An Iteration Model for Identifying Essential Proteins by Combining Comprehensive PPI Network With Biological Information

Shiyuan Li
Changsha University https://orcid.org/0000-0002-4381-7497
Zhen Zhang
Changsha University
Xueyong Li
Changsha University
Yihong Tan
Changsha University
Lei Wang
Changsha University
Zhiping Chen (zpchen@ccsu.edu.cn)
College of Computer Engineering and Applied Mathematics, Changsha University, 410022 Changsha, China 2Hunan Province Key Laboratory of Industrial Internet Technology and Security, Changsha University, 410022 Changsha, China

Research article

Keywords: Essential proteins, TGSO, competitive methods

Posted Date: August 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-54191/v1

Version of Record: A version of this preprint was published at BMC Bioinformatics on September 8th, 2021. See the published version at https://doi.org/10.1186/s12859-021-04300-7.

RESEARCH

An Iteration Model for Identifying Essential Proteins by Combining Comprehensive PPI Network with Biological Information

Shiyuan Li[1,2], Zhen Zhang[3], Xueyong Li[1,2], Yihong Tan[1,2*], Lei Wang[1,2] and Zhiping Chen[1,2*]

*Correspondence: yhtan@ccsu.edu.cn,zpchen@ccsu.edu.cn
[1]College of Computer Engineering and Applied Mathematics, Changsha University, 410022 Changsha, China
[2]Hunan Province Key Laboratory of Industrial Internet Technology and Security, Changsha University, 410022 Changsha, China
Full list of author information is available at the end of the article

Abstract

Background: Essential proteins have great impacts on cell survival and development, and played important roles in disease analysis and new drug design. However, since it is inefficient and costly to identify essential proteins by using biological experiments, then there is an urgent need for automated and accurate detection methods. In recent years, the recognition of essential proteins in protein interaction networks (PPI) has become a research hotspot, and many computational models for predicting essential proteins have been proposed successively.

Results: In order to achieve higher prediction performance, in this paper, a new prediction model called TGSO is proposed. In TGSO, a protein aggregation degree network is constructed first by adopting the node density measurement method for complex networks. And simultaneously, a protein co-expression interactive network is constructed by combining the gene expression information with the network connectivity, and a protein co-localization interaction network is constructed based on the subcellular localization data. And then, through integrating these three kinds of newly constructed networks, a comprehensive protein-protein interaction network will be obtained. Finally, based on the homology information, scores can be calculated out iteratively for different proteins, which can be utilized to estimate the importance of proteins effectively. Moreover, in order to evaluate the identification performance of TGSO, we have compared TGSO with 13 different latest competitive methods based on three kinds of yeast databases. And experimental results show that TGSO can achieve identification accuracies of 94%, 82% and 72% out of the top 1%, 5% and 10% candidate proteins respectively, which are to some degree superior to these state-of-the-art competitive models.

Conclusions: We constructed a comprehensive interactive network based on multi-source data to reduce the noise and errors in the initial PPI, and combined with iterative methods to improve the accuracy of necessary protein prediction, and means that TGSO may be conducive to the future development of essential protein recognition as well.

Background

Numerous studies have shown that essential proteins play important roles in human biological processes. The lack of essential proteins will affect cell growth and development seriously, and the functions of the protein complexes will be lost as well. Essential protein prediction is not only of great significance to the researches on life science, but also able to provide valuable information to the treatment of

diseases and the design of new drugs [1–4]. Traditionally, essential proteins are identified by medical experiments, such as RNA interference and gene knockout [5], [6]. However, these biological experiments are not only time-consuming, but also costly and inefficient. Hence, automated and accurate detection methods become necessary. Up to now, many computational models for identifying essential proteins have been developed successively. For instance, Yu et al found the correlations between bottlenecks and essential proteins, where bottlenecks were defined as proteins with high degrees of centrality [7]. Li Min et al proposed a calculation method to identify essential proteins by adopting a new protein network recognition method based on topological potential and local average connection [8, 9]. Jeong H et al introduced the central lethal rule to estimate the connection between network topology and essential proteins [10]. From then on, based on the concept of centrality, a lot of different methods, including the Degree Centrality (DC) [11], Information Centrality (IC) [12], Eigenvector Centrality (EC) [13], Subgraph Centrality (SC) [14], Betweenness Centrality (BC) [15], Closeness Centrality (CC) [16] and Neighbor Centrality (NC) [17], have been designed successively. However, although these centrality-based methods can improve the efficiency of traditional biological experiments effectively, their recognition abilities are still not very satisfactory, since there are lots of noises such as the false negatives and the false positives existing in the PPI networks [18, 19]. Therefore, in order to further improve the performance of identification models, biological information data including GO (Gene Ontology) statement annotations, gene expression profiles, subcellular data and protein domain data have been integrated with the PPI networks to identify essential proteins. For example, by integrating PPI networks with gene expression data, Li et al established a prediction method called Pec [20]to infer potential essential proteins. Zhang X et al proposed a computational model named CoEWC by combining protein neighborhood clustering characteristics rather than the protein itself with the PPI networks to detect essential proteins [21], and achieved good prediction performance. Zhao et al. designed a model called POEM to predict essential proteins based on overlapping essential modules [22]. Zhao et al. proposed a computational model called RWHN by integrating the subcellular localization and the protein domain information with the PPI networks to identify essential proteins through [23].

