellular localizations is $SC = \{Nucleus, Cytosol, Cytoskeleton, Endoplasmic\ reticulum, Golgi\ apparatus\}$.

After identifying the important subcellular localizations, each subcellular localization is scored based on the number of proteins it contains, to evaluate its role in essential protein identification. Specifically, for each selected subcellular localization SC_i, its score is calculated as $SCS_i = NSC_i/NSC_{max}$, where NSC_i is the number of proteins in SC_i, and NSC_{max} is the maximum number of proteins among all selected subcellular localizations. The resulting score SCS_i ranges from [0, 1] and quantifies the importance of different subcellular localizations. Finally, the cumulative score SSC_v of a protein v is obtained by performing a weighted sum of the scores of all subcellular localizations to which the protein belongs, and its formula is $SSC_v = \sum_{v \in SC_i} SCS_i$. To ensure the comparability of cumulative scores across different proteins, normalization is required. The normalized subcellular localization weighted score SW_v is calculated as:

$$SW_v = \frac{SSC_v}{SSC_{max}} \tag{8}$$

where SSC_{max} is the maximum cumulative score among all proteins, ensuring that the range of SW_v is [0, 1]. This method enhances the accuracy of initial essential protein scores by incorporating subcellular localization information for weighted scoring.

Protein Complexes: The essentiality of a protein is often positively correlated with the number of protein complexes it participates in [40]. Based on this, the protein complex information is utilized to score a protein v, and the scoring formula is:

$$PC_v = \frac{|SPC_v|}{PC_{max}} \tag{9}$$

where SPC_v represents the set of protein complexes to which protein v belongs, and $PC_{max} = \max(|SPC_v|), (v \in V)$ is the maximum value of $|SPC_v|$ across all proteins.

The PC_v value obtained from the above formula ranges from $[0,1]$, and it measures the importance of protein v based on protein complexes.

To integrate various biological information of proteins to accurately reflect their essentiality, the initial score vector P_a^0 combines information from homologous proteins, gene expression data, subcellular localization, and protein complexes. Taking species a as an example, the initial score vector is defined as follows:

$$P_a^0 = OR_v^a \cdot SW_v \cdot (PC_v + GE_v) \tag{10}$$

where OR_v^a(defined in Eq. 1) represents the importance score of protein v in species a based on homologous proteins, GE_v(defined in Eq. 7) is its score based on gene expression data, SW_v(defined in Eq. 8) denotes its weighted score based on subcellular localization, and PC_v(defined in Eq. 9) is its protein complex score. This integration strategy enables a more comprehensive assessment of the essentiality of proteins.

The construction of the intra-layer transition matrix and inter-layer transition matrix in the multi-layer PPI network

A single-layer PPI network is typically represented as a graph structure, denoted as $G_{single} = (V, E)$, where V is the set of nodes (representing proteins) and E is the set of edges (representing interactions between proteins). When constructing a multilayer PPI network, each species' PPI network can be treated as an independent layer. By incorporating homologous protein relationships, interlayer edges are added between protein nodes of different species to establish interspecies connections, forming a multilayer network structure. For example, the PPI network of yeast can be regarded as the first layer, denoted as $G_a = (V_a, E_a)$, where a represents the species yeast; the PPI network of fruitfly is the second layer, denoted as $G_b = (V_b, E_b)$, where b represents the species fruitfly; and the human PPI network is the third layer, denoted as $G_c = (V_c, E_c)$, where c represents the species human. By incorporating homologous protein relationships, interlayer edges are added between these networks. For instance, $E_{a,b}$ denotes the homologous protein relationships between species a (yeast) and b (fruit fly), while $E_{a,c}$ and $E_{b,c}$ represent homologous relationships between yeast and human, and between fruit fly and human, respectively.

Combining the intralayer and interlayer information, the multilayer PPI network can be represented as $G = (V_a, V_b, V_c, E_a, E_b, E_c, E_{a,b}, E_{a,c}, E_{b,c})$. In the multilayer network G, the intralayer edge sets (such as E_a, E_b, E_c) describe the protein interactions within each species, while the interlayer edge sets (such as $E_{a,b}, E_{a,c}, E_{b,c}$) describe the homologous protein relationships between different species. In the process of constructing a multilayer PPI network, various biological data can be utilized to characterize the importance of protein interactions within each layer and the significance of homologous protein relationships between layers. This facilitates the generation of intralayer transition matrices and interlayer transition matrices.

Intra-layer transition matrix

In constructing the intra-layer transition matrix, we integrate weighted edge clustering coefficients, Gene Ontology (GO) semantic similarity, and protein complex information to comprehensively describe the importance of protein interaction edges.

The weighted edge clustering coefficient incorporates homologous protein information into the edge clustering coefficient (ECC). The ECC measures the connectivity tightness between two nodes v and u through the number of their common neighbors $|N_v \cap N_u|$, defined as:

$$ECC_{vu} = \frac{|N_v \cap N_u|}{\min(|N_v| - 1, |N_u| - 1)} \tag{11}$$

To incorporate homologous protein information, we apply homology weighting to common neighbors and define the weighted edge clustering coefficient as:

$$ORECC_{vu} = \frac{1}{N} + \frac{\sum_{N_v \cap N_u}^{k} OR_k}{\min(|N_v| - 1, |N_u| - 1)} \tag{12}$$

where, OR_k (as defined in Eq. 1) represents the homology score of the common neighbor node k, and N is a constant used to avoid division by zero.

GO terms are used to annotate the functional characteristics of proteins. The more similar the GO terms, the closer the functions of the proteins, and the higher the interaction edge weight [41]. The edge weight based on GO annotations is defined as:

$$GOW(v,u) = \frac{|GO_v \cap GO_u|^2}{|GO_v| \cdot |GO_u|} \tag{13}$$

where, GO_v represents the set of GO terms for protein v, and $GOW(v,u)$ is the weight assigned to the edge (v,u).

The edge weight based on protein complex information is defined as:

$$PCW(v,u) = \frac{|SPC_v| \cdot |SPC_u|}{PC_{max}^2} \tag{14}$$

where, SPC_v represents the set of protein complexes to which protein v belongs, and PC_{max} is the maximum $|SPC_v|$ among all proteins, as defined in Eq. 9.

Finally, by integrating the three edge weight components, the weight of the intra-layer transition matrix is defined as:

$$W_{vu} = ORECC_{vu} \cdot (GOW(v,u) + PCW(v,u)) \tag{15}$$

where, $ORECC_{vu}$ is defined in Eq. 12, $GOW(v, u)$ is defined in Eq. 13, and $PCW(v, u)$ is defined in Eq. 14. This formula combines weighted edge clustering coefficients, GO term similarity, and protein complex information to characterize the importance of protein interaction edges from multiple dimensions, providing a biological foundation for the construction of the intra-layer transition matrix.

Inter-layer transition matrix

The interlayer transition matrix constructs cross-layer edges using homologous protein relationships and evaluates the importance of these cross-layer edges based on the criticality scores of homologous proteins and the similarity of their GO annotations.

For homologous proteins v and u from species a and b, the weight of a cross-layer edge is defined as:

$$ORM_{v,u}^{a,b} = OR_v^a \cdot OR_u^b \tag{16}$$

where OR_v^a denotes the criticality score of the homologous protein v in species a, as defined in Eq. 1.

GO annotations provide a standardized language for describing the functions of genes and proteins across different species, facilitating cross-species comparison and annotation. This unified framework highlights the differences in conservation and specificity of proteins in biological processes, particularly since homologous proteins (derived from a common ancestor) often share similar GO annotations [35]. Based on GO annotations, the weight of a cross-layer edge between homologous proteins v and u from species a and b is defined as:

$$GOM_{v,u}^{a,b} = |GO_v^a \cap GO_u^b| \tag{17}$$

where GO_v^a is the set of GO terms for protein v in species a.

