A Node Ranking Method Based on Multiple Layers for Dynamic Protein Interaction Networks

BIN LI, LI PAN, JING SUN, HAOYUE WANG, JUNQIANG JIANG, (Member, IEEE), BO YANG, AND WENBIN LI
Department of Information Science and Engineering, Hunan Institute of Science and Technology, Yueyang 414006, China
Corresponding authors: Li Pan (lipan@hnist.edu.cn) and Wenbin Li (wenbin_lii@163.com)
This work was supported in part by the Scientific Research Fund of Hunan Provincial Education Department of China under Grant 18A312 and Grant 19B231, in part by the Hunan Provincial Postgraduate Research and Innovation Foundation Project of China under Grant CX20211183, and in part by the HNIST Postgraduate Research and Innovation Foundation Project of China.

ABSTRACT Constructing dynamic protein interaction networks (DPIN) is a common way to improve identification accuracy of essential proteins. The existing methods usually aggregate DPIN into a single-layer network where all nodes are sorted by their importance. This treatment makes the dynamic information about proteins in multiple layers lost in the single layer, and thus affects the identification accuracy of essential proteins. This paper proposes a node ranking method based on multiple layers for DPIN to address the problem. First, we calculate the centrality values of all nodes for each time-specific layer, then work out the centrality score of each node by dividing the total of its centrality values across all layers by its layer activity, and finally sort the importance of all nodes by their centrality scores. Different from the methods based on single layer, our method makes full use of centrality values of each protein in time-specific layers, and thus can more effectively utilize the dynamic information of proteins. To evaluate the effectiveness of the node ranking method based on multiple layers, we apply ten network-based centrality methods on multiple layers and compare the results with those on a single layer. Then the predictive performance of the ten centrality methods are validated in terms of sensitivity, specificity, positive predictive value, negative predictive value, F-measure and accuracy. The experimental results for the identification of essential proteins show that the node ranking method based on multiple layers is superior to those based on a single layer and can help to identify essential proteins more accurate.

INDEX TERMS Essential proteins, dynamic protein interaction networks, multiple layers, centrality methods, node ranking.

I. INTRODUCTION
Proteins are divided into essential proteins and non-essential proteins [1], [2]. A protein is said to be essential if its knock-out results in lethality or infertility for an organism. Studies have shown that essential proteins are also related with human disease genes [3]. Therefore, the identification of essential proteins is of great significance for discovering drug targets, developing new drugs, and promoting the development of the biomedical industry. A lot of experimental technologies have been proposed to identify essential proteins, such as gene knockouts [4], RNA interference [5] and conditional knockouts [6]. However, these experiment methods are time-consuming and expensive.

The rapid development of modern high-throughput technologies has accumulated large quantities of omics data [7], [8], [9], which provide new opportunities for predicting essential proteins, protein complexes and functional module from large molecular networks. A series of computational methods have been developed to predict essential proteins based on the properties of protein interaction networks (PIN) [10], [11], [12]. Jeong et al. [10] proposed a centrality-lethality rule and found that the most highly connected proteins in the cell are the most important for...
its survival, and demonstrated that there is a high correlation between the topological characteristics of nodes in PIN and their biological importance. With the further research on computational methods based on topological networks, a large number of centrality methods have proposed to predict essential proteins, such as betweenness centrality (BC) [11], local average connection centrality (LAC) [12] and closeness centrality (CC) [13].

Generally, a static protein interaction network (SPIN) is constructed by using the protein-protein interactions obtained from different experiments or databases. Unfortunately, most of the network-based methods are sensitive to the reliability of the constructed PIN [14]. To improve the reliability of PINs, some researchers tried to construct dynamic protein interaction networks (DPIN) by integrating a variety of biological information into the SPIN [15], [16], [17], [18], [19], [20]. Gene expression data can be used to analyze the activity of genes over time, thus the combination of protein interaction data and gene expression data has become a common method for constructing DPIN [21], [22], [23], [24], [25], [26], [27]. Xiao et al. [22] constructed an active PIN based on the SPIN and gene expression data for the identification of essential proteins. Li et al. [24] integrated gene expression data, subcellular localization and the SPIN to identify essential proteins. Zhang et al. [25] refined the DPIN by using gene expression information and subcellular localization information for predicting essential proteins. Moreover, DPINs have been used for the identification of protein complexes and functional modules [26], [27].

