cells

MDPI

Article

Rule-Based Pruning and In Silico Identification of Essential Proteins in Yeast PPIN

Anik Banik [1], Souvik Podder [1], Sovan Saha [2], Piyali Chatterjee [3], Anup Kumar Halder [4,5], Mita Nasipuri [6], Subhadip Basu [6,*] and Dariusz Plewczynski [4,5,*]

1 Department of Computer Science & Engineering, Dr. Sudhir Chandra Sur Degree Engineering College, 540, Dum Dum Road, Near Dum Dum Jn. Station, Surermath, Kolkata 700074, India
2 Department of Computer Science & Engineering, Institute of Engineering & Management, Salt Lake Electronics Complex, Kolkata 700091, India
3 Department of Computer Science & Engineering, Netaji Subhash Engineering College, Techno City, Panchpota, Garia, Kolkata 700152, India
4 Faculty of Mathematics and Information Sciences, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland
5 Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Banacha 2c Street, 02-097 Warsaw, Poland
6 Department of Computer Science & Engineering, Jadavpur University, 188, Raja S.C. Mallick Road, Kolkata 700032, India
* Correspondence: subhadip.basu@jadavpuruniversity.in (S.B.); d.plewczynski@cent.uw.edu.pl (D.P.)

Abstract: Proteins are vital for the significant cellular activities of living organisms. However, not all of them are essential. Identifying essential proteins through different biological experiments is relatively more laborious and time-consuming than the computational approaches used in recent times. However, practical implementation of conventional scientific methods sometimes becomes challenging due to poor performance impact in specific scenarios. Thus, more developed and efficient computational prediction models are required for essential protein identification. An effective methodology is proposed in this research, capable of predicting essential proteins in a refined yeast protein–protein interaction network (PPIN). The rule-based refinement is done using protein complex and local interaction density information derived from the neighborhood properties of proteins in the network. Identification and pruning of non-essential proteins are equally crucial here. In the initial phase, careful assessment is performed by applying node and edge weights to identify and discard the non-essential proteins from the interaction network. Three cut-off levels are considered for each node and edge weight for pruning the non-essential proteins. Once the PPIN has been filtered out, the second phase starts with two centralities-based approaches: (1) local interaction density (LID) and (2) local interaction density with protein complex (LIDC), which are successively implemented to identify the essential proteins in the yeast PPIN. Our proposed methodology achieves better performance in comparison to the existing state-of-the-art techniques.

Keywords: essential protein; edge weight; node weight; yeast PPIN; local interaction density

Citation: Banik, A.; Podder, S.; Saha, S.; Chatterjee, P.; Halder, A.K.; Nasipuri, M.; Basu, S.; Plewczynski, D. Rule-Based Pruning and In Silico Identification of Essential Proteins in Yeast PPIN. *Cells* **2022**, *11*, 2648. https://doi.org/10.3390/cells11172648

Academic Editor: Yu Xue

Received: 28 July 2022
Accepted: 22 August 2022
Published: 25 August 2022

1. Introduction

Various research areas like protein structure prediction [1,2]; protein function prediction using protein sequences [3,4], protein domains [5,6], and protein–protein interaction networks (PPIN) [7–11]; protein subcellular localization identification [12,13]; and detection of essential proteins [14–16] have significantly been exploited due to the increase in the availability of a large number of proteins/protein sequences in the post-genomic era. In general, essential proteins are the highly connected modules in a PPIN [17]. So, removing any essential protein from the existing network would be fatal, resulting in various functional disorders of living organisms. Most of the research works [18–20] note the fact that deeper analyses of essential proteins in a PPIN will lead to better assimilation of

ideas about the mutation of genes, which is usually considered as the ultimate cause of disease initiation. Thus, essential protein prediction has a significant role in the medical and biological fields of study. Though computational approaches have become the recent trend for establishing the topological relationship between a PPIN and the essentiality of proteins, the previous biological methodologies [21,22] provided the base for the foundation for this research field. Being directed by the centrality–lethality rule [17], centrality measures based on the topological features of biological PPINs have become the center of attraction for most of the existing methodologies [17,23,24] for the identification of essential proteins.

According to Luo et al. [23], computational approaches to essential protein prediction can be broadly classified into two categories: (1) Topological centrality-based approaches at the PPIN level: Centrality measures derived from the topological properties of a PPIN are considered in the topological centrality-based approach. In the work of Li et al. [15], each protein in a PPIN is represented as a material particle. The author estimated the value of each of these particles' topology potential, which gave them a unique ranking. Based on these rankings, the essentiality of proteins is derived. Tang et al. [24] developed a Cytoscape [25] plugin named CytoNCA to evaluate biological PPINs through the computation of various centrality scores. Currently, it supports eight centralities for both unweighted and weighted PPINs: betweenness centrality (BC) [26], closeness centrality (CC) [27], degree centrality (DC) [17], eigenvector centrality (EC) [28], local average connectivity-based method (LAC) [29], network centrality (NC) [14], subgraph centrality (SC) [30], and information centrality (IC) [31]. (2) Heterogeneous feature-based approach: The use of topological centrality measures along with protein-specific features is usually considered a heterogeneous feature-based approach. This can be accomplished by incorporating the gene ontology (GO) terms of proteins [32], protein complexes [33,34], orthologous information [35], subcellular protein localization [36], and gene expression data [37–39] along with a PPIN. Another recent work by Dong et al. [40] considers five relevant features after reviewing several related features in this field of essential protein prediction: (1) domain information [41,42], (2) evolutionary conservation [43,44], (3) sequence components [45,46], (4) network topology [14,33], and (5) expression level [47,48] for essential protein/gene prediction. They have used a support vector machine (SVM) for the same task after splitting the yeast and human data into train and test sets.

Existing computational approaches reveal a relation between protein degree and essentiality. Nevertheless, some experimental analyses, like yeast two-hybrid (Y2H) analyses, have also created conflict, stating that this association may be too fragile for binary or transient PPINs [49,50]. Modular essentiality is highlighted in the work of Ryan et al. [51], where all the proteins in a protein complex are considered to be essential. In contrast, Wang et al. [52] established a strong foundation indicating that essential proteins do have a more significant number of protein complex interactions. They also stated that larger protein complexes are more likely to become essential than smaller ones. Various researchers [53,54] have also shown that essential proteins are usually present in the denser sub-modules of a PPIN formed by a single protein interacting with its adjacent neighbors to perform a specific biological function. Hence, the relation between protein complexes and essentiality must also be considered. In the work of Hart et al. [55], a scoring method is proposed that can yield a subset of observed matrix-model interactions having high confidence scores. Later, these sets are used to infer a yeast's most accurate mapping of protein complexes. The results generated from the proposed work of Hart et al. also established that essentiality depends on a protein complex rather than an individual protein. Ren et al. [33] introduced a centrality-based approach, ECC, which is based on SC [30] and protein complexes. Li et al. [34] also proposed a similar approach to Ren et al., known as united complex centrality (UC). An integrated system of gene expression information and some centralities such as BC [26], PeC [37], DC [17], etc. is used in the work of Zhong et al. [56] for the identification of essential proteins. Other related conventional methodologies in this field of study are range-limited centrality [57], L-index [58], coexpression weighted by clustering coefficient (CoEWC) [59], LeaderRank [60], weighted degree cen-

trality (WDC) [61], an iteration method for predicting essential proteins by integrating orthology with a PPI network (ION) [35], and normalized α-centrality [62]. Among the previously discussed methodologies of essential protein function prediction, a few important ones are highlighted in Table 1.

