PLOS ONE

RESEARCH ARTICLE

A new computational strategy for identifying essential proteins based on network topological properties and biological information

Chao Qin, Yongqi Sun*, Yadong Dong

Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

* yqsun@bjtu.edu.cn

Citation: Qin C, Sun Y, Dong Y (2017) A new computational strategy for identifying essential proteins based on network topological properties and biological information. PLoS ONE 12(7): e0182031. https://doi.org/10.1371/journal.pone.0182031

Editor: Attila Csikász-Nagy, King's College London, UNITED KINGDOM

Received: February 15, 2017

Accepted: July 11, 2017

Published: July 28, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by NO.61572005, National Natural Science Foundation of China, www.nsfc.gov.cn, YQS CQ; NO.61272004, National Natural Science Foundation of China, www.nsfc.gov.cn, YQS; NO. K17JB00220, Fundamental Research Funds for the Central Universities, CQ YQS YDD. The funders had no role in study design, data collection and

Abstract

Essential proteins are the proteins that are indispensable to the survival and development of an organism. Deleting a single essential protein will cause lethality or infertility. Identifying and analysing essential proteins are key to understanding the molecular mechanisms of living cells. There are two types of methods for predicting essential proteins: experimental methods, which require considerable time and resources, and computational methods, which overcome the shortcomings of experimental methods. However, the prediction accuracy of computational methods for essential proteins requires further improvement. In this paper, we propose a new computational strategy named CoTB for identifying essential proteins based on a combination of topological properties, subcellular localization information and orthologous protein information. First, we introduce several topological properties of the protein-protein interaction (PPI) network. Second, we propose new methods for measuring orthologous information and subcellular localization and a new computational strategy that uses a random forest prediction model to obtain a probability score for the proteins being essential. Finally, we conduct experiments on four different *Saccharomyces cerevisiae* datasets. The experimental results demonstrate that our strategy for identifying essential proteins outperforms traditional computational methods and the most recently developed method, SON. In particular, our strategy improves the prediction accuracy to 89, 78, 79, and 85 percent on the YDIP, YMIPS, YMBD and YHQ datasets at the top 100 level, respectively.

Introduction

Essential proteins are the proteins that are indispensable to the survival of an organism; therefore, these proteins are considered to be the basis of life. Deleting any of these proteins will lead to cell death [1]. Thus, identifying essential proteins is of great significance, and it will

analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

help us understand the minimum requirements for cell life and find novel treatments for diseases [2–4].

To date, many traditional biological methods have been proposed for identifying essential proteins, such as gene knockouts [5], conditional knockouts [6] and RNA interference [7]. These traditional biological methods are time consuming, expensive and not always practical. To overcome the shortcomings of these biological methods, a number of computational methods that only consider the topological properties have been proposed, such as degree centrality (DC) [8], betweenness centrality (BC) [9], eigenvector centrality (EC) [10], subgraph centrality (SC) [11], local average connectivity-based method (LAC) [12], and network centrality (NC) [13]. To further improve the prediction accuracy, Li [14] proposed a method called TP that uses topology potential to identify essential proteins. Although these methods facilitate the detection of essential proteins, they only consider the topological properties of the network, and they do not take the intrinsic properties of individual proteins into account. Consequently, many computational methods combined with biological information have been proposed. Such biological information includes protein complex information, gene expression data, subcellular localization information, orthologous protein information, and so on.

A protein complex is a group of proteins that interact with each other and function as a unit at a given time and place in a certain biological process. It has been proven that essential proteins are more likely to gather in protein complexes [15]. Based on this idea, Li [15] proposed a method called united complex centrality (UC) that integrates protein complex information. Luo [16] proposed a method named LIDC combined with the in-degree centrality of complex (IDC), which measures the in-degree value of a protein in the protein complex. Qin [17] proposed a method named LBCC that integrates the local topological features, global topological features, and protein complex information, where the local topological features are Den_1 and Den_2, representing the local densities of networks, and it improved the prediction precision to 75 percent on the YMIPS dataset at the top 100 level.

Gene expression is the process of transcribing and translating genetic information stored in a DNA sequence into functional gene products, which are often proteins, and it is also an important feature for predicting essential proteins. Li [18] proposed a method named PeC that integrates gene expression data for predicting essential proteins. Tang [19] proposed a weighted degree centrality using gene expression data to achieve the reliable prediction of essential proteins. Zhao [20] proposed a method named PEMC that integrates network topology with gene expression profile and protein domain information to construct weighted protein networks for discovering essential proteins.

Some researchers have reported that the locations of proteins are correlated with their essentiality and that essential proteins appear more frequently at specific locations. Zhong [21] proposed a feature selection method for predicting essential proteins, and the results indicated that the subcellular localization information can help increase the prediction accuracy for predicting essential proteins.

Orthologous protein information is another important aspect for identifying essential proteins. Orthologous proteins are proteins that are derived from a common ancestor and generally retain the same or very similar functions. It has been proven that orthologous properties are positively correlated with protein essentiality [22]. Li [23] proposed a method named SON that integrates subcellular localization and orthologous score (OS) information, and this method improved the accuracy of predicting essential proteins to approximately 81 percent on the YDIP dataset at the top 100 level. Li [24] proposed a method named GOS that integrates gene expression, orthology, and subcellular localization information to identify essential proteins.

In this paper, we first introduce several traditional topological properties of protein-protein interaction (PPI) networks, including Laplacian centrality (LC) [25], which is an intermediate measure between global and local properties. We then propose new measures of orthologous score (DOS) and subcellular localization score (SLS), as well as our new prediction method named CoTB, which combines Den_1, Den_2, BC, IDC, LC, DOS and SLS and uses a random forest model to obtain a probability score for the proteins being essential.

We conducted our experiments on four different PPI networks of *Saccharomyces cerevisiae*, namely, YDIP, YMIPS, YMBD, and YHQ, which will be described in the Experimental data section. The experimental results showed that our method, CoTB, obtained superior performance compared to the traditional measures, including DC, BC, SC, EC, NC, and LAC. CoTB exhibited the best performance and obtained prediction precisions of 89, 78, 79, and 85 percent on the YDIP, YMIPS, YMBD and YHQ datasets at the top 100 level, respectively. In particular, compared to the most recently developed method, SON [23], CoTB improved the prediction precisions by at least 9, 10, 8, 8, 7, and 8 percent on the YDIP dataset at the top 100 to top 600 levels, respectively. Compared to GOS [24], CoTB improved the prediction precisions by at least 6, 6, 9, and 10 percent on the YDIP dataset at the top 300 to top 600 levels, respectively. Compared to our LBCC [17], CoTB improved the prediction precisions by at least 20, 4, 21, and 500 percent on the YDIP, YMIPS, YMBD and YHQ datasets at the top 100 level, respectively.

Preliminaries

A PPI network is represented as an undirected simple graph $G(V, E)$ with a set of nodes (proteins) V and a set of edges E (interactions). Let N_v denote the set of neighbour nodes of node v, $|N_v|$ denote the number of neighbours of node v and $G[S]$ denote the induced subgraph of G on node set S. The definitions of several topological properties of a PPI network are as follows.

Degree centrality (*DC*). The DC of a node v is denoted as the total number of its neighbour nodes, and it is denoted as

$$DC(v) = deg(v),$$

where $deg(v)$ is the number of its incident edges.

Betweenness centrality (*BC*). The BC of a node v is calculated based on the shortest paths, and it is denoted as

$$BC(v) = \sum_{s} \sum_{t} \frac{\sigma_{st}(v)}{\sigma_{st}}, s \neq t \neq v \in V,$$

where σ_{st} is the total number of shortest paths from s to t and $\sigma_{st}(v)$ is the total number of shortest paths passing through v from s to t.

Eigenvector centrality (*EC*). The EC of a node v is calculated based on the adjacency matrix of the network, and it is denoted as

$$EC(v) = \alpha_{max}(v),$$

where α_{max} is the eigenvector corresponding to the largest eigenvalue of the adjacency matrix and $\alpha_{max}(v)$ is the vth component of α_{max}.

Local average connectivity centrality (*LAC*). The LAC of a node v describes the closeness of its neighbours, and it is denoted as

$$LAC(v) = \frac{\sum_{u \in N_v} deg^{C_v}(u)}{|N_v|}, u \in N_v,$$

where C_v is the induced subgraph of G on node set N_v and $deg^{C_v}(u)$ is the number of its neighbour nodes in C_v.

Neighbourhood centrality (*NC*). The NC of a node v considers the importance of the relationship between v and its neighbours, and it is denoted as

$$NC(v) = \sum_{u \in N_v} \frac{z_{v,u}}{min(d_v - 1, d_u - 1)},$$

where $z_{v,u}$ is the number of common neighbour vertices of v and u and d_v and d_u are the degrees of nodes v and u, respectively.

Subgraph centrality (*SC*). The SC of a node v measures the participation of a node in all subgraphs of the network, and it is denoted as

$$SC(v) = \sum_{k=0}^{\infty} \frac{\mu_k(v)}{k!},$$

where $\mu_k(v)$ is the number of circles starting and ending at v with length k.

Laplacian centrality (*LC*). For a graph G with n nodes, let $W(G)$ be the adjacency matrix of size n by n, and let $X(G)$ be a matrix as follows:

$$\begin{pmatrix} x_1 & 0 & \dots & 0 \\ 0 & x_2 & \dots & 0 \\ . & . & . & . \\ 0 & 0 & \dots & x_n \end{pmatrix}$$

where x_i is the number of neighbours of node i. The LC $LC(v_i, G)$ of vertex v_i is defined as

$$LC(v_i, G) = \frac{E_L(G) - E_L(G_i)}{E_L(G)},$$

$$E_L(G) = \sum_{i=1}^{n} \lambda_i^2,$$

where G_i is the graph obtained by deleting v_i from G and λ_i is the eigenvalues of the matrix $L(G) = X(G) - W(G)$.

In-degree centrality of complex (*IDC*). The IDC of a node v measures the sum of degrees of node v in different protein complexes, and it is denoted as

$$IDC(v) = \sum_{i \in ComplexSet(v)} IN - Degree(v)_i,$$

where *ComplexSet*(v) is the set of protein complexes with protein v and *IN-Degree*(v)$_i$ is the value of $DC(v)$ for the i–th protein complex in *ComplexSet*(v).

$Den_1(v)$. For a node v, $Den_1(v)$ is the ratio of the number of edges to the number of all possible edges of the induced subgraph in G by the node set $N_v \cup \{v\}$, and it is denoted as

$$Den_1(v) = \frac{2|E(H)|}{|V(H)|(|V(H)| - 1)},$$

where H denotes the induced subgraph of $G[N_v \cup \{v\}]$.

$Den_2(v)$. For a node v, let M_u be the node set for which the distance to v is 2. $Den_2(v)$ is the ratio of the number of edges to the number of all possible edges of the induced subgraph in G by the node set $M_u \cup N_v \cup \{v\}$, and it is denoted as

$$Den_2(v) = \frac{2|E(H)|}{|V(H)|(|V(H)|-1)}, \tag{12}$$

where H denotes the induced subgraph of $G[M_u \cup N_v \cup \{v\}]$.

Methods

New measure of orthologous score

Orthologous proteins are proteins that are derived from a common ancestor and generally retain the same or very similar functions. It has been proven that orthologous properties are positively correlated with protein essentiality [22]. The greater the number of reference organisms in which such a protein appears, the more essential the protein is.

For a protein v, its orthologous score $OS(v)$ from [22] is the number of reference organisms where it appears, and it is denoted as

$$OS(v) = \sum_{i=1}^{s} T_i(v),$$

$$T_i(v) = \begin{cases} 1 & v \in o(i) \\ 0 & v \notin o(i) \end{cases},$$

where s is the number of reference organisms and $o(i)$ is the set of nodes (proteins) in the ith reference organism.

In this paper, we define a new measure of orthologous score $DOS(v)$ as

$$DOS(v) = a * DC(v) + OS(v),$$

where a is a scaling parameter that ranges from 0.1 to 1. Through a large number of experiments for identifying essential proteins on four different testing datasets, we found that our new computational strategy, CoTB, which will be described in the following, obtains the best performance when a is set to 0.1.