A common way of ranking the importance of protein nodes in a DPIN is to aggregate multiple layers of the DPIN into a single-layer network and calculate centrality values of the nodes in the single-layer network [18], [19], [21], [24], [25]. Thus the dynamic information about nodes (or edges) in different layers is lost in the single layer when calculating centrality values. This makes the aggregated single-layer network has less time-specific information than the DPIN, and hence affects the estimation accuracy of centrality methods.

In this paper, we propose a node importance ranking method based on multiple layers for the DPIN to address this problem. We use a centrality method to calculate the centrality value of each node for each layer of the DPIN. Since different layers have different network topologies, a node often has different centrality values for different layers. Second, we compute the activity of layer for each node, which is the number of layers where the node is active. Third, the centrality score of each node is computed by dividing the total of its centrality values across all layers by its layer activity. Finally, the importance of all nodes is sorted by their centrality scores, and the nodes with high placement are regarded as essential. Different from the existing node ranking methods, our method directly calculates the centrality values of nodes based on multiple layers rather than a single layer, which ensures that the time-specific biological knowledge inherent in each layer is used effectively when calculating centrality scores.

To verify the effectiveness of the node ranking method based on multiple layers, we apply ten network-based centrality methods (DC [10], BC [11], LAC [12], CC [13], LC [28], CR [29], LID [30], TP [31], ClusterC [32] and MNC [33]) on multiple layers of the DPIN (M-DPIN) and compare the results with those on a single layer (i.e. the SPIN and the aggregated single layer of the DPIN (A-DPIN)). The predictive performance of the ten centrality methods are evaluated in terms of prediction accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F-measure and accuracy [30], [31], [32], [33], [34]. The experimental results for the identification of essential proteins show that the node ranking method based on multiple layers is superior to those based on a single layer. It has been illustrated that the proposed method can help to identify essential proteins more accurately.

II. METHODS

A. DYNAMIC PROTEIN INTERACTION NETWORKS

A static protein interaction network (SPIN) is an undirected graph $G_S = (V_S, E_S)$, where

1) $V_S$ represents the set of all proteins;
2) $E_S$ represents the set of all interactions between proteins.

Let $T = \{t_1, t_2, \ldots, t_m\}$ be $m$ observation time points, and let $g_{i}^{h}$ denote the gene expression value of protein $v_i$ at time $t_h$. The activity threshold $\tau_i$ of protein $v_i$ is used to judge the activity of a protein according to its gene expression value curve, and is defined in [17]:

$$\tau_i = \mu_i + k \times \frac{\sigma_i^2}{1 + \sigma_i^2}$$  \hspace{1cm} (1)

where $\mu_i$ and $\sigma_i$ respectively are the mean and standard deviation of gene expression values of protein $v_i$, and the parameter $k$ is not constant, which can be adjusted in the range of $(0, 3)$.

The activity $a_i^h$ of protein $v_i$ in the time $t_h$ can be computed by

$$a_i^h = \begin{cases} 1 & g_i^h \geq \tau_i \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (2)

where $a_i^h = 1$ (i.e. the protein $v_i$ is active at time $t_h$) if its gene expression value $g_i^h$ is not less than the activity threshold $\tau_i$; otherwise 0.

Dynamic interaction protein networks (DPIN) is a set of subworknets $G = \{G_1, G_2, \ldots, G_l, \ldots, G_m\}$ on $T$, and $G_l = (V_l, E_l)$ is a subnetwork observed at time $t_l$, where

1) $V_l = V_S$;
2) $E_l = \{(v_i, v_j) \in E_S | a_i^h = a_j^h = 1\}$.

For any edge $(v_i, v_j)$ in the SPIN, if there is a time point $t_h \in T$, such that both $v_i$ and $v_j$ are active at $t_h$, then the edge $(v_i, v_j)$ is in the subnetwork $G_h$.
Let $G_A = (V_A, E_A)$ be the aggregated single-layer network of the DPIN (A-DPIN), where

1) $V_A = V_S$;
2) $E_A = \{(v_i, v_j) \in E_S | \exists t_h \in T, (v_i, v_j) \in E_h\}$.

For any edge $(v_i, v_j)$ in the SPIN, if there is a time point $t_h \in T$, such that $(v_i, v_j) \in E_h$, then the edge $(v_i, v_j)$ is retained in A-DPIN.