The GO database is the largest source of information about gene function in the world [24], which has often been adopted to mine functional similarities between proteins. For instance, Kim et al found that it can improve the prediction performance of models by adopting the informational GO terms to prune the PPI networks [25]. By integrating the GO statement annotation information with the gene expression data, the subcellular localization data and other biological data, Lei et al. proposed a new method for predicting essential proteins in PPI networks based on artificial fish swarm optimization [26]. Zhang et al designed a prediction model called TEGS based on the GO annotations, subcellular localization data, gene expression data and topological information of the PPI networks [27, 28]. Lei et al. designed a model called RSG through combining the RNA-seq data instead of the gene expression data with the GO annotation and subcellular localization to identify essential proteins.

Considering that essential protein is more conservative than non-essential proteins in evolution, Peng et al proposed an iterative method named ION to predict essential

proteins by integrating orthology with PPI network [29]. Zhang et al. introduced a prediction method called OGN through integrating the homology information and the gene expression data with the PPI networks [30]. Lei et al. designed a method called PCSD for identifying essential proteins based on the degree of protein participation in protein complexes and the density of sub graphs [31]. Li et al. developed a prediction model called NCCO to identify potential essential proteins by extending the Pareto optimal consensus model (EPOC) [32]. Zhang et al. designed a dynamic PPI network (FDP) by combining the FDP with homology information to identify essential proteins [33]. In our previous works, an iterative method called CVIM was proposed to determine essential proteins based on the topological and functional features of PPI networks [34].

In this paper, different from above models, a novel centrality-based method called TGSO is proposed by combining biological essence data including the gene expression data, the orthologous information and the subcellular localization data with the topological information in a newly constructed comprehensive PPI network. In TGSO, a new centrality-based method named DBN (Density between nodes) is designed first to calculate the node density in complex networks, which can characterize the physical structure association between nodes in a complex network, and then, based on DBN, a protein aggregation degree interaction network (ADN) can be constructed. Next, by adopting the Pearson correlation coefficient to measure protein co-expressions based on the gene expression data, a protein co-expression interaction network (CEN) can be constructed. Moreover, based on the subcellular localization data, a protein co-localization interaction network (CLN) can be obtained as well. Hence, through integrating these three kinds of interaction networks, a comprehensive PPI network (PCIN) can be constructed. Finally, based on the newly obtained comprehensive PPI network, an iterative method called TGSO is designed to predict potential essential proteins by using the orthology information as the initial scores of proteins. In order to estimate the identification performance of TGSO, intensive experiments have been implemented, and experimental results show that TGSO can achieve more satisfactory prediction performance than state-of-the-art competitive prediction models such as DC [11], IC [12], EC [13], SC [14], BC [15], CC [16], NC [17], PEC [20], CoEWC [21], POEM [22], ION [29], TEGS [28] and CVIM [34] based on two kinds of different databases separately.

Method

As illustrated in Figure 1, the procedure of TGSO mainly includes the following five steps:

Step1: Construction of the ADN (the protein Aggregation Degree interaction Network).

Step2: Construction of the CEN (the protein Co-Expression interaction Network).

Step3: Construction of the CLN (the protein Co-Location interaction Network).

Step4: Construction of the PCIN (the Protein Comprehensive Interaction Network).

Step5: Construction of the TGSO.