New measure of subcellular localization score

The localization of proteins is the location in cells where a protein appears. It has been proven that the localization of proteins is an important factor for determining protein essentiality [21, 23, 26, 27], and statistical results show that essential proteins are more likely to exist in specific cellular locations. For example, many important biological processes, such as DNA replication and mRNA synthesis, usually occur in the nuclear.

In this section, we propose a new measure of subcellular localization score (SLS). For a protein v, we define its SLS as the sum of the subcellular localization coefficient (SLC) of each subcellular localization,

$$SLS(v) = \sum_{v \in s(l)} SLC(l),$$

where $s(l)$ is the set of proteins in the l subcellular location and $SLC(l)$ is the SLC, which is

defined as

$$SLC(l) = \frac{t_l}{t} - \frac{a_l}{a},$$

where a_l is the total number of proteins in the l subcellular location and a is the total number of proteins. The values of t_l and t are obtained after we rank the proteins in descending order by the values of a certain network topology attribute. In [23], Li selected the top 5% proteins as the essential proteins, so we selected the top 5% proteins as essential proteins, that is, t_l is the number of proteins in the l subcellular location from the top 5% proteins, and t is the number of the top 5% proteins. We think that the importance of the l subcellular location is directly proportional to the number of proteins from the top 5% in the l subcellular location. If $SLC(l)$ is greater than 0, it means that there are more essential proteins appearing in the l subcellular location. If $SLC(l)$ is less than 0, it means that essential proteins rarely appear in the l subcellular location.

Because LBCC is one of the most effective methods for identifying essential proteins, we select LBCC [17] to rank proteins in descending order. $LBCC(v)$ is defined as

$$LBCC(v) = \log Den_1(v) + 4 * \log Den_2(v) + 3 * \log IDC(v) + \log BC(v).$$

New computational strategy: CoTB

In this section, we propose our new computational strategy, CoTB. This strategy combines Den_1, Den_2, BC, LC, IDC, SLS and DOS. CoTB is based on the following basic concepts:

1. Den_1 and Den_2, which are two types of densities, represent the local properties of a PPI network. $Den_1(v)$ measures the density of the induced subgraph on the node set of node v and its neighbour nodes. $Den_2(v)$ measures the density of the induced subgraph on the node set of node v and nodes whose distance to node v is less than 3.
2. BC represents a global property of a PPI network. A node with a high BC will have more influence on the transfer of information through the network.
3. LC represents an intermediate attribute between the global and local properties used to measure the importance of a node, and it provides more structural information about the connectivity and density around a node.
4. IDC is another topological property that represents the protein complex information, and it has been proven that essential proteins are more likely to gather in protein complexes.
5. SLS is an intrinsic feature of a protein that represents a correlation between the position of a protein in the cell and the protein being essential.
6. DOS is also an intrinsic feature of a protein, and the larger the value is, the more important it is.

To take advantage of these seven attributes, we use the machine learning method random forest [28], which is an efficient method for investigating classification problems, to obtain the probability scores for predicting essential proteins. This method is implemented using the WEKA software package [29], and the number of generated trees is set to be 1000. We then use three of the four datasets as the training set and the remaining one as the testing set, which will be described in the following section. Finally, the proteins are sorted in descending order according to the values of the probability scores for proteins being essential.

Table 1. Information of the YDIP, YMIPS, YMBD, and YHQ datasets.

Dataset	Proteins	Interactions	Essential proteins
YDIP	5093	24743	1167
YMIPS	4546	12319	1016
YMBD	2559	11835	763
YHQ	4743	23294	1108

https://doi.org/10.1371/journal.pone.0182031.t001

Results and discussion

Experimental data

We performed experiments based on *Saccharomyces cerevisiae* data because its PPI and biological information data were more reliable and complete compared to those of other species, and it has also been widely used in the study of discovering essential proteins. We selected four different datasets from the DIP database [30], the MIPS database (Mammalian Protein-Protein Interaction Database) [31], and the website of the Mark Gerstein Lab (gersteinlab.org), which were denoted as YDIP (S1 Text), YMIPS (S2 Text), YMBD (S3 Text) and YHQ (S4 Text), respectively. The datasets of essential proteins (S1 Excel) were collected from the databases of DEG (Database of Essential Genes) [32], MIPS [31], SGD (Saccharomyces Genome Database) [33], and SGDP (Saccharomyces Genome Deletion Project) [34]. The datasets of protein

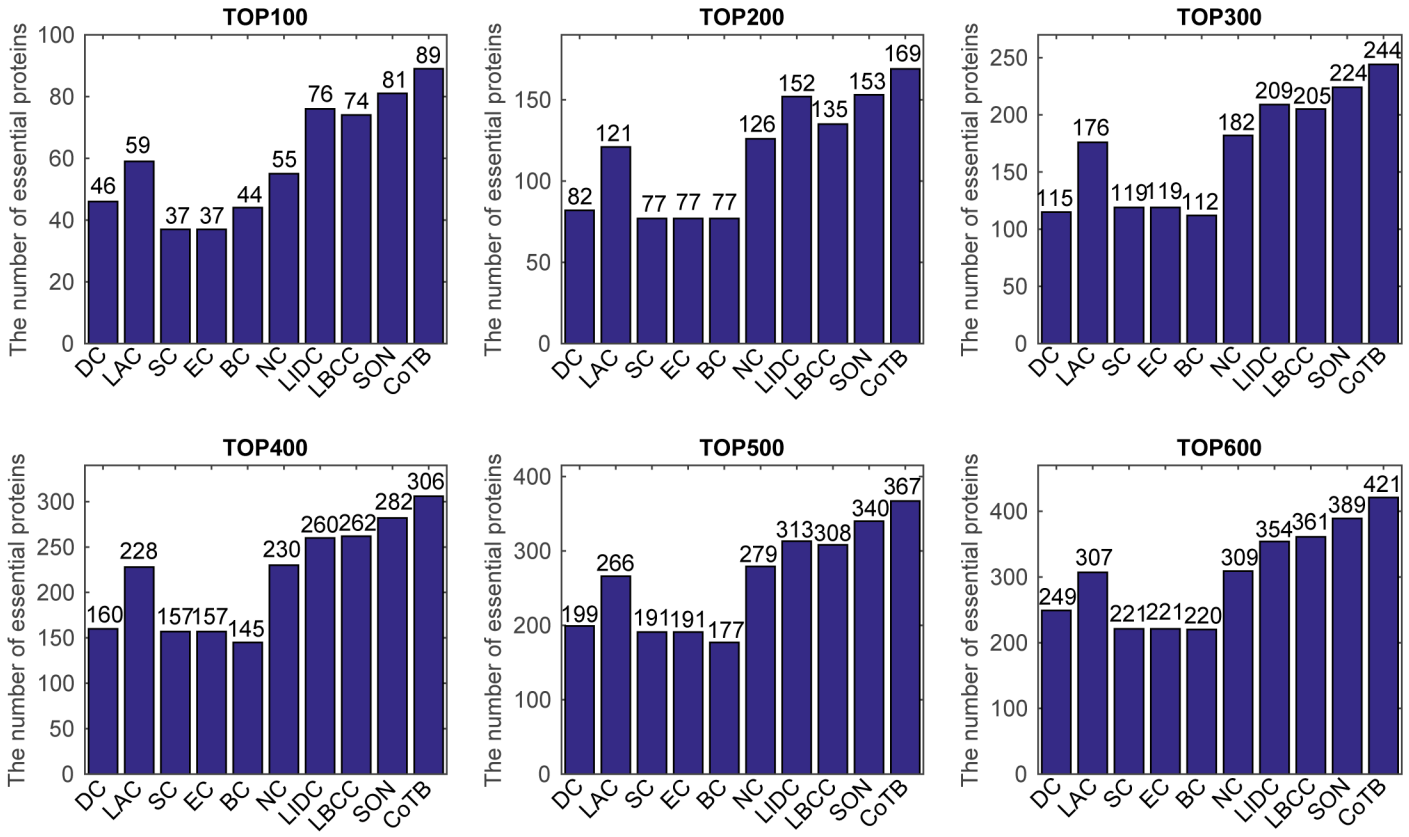

Fig 1. The number of essential proteins predicted by CoTB and the other nine methods at six levels for the YDIP network.

https://doi.org/10.1371/journal.pone.0182031.g001

complexes (S2 Excel) were collected from CM425 [35], CM270 [31], CYC428 and CYC408 [36, 37].

The YDIP dataset contains a total of 5093 proteins, 24743 edges and 1167 essential proteins. The YMIPS dataset contains 4546 proteins, 12319 edges and 1016 essential proteins. YMBD, which was collected from MIPS, BIND and DIP, includes 2559 proteins, 11835 interactions, and 763 essential proteins. The YHQ dataset was constructed by Yu et al. [38], and it contains 4743 proteins, 23294 interactions and 1108 essential proteins. The detailed information of the YDIP, YMIPS, YMBD and YHQ datasets are presented in Table 1.

The dataset of orthologous proteins was downloaded from the InParanoid database [39], and it contains 99 reference organisms of *Saccharomyces cerevisiae*. The subcellular localization dataset of *Saccharomyces cerevisiae* was downloaded from the COMPARTMENTS database [40]. After preprocessing, a total of 4849 different proteins remained, in which there were 1140 essential proteins and 11 different localizations, including cell wall, plasma membrane, cytosol, cytoskeleton, vacuole, peroxisome, Golgi apparatus, endosome, endoplasmic reticulum, nucleus, and mitochondrion.

Comparison with other prediction measures

In this section, we compare CoTB with several existing methods on the four datasets mentioned in the Experimental data section. The algorithms for LIDC, LBCC and SON were implemented according to [16, 17] and [23], respectively, and the other algorithms were

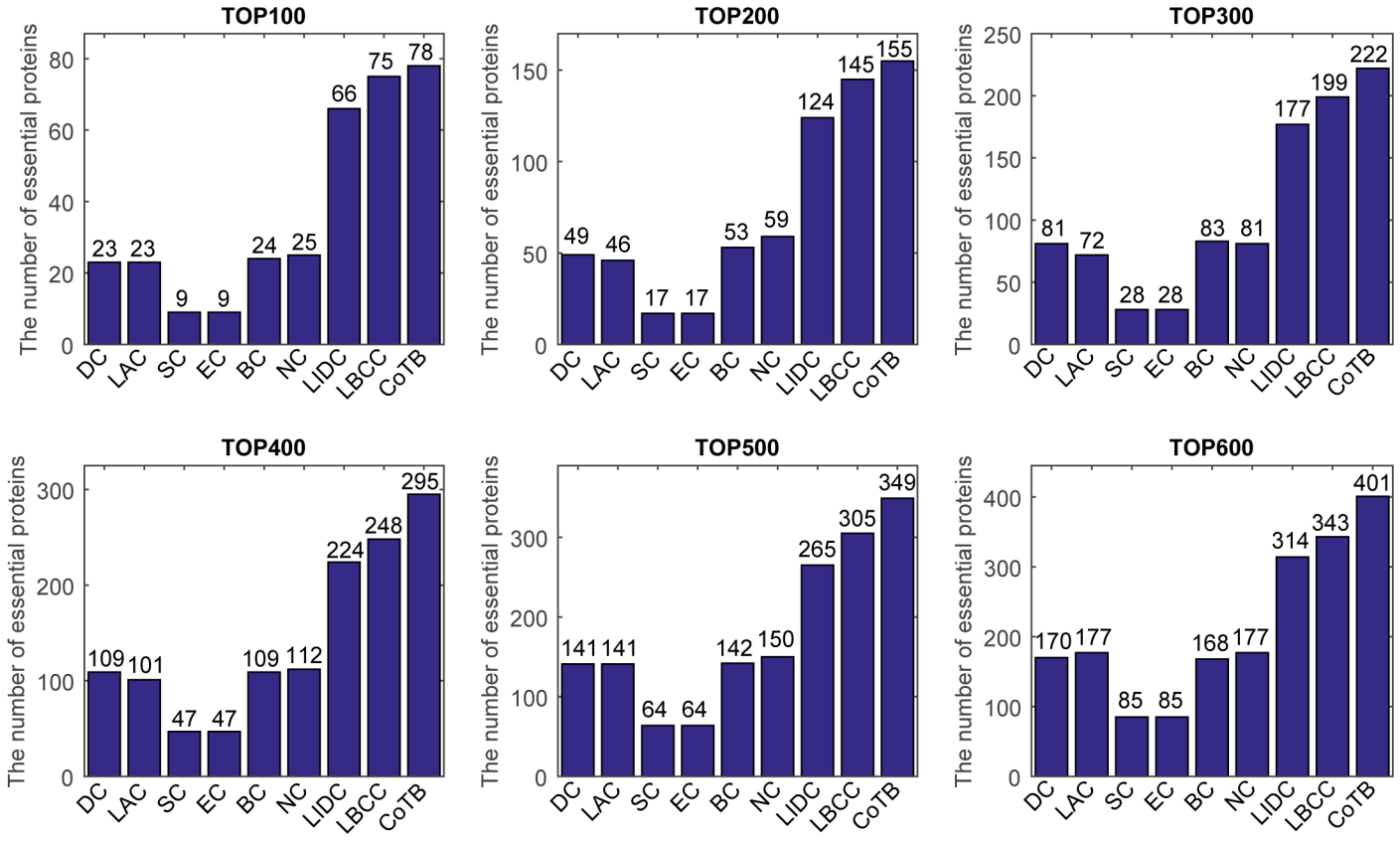

Fig 2. The number of essential proteins predicted by CoTB and the other eight methods at six levels for the YMIPS network.

https://doi.org/10.1371/journal.pone.0182031.g002

implemented using CytoNCA [41], which is a Cytoscape plugin for centrality analysis of biological networks. We selected three of the four datasets as the training set and the remaining one as the testing set, and we selected six levels from the top 100 to top 600 as candidate essential proteins.