Taking species a and b as an example, the interlayer transition matrix is defined as:

$$M_{v,u}^{a,b} = ORM_{v,u}^{a,b} \cdot GOM_{v,u}^{a,b} \tag{18}$$

where, $ORM_{v,u}^{a,b}$ is defined in Eq. 16, and $GOM_{v,u}^{a,b}$ is defined in Eq. 17.

Column normalization

To ensure the stability and convergence of the MLPR algorithm, both the intralayer and interlayer transition matrices need to be column-normalized, ensuring that the sum of elements in each column equals 1.

For the intralayer transition matrix $W(n \times n)$, where each element W_{vu} represents the transition probability from node v to node u within the same layer, if the sum of elements in a column is nonzero, the normalization for each element is defined as:

$$W_{vu} = \frac{W_{vu}}{\sum_{v=1}^{n} W_{vu}}, \quad \text{if} \quad \sum_{v=1}^{n} W_{vu} \neq 0 \tag{19}$$

For the interlayer transition matrix $M(n \times m)$, where each element M_{vu} represents the transition probability from node v in one layer to node u in another layer, if the sum of elements in a column is nonzero, the normalization for each element is defined as:

$$M_{vu} = \frac{M_{vu}}{\sum_{v=1}^{n} M_{vu}}, \quad \text{if} \quad \sum_{v=1}^{n} M_{vu} \neq 0 \tag{20}$$

The above normalization ensures the numerical stability of the transition matrices, providing a solid foundation for the convergence of the MLPR algorithm.

Multiple PageRank model based on multilayer PPI network

The multiple PageRank model aims to integrate the PPI network information of multiple species by constructing a multilayer structure, where each layer corresponds to the PPI network of one species. Intra-layer interactions are represented by the relationship matrix of the PPI network within the species, while inter-layer interactions connect different species through homologous relationships. During each iteration, the model dynamically updates the criticality score of each protein by comprehensively considering intra-layer interactions and inter-layer biases, thus providing comprehensive information for evaluating protein essentiality across species.

The iterative process of the model begins by calculating the bias score Bi_a, as follows:

$$Bi_a = \beta M_{a,b} \cdot P_b^0 + (1-\beta) M_{a,c} \cdot P_c^0 \tag{21}$$

where, Bi_a represents the bias score of species a during the iteration process, which incorporates the homologous relationships between species a and other species (b and c). $M_{a,b}$ and $M_{a,c}$ are the inter-layer transition matrices between species a and species b, c (as defined in Eq. 20), while P_b^0 and P_c^0 are the initial score vectors of species b and c (as defined in Eq. 10). The parameter $\beta \in (0,1)$ controls the weight distribution of homologous relationships from different species on the bias score Bi_a.

Subsequently, the model updates the ranking score of each protein in species a using the following formula:

$$P_a^{t+1} = (1-\lambda)\big(\alpha W_a \cdot P_a^t + (1-\alpha) Bi_a\big) + \lambda P_a^0 \tag{22}$$

where, $\lambda \in (0,1)$ is the damping factor used to simulate random jump behavior, handle isolated nodes, and ensure algorithm stability and convergence. t represents the number of iterations, starting from $t = 0$ and continuing until convergence. P_a^0 is the initial score vector of species a (as defined in Eq. 10), and W_a is the intra-layer transition matrix of species a (as defined in Eq. 19). The parameter $\alpha \in (0,1)$ adjusts the weight ratio between intra-layer interactions and inter-layer biases Bi_a in the score update process.

During each iteration, the model updates the score vector by integrating intra-layer interactions and inter-layer biases. Parameters α and β control the weight distribution for intra-layer and inter-layer biases, as well as the contribution of different species to Bi_a, achieving a balanced utilization of multi-source information. The iteration process continues until the $L1$ norm $\|P_a^{t+1} - P_a^t\|_1$ is smaller than a predefined threshold, ensuring the stability and convergence of the model.

Based on the final converged score vector P_a^t, the model ranks proteins in descending order, with the top-ranked proteins identified as predicted essential proteins. The number of essential proteins varies by species; for instance, the top 25% of proteins in yeast, 10% in fruit fly, and 35% in human are selected. The model integrates diverse biological information within species and homologous information across species, providing more comprehensive and accurate predictions for essential proteins.

The mathematical proof for model convergence

The intra-layer update component of this model is based on the Markov chain concept; however, the overall model is not a strict Markov chain. This is because the model incorporates inter-layer bias terms from multiple species, breaking the closure and pure state-dependency of classical Markov chains. Therefore, the model can be considered a hybrid that extends the Markov chain concept. To prove the model's convergence, it is necessary to analyze the mathematical properties of its iterative formula and demonstrate that the iteration is a contraction mapping, thereby proving convergence based on the Banach fixed-point theorem.

A contraction mapping is defined as follows: If a mapping T has a constant $c \in [0, 1)$ such that for any two vectors x, y, $\| T(x) - T(y) \| \leq c\|x - y\|$, then T is a contraction mapping. Based on this definition, the mapping in the model iteration process can be expressed as:

$$T(P_a^t) = (1 - \lambda)\big(\alpha W_a \cdot P_a^t + (1 - \alpha)Bi_a\big) + \lambda P_a^0 \tag{23}$$

where, $W_a \cdot P_a^t$ is the linear transformation of vector P_a^t by the row-stochastic matrix W_a. Since W_a is row-stochastic, where each row is non-negative and sums to 1, it satisfies $\| W_a \cdot P_a^t\|_1 \leq \|P_a^t\|_1$, meaning it does not amplify vector norms. Additionally, since $\lambda, \alpha \in (0, 1)$, $(1 - \lambda)\alpha W_a \cdot P_a^t$ and other terms collectively form a contraction mapping. The bias term Bi_a is a fixed vector, numerically stable and non-divergent; the term λP_a^0 acts as a stabilizing factor for random jumps, further enhancing model stability. Therefore, the entire mapping T satisfies the conditions of the Banach fixed-point theorem.

According to the Banach fixed-point theorem, the iterative formula $P_a^{t+1} = T(P_a^t)$ will converge to a unique fixed point, ensuring the theoretical convergence of the model. In practical applications, numerical experiments verify the convergence of the model. At each iteration, the change in $L1$ norm $\|P_a^{t+1} - P_a^t\|_1$ decreases as t increases, approaching a minimal value, indicating that the model reaches a stable state.

Influence of parameters

MLPR involves three critical parameters: λ, α, and β, each playing a unique role in the model. The details are as follows:

The parameter λ controls the balance between intra-layer random walk and global jump. During the iterative process, the introduction of λ provides randomness to the model, preventing it from falling into a local optimum. When λ is small, the scores are more influenced by intra-layer interactions, and the model places greater emphasis on direct connections between nodes. Conversely, when λ is large, the model relies more on the initial score vector P_a^0, which helps address the issue of isolated nodes by preserving their initial score influence.

The parameter α governs the balance between intra-layer interactions and inter-layer bias Bi_a. A larger α gives greater weight to intra-layer interactions during the score update, indicating that direct connections between nodes are more critical. In contrast, a smaller α emphasizes the impact of inter-layer bias Bi_a, reflecting the importance of homologous relationships between different species. By adjusting α, the model can strike an appropriate balance between intra-layer random walks and inter-layer bias, thereby improving its performance. The flexibility of this parameter enhances the model's ability to adapt to different network structures and homologous information characteristics.

The parameter β regulates the weight of homologous relationships between different species in the inter-layer bias Bi_a. When β is large, the model gives more importance to the homologous relationship between species a and species b. Conversely, when β is small, the homologous relationship between species a and species c has a more significant influence on the model. By tuning β, the model can balance the influence of different species on Bi_a, thereby capturing homologous information across multiple species more precisely.