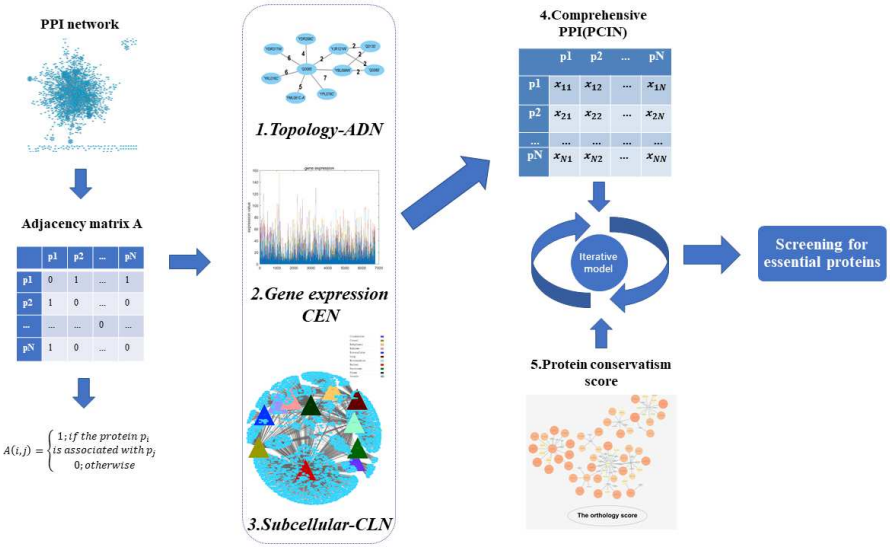


Figure 1: **Flow chart of TGSO.(All the subgraphs in this figure were created by the first author using the open source software Cytoscape in combination with existing experimental data, without any borrowing.)** The initial PPI, combined with subcellular localization and gene expression data as well as network topology information, was integrated to obtain the comprehensive protein interaction network, and the network and protein conservative score were put into the iterative model to obtain the final required protein score.

Let $V = \{p_1, p_2, \ldots, p_N\}$ denote the set of different proteins downloaded from a public database D, and for a pair of proteins p and q in V, if there is a known interaction between them in D, we define that there is an edge $e(p, q)$ between them, then let E represents the set of edges between proteins in V, it is obvious that an original protein-protein interaction (PPI) network $G = (V, E)$ can be obtained. And moreover, based on the original PPI network G, an adjacency matrix $A = (a_{ij})_{N*N}$ can be further constructed, where there is $a_{ij} = 1$, if and only if there exists an edge $e(p_i, p_j)$ between p_i and p_j in E, otherwise there is $a_{ij} = 0$.

Construction of the ADN

Recent researches show that the degrees of connections between essential proteins are often higher than that between non-essential proteins [35], and essential proteins can form tightly connected molecular modules [27]. Hence, based on the modular nature of key proteins, for each edge $e(u, v)$, we can design a local metric called DBN (Density between nodes) to measure the interaction between them in the original PPI network G as follows:

$$DBN(u, v) = \frac{|NG(u) \cap NG(v) + 1|}{min(|NG(u)|, |NG(v)|)} \tag{1}$$

Here, $NG(u) = \{v | \exists e(u, v) \in E, v \in V\}$, represents the set of neighboring nodes of the protein node u in G, and $|NG(u)|$ is the total number of neighboring nodes of the

protein node u in G. According to above formula (1), it is obvious that we can obtain a new matrix DBN, based on which, we can construct a weighted protein-protein interactive network. And for convenience, we define the newly constructed weighted PPI network as the protein Aggregation Degree interactive Network (ADN).

Construction of the CEN

Gene expression refers to the process of synthesizing genetic information from genes into functional gene products. Gene expression products are usually proteins, but the expression products of non-protein coding genes such as transfer RNA (tRNA) or small nuclear RNA (snRNA) genes are functional RNA. Over a period of time, there may be similar expressions between essential proteins. According to the studies of Horyu et al [36], it was found that the Pearson correlation coefficient (PCC) is suitable for measuring the similarities between gene expression profiles. Hence, based on the concept of PCC, for any a pair of proteins u and v, we can calculate the similarity between them as follows:

$$PCC(u,v) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Exp(u,i) - \overline{Exp(u)}}{\sigma(u)}\right)\left(\frac{Exp(v,i) - \overline{Exp(v)}}{\sigma(v)}\right) \quad (2)$$

Here, $Exp(u,i)$ is the expression level of the protein u on the *i-th* time, and it is obvious that for any given protein u, its expression information on a series of n different time nodes constitutes a vector $Exp(u) = \{Exp(u,1), Exp(u,2), \ldots, Exp(u,n)\}$. In addition, $\overline{Exp(u)}$ is the average expression value of the protein u, $\sigma(u)$ is the standard variance for gene expression of the protein u.