The prediction results are shown in Fig 1 for when YDIP was considered as the testing set and the other three datasets were considered as the training set. CoTB improved the prediction precisions to approximately 89, 85, 82, 77, 74, and 71 percent at six levels. CoTB exhibited superior performance compared with the other methods, and it increased the prediction precisions by more than 20, 25, 19, 16, 19, and 16 percent at six levels compared with LBCC. Moreover, CoTB improved the prediction precisions by more than 9, 10, 8, 8, 7, and 8 percent at six levels compared to the most recently developed method, SON.

The prediction results are shown in Fig 2 for when YMIPS was considered as the testing set and the other three datasets were considered as the training set. CoTB improved the prediction precisions to approximately 78, 78, 74, 74, 70, and 67 percent at the six levels. CoTB achieved the best results compared to the other methods, and it increased the prediction precisions by more than 4, 6, 11, 18, 14, and 16 percent at the six levels compared to LBCC, which obtained the best results except for CoTB.

The prediction results are shown in Fig 3 for when YMBD was considered as the testing set and the other three datasets were considered as the training set. CoTB improved the prediction precisions to approximately 79, 75, 76, 74, 72, and 69 percent at six levels from the top 100 to

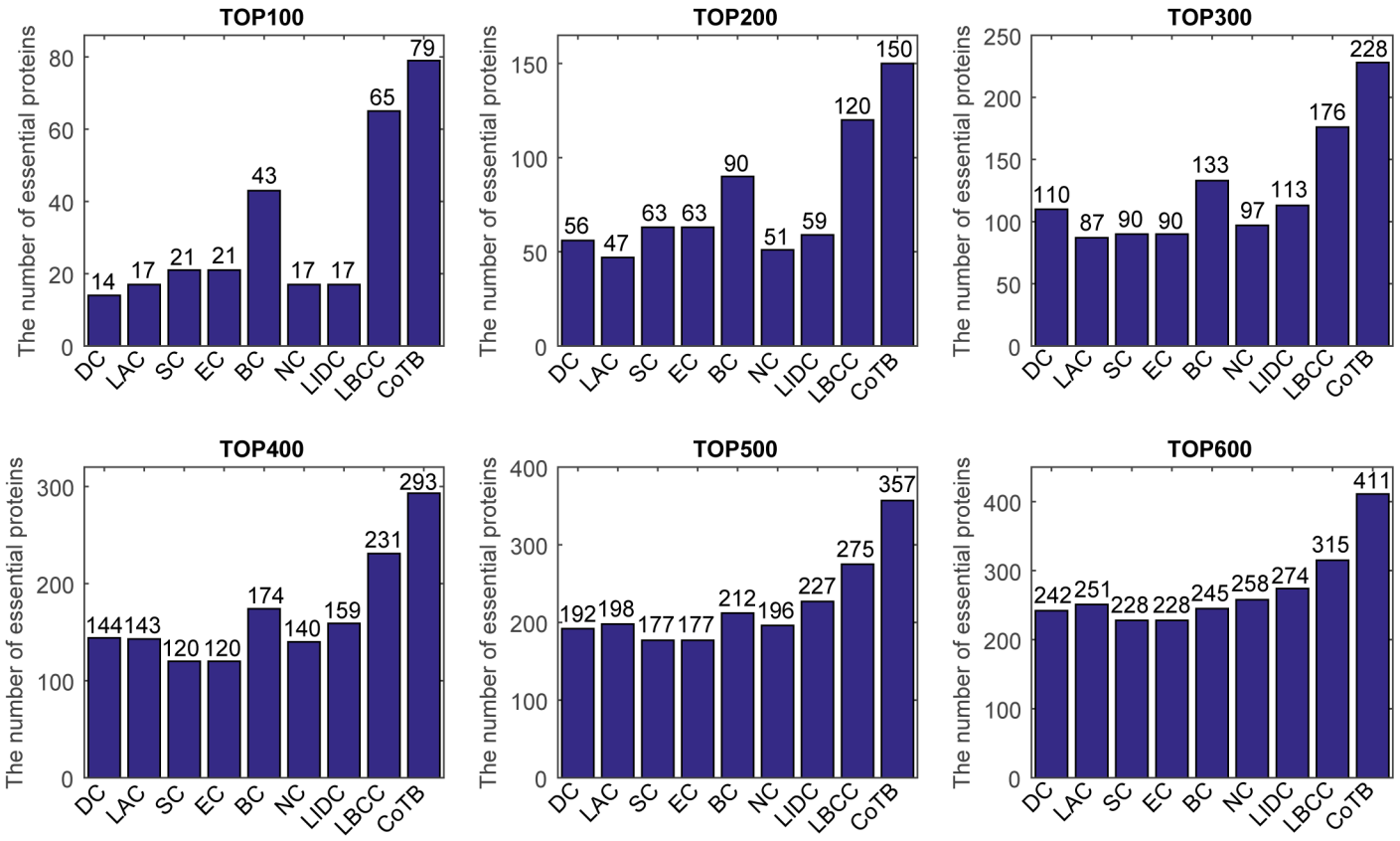

Fig 3. The number of essential proteins predicted by CoTB and the other eight methods at six levels for the YMBD network.

https://doi.org/10.1371/journal.pone.0182031.g003

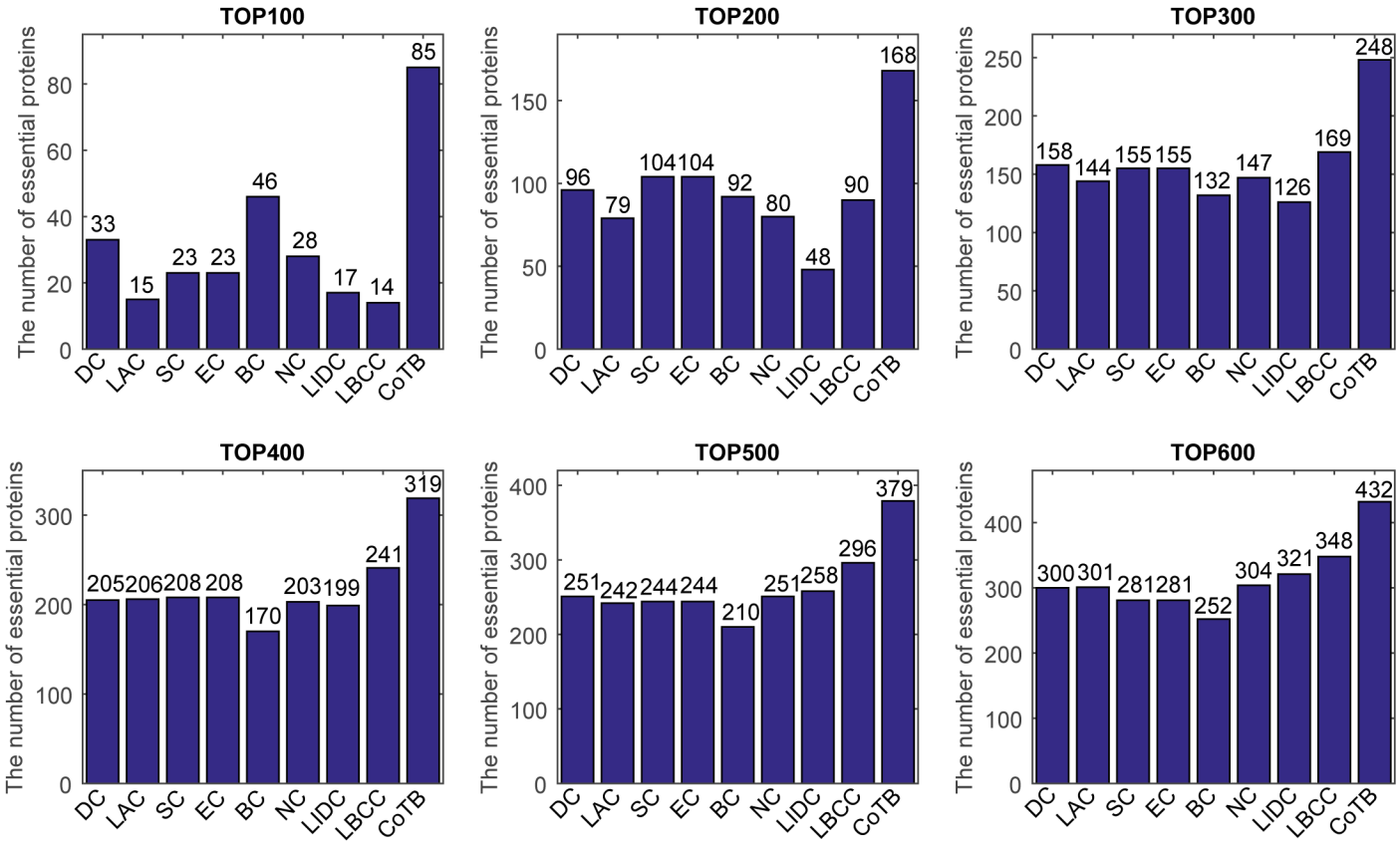

Fig 4. The number of essential proteins predicted by CoTB and the other eight methods at six levels for the YHQ network.

https://doi.org/10.1371/journal.pone.0182031.g004

top 600, respectively. CoTB obtained the best results, and it increased the prediction precisions by more than 21, 25, 29, 26, 29, and 30 percent at six levels compared to LBCC.

The prediction results are shown in Fig 4 for when YHQ was considered as the testing set and the other three datasets were considered as the training set. CoTB improved the prediction precisions to approximately 85, 84, 83, 80, 76, and 72 percent at six levels. Except for CoTB, the largest numbers of true essential proteins predicted at six levels from the top 100 to top 600 were 46 (BC), 104 (SC, EC), 169 (LBCC), 241 (LBCC), 296 (LBCC), and 348 (LBCC). CoTB increased the prediction precisions by more than 84, 61, 46, 32, 28, and 24 percent compared to the largest numbers at the six levels from the top 100 to top 600, respectively.

Validation using six statistical measures and precision-recall curves

In this section, we compare CoTB with the other methods using six statistical measures: sensitivity (SN), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), F-measure, and accuracy (ACC) (see references [13, 16]). Let TP be the number of essential proteins predicted as essential proteins, FP be the number of nonessential proteins predicted as essential proteins, TN be the number of nonessential proteins predicted as nonessential proteins, and FN be the number of essential proteins predicted as nonessential proteins. Then, the

Table 2. Comparative analysis of CoTB and the other methods in terms of SN, SP, PPV, NPV, F-measure, and ACC with four different testing datasets.