**B. NODE RANKING**

The layer activity of a node $v_i$, denoted by $l_i$, represents the number of subnetworks where the node $v_i$ is active, and is computed by

$$l_i = \sum_{h=1}^{m} a_i^h$$  \hspace{1cm} (3)

Let $c_i^h$ be the centrality value of $v_i$ in subnetwork $G_h$, which can be calculated by using a centrality method. Then the centrality score $s_i$ of $v_i$ in the DPIN is defined by

$$s_i = \sum_{h=1}^{m} a_i^h \times c_i^h$$  \hspace{1cm} (4)

where $s_i$ is the quotient of the sum of centrality values of node $v_i$ in all active subnetworks divided by its layer activity.

The steps for ranking nodes based on multiple layers of DPIN (M-DPIN) are as follows:

1) Construct static protein interaction network $G_S$ according to protein-protein interaction data;  
2) Compute the activity threshold of each node and determining its activity at each time point according to gene expression data;  
3) Accord to 1) and 2), deriving the subnetworks $G_1$, $G_2$, ..., $G_m$ from $G_S$;  
4) Calculate the centrality value of each node in each subnetwork by using a centrality method;  
5) Calculate the layer activity of each node by formula (3), and its centrality score for the M-DPIN by formula (4);  
6) Sort centrality scores of all nodes in descending order, then taking the top 100 - top 600 ranked nodes as essential proteins, and finally calculating the identification accuracy.

As shown in Table 1, we calculate the degree centrality values (for the A-DPIN) and degree centrality scores (for the M-DPIN) of proteins YGL004C, YDR335W, YER179W, YPR181C, YJR093C and YKR002W. These protein nodes have the same degree in the A-DPIN, but in fact, the degrees of these nodes in different layers (subnetworks) usually are different. Therefore, the centrality values in the A-DPIN cannot effectively reflect the variation existing in different layers that is caused by the dynamics of gene expression. Nevertheless, in the M-DPIN, the degree centrality scores of these nodes tend to be distinguishable. This is because they are obtained by averaging the centrality values in different layers where they are active.

**III. RESULTS**

**A. EXPERIMENTAL DATA**

All the experiments in this study are based on the protein-protein interaction data of S. cerevisiae, which is now the...
most complete data in all species and has widely been used in the validation of essential protein discovery methods. The protein-protein interaction data used in the experiment was downloaded from the database DIP [35]. Up to now, the known essential proteins are mainly from following databases [36]: DEG, MIPS, SGD, SGDP, containing 1,130 known essential proteins in DIP database. The gene expression data was downloaded from GSE3431 [37], with a total of 6,667 gene expression level data. Each data consists of expression values at 36 observation time points.

B. NETWORK-BASED CENTRALITY METHODS
Up to now, there have been many network-based methods to predict essential proteins. We select ten classic centrality methods and then use them to calculate the centrality values of nodes in all single-layer networks (including the SPIN, the aggregated single-layer network of the DPIN and
TABLE 2. Comparison of the number of essential proteins identification in SPIN, A-DPIN and M-DPIN.

| Centrality | SPIN | A-DPIN | M-DPIN | Improvement at Top 600 compared with |
|------------|------|--------|--------|-----------------------------------|
| Top100-Top600 | Top100-Top600 | Top100-Top600 | SPIN | A-DPIN |
| BC | 48 83 123 164 207 247 | 46 90 132 173 222 262 | 57 112 159 205 250 286 | 15.79% | 9.16% |
| ClusterC | 27 77 121 172 195 250 | 32 91 144 186 222 287 | 61 117 169 212 255 310 | 24.00% | 8.01% |
| LID | 82 142 199 251 303 347 | 79 142 208 257 308 350 | 87 162 226 279 334 375 | 8.07% | 7.14% |
| CC | 49 90 134 178 225 263 | 45 96 135 180 214 255 | 71 133 186 225 265 306 | 16.35% | 20.00% |
| CR | 48 88 126 164 205 248 | 49 91 129 174 210 249 | 57 111 162 208 255 301 | 21.37% | 20.88% |
| DC | 57 103 152 203 256 298 | 56 115 169 217 263 310 | 74 131 194 248 295 341 | 14.43% | 10.00% |
| LC | 49 84 122 163 205 246 | 46 86 132 171 221 256 | 58 111 153 204 249 283 | 15.04% | 10.55% |
| LAC | 82 143 195 251 300 346 | 81 145 204 257 313 359 | 91 159 222 270 324 369 | 6.65% | 2.79% |
| TP | 55 106 150 194 236 282 | 54 111 166 215 259 307 | 75 137 193 247 295 343 | 21.63% | 11.73% |
| MNC | 73 138 194 258 307 344 | 75 140 208 268 317 348 | 79 162 224 287 339 373 | 8.43% | 7.18% |