Existing studies illustrate that the essentiality of proteins is related to the proteins or genes themselves and the molecular modules they belong to [37, 38], and the essential complex biological module consists of a large number of essential proteins that are highly connected and shared between biological functions [39]. Based on these findings, for any a pair of proteins u and v, we can measure the interaction between them in the original PPI network G as follows:

$$Connection(u,v) = PCC(u,v) + \sum_{\varepsilon \in (NG(u) \cap NG(v))} PCC(u,\varepsilon) * PCC(v,\varepsilon) \quad (3)$$

Based on above formula (3), it is obvious that we can construct another weighted protein-protein interactive network. And for convenience, we define the newly constructed weighted PPI network as the protein Co-Expression interaction Network (CEN).

Construction of the CLN

Researches show that protein interactions in human bodies tend to coexist in the same cell compartment or adjacent cell compartments [40]. And it has been demonstrated that the introduction of subcellular localization information is of great help in screening essential proteins [23, 28, 41]. As illustrated in Figure 2 and Figure 3, we have performed a statistical analysis on the number of essential proteins or

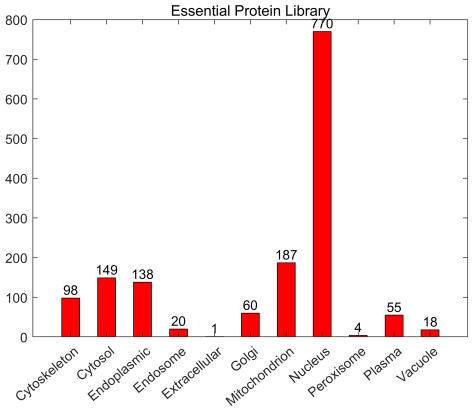
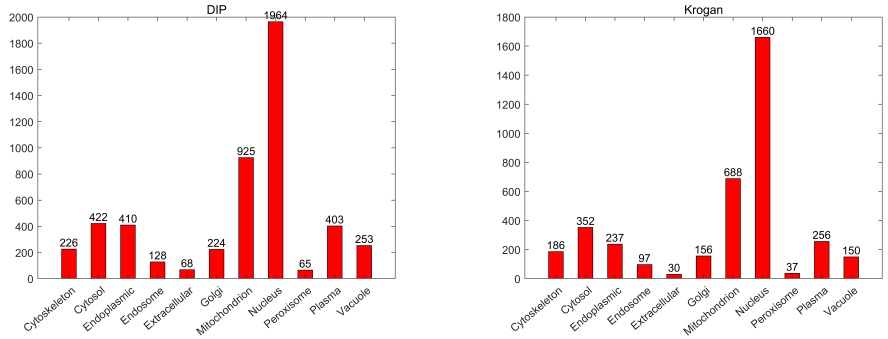


Figure 2: The number of essential proteins about eleven subcellular locations in the Essential Protein Library.



(a) DIP (b) Krogan

Figure 3: The number of proteins about eleven subcellular locations in the DIP and Krogan protein databases.

all proteins about eleven subcellular locations in three kinds of different datasets including the essential protein library, the DIP database and the Krogan protein database respectively. From observing Figure 2 and Figure 3, it is easy to find that the rankings of the number of proteins owned by sub-cells are similar. And there are more essential proteins in the nucleus, while there are only a few essential proteins in the peroxisome. Recent research discover that 76% of protein-protein interactions in yeast cells occur between identical subcells [42]. And in many cases, the product of complex functions is more important than the function of individual proteins, and essential proteins tend to form protein complexes to perform important functions together [37, 38]. Hence, in order to distinguish the importance of different subcellular localizations, for any given subcellular location i, we define the total number of subcellular species related to i as follows:

$$sub\_score(i) = \frac{sub(i)}{\sum_{k=1}^{N} sub(k)} \tag{4}$$

Here, $sub(i)$ represents the number of protein nodes associated with the subcellular location i in the database. Hence, for any give protein u, we can define its self-localization score as follows:

$$S\_score(u) = \sum_{i \in S(u)} sub\_score(i) \tag{5}$$

Here, $S(u)$ is a collection of all subcellular localizations possessed by u. Based on above formula (5), for any a pair of proteins u and v, we can further obtain the co-localization score between them as:

$$colo\_sub(u,v) = \frac{|S(u) \cap S(v)|}{|S(u) \cup S(v)|} * \frac{S\_score(u) + S\_score(v)}{2} \tag{6}$$

According to above formula (6), it is obvious that we can further construct a new weighted protein-protein interactive network. And for convenience, we define the newly constructed weighted PPI network as the protein Co-Localization interaction Network (CLN).