To achieve an optimal balance between intra-layer random walk, global jump, inter-layer bias, and homologous relationships among species, it is necessary to fine-tune λ, α, and β. This parameter tuning process can effectively enhance the model's robustness and performance, enabling it to perform better in complex multilayer networks. To address the high computational cost of traditional parameter-tuning methods, this study proposes a two-step tuning approach that significantly reduces tuning complexity and time cost while ensuring improved model performance. Specifically, in the first step, the parameter λ is optimized independently, as it only controls the balance between intra-layer random walks and global jump, without being influenced by inter-layer bias and homologous relationships, thereby allowing α and β to be temporarily disregarded. Once the optimal value of λ is determined, the second step focuses on tuning α and β, with sensitivity analysis used to find their optimal combination. Compared to the traditional grid search method, which requires conducting $9^3 = 729$ experiments, the two-step tuning approach only requires $9 + 9 \times 9 = 90$ experiments, reducing the number of experiments by nearly 90% and significantly lowering computational cost. The pseudo-code for the MLPR parameter tuning process is presented in Algorithm 1. A detailed description of the two-step parameter tuning process is provided below.

First, analyze the sensitivity of the parameter λ. Since λ is a parameter that controls the balance between intra-layer random walks and global jump, its analysis does not require addressing the control of inter-layer bias and homologous relationships between different species. Therefore, when analyzing the parameter λ, the effects of parameters α and β do not need to be considered. The sensitivity analysis of λ for species a is conducted using the following formula:

$$P_a^{t+1} = (1-\lambda) \cdot W_a \cdot P_a^t + \lambda \cdot P_a^0 \tag{24}$$

We set λ from 0.1 to 0.9 with a step size of 0.1. Table 2 compares the statistical measures under nine different λ values, with the maximum value of each measure for each species highlighted in bold. Figure 2 shows the Jackknife curves for the nine λ values, where the optimal curve for each species is highlighted in bold black.

For the three species–yeast, fruit fly, and human–the results of the statistical measures and Jackknife curves indicate that the optimal λ values are 0.4, 0.9, and 0.9, respectively.

Next, we perform a sensitivity analysis for the parameters α and β. We directly use the optimal λ value obtained in the previous step, denoted as λ_{op}, for tuning α and β. Taking yeast as an example, the tuning process is carried out according to the following formula:

$$\begin{aligned} P_a^{t+1} = (1-\lambda_{op}) * \Big(\alpha * W_a * P_a^t + (1-\alpha) * \Big(\beta * M_{a,b} * P_b^0 \\ + (1-\beta) * M_{a,c} * P_c^0\Big)\Big) + \lambda_{op} * P_a^0 \end{aligned} \tag{25}$$

α and β range from 0.1 to 0.9, with a step size of 0.1. Figure 3 illustrates the ACC values for all combinations of α and β, with the maximum value highlighted using a red

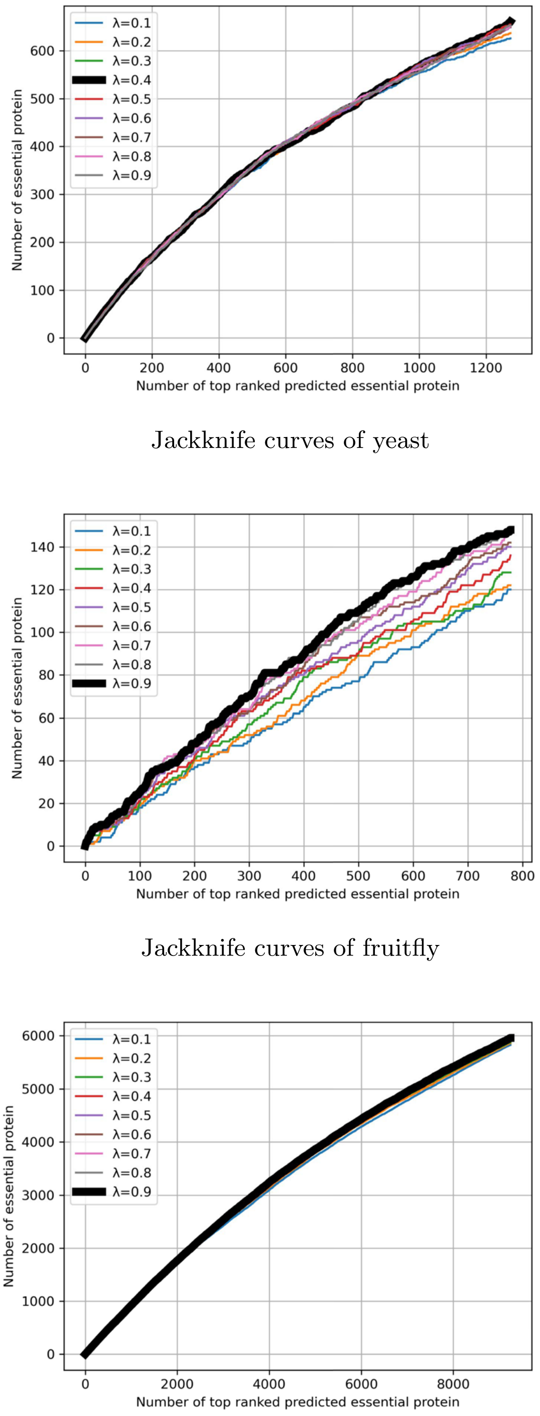
Table 2 Sensitivity analysis of the parameter λ

Dataset	Measures	λ								
		0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Yeast	SN	0.5364	0.5458	0.5587	**0.5664**	0.5613	0.5561	0.5570	0.5587	0.5596
	SP	0.8352	0.8380	0.8418	**0.8441**	0.8426	0.8411	0.8413	0.8418	0.8421
	PPV	0.4918	0.5004	0.5122	**0.5192**	0.5145	0.5098	0.5106	0.5122	0.5130
	NPV	0.8584	0.8613	0.8652	**0.8675**	0.8660	0.8644	0.8647	0.8652	0.8654
	F	0.5131	0.5221	0.5344	**0.5418**	0.5369	0.5320	0.5328	0.5344	0.5352
	ACC	0.7667	0.7711	0.7769	**0.7805**	0.7781	0.7758	0.7762	0.7769	0.7773
Fruitfly	SN	0.2434	0.2475	0.2596	0.2759	0.2840	0.2880	0.2961	0.2961	**0.3002**
	SP	0.9097	0.9100	0.9108	0.9119	0.9125	0.9128	0.9133	0.9133	**0.9136**
	PPV	0.1542	0.1568	0.1645	0.1748	0.1799	0.1825	0.1877	0.1877	**0.1902**
	NPV	0.9468	0.9470	0.9479	0.9490	0.9496	0.9499	0.9505	0.9505	**0.9507**
	F	0.1888	0.1920	0.2014	0.2140	0.2203	0.2234	0.2297	0.2297	**0.2329**
	ACC	0.8675	0.8680	0.8696	0.8716	0.8727	0.8732	0.8742	0.8742	**0.8747**
Human	SN	0.6695	0.6733	0.6752	0.6777	0.6789	0.6801	0.6813	0.6823	**0.6845**
	SP	0.8072	0.8090	0.8100	0.8112	0.8118	0.8124	0.8130	0.8135	**0.8146**
	PPV	0.6306	0.6342	0.6360	0.6384	0.6395	0.6407	0.6418	0.6428	**0.6448**
	NPV	0.8324	0.8343	0.8353	0.8365	0.8371	0.8378	0.8384	0.8389	**0.8400**
	F	0.6495	0.6531	0.6550	0.6575	0.6586	0.6598	0.6609	0.6619	**0.6641**
	ACC	0.7618	0.7643	0.7655	0.7672	0.7680	0.7688	0.7696	0.7702	**0.7717**

dot. For yeast, fruit fly, and human, the optimal values of parameters (α, β) are (0.5, 0.8), (0.1, 0.1), and (0.4, 0.7), respectively.