Dataset	Methods	SN	SP	PPV	NPV	F-measure	ACC
YDIP	DC	0.354	0.846	0.406	0.815	0.378	0.733
	LAC	0.405	0.861	0.465	0.830	0.433	0.757
	SC	0.323	0.837	0.370	0.806	0.345	0.719
	EC	0.323	0.837	0.370	0.806	0.345	0.719
	BC	0.308	0.832	0.354	0.802	0.330	0.712
	NC	0.398	0.859	0.456	0.827	0.425	0.753
	LIDC	0.446	0.873	0.511	0.841	0.476	0.775
	LBCC	0.446	0.873	0.512	0.841	0.477	0.776
	SON	0.497	0.888	0.570	0.856	0.531	0.799
	CoTB	**0.524**	**0.896**	**0.600**	**0.864**	**0.559**	**0.811**
YMIPS	DC	0.252	0.815	0.282	0.791	0.266	0.689
	LAC	0.269	0.820	0.300	0.796	0.284	0.697
	SC	0.139	0.782	0.155	0.759	0.146	0.639
	EC	0.139	0.782	0.155	0.759	0.146	0.639
	BC	0.249	0.814	0.278	0.790	0.263	0.688
	NC	0.281	0.824	0.315	0.799	0.297	0.702
	LIDC	0.423	0.864	0.473	0.839	0.447	0.766
	LBCC	0.430	0.866	0.481	0.841	0.454	0.769
	CoTB	**0.552**	**0.901**	**0.617**	**0.875**	**0.583**	**0.823**
YMBD	DC	0.260	0.825	0.387	0.724	0.311	0.657
	LAC	0.271	0.830	0.404	0.728	0.325	0.664
	SC	0.239	0.816	0.355	0.716	0.285	0.644
	EC	0.239	0.816	0.355	0.716	0.285	0.644
	BC	0.283	0.835	0.422	0.733	0.339	0.671
	NC	0.266	0.828	0.396	0.726	0.318	0.660
	LIDC	0.308	0.846	0.459	0.742	0.369	0.685
	LBCC	0.372	0.873	0.555	0.766	0.445	0.724
	CoTB	**0.480**	**0.919**	**0.715**	**0.806**	**0.574**	**0.788**
YHQ	DC	0.401	0.861	0.468	0.825	0.432	0.754
	LAC	0.431	0.870	0.504	0.834	0.465	0.768
	SC	0.326	0.838	0.380	0.803	0.351	0.719
	EC	0.326	0.838	0.380	0.803	0.351	0.719
	BC	0.330	0.840	0.386	0.804	0.356	0.721
	NC	0.426	0.869	0.497	0.832	0.459	0.765
	LIDC	0.449	0.876	0.524	0.839	0.483	0.776
	LBCC	0.449	0.876	0.524	0.839	0.483	0.776
	CoTB	**0.520**	**0.897**	**0.607**	**0.860**	**0.560**	**0.809**

https://doi.org/10.1371/journal.pone.0182031.t002

six statistical measures are defined as follows:

$$SN = \frac{TP}{TP + FN},$$

$$SP = \frac{TN}{TN + FP},$$

$$PPV = \frac{TP}{TP + FP},$$

$$NPV = \frac{TN}{TN + FN},$$

$$F - measure = \frac{2 * SN * PPV}{SN + PPV},$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}.$$

We sorted the proteins in descending order according to the values of the corresponding measures and chose the top 20 percent proteins as essential proteins, and the other proteins were considered to be nonessential proteins. The results are presented in Table 2, which shows that the values of the six statistical measures for CoTB are consistently higher than those of the other methods on four networks, and CoTB improved the values of *SN*, *SP*, *PPV*, *NPV*, *F – measure*, and *ACC* by more than 5.4, 0.9, 5.2, 0.9, 5.2, and 1.5 percent compared to SON on the YDIP dataset.

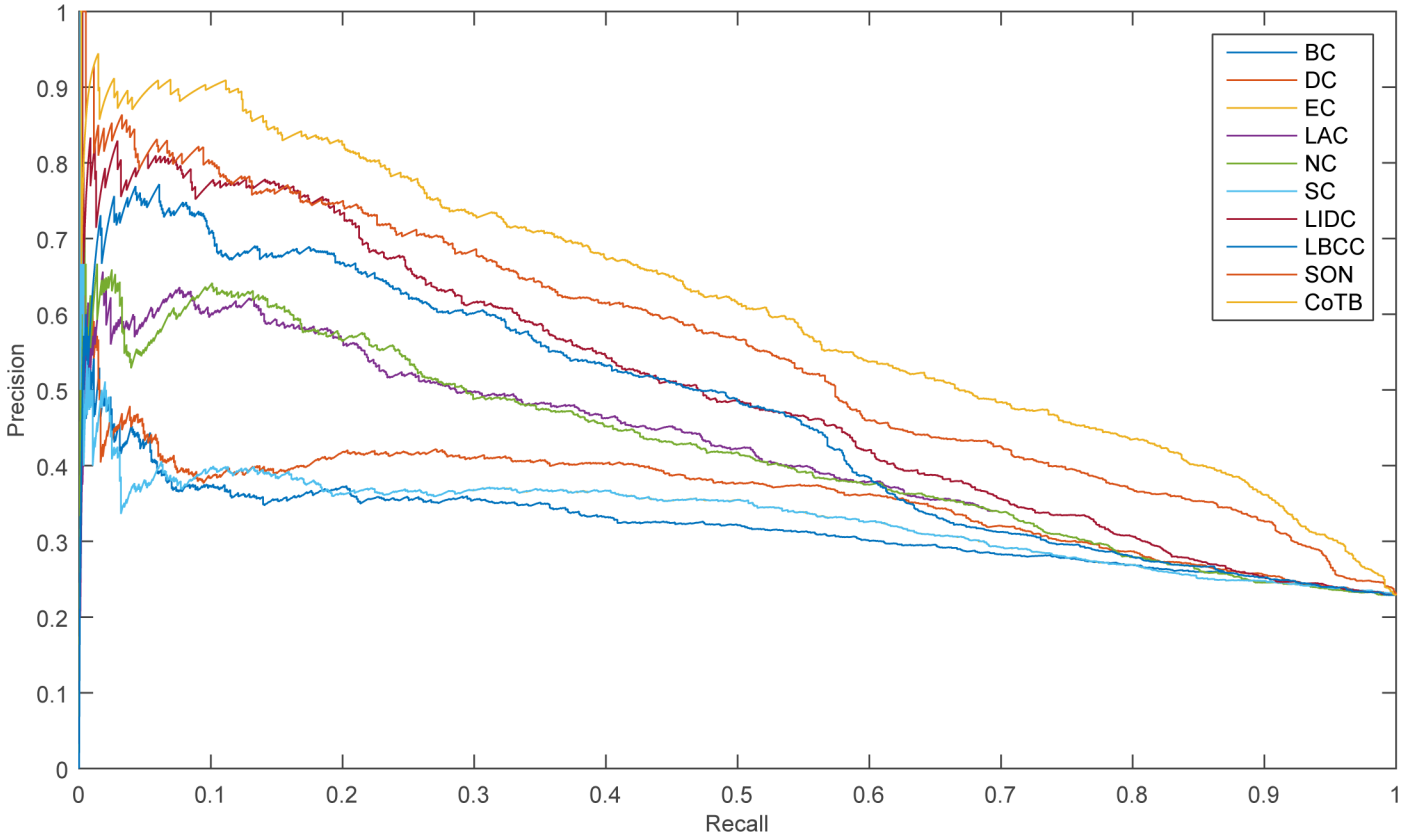

Fig 5. PR curves of CoTB and the other methods for the YDIP network.

https://doi.org/10.1371/journal.pone.0182031.g005

The precision-recall curve is used for assessing the stability of the methods, and it is obtained by plotting

$$Precision(n) = \frac{TP(n)}{TP(n) + FP(n)},$$

$$Recall(n) = \frac{TP(n)}{P},$$

where $TP(n)$ is the number of true essential proteins identified correctly and $FP(n)$ is the number of true essential proteins identified incorrectly among the top n proteins, and P is the number of true essential proteins in total. The results are shown in Figs 5–8. As shown, CoTB performed significantly better than SON, LBCC, and the other methods.

Validation using jackknife methodology

To further investigate the performance of CoTB, we used the jackknife methodology to assess the generality of our method. The x-axis represents the number of proteins ranked in descending order according to the values computed by the corresponding methods, and the y-axis represents the cumulative count of true essential proteins. The area under the curve is always used to measure the generality of a method. As shown in Figs 9–12, CoTB clearly performs better than BC, DC, EC, LAC, NC, SC, LIDC and LBCC on the four datasets, and it also performs better than SON on the YDIP dataset.

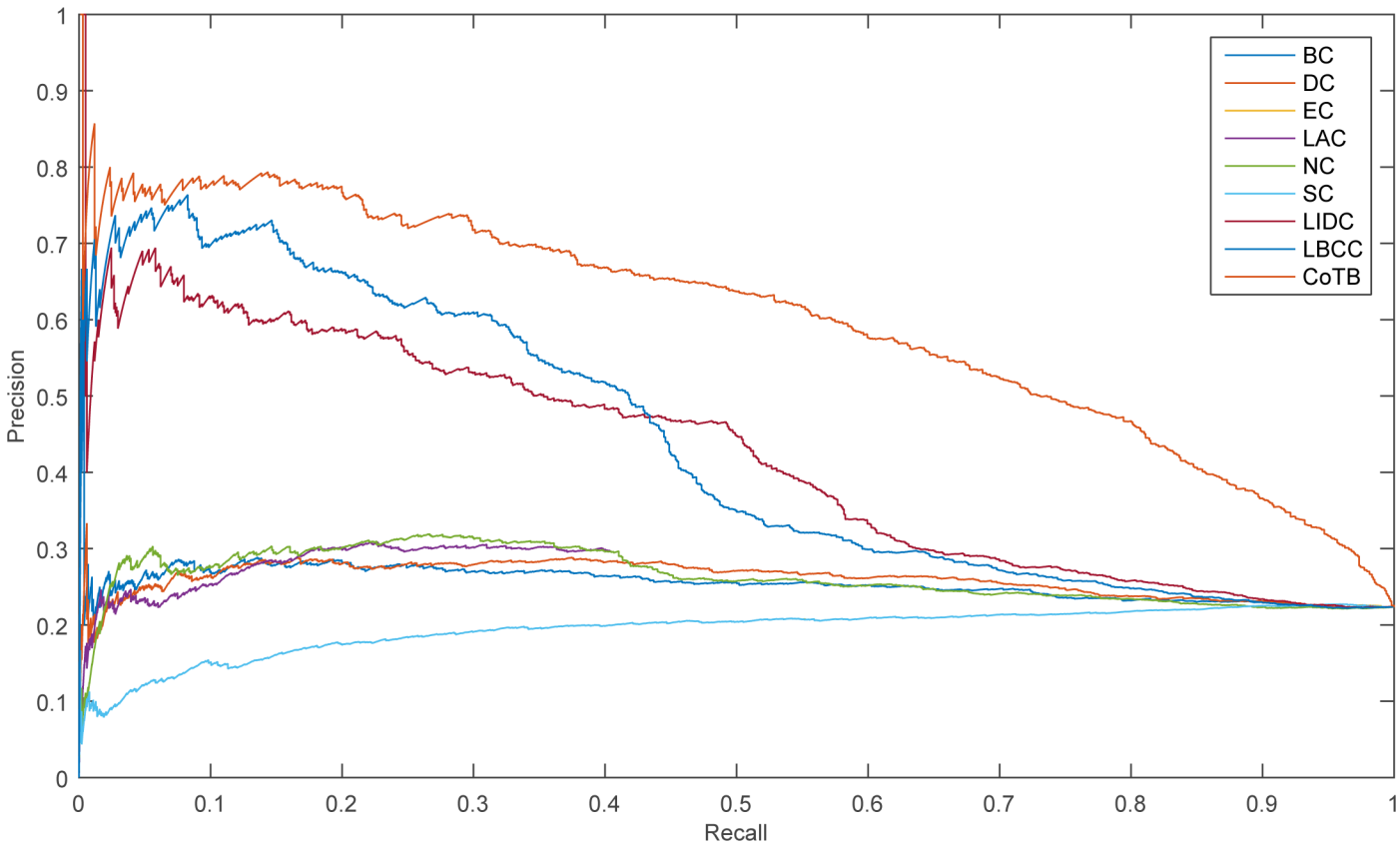

Fig 6. PR curves of CoTB and the other methods for the YMIPS network.