Construction of the PCIN

Based on above three kinds of newly constructed weighted PPI networks such as the AND, CEN and CLN, for any given protein u, we can obtain a unique score for u as follows:

$$LSG(u) = \sum_{v \in NG(u)} DBN(u,v) * (colo\_sub(u,v) + Connection(u,v)) \tag{7}$$

According to above formula (7), for any two given proteins i and j, we can define a comprehensive interaction between them as follows:

$$PCIN(i,j) = \begin{cases} LSG(i) / \sum_{k=1}^{N} LSG(k) & if \quad i = j \\ min(LSG(i), LSG(j)) / \sum_{k=1}^{N} LSG(k) & Otherwise \end{cases} \tag{8}$$

Construction of the TGSO

Peng et al. [29] found that the essentiality of protein is closely related to the degree of protein conservatism. For any given protein u, let $I(u)$ denote its homology score, then we can obtain the conservatism score $O\_score(u)$ corresponding to u based on the original PPI network G as follows:

$$O\_score(u) = \frac{I(u)}{\sum_{k=1}^{N} I(k)} \tag{9}$$

Based on above formula (9), for all N different proteins $p_1, p_2, \ldots, p_N$ in G, then we can obtain their initial scores as follows:

$$P_0 = (O\_score(1), O\_score(2), ..., O\_score(i), ..., O\_score(N)) \tag{10}$$

Finally, based on above newly obtained initial scores and the newly constructed weighted comprehensive PPI network PCIN, we can obtain the criticality scores of all proteins in G iteratively by adopting the following formula:

$$P_{t+1} = (1 - \alpha) * PCIN * P_t + \alpha * P_0 \tag{11}$$

Here, the parameter $\alpha (0 \leqslant \alpha \leqslant 1)$ is used to adjust the proportion of initial scores P_0 and last iteration scores P_t. Based on the above descriptions, the general flowchart of our prediction algorithm TGSO can be mainly described as follows:

Algorithm: TGSO
Input: Original PPI network $G = (V, E)$, subcellular location data, orthologous and gene expression data, the parameters γ and K
Output: Top K percent of proteins sorted by the vector P in descending order
Step1: Constructing the ADN according to the formula (1);
Step2: Constructing the CEN according to the formula (3);
Step3: Constructing the CLN according to the formula (6);
Step4: Constructing the PCIN according to the formula (8);
Step5: Obtaining the initial score vector P_0 according to the formula (10);
Step6: Let $t = 0$; Obtaining P1according to formula (11);
Step7: Let $t = t + 1$; Obtaining P_{t+1} according to formula (11);
Step8: Repeating Step7 until $(||P_{t+1} - P_t||)/|E| < \gamma$;
Step9: Sort proteins by the value of P in the descending order;
Step10: Output top K percent of sorted proteins.

Result and Analysis

Experimental Data

In order to estimate the identification performance of TGSO, in this section, we will compare it with 13 different state-of-the-art competitive prediction models illustrated in the following table 1.

Table 1:A rough introduction to other algorithms

Algorithm	Network topology	Biological information
DC[11]	Degree Centrality	No
IC[12]	Information Centrality	No
EC[13]	Eigenvector Centrality	No
SC[14]	Subgraph Centrality	No
BC[15]	Betweenness Centrality	No
CC[16]	Closeness Centrality	No
NC[17]	Neighbor Centrality	No
Pec[20]	Edge clustering coefficient	Gene expression data
CoEWC[21]	Clustering coefficient	Gene expression data
POEM[22]	Degree Centrality,Subgraph Edge clustering coefficient,Closeness Centrality	Gene expression data
ION[29]	Edge clustering coefficient	Orthologous data
CVIM[34]	Average triangle,neighbor average triangle	Orthologous data,Gene expression data
TEGS[28]	Edge clustering coefficient	Gene Ontology,subcellular localization Gene expression data

Since saccharomyces cerevisiae includes the most complete PPI data and rich biological information data, and is widely used to evaluate essential protein prediction models, we will first evaluate the performance of TGSO based on three saccharomyces cerevisiae related databases such as the DIP database [43], the Krogan database [44], and the Gavin database [45]. After filtering out repetitive interactions and self-interactions, as shown in the table 2, we finally obtained a total of 5,093 proteins and 24,743 interactions from the DIP database, 14,317 pairs of interactions between 3672 proteins from the Krogan database, and 1855 proteins and 7669 interactions from the Gavin database respectively.