Parameters α and β further improve the model's precision in identifying essential proteins by regulating inter-layer bias and homologous relationships within inter-layer species. To demonstrate the effectiveness of parameters α and β on the final results of the model, Table 3 compares the statistical metrics of the model's performance with all parameters λ, α, and β included versus with only λ included. The results indicate that for yeast, fruit fly, and human, the model achieves the best outcomes when all parameters are included. Figure 4 presents the Jackknife curves comparing the model's performance using all parameters versus using only λ. The optimal curve for each species is highlighted in bold black. In the fruit fly and human datasets, the Jackknife curves with all parameters consistently outperform those with only λ. In the yeast dataset, the full-parameter model performs slightly worse than the λ-only model for ranks up to 1200 but surpasses it thereafter, achieving superior overall statistical measures. This further confirms the effectiveness of introducing inter-layer bias Bi_a and homologous relationships within Bi_a across different species as described in Eq. 22.

Algorithm 1 MLPR

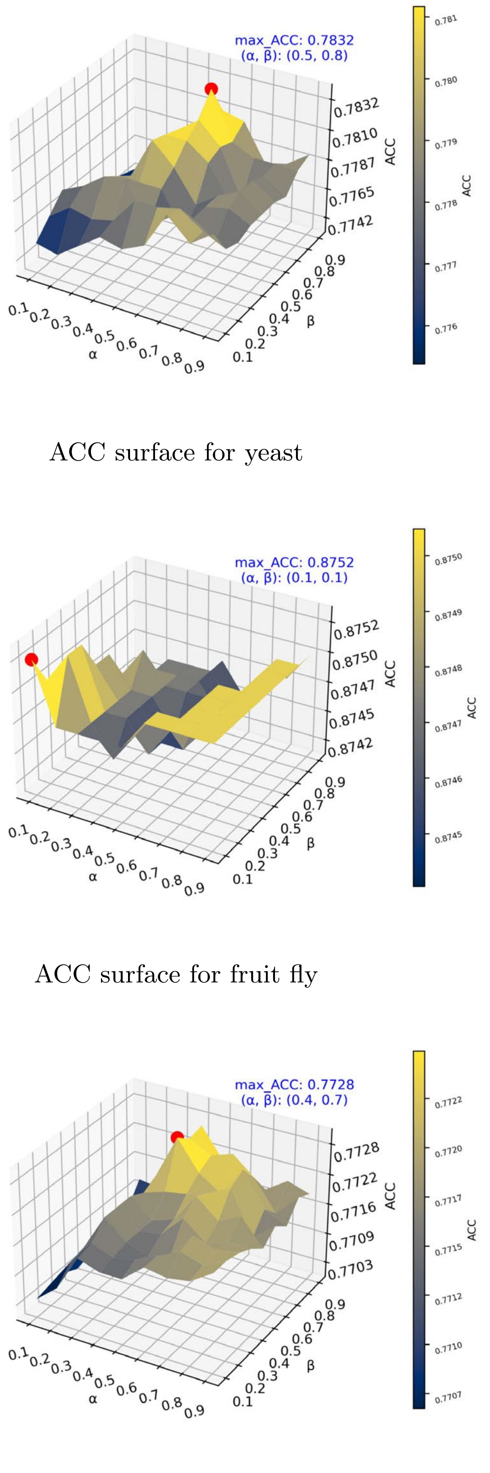


Jackknife curves of yeast

Jackknife curves of fruitfly

Jackknife curves of human

Fig. 2 Jackknife curves the nine selected *lambda* values



ACC surface for yeast

ACC surface for fruit fly

ACC surface for human

Fig. 3 ACC surfaces for all combinations of parameters α and β

Table 3 Sensitivity analysis of parameters α and β

Dataset	Parameters	SN	SP	PPV	NPV	F	ACC
Yeast	$\lambda = 0.4$	0.5664	0.8441	0.5192	0.8675	0.5418	0.7805
	$(\lambda,\alpha,\beta) = (0.4, 0.5, 0.8)$	**0.5724**	**0.8459**	**0.5247**	**0.8694**	**0.5475**	**0.7832**
Fruitfly	$\lambda = 0.9$	0.3002	0.9136	0.1902	0.9507	0.2329	0.8747
	$(\lambda,\alpha,\beta) = (0.9, 0.1, 0.1)$	**0.3043**	**0.9139**	**0.1928**	**0.9510**	**0.2360**	**0.8752**
Human	$\lambda = 0.9$	0.6845	0.8146	0.6448	0.8400	0.6641	0.7717
	$(\lambda,\alpha,\beta) = (0.9, 0.4, 0.7)$	**0.6862**	**0.8154**	**0.6464**	**0.8409**	**0.6657**	**0.7728**

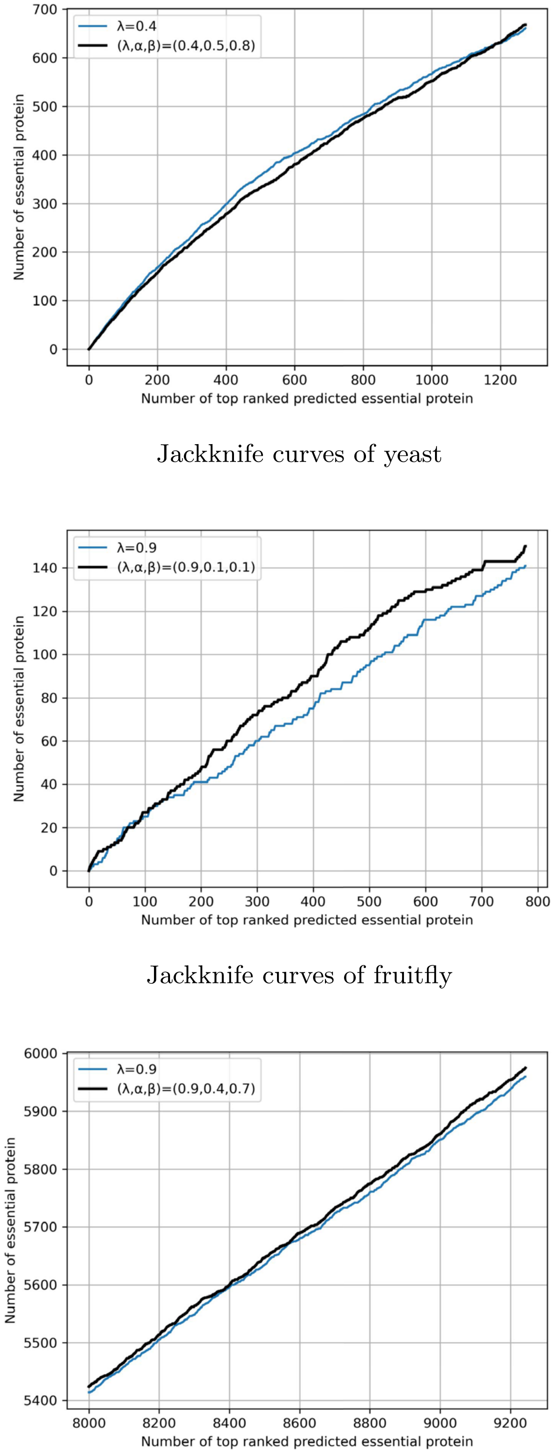
```
Input: The multilayer PPI networks $G = (V_a, V_b, V_c, E_a, E_b, E_c, E_{a,b}, E_{a,c}, E_{b,c})$ of species $a$, $b$ and $c$; Multi-biological data; Parameter $\lambda$, $\alpha$ and $\beta$.
Output: The predicted essential proteins.
Begin
Initialize initial score $P_a^0$, $P_b^0$, $P_c^0$ by Eq. 10;
Take species $a$ as an example to illustrate the MLPR algorithm;
for $\lambda$ ranges from $0.1$ to $0.9$ with a step size of $0.1$ do
    repeat
        Compute $P_a^t$ by Eq. 24, set $t = t + 1$;
    until $||P_a^{t+1} - P_a^t|| < 10^{-6}$
    Find the value of $\lambda$ that yields the best result, denoted as $\lambda_{best}$;
end for
for $\alpha$ and $\beta$ range from $0.1$ to $0.9$ with a step size of $0.1$ do
    repeat
        Compute $P_a^t$ by Eq. 25, set $t = t + 1$, $\lambda = \lambda_{best}$;
    until $||P_a^{t+1} - P_a^t|| < 10^{-6}$
    Adjust the values of $\alpha$ and $\beta$ to find the optimal $P_a^t$;
end for
Sort the proteins in descending order based on the values of optimal $P_a^t$, and output the
    top-ranked proteins as the predicted essential proteins.
End
```