Table 1. Computational studies based on essential protein prediction.

Utilized Features	Description	Database	References
Subcellular localization	An efficient method to identify essential proteins for different species by integrating protein subcellular localization information.	PPIN of *Saccharomyces cerevisiae*, *Homo sapiens*, *Mus musculus* and *Drosophila melanogaster*	[36]
Protein complex, degree, subgraph	A new method for predicting essential proteins based on participation degree in protein complex and subgraph Density.	PPIN of *Saccharomyces cerevisiae*	[54]
Orthology, gene expression, PPIN	Predicting essential proteins by integrating orthology, gene expressions, and PPIN.	PPIN of *Saccharomyces cerevisiae*	[39]
CC and orthology	United neighborhood closeness centrality and orthology for predicting essential proteins.	PPIN of *Saccharomyces cerevisiae*	[63]
Node, edge clustering coefficient	Identification of essential proteins using improved node and edge clustering coefficient.	PPIN of *Saccharomyces cerevisiae* and *Drosophila melanogaster*	[22]
Centrality scores	CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks.	_	[24]
Protein complex	Identification of essential proteins based on a new combination of local interaction density and protein complexes.	PPIN of *Saccharomyces cerevisiae*	[23]
PPIN, protein complex	Prediction of essential proteins by integration of PPI network topology and protein complex information.	PPIN of *Saccharomyces cerevisiae*	[33]

Though the existing computational approaches can identify essential proteins efficiently, these methods produce more false positives. To overcome this, a new methodology for essential protein identification is proposed in this work. This method works in two phases: (1) the first phase deals with the non-essential proteins present in the PPIN using two topological features, node and edge weight [64], which ensure the presence of only the reliable nodes and edges in the PPIN—in other words, they focus only on the densely connected modules in the PPIN [7]. (2) In the next phase, local interaction density (LID) [23] and local interaction density with protein complex (LIDC) [23] are used for the identification of essential proteins in the PPIN. All the required data supporting the proposed methodology, including basic terminologies like node weight, edge weight, LID, and LIDC centralities, are given in the Supplementary Materials, available here.

In the upcoming section, the dataset of *Yeast* PPIN used for the proposed methodology will be discussed. Following that, the detailed implementation of our rule-based pruning research and the application of LID and LIDC will be highlighted, along with the pictorial representation of PPIN-related terminologies. Finally, the paper will be ended with a results and discussion section, followed by the conclusion.

2. Dataset

For the proposed work, the PPIN database of yeast, i.e., *Saccharomyces cerevisiae*, is used. It was downloaded from the DIP database [65,66] (named YDIP_5093 in the work of Luo et al. [23]), which includes 5093 proteins and 24,743 interactions. The PPIN of yeast is highlighted in Figure S1 in Supplementary Materials. Moreover, a protein complex, marked as Complex_745 [23], is also used along with LIDC [23] in the second phase of our proposed methodology. It contains about 745 protein complexes involving 2167 proteins.

This protein complex is a combination of four natural protein complex datasets: (1) CM270 is obtained from the MIPS database [67]; (2) CM425 [68] is obtained from MIPS (Mewes 2005), Aloy et al. [69], and the SGD database [70]; (3) the last two, CYC408 and CYC428, are obtained from CYC2008 of the Wodak Laboratory [71,72].

3. Methodology

This section proposes a methodology that identifies proteins as topologically more connected by applying a network-based scoring technique to the processed and rule-based pruned network. The network is pruned by removing some nodes and edges having less node weight and edge weight than the specified cut-off value. Thus, less interconnected proteins are identified based on their degree and other parameters and removed, as they are not very topologically significant. The entire working mechanism of the proposed methodology in this research work is highlighted in Algorithm 1.

The PPIN of yeast contains some topologically less important proteins, i.e., proteins having degree 0 or 1 or fewer interconnections between their neighbors than the rest of the proteins, representing their non-essentiality. Edge reliability is another factor that must be considered for identifying essential proteins. Thus, the reliability of every node and edge is investigated by calculating node and edge weights [64] in the first phase of the proposed methodology. The node weight W_v of a node $v \in V$ in PPI networks [64] is the average degree of all nodes in G'_v, a sub-graph of the network G_v. It is represented by

$$W_v = \sum_{u \in V''} deg(u)/|V''|$$

where V'' is the set of nodes in G'_v. $|V''|$ is the number of nodes in G'_v, and $deg(u)$ is the degree of a node $u \in V''$ in W_v. The edge weight W_{uv} [64] of nodes u and v is represented by

$$W_{uv} = (\Gamma(u) \cap \Gamma(v))/(\Gamma(u) \cup \Gamma(v))$$

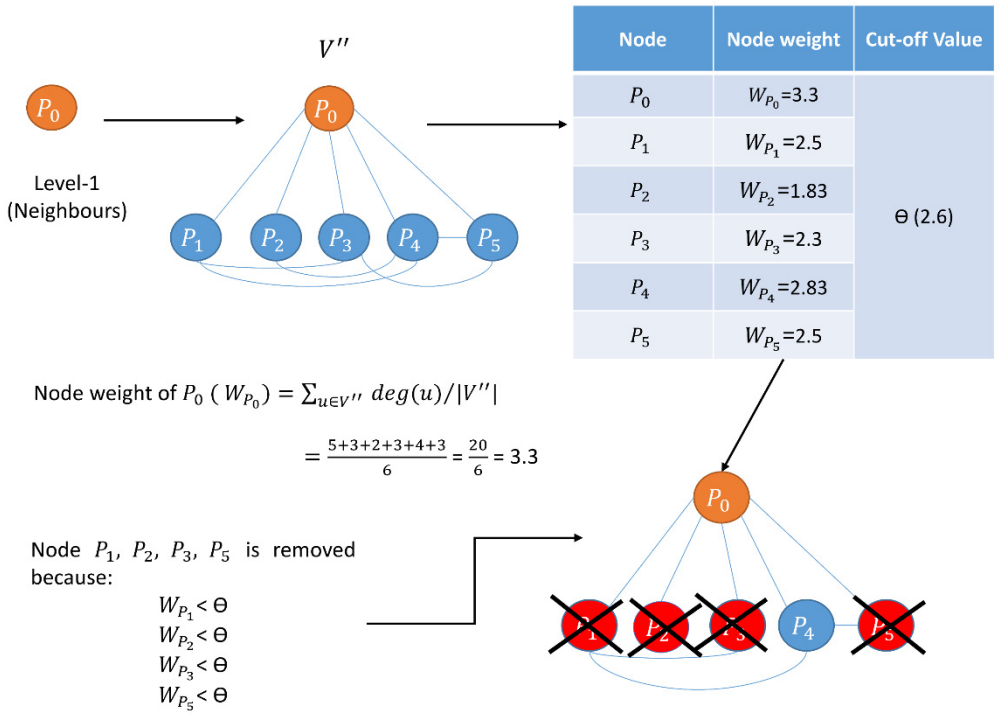
where $\Gamma(u)$ and $\Gamma(v)$ are neighbors of u and v, respectively. $\Gamma(u) \cap \Gamma(v)$ represents all common neighbors of u and v, and $\Gamma(u) \cup \Gamma(v)$ means all distinct neighbors of u and v.