All subnetworks of the DPIN). The centrality values of nodes from all subnetworks will be further processed to obtain the centrality scores of nodes by using the proposed method. The selected ten centrality methods based on network separately are degree centrality (DC) [10], betweenness centrality (BC) [11], close centrality (CC) [13], load centrality (LC) [28], local average connection centrality (LAC) [12], cycles ration (CR) [29], local interaction density centrality (LID) [30], topological centrality (TP) [31], cluster coefficient centrality (ClusterC) [32] and maximum neighbor component centrality (MNC) [33], which are defined respectively as follows:

\[ DC(i) = \text{deg}(i) \] (5)

\[ BC(i) = \sum_{k \in V, k \neq i} \frac{\rho(k, i, j)}{\rho(k, j)}, \quad k \neq i \neq j \] (6)

\[ CC(i) = \frac{n - 1}{\sum_{i \neq j} \rho(i, j)} \] (7)

\[ LC(i) = \sum_{s, d \in V} \theta_{s, d}(i) \] (8)

\[ LAC(i) = \frac{\sum_{i \in S} \text{deg}_{C_i}(j)}{\text{deg}(i)} \] (9)

\[ CR(i) = \begin{cases} 0, & c_{ij} = 0 \\ \sum_{j, c_{ij} > 0} c_{ij}, & c_{ij} > 0 \end{cases} \] (10)

\[ LID(i) = \frac{|E_{ab, int}(i)|}{|V_{nb, int}(i)|} \] (11)

\[ TP(i) = \sum_{j=1}^{N} \left( m_j \times e^{-\frac{(\rho(i, j))^2}{\sigma^2}} \right) \] (12)

\[ ClusterC(i) = \frac{E(i)}{\text{deg}(i) \times (\text{deg}(i) - 1)/2} \] (13)

\[ MNC(i) = |\text{maximum connected component of } v_i| \] (14)

Among them, \( \text{deg}(i) \) is the degree of vertex \( v_i \). Where \( \rho(k, i, j) \) is the number of shortest paths from \( v_k \) through \( v_i \) to \( v_j \). Let \( \rho(k, j) \) be the number of shortest paths from \( v_k \) to \( v_j \). \( \theta_{s, d} \) is a quantity of a commodity that is sent from \( v_s \) to \( v_d \). \( \theta_{s, d}(i) \) is the amount of commodity forwarded by \( v_i \). Where \( S_i \) denotes the neighbor vertex set of \( v_i \). Let \( Z_{ij} = |S_i \cap S_j| \) be the number of triangles containing the edge \((v_i, v_j)\), that is the intersection of the neighbor set of \( v_i \) and the neighbor set of \( v_j \) the number of elements, and \( C_i \) be the derived subgraph \( G[S_i] \) of \( G \) with respect to \( S_i \), \( \text{deg}^C(i) \) is the degree of \( v_i \) in \( C_i \). Where \( S \_C \) is the set of the shortest cycles associated with \( v_i \), and \( S \_C = \bigcup_{i \in V} S \_C \) is the set of all shortest cycles of \( G \). Where \( C_{ij} \) denotes the number of cycles in \( S \_C \) that contain the vertex \( v_{ij} \). \( C_{ij} \) denotes the number of cycles in \( S \_C \) that pass through both \( v_i \) and \( v_j \). \( |E_{NB, INT}(i)| \) is the composition of the neighbor vertices of \( v_i \) the number of edges, \( |V_{NB, INT}(i)| \)
is the number of non-isolated vertices among the neighbor vertices of \( v_i \). Where \( m_j = 1 \), and \( \sigma \) is the constant value 0.9428. Where \( E(\cdot) \) is the number of neighbors of \( v_i \). Where \( |\cdot| \) is an operator to count the number of elements in a set.