https://doi.org/10.1371/journal.pone.0182031.g006

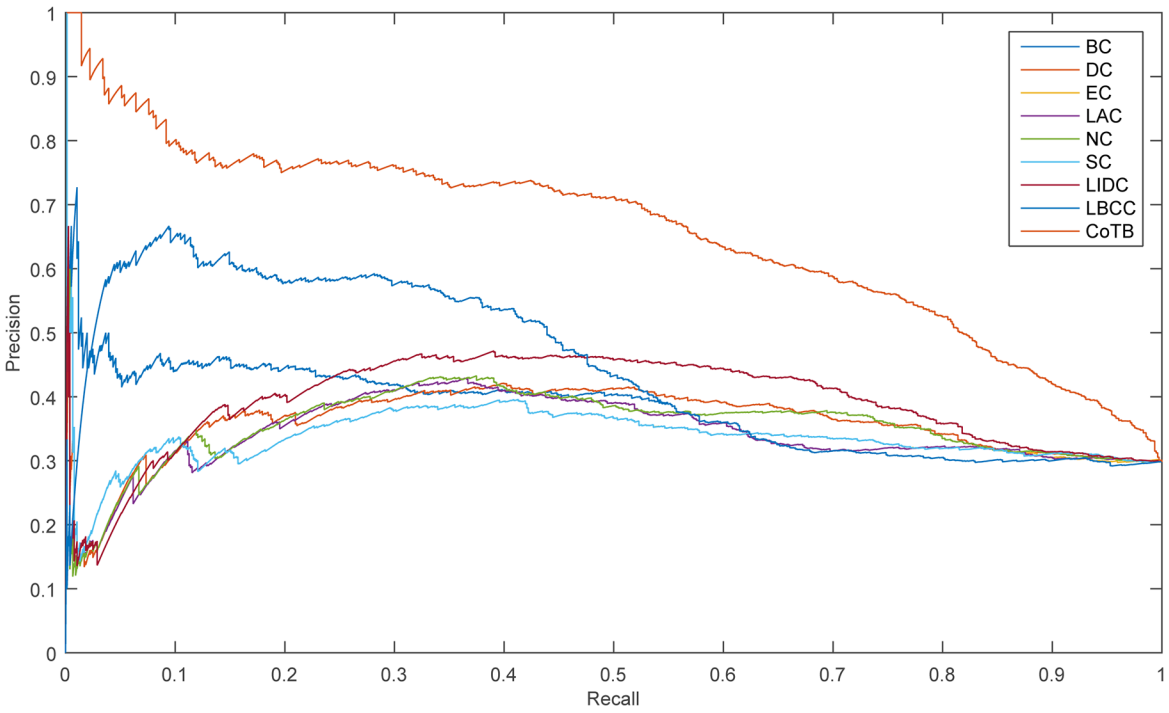

Fig 7. PR curves of CoTB and the other methods for the YMBD network.

https://doi.org/10.1371/journal.pone.0182031.g007

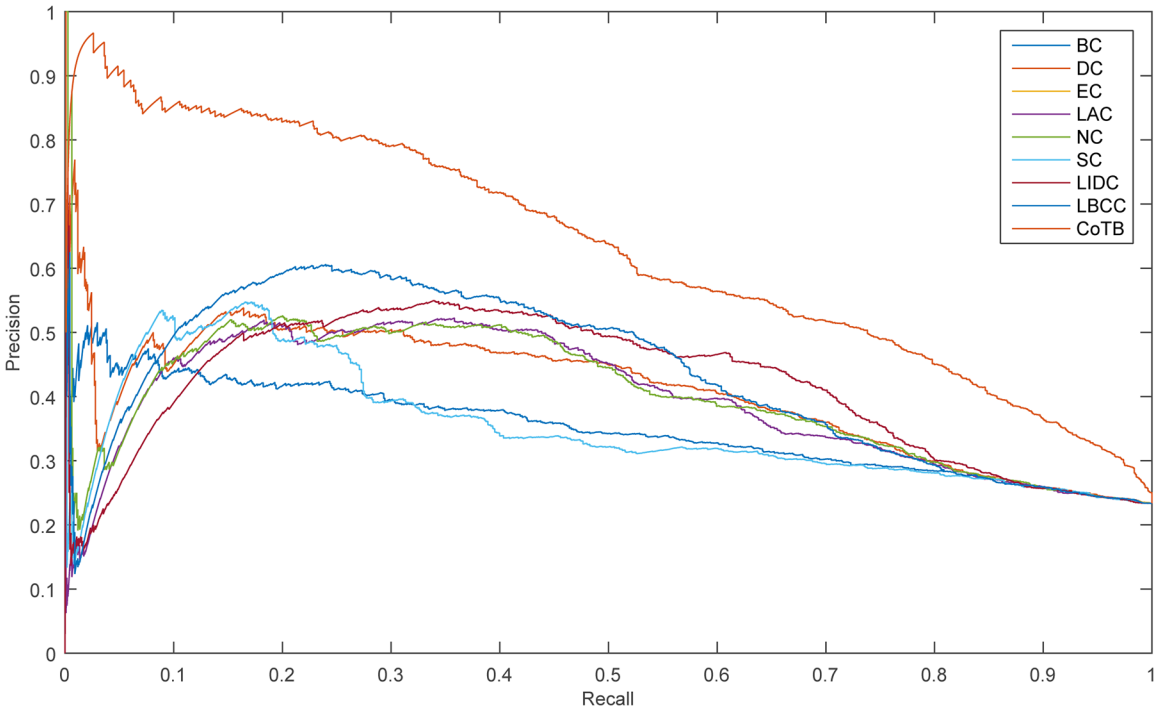

Fig 8. PR curves of CoTB and the other methods for the YHQ network.

https://doi.org/10.1371/journal.pone.0182031.g008

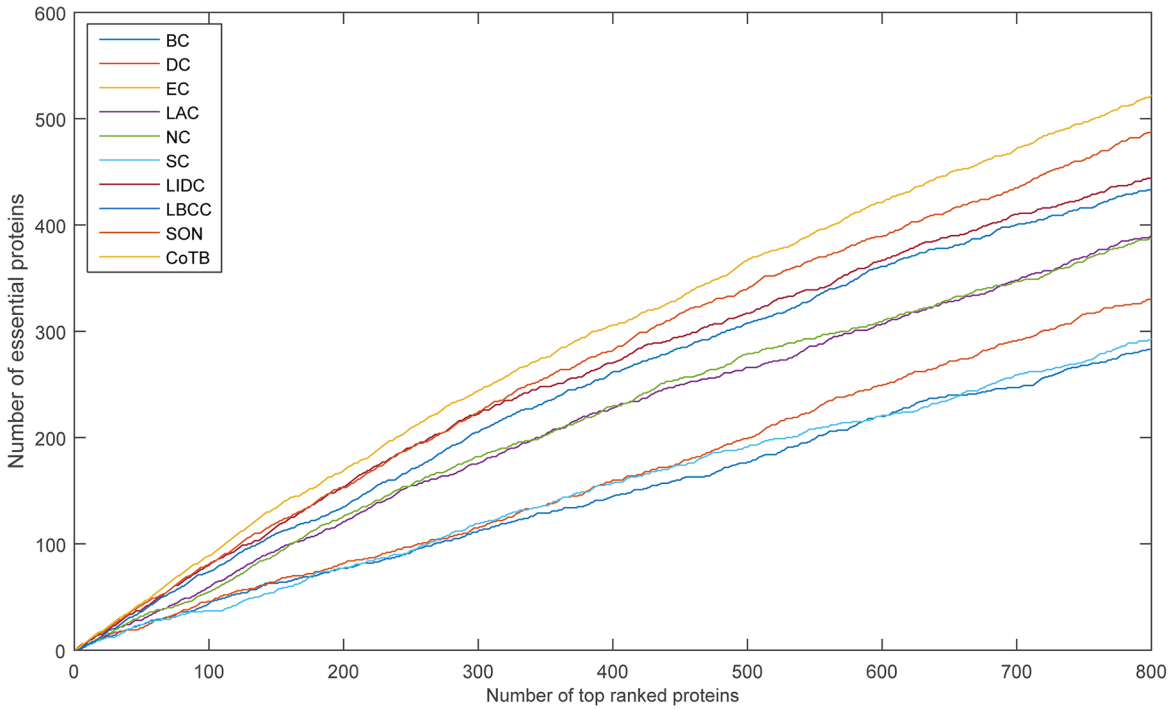

Fig 9. Jackknife curves of CoTB and the other methods for the YDIP network.

https://doi.org/10.1371/journal.pone.0182031.g009

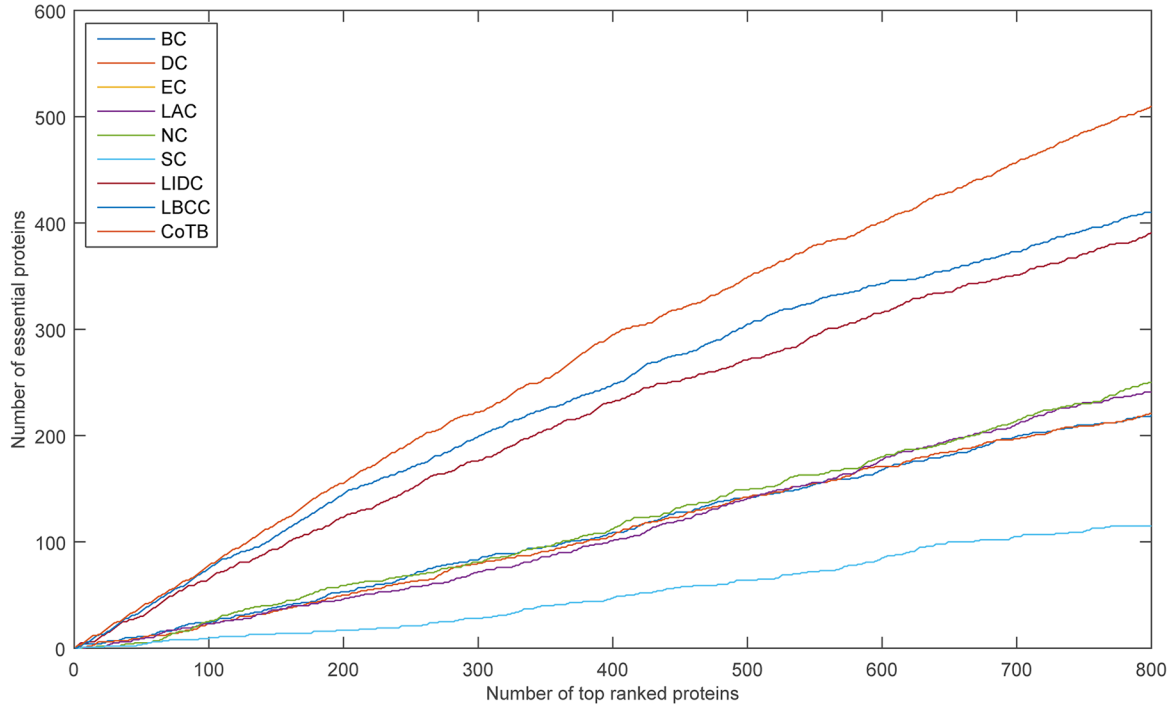

Fig 10. Jackknife curves of CoTB and the other methods for the YMIPS network.