Less reliable nodes and interconnections are pruned. Thus, in an interaction network, a protein's interconnectivity with other proteins and the reliability of those interactions make the pruning strategy stronger. Moreover, setting various cut-off levels for node and edge weights is integral to this phase. So, three cut-off levels, i.e., high, medium, and low [73] (see Algorithm 1), are evaluated to see the changes in the prediction accuracy level in the second phase of essential protein identification. The cut-off (θ_k) is calculated by the following mathematical equation:

$$\theta_k = \alpha + k \times \sigma \times \left(1 - \frac{1}{1+\sigma^2}\right)$$

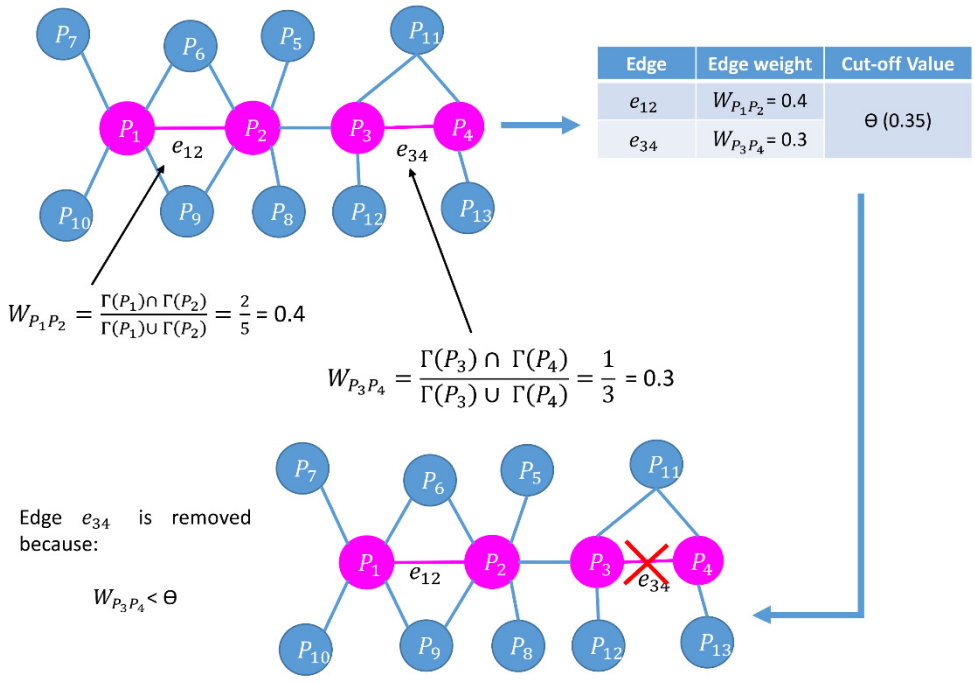
where $k \in \{1, 2, 3\}$ defines low, medium, and high cut-offs, respectively. α is determined to be the mean of the node weight/edge weight values, while σ is considered to be the standard deviation of the node weight/edge weight values.

This approach filters out a refined PPIN of yeast containing denser sub-modules [7]. Moreover, as discussed in the introduction, essential proteins tend to lie in the denser sub-modules or protein complexes of a PPIN. Thus, the first phase plays a significant role in this research. The computation of the node and edge weights of two different synthetic networks are highlighted in Figures 1 and 2, respectively.



Node	Node weight	Cut-off Value
P_0	W_{P_0}=3.3	Θ (2.6)
P_1	W_{P_1}=2.5	
P_2	W_{P_2}=1.83	
P_3	W_{P_3}=2.3	
P_4	W_{P_4}=2.83	
P_5	W_{P_5}=2.5	

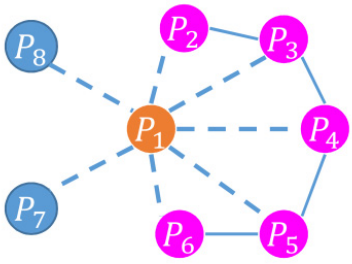
Figure 1. Schematic diagram of computation of node weight. It retains proteins having maximum connectivity. Root node (protein) is denoted by orange while its corresponding neighbors (proteins) are highlighted in blue. The filtered-out nodes (proteins) are represented in red.



Edge	Edge weight	Cut-off Value
e_{12}	$W_{P_1P_2}$= 0.4	Θ (0.35)
e_{34}	$W_{P_3P_4}$= 0.3	

Figure 2. Schematic diagram of computation of edge weight. Edge weight retains only the reliable edges in a PPIN. Edge weight has been calculated for the edges connected with the nodes (proteins) marked with pink color whereas the neighbors (proteins) and their connected edges are highlighted in blue color.

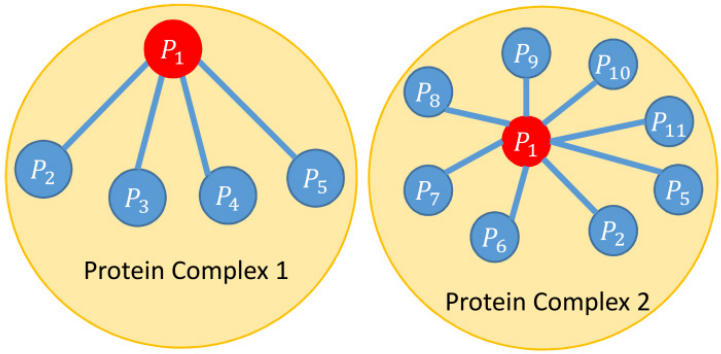
As discussed in the introduction, computational approaches to essential protein prediction can be of two types: (1) topological centrality-based approaches and (2) heterogeneous feature-based approaches. Experimental data [23] show the topology network centrality-based scoring technique, LID [23], and the heterogeneous feature-based approach, LIDC [23], perform better than the other existing approaches to essential protein identification. So, for each node and edge weight cut-off level in the second phase, LID (Luo and Qi 2015) and LIDC [23] are computed for each protein. LIDC combines heterogeneous values obtained from LID, in-degree centrality of complex (IDC) derived from protein complex Complex_745 [23], and ranking of an individual protein. The procedure for computing LIDC is shown in Figure 3. Finally, the proteins are sorted in descending order according to their computed LIDC values. Protein sets are selected as essential in two different ranking ranges (top 100–200 proteins). This selection strategy is the same as in Luo et al.'s work [23].