C. PREDICTION OF ESSENTIAL PROTEINS FOR SPIN, A-DPIN AND M-DPIN

We rank the importance of nodes for the SPIN and the aggregated single layer of the DPIN (A-DPIN) by using
ten known centrality methods and for multiple layers of DPIN (M-DPIN) by using the proposed node ranking method. Then we calculate and compare the number of essential proteins at the Top 100, Top 200, Top 300, Top 400, Top 500 and Top 600. The experimental results on the DIP database are shown in Figure 1 and Table 2.

It can be seen from Figure 1 that compared with the SPIN and A-DPIN, the M-DPIN can improve the number of essential proteins identification of these centrality methods, especially for the path-based and degree-based centrality methods (such as BC, CC and DC). This indicates that the node ranking method based on multiple layers is superior to the existing methods based on the single layer.

**TABLE 3.** The comparison of SN, SP, F, PPV, NPV and ACC of ten centrality methods applied on SPIN, A-DPIN and M-DPIN.

| Network | Centrality | SN  | SP  | PPV | NPV  | F   | ACC  |
|---------|------------|-----|-----|-----|------|-----|------|
| SPIN    | BC         | 0.389 | 0.809 | 0.389 | 0.809 | 0.525 | 0.709 |
|         | ClusterC   | 0.422 | 0.819 | 0.422 | 0.819 | 0.557 | 0.725 |
|         | LID        | 0.467 | 0.834 | 0.467 | 0.834 | 0.599 | 0.746 |
|         | CC         | 0.386 | 0.808 | 0.386 | 0.808 | 0.522 | 0.708 |
|         | CR         | 0.368 | 0.803 | 0.368 | 0.803 | 0.505 | 0.699 |
|         | DC         | 0.441 | 0.825 | 0.441 | 0.825 | 0.575 | 0.734 |
|         | LC         | 0.386 | 0.808 | 0.386 | 0.808 | 0.522 | 0.708 |
|         | LAC        | 0.474 | 0.836 | 0.474 | 0.836 | 0.604 | 0.749 |
|         | TP         | 0.416 | 0.818 | 0.416 | 0.818 | 0.551 | 0.722 |
|         | MNC        | 0.463 | 0.832 | 0.463 | 0.832 | 0.595 | 0.744 |

|         | BC         | 0.367 | 0.802 | 0.367 | 0.802 | 0.504 | 0.699 |
|         | ClusterC   | 0.458 | 0.831 | 0.458 | 0.831 | 0.590 | 0.742 |
|         | LID        | 0.467 | 0.834 | 0.467 | 0.834 | 0.599 | 0.746 |
|         | CC         | 0.355 | 0.798 | 0.355 | 0.798 | 0.491 | 0.693 |
|         | CR         | 0.383 | 0.807 | 0.383 | 0.807 | 0.520 | 0.706 |
|         | DC         | 0.451 | 0.829 | 0.451 | 0.829 | 0.584 | 0.739 |
|         | LC         | 0.367 | 0.802 | 0.367 | 0.802 | 0.504 | 0.699 |
|         | LAC        | 0.470 | 0.834 | 0.470 | 0.834 | 0.601 | 0.748 |
|         | TP         | 0.420 | 0.819 | 0.420 | 0.819 | 0.555 | 0.724 |
|         | MNC        | 0.467 | 0.834 | 0.467 | 0.834 | 0.599 | 0.746 |

|         | BC         | 0.411 | 0.816 | 0.411 | 0.816 | 0.546 | 0.719 |
|         | ClusterC   | 0.466 | 0.833 | 0.466 | 0.833 | 0.598 | 0.746 |
|         | LID        | 0.480 | 0.837 | 0.480 | 0.837 | 0.610 | 0.752 |
|         | CC         | 0.440 | 0.825 | 0.440 | 0.825 | 0.574 | 0.733 |
|         | CR         | 0.445 | 0.827 | 0.445 | 0.827 | 0.579 | 0.736 |
|         | DC         | 0.479 | 0.837 | 0.479 | 0.837 | 0.609 | 0.752 |
|         | LC         | 0.408 | 0.815 | 0.408 | 0.815 | 0.544 | 0.718 |
|         | LAC        | 0.477 | 0.837 | 0.477 | 0.837 | 0.608 | 0.751 |
|         | TP         | 0.477 | 0.837 | 0.477 | 0.837 | 0.608 | 0.751 |
|         | MNC        | 0.485 | 0.839 | 0.485 | 0.839 | 0.615 | 0.755 |
As shown in Table 2, among the ten centrality methods, the predictive performance of BC, ClusterC, CC, CR, DC, LC and TP at Top 600 in the M-DPIN increased over 10%, compared with those in the SPIN and A-DPIN. Taking the typical centrality measure DC as an example, its application on M-DPIN are added 43 and 31 essential proteins respectively, compared to the case applied on SPIN and A-DPIN when predicting the Top 600 essential candidates. These indicate that node ranking method can improve the path-based and degree-based centrality methods better.