Experimental results and discussion

Statistical measures and jackknife curves

We evaluate and demonstrate the superiority of the MLPR model using six statistical measures: Sensitivity (SN), Specificity (SP), Positive Predictive Value (PPV), Negative Predictive Value (NPV), F-measure (F), and Accuracy (ACC). These measures are defined as follows: $SN = TP/(TP + FN)$, $SP = TN/(TN + FP)$, $PPV = TP/(TP + FP)$, $NPV = TN/(TN + FN)$, $F = 2 \times SN \times PPV/(SN + PPV)$, $ACC = (TP + TN)/(TP + FP + TN + FN)$, where, TP represents True Positives, FP represents False Positives, TN represents True Negatives, and FN represents False Negatives. Higher values of these measures indicate greater accuracy of the essential protein identification method.

In addition, we plot the Jackknife curve to illustrate the change in the number of true positives (TP) in the predicted set of essential proteins as the ranking increases. A higher cumulative curve indicates better algorithm performance. This visualization intuitively reflects the model's prediction effectiveness across different ranking ranges.



Jackknife curves of yeast

Jackknife curves of fruitfly

Jackknife curves of human

Fig. 4 Jackknife curves: considering all Parameters vs. using only λ

Ablation experiment

To validate the effectiveness of constructing multilayer PPI networks based on homologous relationships among three species and demonstrate the advantages of the multiple PageRank model in identifying essential proteins, we designed the following ablation experiments: 1. Use the initial scores defined in defined in Eq. 10 to assess the essentiality of proteins. 2. Construct a single-layer PPI network based on a single species and identify essential proteins using the traditional PageRank model. 3. Construct a two-layer PPI network based on two species and identify essential proteins using the dual PageRank model. 4. Construct a three-layer PPI network based on three species and identify essential proteins using the MLPR algorithm proposed in this paper.

Initial scores: To demonstrate the advantages of the multiple PageRank model in identifying essential proteins, we use the initial scores defined in Eq. 10 to assess the essentiality of proteins.

Single specie: Single-species experiments use only single-species data, do not use homologous data, nor do it require constructing inter-layer transition matrix. The initial protein score vector is defined as $P^0 = SW_v \cdot PC_v$, where SW_v and PC_v are defined by Eqs. 8 and 9, respectively. The transition probability matrix is defined as $W_{vu} = ECC_{vu} \cdot \left(GOW(v,u) + PCW(v,u)\right)$, where ECC_{vu}, $GOW(v,u)$, and $PCW(v,u)$ are defined by Eqs. 11, 13, and 14, respectively.

The traditional PageRank model iterates based on the initial score vector P^0 and the transition probability matrix W, using the formula: $P^{t+1} = (1-\lambda) \cdot W \cdot P^t + \lambda \cdot P^0$.

Two species: The two-species experiments simplify the MLPR algorithm. Each species undergoes two experiments. Taking species a as an example: 1. Construct a two-layer PPI network based on the homologous relationships between species a and species b and identify essential proteins using the dual PageRank model. 2. Conduct the same experiment for species a and species c. For the homologous relationship experiment between species a and species b, the initial protein score vector is defined as $P_a^0 = OR_v^a \cdot SW_v \cdot PC_v$, where SW_v and PC_v are defined by Eqs. 8 and 9, and the protein homologous score is defined as $OR_v^a = \left|orth_v^b\right| / OR_{max}^a$, where, $orth_v^b$ represents the set of homologous proteins of protein v in species b, and $OR_{max}^a = \max\left(OR_v^a\right), (v \in V_a)$. The intra-layer transition probability matrix is defined as $W_{vu} = ORECC_{vu} \cdot \left(GOW(v,u) + PCW(v,u)\right)$, where $ORECC_{vu}$ is given by Eq. 12. In Eq. 12, $OR_v^a = \left|orth_v^b\right| / OR_{max}^a$. The terms $GOW(v,u)$ and $PCW(v,u)$ are defined in Eqs. 13 and 14, respectively. The inter-layer transition probability matrix is defined as $M_{vu}^{a,b} = ORM_{vu}^{a,b} \cdot GOM_{vu}^{a,b}$, where $ORM_{vu}^{a,b}$ and $GOM_{vu}^{a,b}$ are defined by Eqs. 16 and 17, respectively. In Eq. 16, $OR_v^a = \left|orth_v^b\right| / OR_{max}^a$ and $OR_u^b = \left|orth_u^a\right| / OR_{max}^b$.

The dual PageRank model iterates based on the initial score vector P_a^0, the intra-layer transition probability matrix W_a, the inter-layer transition probability matrix $M_{a,b}$, and the initial score vector P_b^0 of species b, using the formula:

$$P_a^{t+1} = (1-\lambda) \cdot \left(\alpha \cdot W_a \cdot P_a^t + (1-\alpha) \cdot M_{a,b} \cdot P_b^0\right) + \lambda \cdot P_a^0. \tag{26}$$

Three species: The three-species experiment adopts the MLPR algorithm proposed in this paper for analysis.

Results and analysis: All experiments are conducted using optimal parameter configurations, and the results are shown in Table 4. The highest value for each statistical