P_7 and P_8 are disconnected neighbors of P_1
P_2, P_3, P_4, P_5 and P_6 are the connected neighbours of P_1

$$LID\ (P_1) = \frac{|E(P_1)|}{|V(P_1)|} = \frac{4}{5}$$

$|E(P_1)|$ is the number of connections (edges) between neighbours of P_1 and $|V(P_1)|$ are the number of neighbours connected with each other.



IDC value of protein P_1 is 4 + 8 = 12.

Protein	LID	Rank
P_1	0.8	1
P_3	0.7	2
P_5	0.6	3
P_2	0.5	4
P_4	0.45	5
P_6	0.40	6

$$LIDC(P_1) = LID(P_1) \times \left(1 - \frac{RANK(P_1)}{N}\right) + IDC(P_1) \times \frac{RANK(P_1)}{N}$$

$$LIDC(P_1) = 0.8 \times \left(1 - \frac{1}{11}\right) + 12 \times \frac{1}{11} = 1.81$$

- N is the number of proteins in the current network
- $RANK(P_1)$ is the order number of the descending sort of protein P_1 according to $LID(P_1)$ in the current network

Figure 3. Schematic diagram of computation of LIDC. It is a combination of 3 scores: (1) LID, (2) IDC, and (3) ranking score. Disconnected neighbors (proteins) are highlighted in blue color whereas inter-connected neighbors (proteins) are represented in pink color. Protein complex is represented in yellow.

Algorithm 1 (Essential Protein Prediction)

Input: PPIN of yeast
Output: List of Essential and Non-essential Protein

Begin
//**calculating node weight**
for every node ***P*** in the network
Calculate the node weight, $W_p = \frac{\sum_{u \in V'}(\deg(u))}{|V'|}$
//V' is the set of neighbors of node P, and $|V'|$ is the number of proteins in V'
//$deg(u)$ is the degree of a node $u \in V'$
//**end of calculating node weight**
Compute $\theta_k = \alpha + k \times \sigma \times \left(1 - \frac{1}{1+\sigma^2}\right)$// **Cut-off calculation of node weight**
//α is the mean of node weight, σ is the standard deviation of node weight, $k \in \{1, 2, 3\}$ denotes three different
//cut-offs, i.e., low, medium, and high, respectively.
//**reduction of network based on Th_k of node weights**
for every node P in the network
if node weight of $P < \theta_k$
remove P from the network
//**end of reduction of network based on Th_k of node weights**
//**edge weight calculation**
for every edge ***E*** in the network
Calculate edge weight, $W_{uv} = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$
//$\Gamma(u)$ and $\Gamma(v)$ are the neighbors of u and v, respectively
//$\Gamma(u) \cap \Gamma(v)$ represents all common neighbors of u and v
//$\Gamma(u) \cup \Gamma(v)$ represents all distinct neighbors of u and v
//**end of edge weight calculation**
Compute $\theta_k = \alpha + k \times \sigma \times \left(1 - \frac{1}{1+\sigma^2}\right)$//**Cut-off calculation of edge weight**
//α is the mean of edge weight, σ is the standard deviation of edge weight, $k \in \{1, 2, 3\}$ denotes three different
//cut-offs, i.e., low, medium, and high, respectively.
//**reduction of network based on Th_k of edge weights**
for every edge E in the network
if edge weight of $E < \theta_k$
remove E from the network
//**end of reduction of network based on Th_k of edge weights**
//**calculate LIDC for low, medium, and high node edge weight**
//**calculation of LIDC**
for every node u in the pruned network, compute
$LID(u) = \frac{|E(u)|}{|V(u)|}$
//$|E(u)|$ is the number of connections (edges) between neighbors of u, and $|V(u)|$ are the number of neighbors
//connected with each other
//**end of calculation of LID**
$IDC(u) = \sum_{i \in ComplexSet(u)} IN - Degree(u)_i$
//$ComplexSet(u)$ denotes a set of protein complexes that include protein u
//$IN - Degree(u)_i$ is the degree of protein u in ith protein complex that belongs to $ComplexSet(u)$
//**end of calculation of IDC**
$LIDC(u) = LID(u) \times \left(1 - \frac{RANK(u)}{N}\right) + IDC(u) \times \frac{RANK(u)}{N}$
//$LID(u)$ is the value of the LID, $IDC(u)$ is the value of IDC of the protein complex of protein u,
//N is the number of proteins in the current network,
//$RANK(u)$ is the order number of the descending sort of protein u according to $LID(u)$ in the current network
//**end of calculation of LIDC**
Choose proteins in six ranking ranges (top 100–600) as essential protein sets.
End

4. Result and Discussion

As mentioned earlier, in this proposed work, an LIDC-based scoring technique [23] is used to mark proteins as essential in the topologically processed PPIN, and six different ranking ranges (top 100–600 proteins) are considered. The PPIN of yeast after predicting essential and non-essential proteins at ranking 100 is highlighted in Figure 4. The essentialness of protein sets in the different ranking ranges (top 100–600) at three different cut-offs, i.e., low node and edge weight, medium node and edge weight, and high node and edge weight, are validated against the essential protein set [23] (containing 1285 essential and 4394 non-essential proteins) formed from different databases like MIPS [67], SGD [70], DEG [74], and SGDP [75]. The comparison of the number of predicted essential proteins by our proposed method and several other existing methods like DC [17], BC [26], NC [14], LID [23], PeC [37], CoEWC [59], WDC [61], ION [35], LIDC [23], UC [34], etc. at the three cut-off levels are highlighted in the Supplementary Figures, i.e., Figures S2, S3, and S5–S8. From these figures, it is clear that our method generates an almost equal or greater number of essential proteins compared to LIDC [23] in most cases of the cut-off. This number is comparatively higher when compared to the other methods except for ION. The same observation has also been noted when the jackknife methodology is used to evaluate the proposed method against the others (see Figure 5). Though 20 percent of proteins are considered for evaluating precision, recall, and F-Score, our proposed methodology surpasses the others (see Table 2).