D. VALIDATED BY USING THE JACKKNIFING METHODOLOGY

To evaluate the effectiveness of M-DPIN more generally, we further use the jackknifing methodology. It can be seen from Figure 2 that as the number of essential candidates...
The comparison of the number of essential proteins identified by ten centrality methods in SPIN, A-DPIN and M-DPIN under the database BioGRID.

Increases, we can see that the essential proteins’ recognition rate of the six centrality methods of BC, CC, CR, DC, LC and TP are gradually widening the gap with SPIN and A-DPIN. And compared with the SPIN and A-DPIN, the M-DPIN can improve the number of essential proteins identification of these centrality methods, especially for the path-based and degree-based centrality methods (such as BC, CC and DC). These indicate that the proposed method based on multiple layers is superior to the existing methods based on the single layer.

E. Validated by Six Evaluation Methods

We use the sensitivity (SN), specificity (SP), F-measure (F), positive predictive value (PPV), negative predictive value (NPV), and accuracy rate (ACC) for the validation of essential protein discovery methods in the SPIN, A-DPIN and M-DPIN, respectively.

Let TP denote the number of true positives, TN denote the number of true negatives. FP is the number of false positives and FN is the number of false negatives. In this study, true positives means essential proteins correctly predicted as
essential. True negatives means non-essential proteins correctly predicted as non-essential. False positives represents non-essential proteins incorrectly predicted as essential. False negatives denotes essential proteins incorrectly predicted as non-essential.

The SN, SP, F, PPV, NPV, and ACC are separately defined as follows:

\[
SN = \frac{TP}{TP + FN} \quad (15)
\]
\[
SP = \frac{TN + FP}{2 \times SN \times SP} \quad (16)
\]
\[
F = \frac{SN + SP}{2} \quad (17)
\]
\[
PPV = \frac{TP}{TP + FP} \quad (18)
\]
\[
NPV = \frac{TN + FN}{TN + FP} \quad (19)
\]
\[
ACC = \frac{TP + TN + FN + FP}{TP} \quad (20)
\]

In the DIP, there are 4,746 proteins, including 1,130 essential proteins and 3,616 non-essential proteins. So we let TP denotes the number of essential proteins in the top 1,130 proteins. FN is the number of essential proteins in the bottom 3,616 proteins when the 4,746 proteins are ranked in descending order by centralities. Therefore, we can get that TN is the number of non-essential proteins in the bottom 3,616 proteins, and FP is the number of non-essential proteins in the top of 1,130 proteins. At this time, the SN and PPV values of each centrality method are the same, and the SP and NPV values are the same. The six evaluated methods’ value of the ten methods applied on SPIN, A-DPIN and M-DPIN were calculated as shown in Table 3.

It can be seen from Table 3, we can see that the sensitivity, specificity, F-measure, positive predictive value, negative predictive value, and accuracy rate of the ten methods applied on our method based on multi-layer protein interaction networks are consistently higher than those on SPIN and A-DPIN, which indicates that the node ranking method can improve the identification accuracy of these centrality methods better.

F. ANALYSIS OF THE PARAMETER K

In order to analyze the effect of parameter \(k\) on the A-DPIN and M-DPIN, we set the range of parameter \(k\) to \([0, 3]\), and test the number of essential proteins at the Top 1,130 (for database DIP) to select the best value of \(k\) for most of centrality methods. The experimental results are shown in Figure 3.

In Figure 3, the green line is always straight, because the SPIN is independent of \(k\). So we use it only as a reference. It is easy to find that in most cases, the curve of the number of identified essential proteins for the M-DPIN is higher than those for the SPIN and A-DPIN. When \(k = 2\), the most of centrality methods can reach a peak for the M-DPIN, such as ClusterC, LID, CR, DC, LAC and TP. Therefore, we take \(k = 2\) as the default value of the parameter for the M-DPIN.