https://doi.org/10.1371/journal.pone.0182031.g010

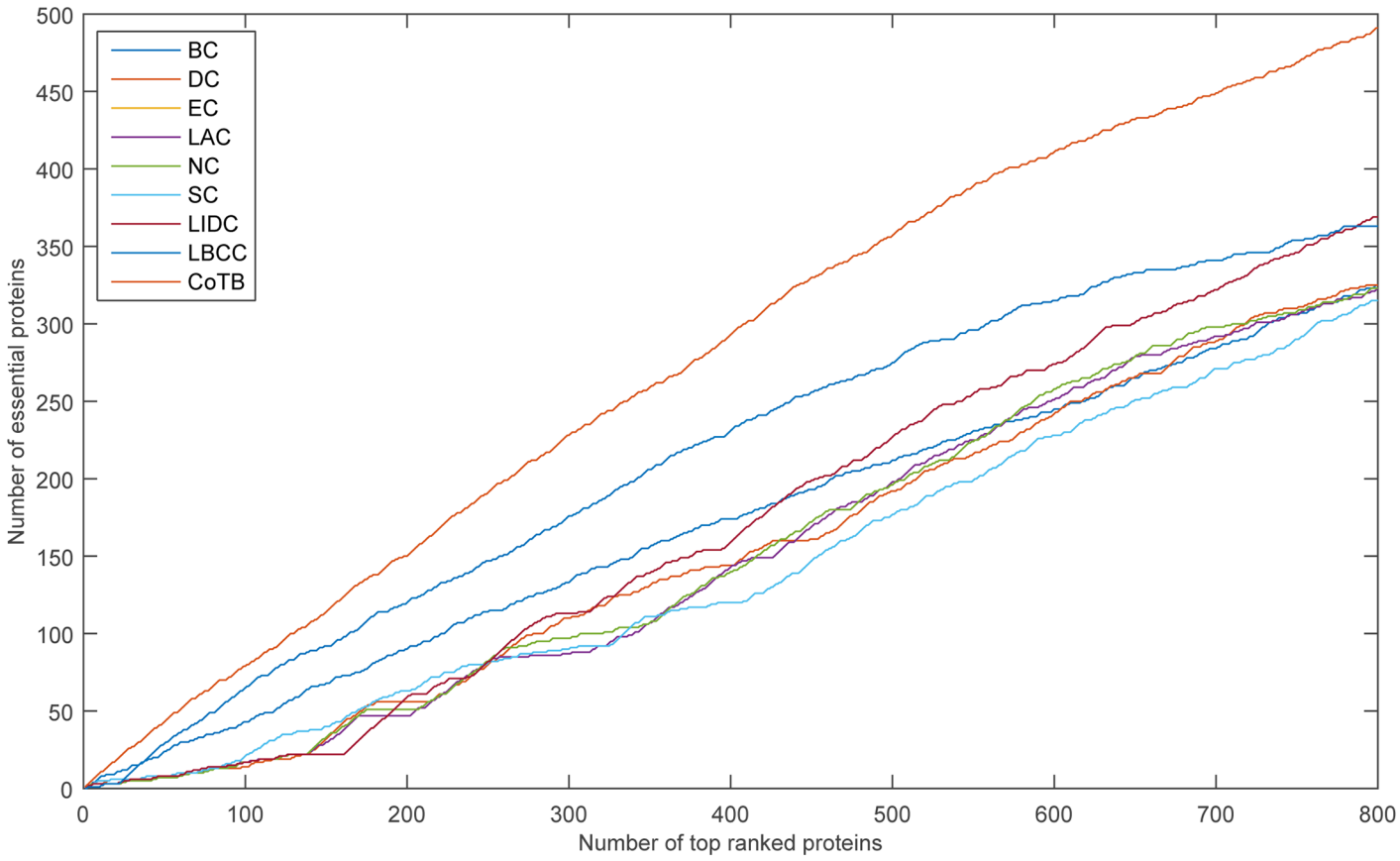

Fig 11. Jackknife curves of CoTB and the other methods for the YMBD network.

https://doi.org/10.1371/journal.pone.0182031.g011

Differences between CoTB and the other nine existing methods

To further analyse the differences between CoTB and the other nine existing methods, we compared the performances of the methods in predicting the top 100 proteins ranked by the corresponding methods on the YDIP dataset (see the supplementary S3 Excel). The overlapping rates of proteins predicted by CoTB and the other nine methods are presented in Table 3. Compared to the traditional methods that use topological features, such as BC, DC, EC, LAC, NC and SC, the overlapping rates are less than 19 percent. Compared to LIDC, LBCC and SON, the overlapping rates are 46, 37 and 25 percent, respectively. It is clear that CoTB significantly differs from the traditional methods, and it takes more biological knowledge into account, which helps locate essential proteins more accurately and stably.

Subsequently, we analysed the top 100 proteins identified by LIDC, LBCC, SON and CoTB on the YDIP dataset. For LIDC and CoTB, 46 of the same proteins are identified by these methods, and for the remaining 54 proteins, CoTB identified 45 true essential proteins, whereas LIDC identified 36 true essential proteins. For LBCC and CoTB, 37 of the same

Table 3. The overlapping rates of the proteins predicted by CoTB and the other nine methods for the YDIP network.

Method	BC	DC	EC	LAC	NC	SC	LIDC	LBCC	SON
Overlapping rate	3%	4%	3%	19%	13%	3%	46%	37%	25%

https://doi.org/10.1371/journal.pone.0182031.t003

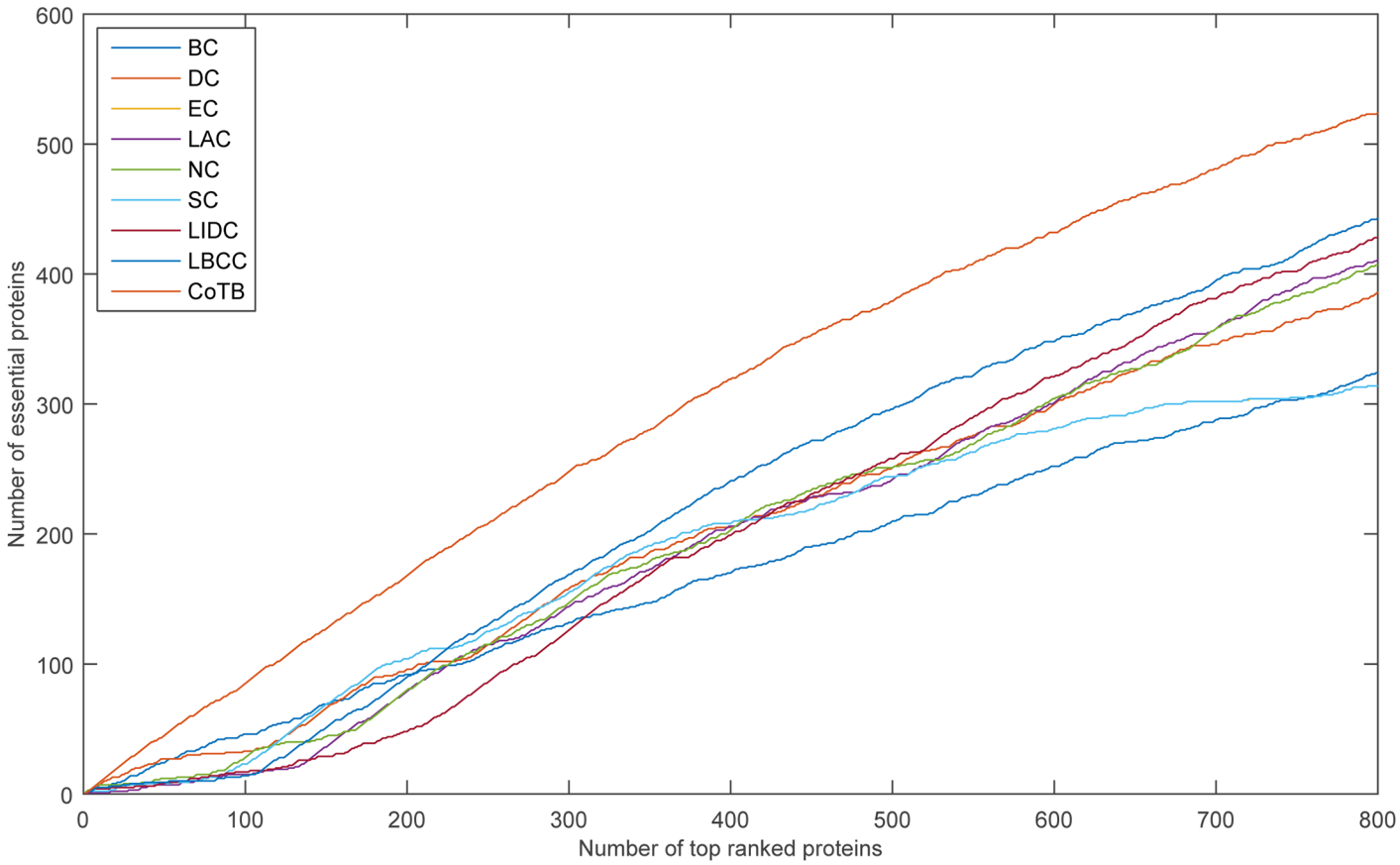

Fig 12. Jackknife curves of CoTB and the other methods for the YHQ network.

https://doi.org/10.1371/journal.pone.0182031.g012

proteins are identified by these methods, and for the remaining 63 proteins, CoTB identified 54 true essential proteins, whereas LBCC identified 39 true essential proteins. For SON and CoTB, 25 of the same proteins are identified by these methods, and for the remaining 75 proteins, CoTB identified 66 true essential proteins, whereas SON identified 58 true essential proteins.

Therefore, the comparative results demonstrate that CoTB is distinctly different from the other methods, and it can identify more true essential proteins.

Moreover, we also conducted experiments on each of the four datasets (YDIP, YMIPS, YMBD, and YHQ) through 10-fold cross-validation. The average AUC values are listed in Table 4. For the YDIP dataset, we have listed the AUC values of GEP and Acencio as

Table 4. Comparison of CoTB with other methods.

Dataset	Methods	AUC
YDIP	**CoTB**	**0.788**
	GEP	0.773
	Acencio	0.778
YMIPS	CoTB	0.808
YMBD	CoTB	0.788
YHQ	CoTB	0.802

https://doi.org/10.1371/journal.pone.0182031.t004

mentioned in the paper [42]. CoTB obtains the best performance among GEP and Acencio. For the other datasets, we have listed the average AUC of CoTB. The results further demonstrate that CoTB is an effective method for identifying essential proteins.

Results on human PPI network

To further assess the performance of the CoTB method, we also conducted experiments on a human PPI network. The human PPI network data, denoted HDIP, were downloaded from the DIP database [30]. The protein complex set, denoted HCOM, was downloaded from CORUM [43]. The essential proteins were downloaded from DEG [32]. The dataset of orthologous proteins was downloaded from the InParanoid database [39] containing 71 reference organisms. Finally, the subcellular localization information was downloaded from the COMPARTMENTS database [40]. HDIP consists of 4647 interactions and 2914 proteins, including 1887 essential proteins, and HCOM contains 1283 protein complexes.

We used the four different YDIP, YMIPS, YMBD and YHQ datasets as the training set and the HDIP dataset as the testing set. First, we compared the performances of CoTB and the other eight methods at six levels from the top 100 to top 600. As shown in Fig 13, CoTB achieved the best results at the top 300-500 levels and exhibited performance similar to that of the methods attaining the best results at the top 100, 200 and 600 levels.