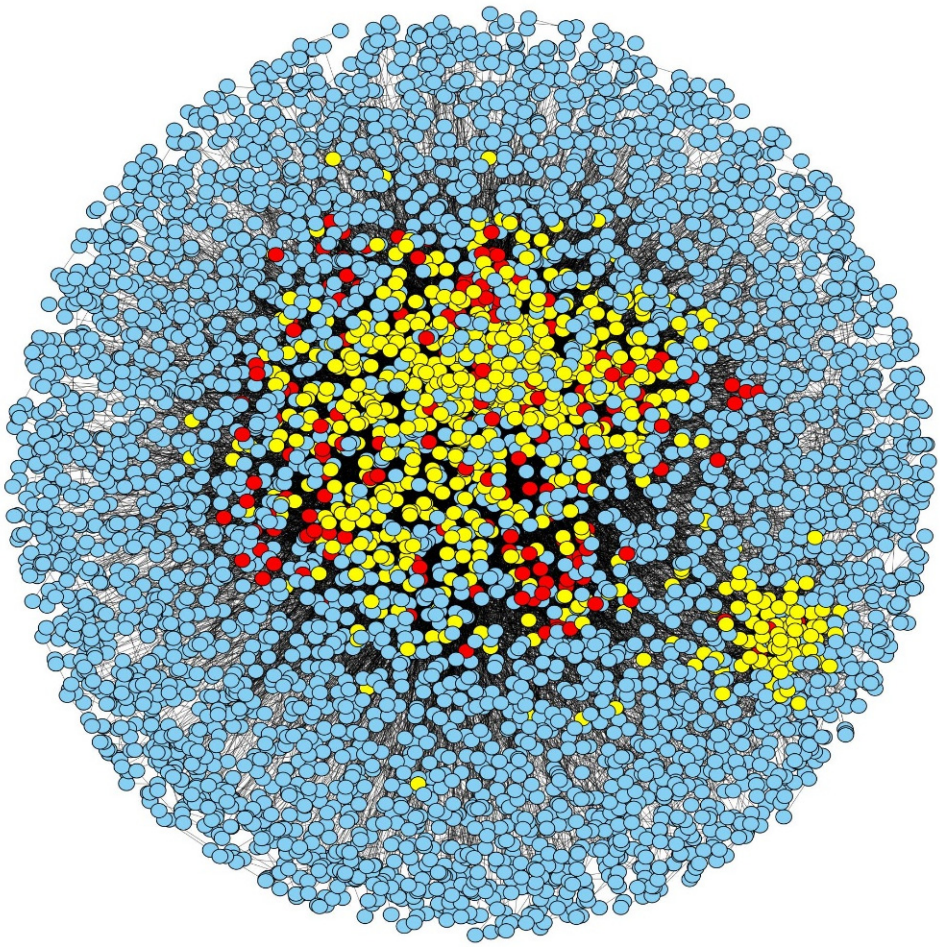


Figure 4. Essential and non-essential proteins in PPIN of yeast at a low cut-off. The yellow-colored proteins are the predicted non-essential ones, while the red ones are the predicted essential proteins. The blue-colored nodes represent proteins that are filtered out in the pre-filtering stage.

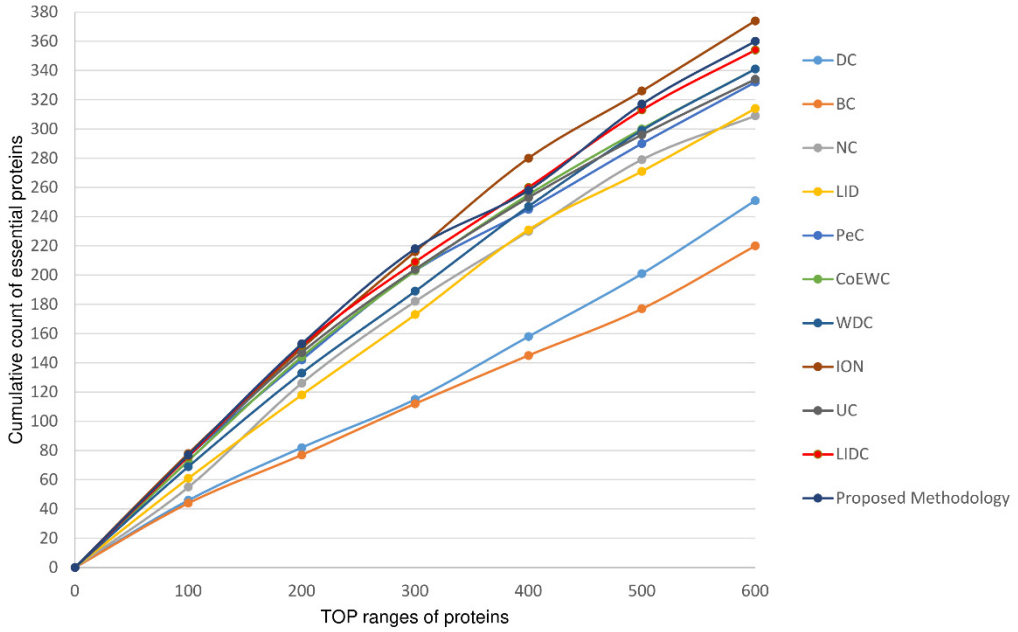


Figure 5. Validation of proposed methodology. All the methods are compared using the jackknife methodology for six different ranking ranges (top 100–600 proteins).

Table 2. Performance analysis of proposed method with other methodologies.

Methods	Precision	Recall	F-Score
DC (Jeong et al. 2001)	0.41	0.35	0.38
BC (Joy et al. 2005)	0.35	0.31	0.33
NC (Jianxin Wang et al. 2012)	0.46	0.40	0.43
LID (Luo and Qi 2015)	0.45	0.39	0.42
PeC (Li et al. 2012)	0.46	0.40	0.43
CoEWC (Zhang et al. 2013)	0.47	0.41	0.44
WDC (Xiwei et al. 2014)	0.48	0.42	0.45
ION (Peng et al. 2012)	0.53	0.41	0.46
UC (Li et al. 2017)	0.48	0.42	0.45
LIDC (Luo and Qi 2015)	0.50	0.44	0.47
Proposed Methodology	0.77	0.44	0.56

To compare and validate the performance of the proposed method, the top 20 percent of proteins [23] from the ranking result are selected as essential, while the remaining proteins are designated as non-essential. This selection strategy is the same as in Luo et al.'s work [23]. Precision, recall, and F-score are considered performance evaluation metrics. The performance analysis is highlighted in Table 2. It can be derived from Table 2 that our proposed method performs better than the others in terms of precision, recall, and F-score. This signifies that it succeeds in returning most of the relevant proteins compared to the training set of essential proteins. High precision also indicates a low false positive rate. Removing less important nodes and edges and working on the pruned network makes our proposed method worthy and superior to the methods listed in Table 2 and enables us to get high precision, recall, and F-score values.

Our proposed method's satisfactory performance is achieved using node and edge weights with three proper levels of cut-offs. The pruned PPIN network of yeast at ranking 100 is shown in Figure S4 in the Supplementary Materials. It should also be noted here that though the working mechanisms of LIDC [23] and our proposed method are almost the same, LIDC [23] is applied to the entire PPIN database of yeast, while our proposed method works on a filtered PPIN generated by using three levels of cut-offs on both node and edge weights. The statistics of predicted essential proteins in a filtered PPIN of yeast at

three cut-off levels—low node and edge weight, medium node and edge weight, and high node and edge weight—are displayed in Table 3. The overall precision, recall, and F-score at three levels of cut-offs are shown in Table 4.

Table 3. Network statistics of pruned PPIN of yeast at three levels of cut-offs.

Cut-Off Levels	Proteins after Node Reduction	Interactions after Node Reduction	Proteins after Edge Reduction	Interactions after Node Reduction	Essential Protein	Non-Essential Protein
Low	1393	14,063	985	3907	198	787
Medium	1374	13,924	969	3847	194	775
High	1340	13,714	931	3733	187	744

Table 4. Performance analysis of our proposed method at three levels of cut-offs.