Table 4 Ablation experiment

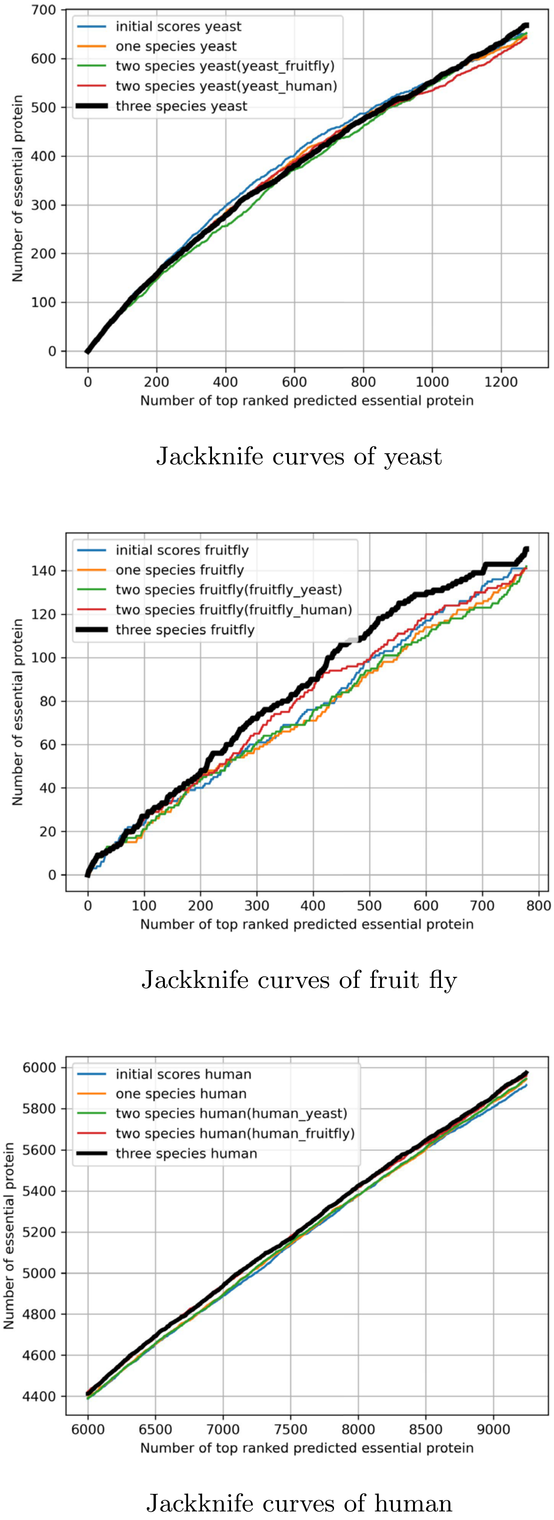
Dataset	Species number	SN	SP	PPV	NPV	F	ACC
Yeast	Initial scores yeast	0.5578	0.8416	0.5114	0.8649	0.5336	0.7766
	One species	0.5536	0.8403	0.5075	0.8636	0.5295	0.7746
	Two species(yeast fruitfly)	0.5587	0.8418	0.5122	0.8652	0.5344	0.7769
	Two species(yeast human)	0.5501	0.8393	0.5043	0.8626	0.5262	0.7730
	Three species(MLPR)	**0.5724**	**0.8459**	**0.5247**	**0.8694**	**0.5475**	**0.7832**
Fruitfly	Initial scores fruitfly	0.2880	0.9128	0.1825	0.9499	0.2234	0.8732
	One species	0.2860	0.9126	0.1812	0.9498	0.2219	0.8729
	Two species (fruitfly yeast)	0.2880	0.9128	0.1825	0.9499	0.2234	0.8732
	Two species (fruitfly human)	0.2860	0.9126	0.1812	0.9498	0.2219	0.8729
	Three species (MLPR)	**0.3043**	**0.9139**	**0.1928**	**0.9510**	**0.2360**	**0.8752**
Human	Initial scores human	0.6793	0.8120	0.6399	0.8374	0.6591	0.7683
	One species	0.6826	0.8136	0.6430	0.8390	0.6622	0.7704
	Two species (human yeast)	0.6829	0.8138	0.6433	0.8392	0.6625	0.7706
	Two species (human fruitfly)	0.6845	0.8146	0.6448	0.8400	0.6641	0.7717
	Three species (MLPR)	**0.6862**	**0.8154**	**0.6464**	**0.8409**	**0.6657**	**0.7728**

measure across species is highlighted in bold. Figure 5 illustrates the Jackknife curves for all experiments, with the best-performing curve for each species highlighted in bold black. As shown in Table 4 and Fig. 5, the MLPR algorithm consistently outperforms other ablation methods. These results demonstrate that incorporating homologous relationships among the three species effectively enhances the overall performance of the MLPR algorithm and highlight the advantages of the multiple PageRank model in identifying essential proteins.

Analysis of the performance of MLPR and other methods

To validate the performance of the proposed MLPR method, we conduct a comprehensive comparison with traditional methods, including SIGEP, TS-PIN, RWEP, RWO, and SESN, as well as two deep learning models, DeepEP and MBIEP. The experiments use the same datasets and consistent evaluation metrics to ensure the fairness and reliability of the results. Tables 5 and 6 present the statistical measures for traditional and deep learning methods, respectively, with the highest measure for each species highlighted in bold. Figure 6 illustrates the Jackknife curves for traditional methods. However, due to differences in output formats between MLPR and deep learning models, the Jackknife curve is not used in the comparative analysis.

In comparisons with traditional methods, SIGEP, RWEP, and SESN only focus on a single species and do not utilize homologous relationships across species. Experimental results show that MLPR significantly outperforms these methods on datasets from all three species. SIGEP, which does not integrate biological data, performs significantly worse than MLPR, demonstrating that integrating diverse biological data effectively enhances the identification of essential proteins. Although RWEP and SESN use multiple biological datasets, they do not account for interspecies homologous relationships, resulting in inferior performance compared to MLPR. Notably, in datasets of fruit fly and human, the Jackknife curve of MLPR consistently exceeds those of



Jackknife curves of yeast

Jackknife curves of fruit fly

Jackknife curves of human

Fig. 5 Jackknife curves of ablation experiment

Table 5 Comparison of statistical measures between MLPR and other methods

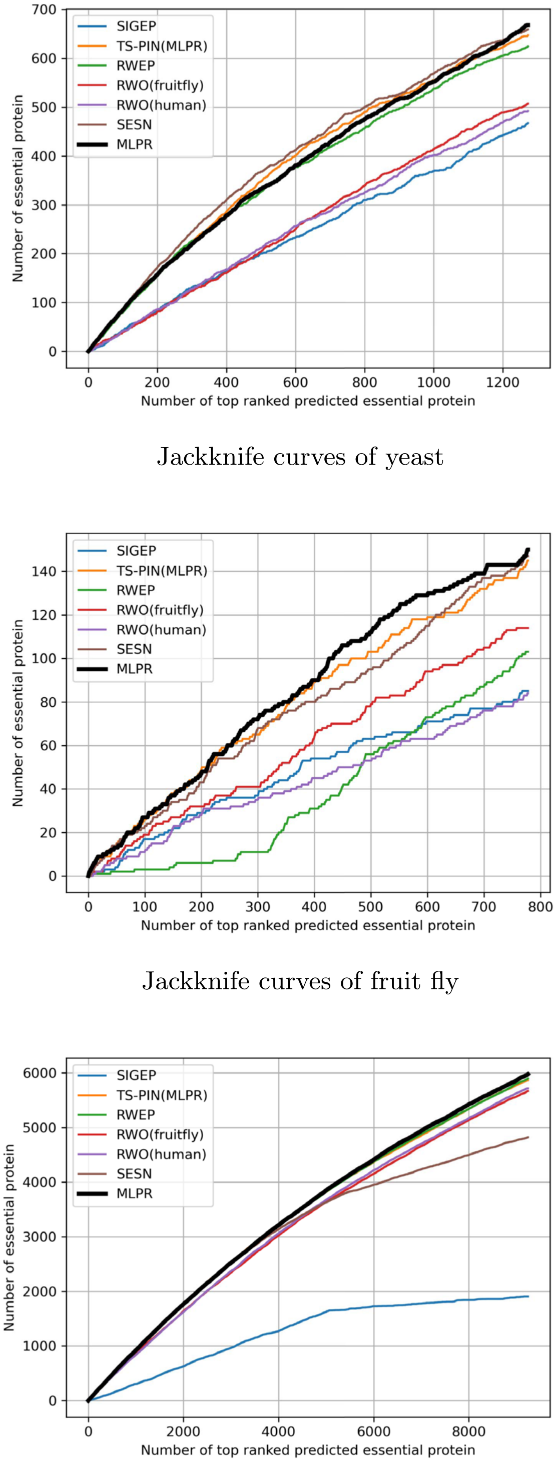
Dataset	Methods	SN	SP	PPV	NPV	F	ACC
Yeast	SIGEP	0.4002	0.7947	0.3668	0.8168	0.3828	0.7043
	TS-PIN(MLPR)	0.5544	0.8406	0.5082	0.8639	0.5303	0.7750
	RWEP	0.5347	0.8347	0.4902	0.8579	0.5115	0.7660
	RWO(fruitfly)	0.4344	0.8049	0.3983	0.8272	0.4156	0.7200
	RWO(human)	0.4216	0.8011	0.3865	0.8233	0.4033	0.7141
	SESN	0.5647	0.8436	0.5177	0.8670	0.5402	0.7797
	MLPR	**0.5724**	**0.8459**	**0.5247**	**0.8694**	**0.5475**	**0.7832**
Fruitfly	SIGEP	0.1724	0.9049	0.1093	0.9418	0.1338	0.8585
	TS-PIN(MLPR)	0.2941	0.9132	0.1864	0.9503	0.2282	0.8740
	RWEP	0.2089	0.9074	0.1324	0.9443	0.1621	0.8632
	RWO(yeast)	0.2312	0.9089	0.1465	0.9459	0.1794	0.8660
	RWO(human)	0.1704	0.9048	0.1080	0.9416	0.1322	0.8583
	SESN	0.2982	0.9134	0.1889	0.9506	0.2313	0.8745
	MLPR	**0.3043**	**0.9139**	**0.1928**	**0.9510**	**0.2360**	**0.8752**
Human	SIGEP	0.2190	0.5856	0.2063	0.6039	0.2125	0.4647
	TS-PIN(MLPR)	0.6734	0.8091	0.6343	0.8343	0.6533	0.7643
	RWEP	0.6772	0.8109	0.6379	0.8363	0.6569	0.7668
	RWO(yeast)	0.6507	0.7979	0.6130	0.8229	0.6313	0.7494
	RWO(fruitfly)	0.6565	0.8008	0.6184	0.8258	0.6369	0.7532
	SESN	0.5537	0.7502	0.5216	0.7736	0.5372	0.6854
	MLPR	**0.6862**	**0.8154**	**0.6464**	**0.8409**	**0.6657**	**0.7728**