Table 2: The detail information of the three PPI datasets

Dataset	Proteins	Interactions	essential	Gene expression covers
DIP	5093	24743	1167	4981
Krogan	3672	14317	929	3610
Gavin	1855	7669	714	1827

Moreover, as a benchmark dataset for testing the accuracy of different identification models, a set of 1293 essential genes is derived from the MIPS[46], the Saccharomyces Genome Database(SGD)[47], the Saccharomyces Genome Deletion Project Database (SGDP)[48], and the Database of Essential Genes (DEG)[49] simultaneously. In addition, the gene expression data of Saccharomyces cerevisiae is obtained from the work proposed by Tu et al [50], which contains 6777 gene products and 36 samples. The orthologous information is downloaded from the InParanoid database (Version 7) [51]. Besides, as illustrated in above Figure 2 and Figure 3, we derived eleven subcellular locations related to eukaryotic cells from the COMPARTMENTS database [52, 53] as well.

Finally, in order to evaluate the uniqueness and efficiency of TGSO, in this section, we will first adopt different measurements such as accuracy, jackknife, Precision Recall regression curve (PR-curves) and Receiver Operating Characteristic curve (ROC) to compare TGSO with 13 competitive prediction models shown in Table 1 comprehensively. And then, we will further estimate the effect of the parameter α on the performance of TGSO.

Comparisons between TGSO and 13 representative methods

In this section, two kinds of datasets downloaded from the DIP database and the Krogan database separately are adopted to compare TGSO with 13 competitive prediction models illustrated in Table 1. And as a result, Figure 4 and Figure 5 show the comparison results based on the DIP database and the Krogan database respectively.

From observing Figure 4, it is not difficult to see that in the top 1% (51) potential key proteins, TGSO has screened out 48 true essential proteins, with an accuracy rate of 94%. Among 5% (255) and 10% (510) candidate critical proteins, there are 208 and 368 true essential proteins having been identified by TGSO separately, with an accuracy rate of 82% and 72% as well.

Comparing with traditional centrality-based methods such as DC, IC, EC, SC, BC, CC and NC, the number of true essential proteins detected by TGSO has obvious advantages. Especially except NC, TGSO predicted 100% more accurately than other centrality methods in the top 1% and 5% of candidate essential proteins. And simultaneously, in the top 10% predicted essential proteins, while comparing with DC, IC, EC, SC, BC, CC and NC, the prediction accuracy of TGSO has increased by 77.78%, 75.24%, 88.72%, 88.72%, 102.2%, 90.67% and 30.5% respectively. Moreover, while comparing with methods that combined PPI networks with multiple biological data, such as Pec, CoEWC, ION, POEM and CVIM, TGSO can still achieve the highest prediction accuracy in any range from the top 1% to 25% of potential key proteins. Therefore, it is easy to draw the conclusion that TGSO can achieve the best prediction performance based on the DIP database.

From observing Figure 5, it can be found that TGSO can achieve similar prediction performance based on the Krogan database. For instance, among the top 1% (37) candidate critical proteins, 35 true essential proteins have been detected by TGSO, with the accuracy rate of 95%, while in the top 15% (551) potential essential proteins, TGSO can still achieve the accuracy rate of 66.06%, which is 76.70% higher than that of the worst-performing CC, and 11.31% and 13.40% higher than that of the best-performing CVIM and TEGS respectively in these 13 tradition competitive models. Furthermore, with the increasing of candidate key proteins, the accuracy rate of all kinds of prediction models will decrease inevitably, but in the top 25%, the number of true essential proteins detected by TGSO has reached 515, which is still much higher than 479 detected by CVIM and 480 discovered by ION. Hence, we can draw the conclusion that TGSO can achieve the best identification performance based on both the Krogan database and the DIP database while comparing with these 13 competitive state-of-the-art prediction models.

Validation with jackknife methodology

In order to evaluate the TGSO model more comprehensively and specifically, we will further adopt the jackknife method [54] to compare TGSO with 13 competitive methods in this section. And as a result, Figure 6 and Figure 7 illustrate the comparison results. From observing Figure 6(a), it is not difficult to see that TGSO can achieve better performance than these centrality-based methods including DC, IC, EC, SC, BC, CC and NC. Moreover, from observing Figure 6(b), it is obvious that the prediction performance of TGSO is significantly better than those multiple biological data based methods such as Pec, CoEWC, POEM and ION as well.