Cut-Off Levels	Recall	Precision	F-Score
Low	0.41	0.75	0.53
Medium	0.42	0.76	0.54
High	0.44	0.77	0.56

5. Conclusions

Identifying essential proteins is considered one of the most challenging research areas. It helps us identify the significant proteins that are biologically active and play a crucial part in performing vital specific functions of the human body. These proteins might also be essential in transmitting disease or infection when the body is exposed to pathogens. Thus, the computational methods developed for identifying essential proteins should be very effective. PPIN is one of the resources through which this can be done. However, it should be borne in mind that all the network features must be adequately assessed, and the presence of reliable nodes and edges must be ensured. The proposed methodology efficiently identifies essential proteins from a pruned network using local interaction density and local interaction density with a protein complex. The rule-based network pruning is based on specific cut-off edge and node weight values. A detailed comparative study on the performance evaluation of the proposed method and other methods reveals the superiority of this method over others. Because this method solely depends on topological attributes, care should be taken to use a noise-free protein–protein interaction network. This work may be extended to the protein interaction network of any other organism in our future work. However, it should be kept in mind that the essentiality of genes is dynamic. It depends upon the surrounding environment. So, even if several PPIN data of yeast are used for the computational identification of essential proteins/genes, it cannot be assured that the genetic backgrounds set as an experimental environment for all the yeast strains are similar or not [76].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cells11172648/s1. Figure S1: PPIN Network of Yeast of YDIP_5093. It contains 5093 proteins and 24743 interactions; Figure S2: Prediction comparison. Comparison of number of predicted essential proteins for low node and edge weight threshold; Figure S3: Prediction comparison. Comparison of top 100 and top 200 predicted essential proteins for medium node and edge weight threshold; Figure S4: Pruned PPIN of yeast at Low Threshold. Yellow colored nodes are non-essential proteins while the green colored nodes are the essential ones; Figure S5: Prediction comparison. Comparison of top 100 and top 200 predicted essential proteins for high node and edge weight threshold; Figure S6: Prediction comparison. Comparison of number of top 300, top 400, top 500 and top 600 predicted essential proteins for low node and edge weight threshold; Figure S7: Prediction comparison. Comparison of number of top 300, top 400, top 500 and top 600 predicted essential proteins for medium node and edge weight threshold; Figure S8: Prediction comparison. Comparison of number of top 300, top 400, top 500 and top 600 predicted essential proteins for high node and edge weight threshold.

Author Contributions: Conceptualization, A.B., S.P., S.S., P.C., A.K.H. and M.N.; data curation, A.B., S.P. and S.S.; formal analysis, A.B., S.P., S.S., P.C., A.K.H., M.N., S.B. and D.P.; funding acquisition, S.B. and D.P.; investigation, P.C., M.N., S.B. and D.P.; methodology, A.B., S.P., S.S., P.C., A.K.H. and S.B.; project administration, P.C., M.N., S.B. and D.P.; resources, S.S., M.N. and S.B.; software, A.B., S.P. and S.S.; supervision, P.C., M.N. and S.B.; validation, M.N., S.B. and D.P.; visualization, S.S.; writing—original draft, A.B., S.P., S.S. and A.K.H.; writing—review and editing, P.C., M.N., S.B. and D.P. All authors have read and agreed to the published version of the manuscript.

Funding: 1. UGC Research Award (F.30-31/2016(SA-II)) from UGC, Government of India, and DBT project (No.BT/PR16356/BID/7/596/2016), Ministry of Science and Technology, Government of India. 2. Excellence Initiative: Research University (IDUB) IDUB grant BOB-IDUB-622-197/2021. 3. Polish National Science Center (2019/35/O/ST6/02484 and 2020/37/B/NZ2/03757), Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. 4. European Commission Horizon 2020 Marie Skłodowska-Curie ITN Enpathy grant "Molecular Basis of Human enhanceropathies"; and U.S. National Institutes of Health 4DNucleome grant 1U54DK107967-01 "Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation" 5. Marie Skłodowska-Curie Action: co-supported as RENOIR Project by the European Union Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 691152 and by the Ministry of Science and Higher Education (Poland), grant Nos. W34/H2020/2016, 329025/PnH/2016. 6. Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme. 7. Polish Ministry of Science and Higher Education (decision no. 7054/IA/SP/2020 of 2020-08-28).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code of the work is available on GitHub at the following link for free academic use.

Acknowledgments: The authors are thankful to the CMATER research laboratory of the Computer Science Department, Jadavpur University, India, for providing infrastructure facilities during the progress of the work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Deng, H.; Jia, Y.; Zhang, Y. Protein structure prediction. *Int. J. Mod. Phys. B* **2018**, *32*, 1840009. [CrossRef] [PubMed]
2. Krupa, P.; Mozolewska, M.A.; Joo, K.; Lee, J.; Czaplewski, C.; Liwo, A. Prediction of Protein Structure by Template-Based Modeling Combined with the UNRES Force Field. *J. Chem. Inf. Model.* **2015**, *55*, 1271–1281. [CrossRef] [PubMed]
3. Makrodimitris, S.; van Ham, R.C.H.J.; Reinders, M.J.T. Improving protein function prediction using protein sequence and GO-term similarities. *Bioinformatics* **2018**, *35*, 1116–1124. [CrossRef]
4. Koskinen, P.; Törönen, P.; Nokso-Koivisto, J.; Holm, L. PANNZER: High-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics* **2015**, *31*, 1544–1552. [CrossRef] [PubMed]
5. Das, S.; Orengo, C.A. Protein function annotation using protein domain family resources. *Methods* **2016**, *93*, 24–34. [CrossRef]
6. Rentzsch, R.; Orengo, C.A. Protein function prediction using domain families. *BMC Bioinform.* **2013**, *14*, S5. [CrossRef]
7. Saha, S.; Chatterjee, P.; Basu, S.; Nasipuri, M.; Plewczynski, D. FunPred 3.0: Improved protein function prediction using protein interaction network. *PeerJ* **2019**, *7*, e6830. [CrossRef]
8. Saha, S.; Chatterjee, P.; Basu, S.; Kundu, M.; Nasipuri, M. FunPred-1: Protein function prediction from a protein interaction network using neighborhood analysis. *Cell. Mol. Biol. Lett.* **2014**, *19*, 675–691. [CrossRef]
9. Basak, S.N.; Biswas, A.K.; Saha, S.; Chatterjee, P.; Basu, S.; Nasipuri, M. Target Protein Function Prediction by Identification of Essential Proteins in Protein-Protein Interaction Network. In Proceedings of the Computational Intelligence, Communications, and Business Analytics, Singapore, 26 June 2019; pp. 219–231.
10. Saha, S.; Prasad, A.; Chatterjee, P.; Basu, S.; Nasipuri, M. Protein function prediction from protein–protein interaction network using gene ontology based neighborhood analysis and physico-chemical features. *J. Bioinform. Comput. Biol.* **2018**, *16*, 1850025. [CrossRef]
11. Zhao, B.; Hu, S.; Li, X.; Zhang, F.; Tian, Q.; Ni, W. An efficient method for protein function annotation based on multilayer protein networks. *Hum. Genom.* **2016**, *10*, 33. [CrossRef]
12. Savojardo, C.; Martelli Pier, L.; Fariselli, P.; Profiti, G.; Casadio, R. BUSCA: An integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res.* **2018**, *46*, W459–W466. [CrossRef] [PubMed]