RWEP and SESN. In the yeast dataset, MLPR slightly underperforms SESN for ranks up to 1200 but surpasses SESN thereafter, with superior statistical measures overall.

For the TS-PIN method, we input its refined networks into MLPR to form the TS-PIN(MLPR) method. Experimental results show that MLPR significantly outperforms TS-PIN(MLPR) across datasets of all three species. As shown in Fig. 6, the Jackknife curve of MLPR consistently remains above that of TS-PIN(MLPR) in fruit fly and human datasets. In the yeast dataset, while MLPR's Jackknife curve is slightly lower than TS-PIN(MLPR)'s for the top 1000 ranks, it surpasses TS-PIN(MLPR) beyond rank 1000. Additionally, MLPR exhibits superior statistical measures compared to TS-PIN(MLPR). These results indicate that the TS-PIN algorithm does not provide substantial improvement to MLPR's performance.

Compared to the RWO method, MLPR integrates homologous relationships among three species and cross-species GO annotations, assigning weights to interlayer edges and demonstrating stronger performance advantages. To validate the effectiveness of MLPR in incorporating homologous relationships (e.g., yeast and fruit fly, yeast and human, fruit fly and human), RWO conducts two experiments for each species, each utilizing the homologous relationships between that species and the other two species. For example, for yeast, RWO experiments are based on the homologous relationships between yeast and fruit fly (RWO(fruitfly)) and between yeast and human (RWO(human)). As shown in Table 5 and Fig. 6, MLPR consistently outperforms RWO, regardless of the interspecies homologous relationship used by RWO.

In comparisons with deep learning models, since MLPR outputs essentiality ranking scores while deep learning models provide probabilities for positive (minority) classes,



Jackknife curves of yeast

Jackknife curves of fruit fly

Jackknife curves of human

Fig. 6 Jackknife curves of MLPR and other methods

Table 6 Comparison between MLPR and deep learning methods

Dataset	Methods	SN	SP	PPV	NPV	F	ACC
Yeast	DeepEP	0.5526	0.8000	0.4599	0.8530	0.5020	0.6046
	MBIEP	**0.6228**	0.6459	0.3515	0.8475	0.4494	0.6400
	MLPR	0.5614	**0.8514**	**0.5378**	**0.8630**	**0.5494**	**0.7831**
Fruitfly	DeepEP	**0.2857**	0.7723	0.0778	0.9415	0.1223	0.7416
	MBIEP	0.2449	0.8999	0.1412	0.9426	0.1691	0.8586
	MLPR	0.1633	**0.9492**	**0.1778**	**0.9441**	**0.1702**	**0.8997**
Human	DeepEP	0.3851	0.9780	0.8957	0.7639	0.5386	0.7826
	MBIEP	0.3713	0.9706	0.8613	0.7585	0.5189	0.7731
	MLPR	**0.4069**	**0.9785**	**0.9031**	**0.7705**	**0.5610**	**0.7902**

we calculate MLPR's statistical measures based on the test set of the deep learning models to ensure fairness. All methods are tested on the same datasets using consistent statistical measures. While DeepEP and MBIEP show higher sensitivity (SN) on certain datasets, they achieve the lowest scores on all other measures. This is primarily due to the imbalanced nature of the datasets, which causes the models to favor predicting samples as positive (minority class) during training. This bias significantly increases false positives (FP) and, due to insufficient focus on the negative (majority) class, reduces the counts of true negatives (TN) and false negatives (FN). These factors collectively result in lower specificity (SP), accuracy (ACC), positive predictive value (PPV), negative predictive value (NPV), and F1-score. In contrast, MLPR demonstrates stronger robustness and comprehensiveness in handling imbalanced datasets, effectively avoiding these biases and achieving superior performance across all measures.

In summary, MLPR leverages homologous protein relationships, multi-biological data, and multiple PageRank model based on multilayer PPI network to significantly improve the performance of essential protein identification. It outperforms both traditional and deep learning methods across statistical measures, showcasing exceptional overall advantages.

Conclusions

The prediction and study of essential proteins not only help to reveal the fundamental requirements for cell survival and growth regulation mechanisms but also deepen our understanding of disease mechanisms and provide significant insights for drug development. Currently, most essential protein identification methods focus on the PPI networks of a single species, failing to fully exploit the homologous relationships across species. However, homologous relationships can connect proteins from different species into multilayer PPI networks. Existing methods typically construct cross-layer edges based on homologous relationships between two species but fail to incorporate biological attributes to evaluate the biological importance of these edges. Furthermore, since homologous proteins are often highly conserved across multiple species, extending homologous relationships to more species can better assess the significance of cross-layer edges.

To address these issues, we proposed a novel model, MLPR, which utilizes homologous proteins to construct multilayer PPI networks and combines the multiple PageRank model to identify essential proteins. In this study, we integrated homologous protein data from three species to construct inter-layer transition matrices and assigned biological weights to cross-layer edges by incorporating the biological attributes of homologous proteins and cross-species GO annotations. The MLPR model comprehensively considers homologous relationships across multiple species, integrates various biological data to initialize protein scores, and introduces three critical parameters to optimize the balance among intralayer random walks, global jumps, interlayer biases, and interspecies homologous relationships. After model convergence, protein scores are ranked in descending order, and the top-ranked proteins are identified as the predicted essential proteins. Experimental results demonstrate that MLPR outperforms other comparative methods in performance. Ablation experiments further verify the contribution of integrating homologous relationships from three species to the overall performance improvement of MLPR.

In future studies, we plan to design new models to automatically learn the features of homologous relationships across multiple species and develop algorithms capable of handling multi-type biological data to further enhance the performance of essential protein identification.

Acknowledgements

The author thanks the anonymous reviewers for their comments and suggestions. Additionally, the author would like to thank all the teachers and students who participated in this research for their guidance and assistance.