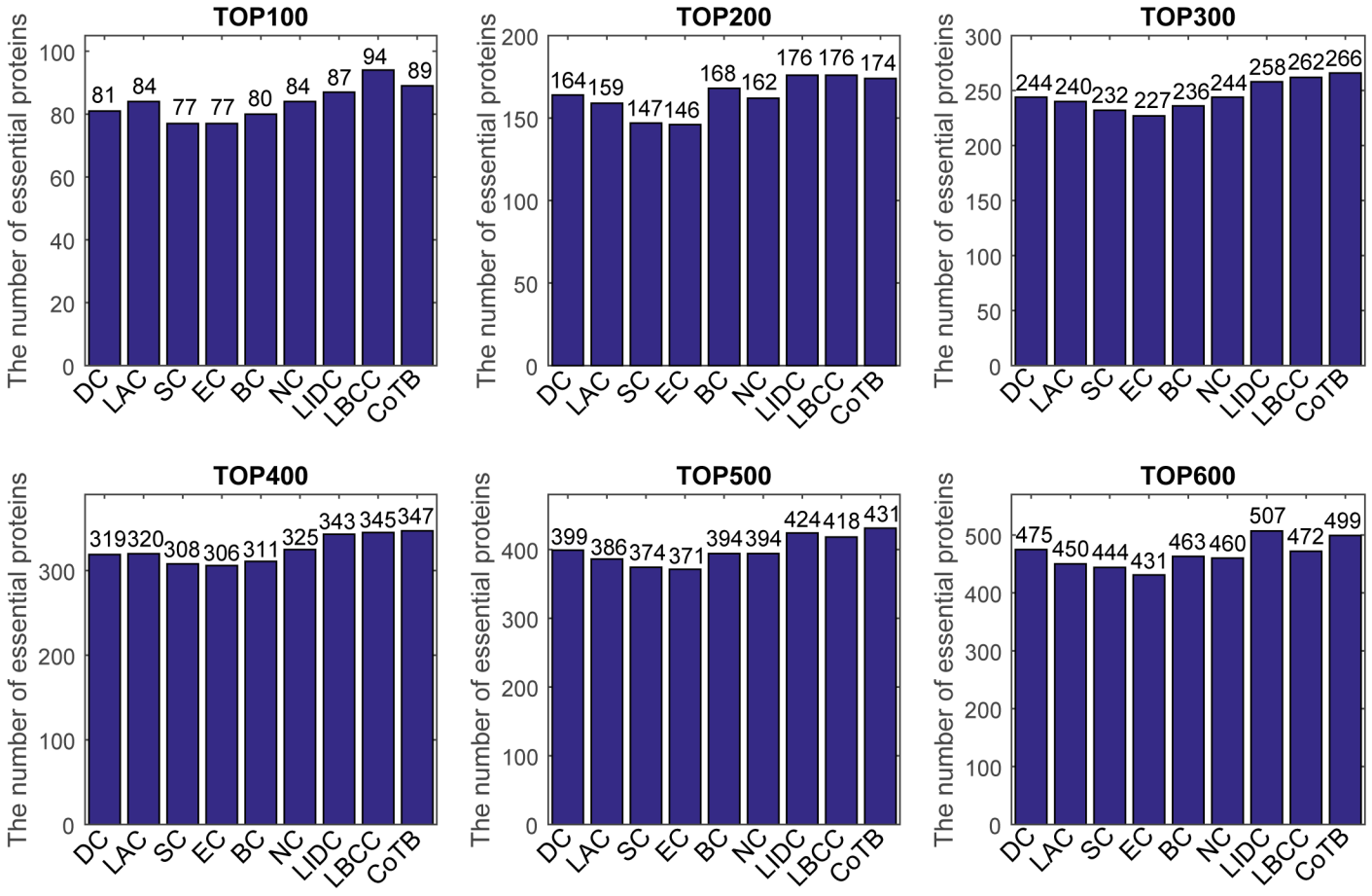

Fig 13. The number of essential proteins predicted by CoTB and the other eight methods at six levels for the HDIP network.

https://doi.org/10.1371/journal.pone.0182031.g013

Table 5. Comparative analysis of CoTB and the other methods in terms of SN, SP, PPV, NPV, F-measure, and ACC on the HDIP dataset.

Dataset	Methods	SN	SP	PPV	NPV	F-measure	ACC
HDIP	DC	0.244	0.882	0.792	0.389	0.373	0.469
	LAC	0.232	0.860	0.753	0.379	0.355	0.453
	SC	0.230	0.856	0.746	0.377	0.352	0.451
	EC	0.223	0.843	0.723	0.371	0.341	0.442
	BC	0.240	0.873	0.777	0.385	0.366	0.463
	NC	0.235	0.866	0.763	0.381	0.360	0.457
	LIDC	**0.262**	**0.914**	**0.849**	**0.403**	**0.400**	**0.492**
	LBCC	0.245	0.884	0.796	0.389	0.375	0.470
	CoTB	0.257	0.906	0.833	0.399	0.393	0.486

https://doi.org/10.1371/journal.pone.0182031.t005

Then, we used six statistical measures, precision-recall curves and jackknife curves to evaluate the performance of the proposed CoTB method and the other eight methods. As shown in Table 5, the values of the six statistical measures for CoTB were slightly lower than for LIDC. From the precision-recall curves shown in Fig 14, CoTB obtained better performance between the recall levels of 0.11-0.22 and 0.38-0.82. From the jackknife curves shown in Fig 15, CoTB exhibited performance similar to that of LIDC and LBCC before the top 280 and achieved

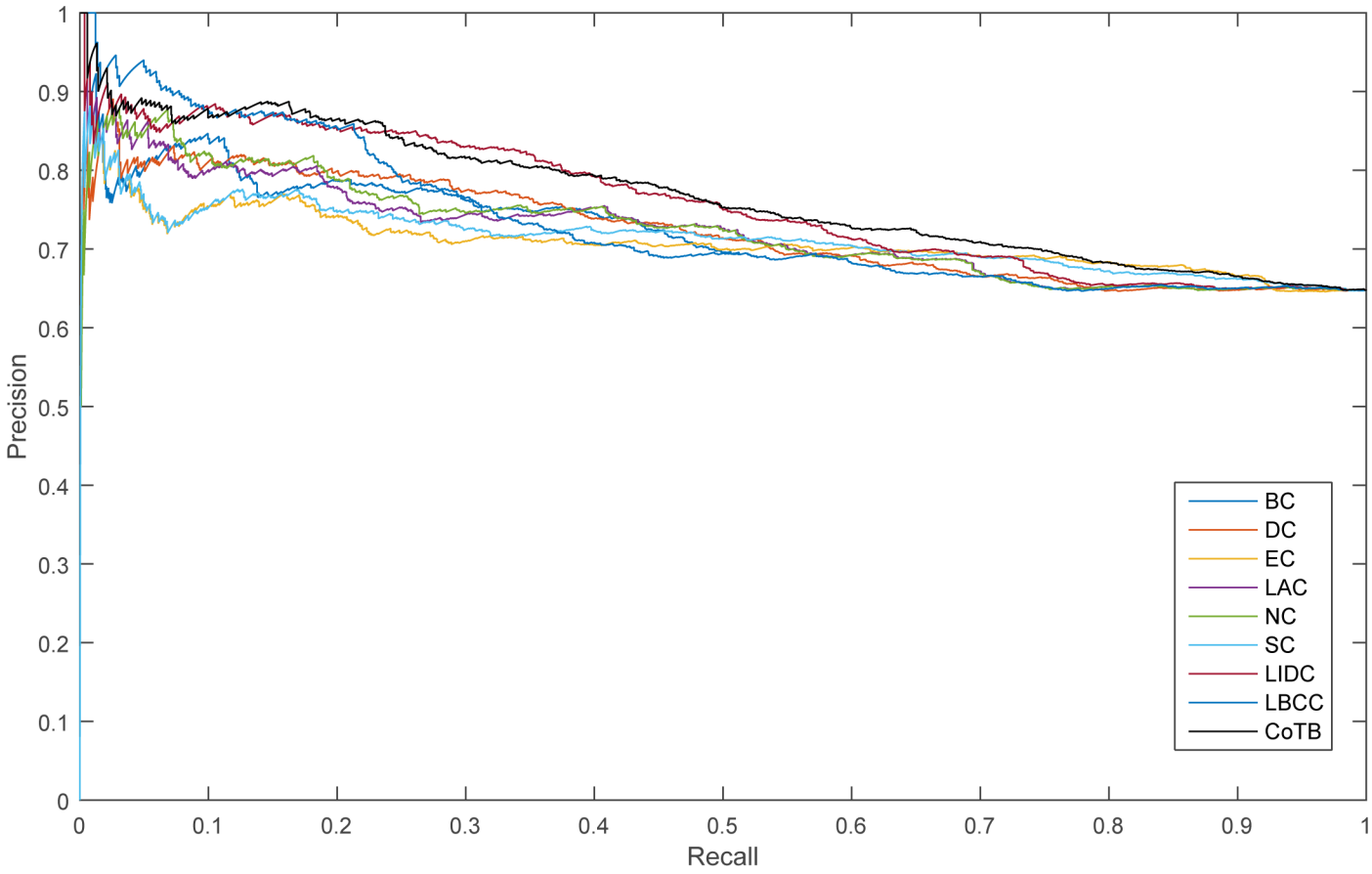

Fig 14. PR curves of CoTB and the other eight methods for the HDIP network.

https://doi.org/10.1371/journal.pone.0182031.g014

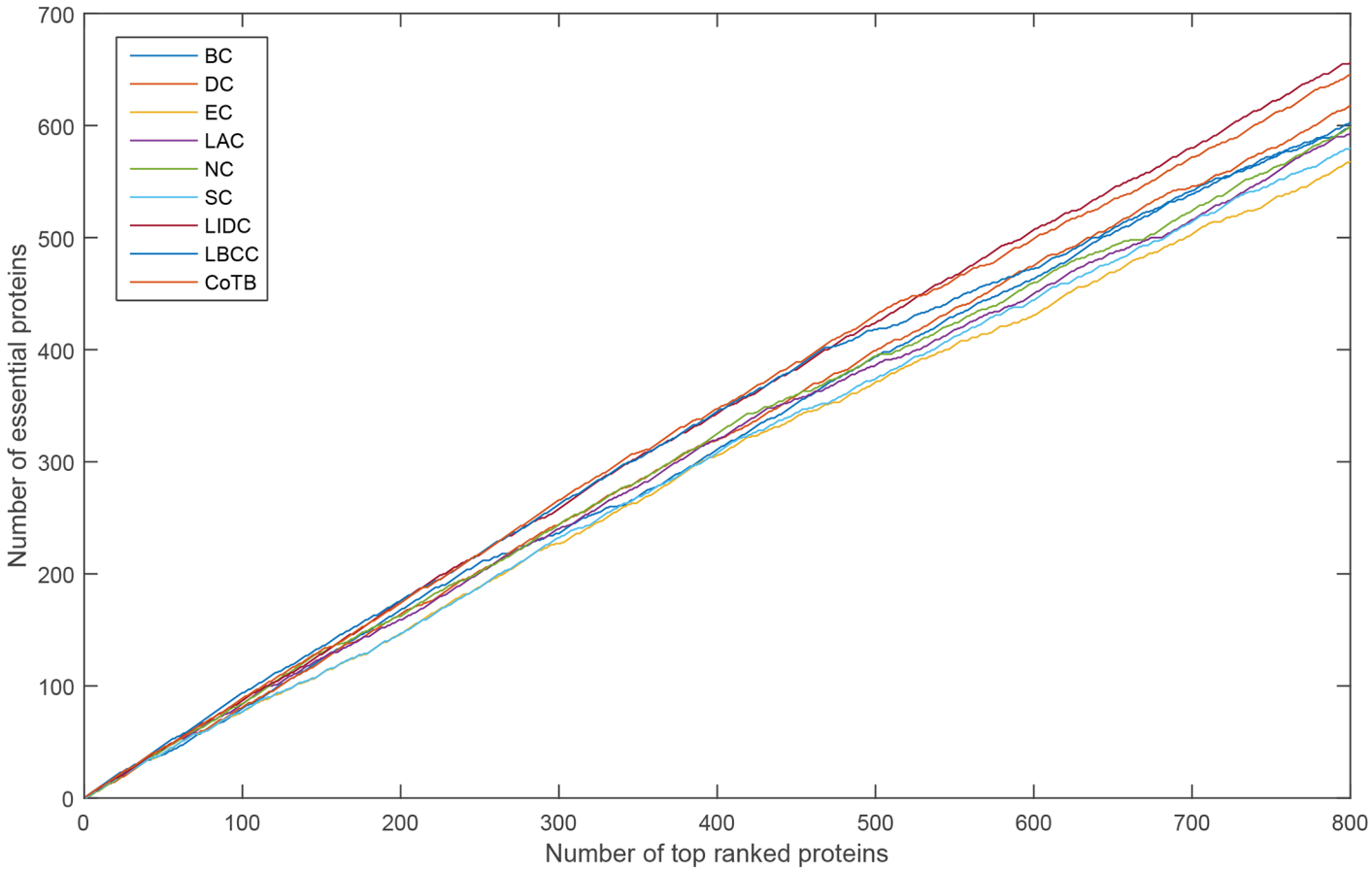

Fig 15. Jackknife curves of CoTB and the other eight methods for the HDIP network.

https://doi.org/10.1371/journal.pone.0182031.g015

better performance between the top 280 to top 530. Hence, CoTB is also an effective method for discovering essential proteins for the human PPI network HDIP.