13. Nebenführ, A. Identifying Subcellular Protein Localization with Fluorescent Protein Fusions After Transient Expression in Onion Epidermal Cells. In *Plant Cell Morphogenesis: Methods and Protocols*; Žárský, V., Cvrčková, F., Eds.; Humana Press: Totowa, NJ, USA, 2014; pp. 77–85.
14. Jianxin Wang, M.L.H.W.Y.P.; Min, L.; Huan, W.; Yi, P.; Wang, J.; Li, M.; Wang, H.; Pan, Y.; Jianxin, W.; Min, L.; et al. Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1070–1080. [CrossRef] [PubMed]
15. Li, M.; Lu, Y.; Wang, J.; Wu, F.; Pan, Y. A Topology Potential-Based Method for Identifying Essential Proteins from PPI Networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *12*, 372–383. [CrossRef] [PubMed]
16. Li, M.; Wang, J.; Wang, H.; Pan, Y. Essential Proteins Discovery from Weighted Protein Interaction Networks. In Proceedings of the Bioinformatics Research and Applications, Berlin, Germany, 26 June 2010; pp. 89–100.
17. Jeong, H.; Mason, S.P.; Barabási, A.L.; Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **2001**, *411*, 41–42. [CrossRef]
18. Jimenez-Sanchez, G.; Childs, B.; Valle, D. Human disease genes. *Nature* **2001**, *409*, 853–855. [CrossRef]
19. Gill, N.; Singh, S.; Aseri, T.C. Computational Disease Gene Prioritization: An Appraisal. *J. Comput. Biol.* **2014**, *21*, 456–465. [CrossRef]
20. Zhu, C.; Wu, C.; Aronow, B.J.; Jegga, A.G. Computational approaches for human disease gene prediction and ranking. *Adv. Exp. Med. Biol.* **2014**, *799*, 69–84. [CrossRef]
21. Giaever, G.; Chu, A.M.; Ni, L.; Connelly, C.; Riles, L.; Véronneau, S.; Dow, S.; Lucau-Danila, A.; Anderson, K.; André, B.; et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **2002**, *418*, 387–391. [CrossRef]
22. Yuan, Z.; Chong, W. Identification of Essential Proteins Using Improved Node and Edge Clustering Coefficient. In Proceedings of the 2018 37th Chinese Control Conference (CCC), Wuhan, China, 25–27 July 2018; pp. 3258–3262.
23. Luo, J.; Qi, Y. Identification of Essential Proteins Based on a New Combination of Local Interaction Density and Protein Complexes. *PLoS ONE* **2015**, *10*, e0131418.
24. Tang, Y.; Li, M.; Wang, J.; Pan, Y.; Wu, F.-X. CytoNCA: A cytoscape plugin for centrality analysis and evaluation of protein interaction networks. *Bio Syst.* **2015**, *127*, 67–72. [CrossRef]
25. Smoot, M.E.; Ono, K.; Ruscheinski, J.; Wang, P.L.; Ideker, T. Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics* **2011**, *27*, 431–432. [CrossRef] [PubMed]
26. Joy, M.P.; Brock, A.; Ingber, D.E.; Huang, S. High-Betweenness Proteins in the Yeast Protein Interaction Network. *J. Biomed. Biotechnol.* **2005**, *2005*, 96–103. [CrossRef] [PubMed]
27. Wuchty, S.; Stadler, P.F. Centers of complex networks. *J. Theor. Biol.* **2003**, *223*, 45–53. [CrossRef]
28. Bonacich, P. Power and Centrality: A Family of Measures. *Am. J. Sociol.* **1987**, *92*, 1170–1182. [CrossRef]
29. 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*, 143–150. [CrossRef]
30. Estrada, E.; Rodríguez-Velázquez, J.A. Subgraph centrality in complex networks. *Phys. Rev. E* **2005**, *71*, 056103. [CrossRef]
31. S Karen, M.Z.; Stephenson, K.; Zelen, M. Rethinking centrality: Methods and examples. *Soc. Netw.* **1989**, *11*, 1–37. [CrossRef]
32. Hsing, M.; Byler, K.G.; Cherkasov, A. The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks. *BMC Syst. Biol.* **2008**, *2*, 80. [CrossRef]
33. Ren, J.; Wang, J.; Li, M.; Wang, H.; Liu, B. Prediction of Essential Proteins by Integration of PPI Network Topology and Protein Complexes Information. In Proceedings of the International Symposium on Bioinformatics Research and Applications, Changsha, China, 27–29 May 2011; pp. 12–24.
34. Li, M.; Lu, Y.; Niu, Z.; Wu, F.-X. United Complex Centrality for Identification of Essential Proteins from PPI Networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *14*, 370–380. [CrossRef]
35. Peng, W.; Wang, J.; Wang, W.; Liu, Q.; Wu, F.-X.; Pan, Y. Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. *BMC Syst. Biol.* **2012**, *6*, 87. [CrossRef]
36. Peng, X.; Wang, J.; Zhong, J.; Luo, J.; Pan, Y. An efficient method to identify essential proteins for different species by integrating protein subcellular localization information. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Washington, DC, USA, 9–12 November 2015; pp. 277–280.
37. 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*, 15. [CrossRef]
38. Xiao, Q.; Wang, J.; Peng, X.; Wu, F.-x.; Pan, Y. Identifying essential proteins from active PPI networks constructed with dynamic gene expression. *BMC Genom.* **2015**, *16* (Suppl. S3), S1. [CrossRef] [PubMed]
39. Zhang, X.; Xiao, W.; Hu, X. Predicting essential proteins by integrating orthology, gene expressions, and PPI networks. *PLoS ONE* **2018**, *13*, e0195410. [CrossRef] [PubMed]
40. Dong, C.; Jin, Y.-T.; Hua, H.-L.; Wen, Q.-F.; Luo, S.; Zheng, W.-X.; Guo, F.-B. Comprehensive review of the identification of essential genes using computational methods: Focusing on feature implementation and assessment. *Brief. Bioinform.* **2020**, *21*, 171–181. [CrossRef] [PubMed]
41. Cheng, J.; Wu, W.; Zhang, Y.; Li, X.; Jiang, X.; Wei, G.; Tao, S. A new computational strategy for predicting essential genes. *BMC Genom.* **2013**, *14*, 910. [CrossRef]
42. Cheng, J.; Xu, Z.; Wu, W.; Zhao, L.; Li, X.; Liu, Y.; Tao, S. Training Set Selection for the Prediction of Essential Genes. *PLoS ONE* **2014**, *9*, e86805. [CrossRef]