Author contributions

HZ and TW obtained and processed datasets. HZ and GL designed the new method, MLPR. GL, and HX provided suggestions and analyzed the results. HZ wrote the manuscript. HZ, GL, and HX reviewed and edited this manuscript. All authors contributed to this work and approved the submitted version.

Funding

This work was supported by the National Nature Science Foundation of China [Grant Number 62372208]; Science and Technology Development Program of Jilin Province [Grant Number:YD ZJ202501ZYTS325]; Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China, Jilin University.

Availability of data and materials

The processed dataset and source codes are available in https://github.com/zhaohe555/MLPR

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 22 January 2025 Accepted: 21 February 2025
Published online: 10 March 2025

References

1. Yang Y-M, Jung Y, Abegg D, Adibekian A, Carroll KS, Karbstein K. Chaperone-directed ribosome repair after oxidative damage. Mol Cell. 2023;83(9):1527–37.
2. Li M, Zheng R, Li Q, Wang J, Wu F-X, Zhang Z. Prioritizing disease genes by using search engine algorithm. Curr Bioinform. 2016;11(2):195–202.
3. Menor-Flores M, Vega-Rodríguez MA. Decomposition-based multi-objective optimization approach for ppi network alignment. Knowl Based Syst. 2022;243:108527.
4. Li X, Li W, Zeng M, Zheng R, Li M. Network-based methods for predicting essential genes or proteins: a survey. Brief Bioinform. 2020;21(2):566–83.
5. Li M, Wang J, Chen X, Wang H, Pan Y. A local average connectivity-based method for identifying essential proteins from the network level. Comput Biol Chem. 2011;35(3):143–50.
6. Wang J, Li M, Wang H, Pan Y. Identification of essential proteins based on edge clustering coefficient. IEEE/ACM Trans Comput Biol Bioinform. 2011;9(4):1070–80.
7. Estrada E, Rodriguez-Velazquez JA. Subgraph centrality in complex networks. Phys Rev E. 2005;71(5):056103.
8. Tang Y, Li M, Wang J, Pan Y, Wu F-X. Cytonca: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks. Biosystems. 2015;127:67–72.
9. Liu Y, Liang H, Zou Q, He Z. Significance-based essential protein discovery. IEEE/ACM Trans Comput Biol Bioinform 2020;19(1):633–42.

10. Li M, Ni P, Chen X, Wang J, Wu F-X, Pan Y. Construction of refined protein interaction network for predicting essential proteins. IEEE/ACM Trans Comput Biol Bioinform. 2017;16(4):1386–97.
11. Lei X, Yang X, Fujita H. Random walk based method to identify essential proteins by integrating network topology and biological characteristics. Knowl Based Syst. 2019;167:53–67.
12. Zhao H, Liu G, Cao X. A seed expansion-based method to identify essential proteins by integrating protein-protein interaction sub-networks and multiple biological characteristics. BMC Bioinform. 2023;24(1):452.
13. Tan J, Kuang L, Wang L. Method for essential protein prediction based on the naíve bayesian classifier and bioinformation fusion. In: Proceedings of the 2022 11th International Conference on Bioinformatics and Biomedical Science, 2022;1–7.
14. Zeng M, Li M, Wu F-X, Li Y, Pan Y. Deepep: a deep learning framework for identifying essential proteins. BMC Bioinform. 2019;20:1–10.
15. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016;855–864.
16. Lu P, Tian J. Acdmbi: a deep learning model based on community division and multi-source biological information fusion predicts essential proteins. Comput Biol Chem. 2024;2024:108115.
17. Yue Y, Ye C, Peng P-Y, Zhai H-X, Ahmad I, Xia C, Wu Y-Z, Zhang Y-H. A deep learning framework for identifying essential proteins based on multiple biological information. BMC Bioinform. 2022;23(1):318.
18. Wang B, Ma X, Wang C, Zhang M, Gong Q, Gao L. Conserved control path in multilayer networks. Entropy. 2022;24(7):979.
19. Tortosa L, Vicent JF, Yeghikyan G. An algorithm for ranking the nodes of multiplex networks with data based on the pagerank concept. Appl Math Comput. 2021;392:125676.
20. Cheriyan J, Sajeev G. An improved pagerank algorithm for multilayer networks. In: 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2020;1–6. IEEE.
21. Jin H, Zhang C, Ma M, Gong Q, Yu L, Guo X, Gao L, Wang B. Inferring essential proteins from centrality in interconnected multilayer networks. Physica A: Stat Mech Appl. 2020;557:124853.
22. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S-M, Eisenberg D. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 2002;30(1):303–5.
23. Chatr-Aryamontri A, Breitkreutz B-J, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, et al. The biogrid interaction database: 2015 update. Nucleic Acids Res. 2015;43(D1):470–8.
24. Mewes H-W, Amid C, Arnold R. Frishman: Mips: analysis and annotation of proteins from whole genomes. Nucleic Acids Res. 2004;32(suppl 1):41–4.
25. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, et al. Sgd: Saccharomyces genome database. Nucleic Acids Res. 1998;26(1):73–9.
26. Zhang R, Lin Y. Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res. 2009;37(suppl 1):455–8.
27. Chen W-H, Minguez P, Lercher MJ, Bork P. Ogee: an online gene essentiality database. Nucleic Acids Res. 2012;40(D1):901–6.
28. Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin A-C, Bork P, Superti-Furga G, Serrano L, et al. Structure-based assembly of protein complexes in yeast. Science. 2004;303(5666):2026–9.
29. Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res. 2009;37(3):825–31.
30. Pu S, Vlasblom J, Emili A, Greenblatt J, Wodak SJ. Identifying functional modules in the physical interactome of saccharomyces cerevisiae. Proteomics. 2007;7(6):944–60.
31. Guruharsha K, Rual J-F, Zhai B, Mintseris J, Vaidya P, Vaidya N, Beekman C, Wong C, Rhee DY, Cenaj O, et al. A protein complex network of drosophila melanogaster. Cell. 2011;147(3):690–703.
32. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes H-W. Corum: the comprehensive resource of mammalian protein complexes-2009. Nucleic Acids Res. 2010;38(suppl 1):497–501.
33. Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R, Jensen LJ. Compartments: unification and visualization of protein subcellular localization evidence. Database. 2014;2014:bau012.
34. Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer EL. Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 2010;38(suppl 1):196–203.
35. Wang W, Meng X, Xiang J, Shuai Y, Bedru HD, Li M. Caco: a core-attachment method with cross-species functional ortholog information to detect human protein complexes. IEEE J Biomed Health Inform. 2023;27:4569–78.
36. Cosentino S, Sriswasdi S, Iwasaki W. Sonicparanoid2: fast, accurate, and comprehensive orthology inference with machine learning and language models. Genome Biol. 2024;25(1):195.
37. Laurent JM, Garge RK, Teufel AI, Wilke CO, Kachroo AH, Marcotte EM. Humanization of yeast genes with multiple human orthologs reveals functional divergence between paralogs. PLoS Biol. 2020;18(5):3000627.
38. Li M, Zhang H, Wang J-X, Pan Y. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. BMC Syst Biol. 2012;6(1):1–9.
39. Lei X, Ding Y, Fujita H, Zhang A. Identification of dynamic protein complexes based on fruit fly optimization algorithm. Knowl Based Syst. 2016;105:270–7.
40. Lu P, Yu J. Two new methods for identifying essential proteins based on the protein complexes and topological properties. IEEE Access. 2020;8:9578–86.
41. Lei X, Zhang Y, Cheng S, Wu F-X, Pedrycz W. Topology potential based seed-growth method to identify protein complexes on dynamic ppi data. Inf Sci. 2018;425:140–53.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.