Conclusion

Identifying essential proteins is of great importance for understanding the molecular mechanisms of cellular life. Many computational methods combined with biological information have recently been proposed for this purpose. In 2016, Qin [17] proposed a method named LBCC based on the combination of topological features and protein complex information. This method improved the prediction accuracy to 74 percent on the YDIP dataset. Li [23] proposed a method named SON that uses a combination of topological features and biological information. This method improved the prediction accuracy to 81 percent on the YDIP dataset.

In this paper, we propose a new computational strategy named CoTB to predict essential proteins. First, we introduce several topological properties. Second, we propose new measures of orthologous score (DOS) and subcellular localization score (SLS), as well as a new computational strategy that combines Den_1, Den_2, BC, IDC, LC, DOS and SLS and uses a random forest prediction model to obtain a probability score for the proteins being essential. Finally, we apply CoTB on four networks of *Saccharomyces cerevisiae* and perform comprehensive comparisons of CoTB with nine other previously proposed methods. The results at six levels from

the top 100 to top 600 demonstrate that our new method, CoTB, is more accurate than the other methods. Compared to the recently developed method SON, CoTB improves the prediction precisions by more than 9, 10, 8, 8, 7, and 8 percent at six levels on the YDIP dataset. Compared to the recently developed method LBCC, CoTB increases the prediction precisions by more than 4, 6, 11, 18, 14, and 16 percent at six levels on the other three datasets. In particular, CoTB improves the prediction precisions to 89, 78, 79, and 85 percent at the top 100 level on the YDIP, YMIPS, YMBD, and YHQ datasets, respectively. From the analysis of the six statistical measures, PR curves and jackknife curves, we find that CoTB is significantly superior to the other methods. Moreover, we also applied CoTB to a human PPI network, HDIP. The experimental results show that CoTB is also an effective method for predicting essential proteins for the HDIP network. There are two reasons leading to the outstanding performance of CoTB: the first reason is that it combines topological properties (both local and global properties) unlike SON and biological information (both subcellular localization and orthologous proteins) unlike LBCC. The second reason is that the machine learning method, random forest, plays an important role in the process of using these attributes to predict essential proteins. In conclusion, CoTB is a more effective, stable, and accurate method for predicting essential proteins.

Supporting information

S1 Text. Protein interaction data in YDIP.
(TXT)

S2 Text. Protein interaction data in YMIPS.
(TXT)

S3 Text. Protein interaction data in YMBD.
(TXT)

S4 Text. Protein interaction data in YHQ.
(TXT)

S1 Excel. Essential protein and nonessential protein data.
(XLS)

S2 Excel. Protein complex data.
(XLSX)

S3 Excel. The top 100 proteins obtained by the corresponding methods.
(XLSX)

Acknowledgments

We would like to thank the the editors and referees for their valuable comments and suggestions, which led to the improvement of the present version.

Author Contributions

Conceptualization: Chao Qin.

Data curation: Chao Qin, Yadong Dong.

Formal analysis: Chao Qin, Yongqi Sun, Yadong Dong.

Funding acquisition: Yongqi Sun.

Investigation: Chao Qin.

Methodology: Chao Qin, Yadong Dong.

Project administration: Yongqi Sun.

Supervision: Yongqi Sun.

Validation: Chao Qin.

Writing – original draft: Chao Qin.

Writing – review & editing: Chao Qin, Yongqi Sun, Yadong Dong.

References

1. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, et al. Functional Characterization of the S. cerevisiae Genome by Gene Deletion and Parallel Analysis. Science. 1999; 285(5429):901–906.
2. Furney SJ, Mar AM, Lópezbigas N. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics. 2006; 7(1):1–11.
3. Li M, Zheng R, Li Q, Wang J, Wu FX, Zhang Z. Prioritizing Disease Genes by Using Search Engine Algorithm. Curr Bioinforma. 2016; 11(2):195–202. https://doi.org/10.2174/1574893611666160125220905
4. Hu W, Sillaots S, Lemieux S, Davison J, Kauffman S, Breton A, et al. Essential Gene Identification and Drug Target Prioritization in Aspergillus fumigatus. Plos Pathog. 2007; 3(3):e24–e24. https://doi.org/10.1371/journal.ppat.0030024 PMID: 17352532
5. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, et al. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature. 2002; 418(6896):387–391. https://doi.org/10.1038/nature00935 PMID: 12140549
6. Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, et al. Large-scale essential gene identification in *Candida albicans* and applications to antifungal drug discovery. Mol Microbiol. 2003; 50(1): 167–181. https://doi.org/10.1046/j.1365-2958.2003.03697.x PMID: 14507372
7. Cullen LM, Arndt GM. Genome-wide screening for gene function using RNAi in mammalian cells. Immunol Cell Biol. 2005; 83(3):217–223. https://doi.org/10.1111/j.1440-1711.2005.01332.x PMID: 15877598
8. Proctor CH, Loomis CP. Analysis of sociometric data. Res Methods Soc Relat. 1951; 2:561–585.
9. Freeman LC. A set of measures of centrality based on betweenness. Sociometry. 1977; 40(1):35–41. https://doi.org/10.2307/3033543
10. Bonacich P. Power and centrality: A family of measures. Amer J Sociol. 1987; 92(5):1170–1182. https://doi.org/10.1086/228631
11. Estrada E, Rodriguez-Velazquez JA. Subgraph centrality in complex networks. Phys Rev E. 2005; 71(5):056103. https://doi.org/10.1103/PhysRevE.71.056103
12. 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):143–150. https://doi.org/10.1016/j.compbiolchem.2011.04.002 PMID: 21704260
13. Wang J, Li M, Wang H, Pan Y. Identification of essential proteins based on edge clustering coefficient. IEEE/ACM Trans Comput Biol Bioinformat. 2012; 9(4):1070–1080. https://doi.org/10.1109/TCBB.2011.147
14. Li M, Lu Y, Wang J, Wu F. A Topology Potential-Based Method for Identifying Essential Proteins from PPI Networks. IEEE/ACM Trans Comput Biol Bioinformat. 2014; 12(2):372–383. https://doi.org/10.1109/TCBB.2014.2361350
15. Li M, Lu Y, Niu Z, Wu FX. United Complex Centrality for Identification of Essential Proteins from PPI Networks. IEEE/ACM Trans Comput Biol Bioinformat. 2017; 14(2):370–380. https://doi.org/10.1109/TCBB.2015.2394487
16. Luo J, Qi Y. Identification of essential proteins based on a new combination of local interaction density and protein complexes. PloS ONE. 2015; 10(6):e0131418. https://doi.org/10.1371/journal.pone.0131418 PMID: 26125187
17. Qin C, Sun Y, Dong Y. A New Method for Identifying Essential Proteins Based on Network Topology Properties and Protein Complexes. PloS ONE. 2016; 11(8):e0161042. https://doi.org/10.1371/journal.pone.0161042 PMID: 27529423

18. Li M, Zhang H, Wang J, 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):15. https://doi.org/10.1186/1752-0509-6-15 PMID: 22405054

19. Tang X, Wang J, Zhong J, Pan Y. Predicting essential proteins based on weighted degree centrality. IEEE/ACM Trans Comput Biol Bioinformat. 2014; 11(2):407–418. https://doi.org/10.1109/TCBB.2013.2295318

20. Zhao B, Wang J, Li X, Wu F. Essential protein discovery based on a combination of modularity and conservatism. Methods. 2016; 110:54–63. https://doi.org/10.1016/j.ymeth.2016.07.005 PMID: 27402354

21. Zhong J, Wang J, Peng W, Zhang Z, Li M. A Feature Selection Method for Prediction Essential Protein. Tsinghua Sci Technol. 2015; 20(5):491–499. https://doi.org/10.1109/TST.2015.7297748

22. Peng W, Wang J, Wang W, Liu Q, Wu FX, Pan Y. Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. BMC Syst Biol. 2012; 6(1):87. https://doi.org/10.1186/1752-0509-6-87 PMID: 22808943

23. Li G, Li M, Wang J, Wu J, Wu FX, Pan Y. Predicting essential proteins based on subcellular localization, orthology and PPI networks. BMC Bioinformatics. 2016; 17(8):279. https://doi.org/10.1186/s12859-016-1115-5 PMID: 27586883

24. Li M, Niu Z, Chen X, Zhong P, Wu F, Pan y. A Reliable Neighbor-Based Method for Identifying Essential Proteins by Integrating Gene Expressions, Orthology, and Subcellular Localization Information. Tsinghua Sci Technol. 2016; 21(6):668–677. https://doi.org/10.1109/TST.2016.7787009

25. Qi X, Fuller E, Wu Q, Wu Y, Zhang C. Laplacian centrality: A new centrality measure for weighted networks. Inf Sci. 2012; 194(5):240–253. https://doi.org/10.1016/j.ins.2011.12.027

26. Acencio ML, Lemke N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinformatics. 2009; 10(1):290. https://doi.org/10.1186/1471-2105-10-290 PMID: 19758426

27. Peng X, Wang J, Wang J, Wu FX, Pan Y. Rechecking the Centrality-Lethality Rule in the Scope of Protein Subcellular Localization Interaction Networks. PloS ONE. 2015; 10(6):e0130743. https://doi.org/10.1371/journal.pone.0130743 PMID: 26115027

28. Breiman L. Random Forest. Mach Learn. 2001; 45:5–32.

29. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor. 2008; 11(1):10–18. https://doi.org/10.1145/1656274.1656278

30. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: The database of interacting proteins. Nucleic Acids Res. 2000; 28(1):289–291. https://doi.org/10.1093/nar/28.1.289 PMID: 10592249

31. Mewes HW, Frishman D, Mayer KF, Münsterkötter M, Noubibou O, Pagel P, et al. MIPS: Analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Res. 2006; 34(suppl 1):D169–D172. https://doi.org/10.1093/nar/gkj148 PMID: 16381839

32. Zhang R, Lin Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res. 2009; 37(suppl 1):D455–D458. https://doi.org/10.1093/nar/gkn858 PMID: 18974178

33. Issel-Tarver L, Christie KR, Dolinski K, Andrada R, Balakrishnan R, Ball CA, et al. *Saccharomyces* genome database. Methods Enzymol. 2002; 350:329–346. https://doi.org/10.1016/S0076-6879(02)50972-1 PMID: 12073322

34. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science. 1999; 285(5429):901–906. https://doi.org/10.1126/science.285.5429.901 PMID: 10436161

35. Friedel CC, Krumsiek J, Zimmer R. Bootstrapping the interactome: Unsupervised identification of protein complexes in yeast. J Comput Biol. 2009; 16(8):971–987. https://doi.org/10.1089/cmb.2009.0023 PMID: 19630542

36. Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res. 2009; 37(3):825–831. https://doi.org/10.1093/nar/gkn1005 PMID: 19095691

37. Pu S, Vlasblom J, Emili A, Greenblatt J, Wodak SJ. Identifying functional modules in the physical interactome of *Saccharomyces cerevisiae*. Proteomics. 2007; 7(6):944–960. https://doi.org/10.1002/pmic.200600636 PMID: 17370254

38. Yu H, Greenbaum D, Lu HX, Zhu X, Gerstein M. Genomic analysis of essentiality within protein networks. TRENDS Genet. 2004; 20(6):227–231. https://doi.org/10.1016/j.tig.2004.04.008 PMID: 15145574

39. Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 2010; 38(suppl 1):D196–D203. https://doi.org/10.1093/nar/gkp931 PMID: 19892828

40. Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R, et al. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database. 2014; 2014:900–900. https://doi.org/10.1093/database/bau012

41. Tang Y, Li M, Wang J, Pan Y, Wu FX. CytoNCA: A cytoscape plugin for centrality analysis and evaluation of protein interaction networks. Biosystems. 2015; 127:67–72. https://doi.org/10.1016/j.biosystems.2014.11.005 PMID: 25451770

42. Zhong J, Wang J, Peng W, Zhang Z, Pan Y. Prediction of essential proteins based on gene expression programming. BMC Genomics. 2013; 14(4):S7. PMID: 24267033

43. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: The comprehensive resource of mammalian protein complexes-2009. Nucleic Acids Res. 2010; 38(suppl 1): D497–D501. https://doi.org/10.1093/nar/gkp914 PMID: 19884131