G. VALIDATED BY THE DATABASE BIOGRID

In order to further validate the effectiveness of our method, we conducted experiments in the BioGRID [38] database, which contains 5,616 proteins and 52,833 pairs of protein interactions, and 1,199 are known to be essential proteins. We apply ten network-based centrality methods on the M-DPIN and compare the results with these of the same methods applied on the static network (SPIN) and the aggregated single-layer network (A-DPIN). The number of essential proteins identified in SPIN, A-DPIN and M-DPIN is counted as the Top 1%, Top 5%, Top 10%, Top 15%, Top 20% and Top 25%, as shown in Figure 4.

It can be seen from Figure 4, the number of essential proteins at the Top 1%, Top 5%, Top 10%, Top 15%, Top 20% and Top 25% in M-DPIN is almost higher than SPIN and A-DPIN, which indicates the effectiveness of our method based on the multiple layers for identifying essential proteins.

IV. CONCLUSION

The construction of the DPIN by integrating gene expression data into the SPIN is a common method for the improvement of the identification accuracy of essential proteins. This method usually needs to aggregate all layers of the DPIN into a single-layer network, and then calculate centrality values in the single-layer network by using a centrality method. Since the different states in different layers of the DPIN are aggregated into a state in the single-layer network, and thus the dynamic information about the node (or the edge) in different layers is lost in the single layer. This makes the aggregated single-layer network (A-DPIN) has less accurate than the M-DPIN, and hence affects the estimation accuracy of centrality methods.

In this paper, we propose a node ranking method based on multiple layers for the DPIN to address this problem. Different from the existing node ranking methods, our method calculates the centrality score of each node by averaging its centrality values in all active layers, which ensures that the time-specific biological knowledge inherent in each layer is used effectively when calculating centrality scores. To evaluate that the node ranking method based on multiple layers for the DPIN is more suitable to be used in the identification of essential proteins, we apply ten network-based essential protein discovery methods on multiple layers of the DPIN and compare the results with those on a single layer (i.e. the SPIN and the aggregated single layer of the DPIN). The experimental results for the identification of essential proteins show that the node ranking method based on multiple layers is superior to those based on a single layer. This shows our method can more effectively utilize the dynamic information of proteins than those methods based on single layer.

As future work, it would be interesting to evaluate the importance of a node in the DPIN by utilizing dynamic information of each time-specific layer more effectively.
Acknowledgment
(Bin Li and Li Pan contributed equally to this work.)

References

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

[2] X. Lu, J. Dong, J. C. Rhodes, H. Lu, and L. J. Lu, “Predicting essential genes for identifying potential drug targets in Aspergillus fumigatus,” Comput. Biol. Chem., vol. 50, pp. 29–40, Jun. 2014.

[3] S. J. Furney, M. M. Albá, and N. López-Bigas, “Differences in the evolutionary history of disease genes affected by dominant or recessive mutations,” BMC Genomics, vol. 7, no. 1, p. 165, Dec. 2006.

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

[5] L. M. Cullen and G. M. Arnett, “Genome-wide screening for gene function using RNAi in mammalian cells,” Immunity, vol. 83, no. 3, pp. 217–223, Jun. 2005.

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

[7] S. Fields and O.-K. Song, “A novel genetic system to detect protein–protein interactions,” Nature, vol. 340, no. 6230, pp. 245–246, Jul. 1989.

[8] H. Zhu, M. Bilgini, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Billingmaier, T. Houleik, T. Mitchell, P. Miller, R. A. Dean, M. Gerstein, and M. Snyder, “Global analysis of protein activities using proteome chips,” Science, vol. 293, no. 5537, pp. 2101–2105, Sep. 2001.

[9] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, “A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae,” Nature, vol. 403, no. 6770, pp. 623–627, 2000.

[10] M. Li, P. Ni, X. Chen, J. Wang, F. Wu, and Y. Pan, “Construction of refined protein interaction networks and RNA-seq datasets,” BMC Genomics, vol. 16, no. 4, pp. 1386–1397, Jul/Aug. 2015.

[11] Z. Zhang, J. Ruan, J. Gao, and F.-X. Wu, “Predicting essential proteins from protein-protein interactions using order statistics,” J. Theor. Biol., vol. 480, pp. 274–283, Nov. 2019.

[12] X. Tang, J. Wang, B. Liu, M. Li, G. Chen, and Y. Pan, “A comparison of the functional modules identified from time course and static PPI network data,” BMC Bioinf., vol. 12, no. 1, p. 339, Dec. 2011.