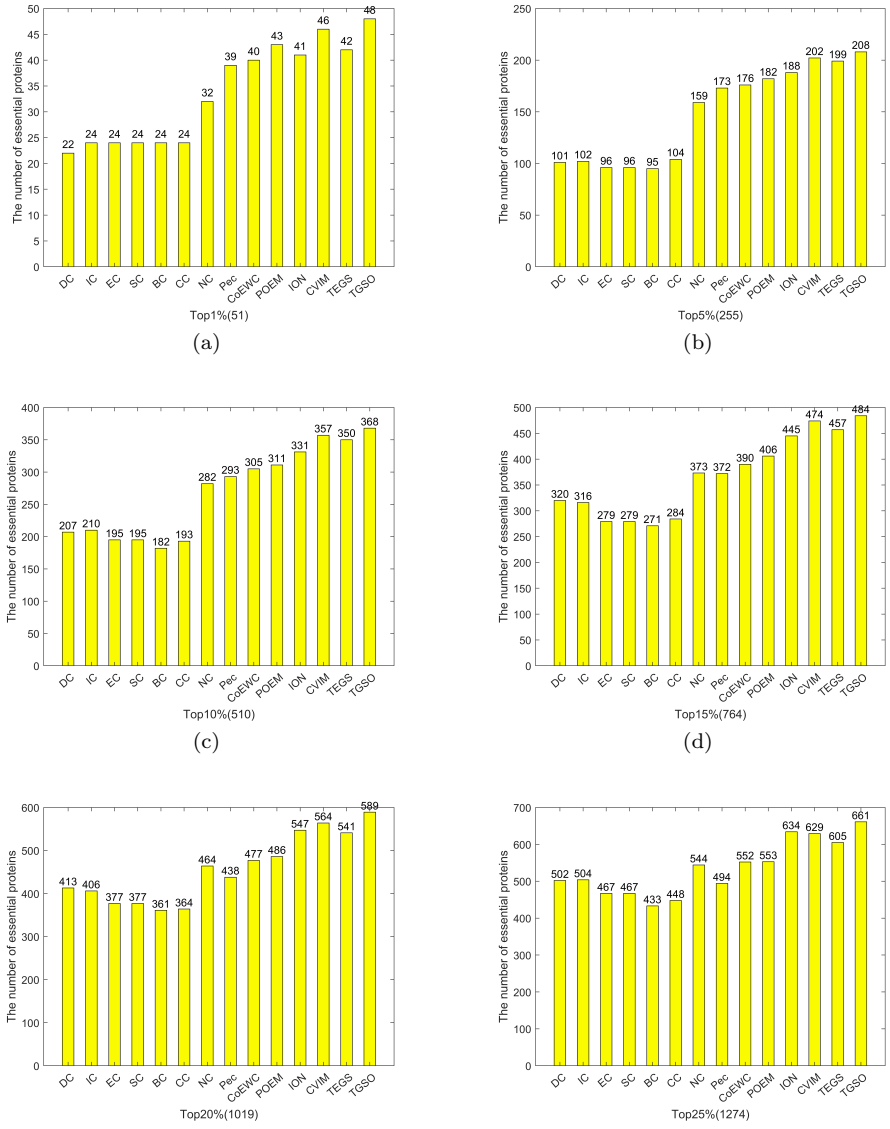


Figure 4: (a) Top 1% ranked proteins. (b) Top 5% ranked proteins. (c) Top 10% ranked proteins. (d) Top 15% ranked proteins. (e) Top 20% ranked proteins. (f) Top 25% ranked proteins. This figure illustrates the comparison of the number of essential proteins predicted by TGSO and 13 competing methods on the DIP dataset. The graph shows the number of truly essential proteins found by each method. The numbers in parentheses indicate the number of proteins ranked in each highest percentage.