43. Sakharkar, K.R.; Sakharkar, M.K.; Chow, V.T. A novel genomics approach for the identification of drug targets in pathogens, with special reference to Pseudomonas aeruginosa. *Silico Biol.* **2004**, *4*, 355–360.
44. Song, J.H.; Ko, K.S.; Lee, J.Y.; Baek, J.Y.; Oh, W.S.; Yoon, H.S.; Jeong, J.Y.; Chun, J. Identification of essential genes in Streptococcus pneumoniae by allelic replacement mutagenesis. *Mol. Cells* **2005**, *19*, 365–374.
45. Sarangi, A.N.; Lohani, M.; Aggarwal, R. Prediction of essential proteins in prokaryotes by incorporating various physico-chemical features into the general form of Chou's pseudo amino acid composition. *Protein Pept. Lett.* **2013**, *20*, 781–795. [CrossRef]
46. Ning, L.W.; Lin, H.; Ding, H.; Huang, J.; Rao, N.; Guo, F.B. Predicting bacterial essential genes using only sequence composition information. *Genet. Mol. Res. GMR* **2014**, *13*, 4564–4572. [CrossRef]
47. Jeong, H.; Oltvai, Z.N.; Barabási, A.L. Prediction of Protein Essentiality Based on Genomic Data. *Complexus* **2003**, *1*, 19–28. [CrossRef]
48. Chen, Y.; Xu, D. Understanding protein dispensability through machine-learning analysis of high-throughput data. *Bioinformatics* **2005**, *21*, 575–581. [CrossRef] [PubMed]
49. Zotenko, E.; Mestre, J.; O'Leary, D.P.; Przytycka, T.M. Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.* **2008**, *4*, e1000140. [CrossRef]
50. Yu, H.; Braun, P.; Yildirim, M.A.; Lemmens, I.; Venkatesan, K.; Sahalie, J.; Hirozane-Kishikawa, T.; Gebreab, F.; Li, N.N.; Simonis, N.; et al. High-quality binary protein interaction map of the yeast interactome network. *Science* **2008**, *322*, 104–110. [CrossRef] [PubMed]
51. Ryan, C.J.; Krogan, N.J.; Cunningham, P.; Cagney, G. All or nothing: Protein complexes flip essentiality between distantly related eukaryotes. *Genome Biol. Evol.* **2013**, *5*, 1049–1059. [CrossRef]
52. Wang, H.; Kakaradov, B.; Collins, S.R.; Karotki, L.; Fiedler, D.; Shales, M.; Shokat, K.M.; Walther, T.C.; Krogan, N.J.; Koller, D. A Complex-based Reconstruction of the Saccharomyces cerevisiae Interactome. *Mol. Cell. Proteom.* **2009**, *8*, 1361–1381. [CrossRef]
53. He, X.; Zhang, J. Why do hubs tend to be essential in protein networks? *PLoS Genet.* **2006**, *2*, e88. [CrossRef]
54. Lei, X.; Yang, X. A new method for predicting essential proteins based on participation degree in protein complex and subgraph density. *PLoS ONE* **2018**, *13*, e0198998. [CrossRef]
55. Hart, G.T.; Lee, I.; Marcotte, E.R. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinform.* **2007**, *8*, 236. [CrossRef]
56. Zhong, J.; Wang, J.; Peng, W.; Zhang, Z.; Pan, Y. Prediction of essential proteins based on gene expression programming. *BMC Genom.* **2013**, *14*, s4–s7. [CrossRef]
57. Ercsey-Ravasz, M.; Lichtenwalter, R.N.; Chawla, N.V.; Toroczkai, Z. Range-limited centrality measures in complex networks. *Phys. Rev. E* **2012**, *85*, 066103. [CrossRef] [PubMed]
58. Korn, A.; Schubert, A.; Telcs, A. Lobby index in networks. *Phys. A Stat. Mech. Its Appl.* **2009**, *388*, 2221–2226. [CrossRef]
59. Zhang, X.; Xu, J.; Xiao, W.X. A New Method for the Discovery of Essential Proteins. *PLoS ONE* **2013**, *8*, e58763. [CrossRef] [PubMed]
60. Lü, L.; Zhang, Y.-C.; Yeung, C.H.; Zhou, T. Leaders in social networks, the Delicious case. *PLoS ONE* **2011**, *6*, e21202. [CrossRef]
61. Tang, X.; Wang, J.; Zhong, J.; Pan, Y. Predicting essential proteins basedon weighted degree centrality. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2014**, *11*, 407–418. [CrossRef]
62. Ghosh, R.; Lerman, K. Parameterized centrality metric for network analysis. *Phys. Rev. E—Stat. Nonlinear Soft Matter Phys.* **2011**, *83*, 066118. [CrossRef]
63. Li, G.; Li, M.; Wang, J.; Li, Y.; Pan, Y. United neighborhood closeness centrality and orthology for predicting essential proteins. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, *17*, 1451–1458. [CrossRef]
64. Wang, S.; Wu, F. Detecting overlapping protein complexes in PPI networks based on robustness. *Proteome Sci.* **2013**, *11*, S18. [CrossRef]
65. Xenarios, I.; Rice, D.W.; Salwinski, L.; Baron, M.K.; Marcotte, E.M.; Eisenberg, D. DIP: The Database of Interacting Proteins. *Nucleic Acids Res.* **2000**, *28*, 289–291. [CrossRef]
66. Xenarios, I. DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **2002**, *30*, 303–305. [CrossRef]
67. Mewes, H.W. MIPS: Analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* **2005**, *34*, D169–D172. [CrossRef] [PubMed]
68. Friedel, C.C.; Krumsiek, J.; Zimmer, R. Bootstrapping the interactome: Unsupervised identification of protein complexes in yeast. In *Research in Computational Molecular Biology*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 3–16.
69. Aloy, P.; Böttcher, 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*, 2026–2029. [CrossRef] [PubMed]
70. Cherry, J.M.; Adler, C.; Ball, C.; Chervitz, S.A.; Dwight, S.S.; Hester, E.T.; Jia, Y.; Juvik, G.; Roe, T.; Schroeder, M.; et al. SGD: Saccharomyces genome database. *Nucleic Acids Res.* **1998**, *26*, 73–79. [CrossRef] [PubMed]
71. Pu, S.; Wong, J.; Turner, B.; Cho, E.; Wodak, S.J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* **2009**, *37*, 825–831. [CrossRef] [PubMed]
72. Pu, S.; Vlasblom, J.; Emili, A.; Greenblatt, J.; Wodak, S.J. Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics* **2007**, *7*, 944–960. [CrossRef] [PubMed]

73. Zhang, Y.; Lin, H.; Yang, Z.; Wang, J.; Liu, Y.; Sang, S. A method for predicting protein complex in dynamic PPI networks. *BMC Bioinform.* **2016**, *17*, 229. [CrossRef] [PubMed]
74. Zhang, R.; Lin, Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.* **2009**, *37*, D455–D458. [CrossRef]
75. Winzeler, E.A.; Shoemaker, D.D.; Astromoff, A.; Liang, H.; Anderson, K.; Andre, B.; Bangham, R.; Benito, R.; Boeke, J.D.; Bussey, H.; et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **1999**, *285*, 901–906. [CrossRef]
76. Gurumayum, S.; Jiang, P.; Hao, X.; Campos, T.L.; Young, N.D.; Korhonen, P.K.; Gasser, R.B.; Bork, P.; Zhao, X.-M.; He, L.-j.; et al. OGEE v3: Online GEne Essentiality database with increased coverage of organisms and human cell lines. *Nucleic Acids Res.* **2021**, *49*, D998–D1003. [CrossRef]