[13] X. Shen, L. Y, X. Jiang, T. He, X. Hu, and J. Yang, “Mining temporal protein complex based on the dynamic PIN weighted with connected affinity and gene co-expression,” PLoS ONE, vol. 11, no. 4, Apr. 2016, Art. no. e0153967.

[14] L. Maccari, L. Ghiora, A. Guerrieri, A. Montresor, and R. L. Cigno, “Exact distributed load centrality computation: Algorithms, convergence, and applications to distance vector routing,” IEEE/ACM Trans. Parallel Distrib. Syst., vol. 31, no. 7, pp. 1693–1706, Jul. 2020.

[15] T. Fan, L. Lü, D. Shi, and T. Zhou, “Characterizing cycle structure in complex networks,” Commun. Phys., vol. 4, no. 1, p. 272, Dec. 2021.

[16] Y. Qi and J. Luo, “Prediction of essential proteins based on local interaction density,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 13, no. 6, pp. 1170–1182, Nov. 2016.

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

[18] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” Nature, vol. 393, no. 6684, pp. 440–442, 1998.

[19] C.-Y. Lin, C.-H. Chin, H.-H. Wu, S.-H. Chen, C.-W. Ho, and M.-T. Ko, “Hubba: Hub objects analyzer—a framework of interactome hubs identification for network biology,” Nucleic Acids Res., vol. 36, no. W438–W443, Jul. 2008.

[20] Y. Tang, M. Li, J. Wang, Y. Pan, and F.-X. Wu, “CytoNCA: A cytoscape plugin for centrality analysis and evaluation of protein interaction networks,” Biosystems, vol. 127, pp. 67–72, Jan. 2015.

[21] X. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, “DIP: The Database of interacting proteins,” Nucleic Acids Res., vol. 28, no. 1, pp. 289–291, 2000.

[22] M. Li, J.-T. Hsiang, C.-Y. Wu, Y.-J. Oyang, H.-F. Juan, and A.-L. Barabási, “Identifying essential proteins based on sub-network partition and prioritization by integrating subcellular localization information,” J. Theor. Biol., vol. 447, pp. 65–73, Jun. 2018.

[23] W. Zhang, J. Xu, and X. Zou, “Predicting essential proteins by integrating network topology, subcellular localization information, gene expression profile and GO annotation data,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 17, no. 6, pp. 2053–2061, Nov./Dec. 2020.
LI PAN received the Ph.D. degree in computer science from Tongji University, China, in 2009. He is currently a Professor with the School of Information Science and Engineering, Hunan Institute of Science and Technology, Yueyang, Hunan, China. His research interests include bioinformatics and Petri nets.

JING SUN is currently pursuing the Graduate degree with the Hunan Institute of Science and Technology. Her research interests include bioinformatics and complex networks.

HAOYUE WANG is currently pursuing the Graduate degree with the Hunan Institute of Science and Technology. His research interests include bioinformatics and complex networks.

JUNQIANG JIANG (Member, IEEE) received the Ph.D. degree in software engineering from Hunan University, Changsha, China, in 2017. He is currently an Associate Professor with the School of Information Science and Engineering, Hunan Institute of Science and Technology, China. His research interests include cloud computing, parallel computing, and workflow scheduling. He has been awarded the Young Elite Teacher of Hunan province general universities. His research has been supported by the Natural Science Foundation of Hunan Province and the Scientific Research Fund of the Hunan Provincial Education Department.

BO YANG received the B.Sc. degree in mechanical engineering from Zhengzhou University, China, in 1996, the M.Sc. degree in computer applications technology from Xiangtan University, China, in 2004, and the Ph.D. degree in mechanical and electrical engineering from Central South University, China, in 2010. Since 2012, he has been an Associate Professor with the College of Information Science and Technology, Hunan Institute of Science and Technology. His research interests include MR brain image analysis, statistical pattern recognition, and machine learning.

WENBIN LI received the B.S. degree in computer science and technology from Hunan Normal University, Changsha, China, in 2003, the M.S. degree in computer applications technology from the Changsha University of Science and Technology, Changsha, in 2006, and the Ph.D. degree in control engineering from Central South University, Changsha, in 2020. His research interests include industrial process control, evolutionary computation, and multi-objective optimization.