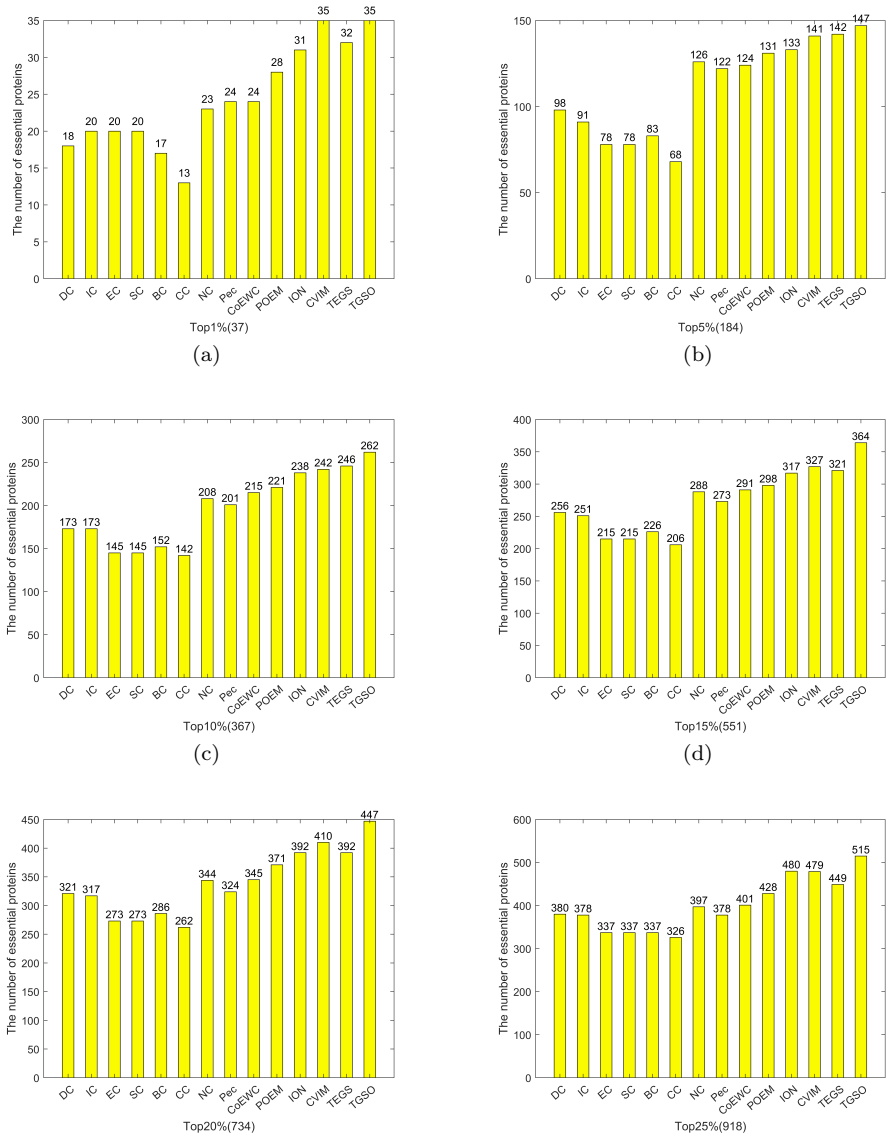
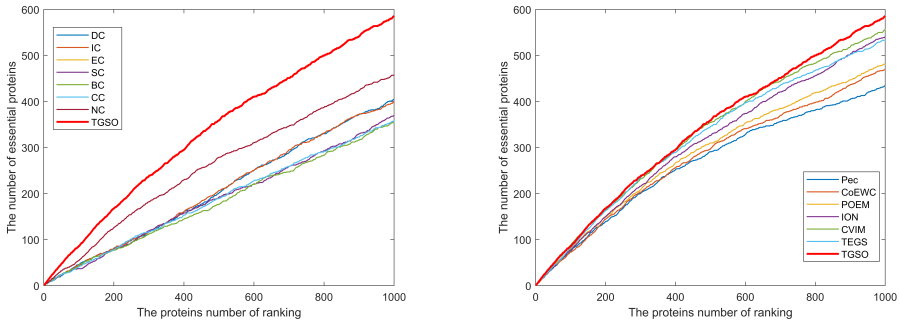


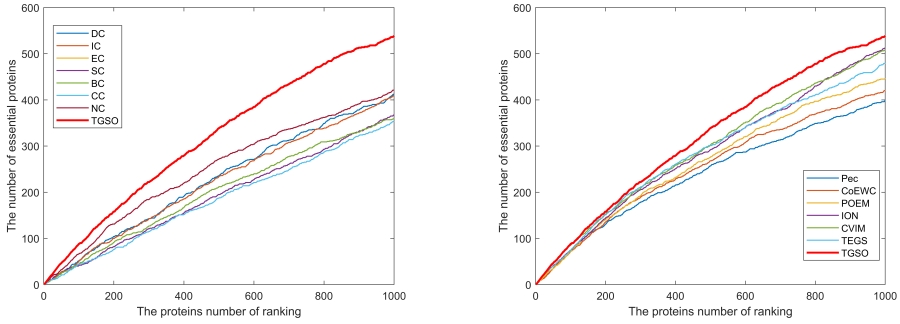
Figure 5: (a) Top 1% ranked proteins. (b) Top 5% ranked proteins. (c) Top 10% ranked proteins. (d) Top 15% ranked proteins. (e) Top 20% ranked proteins. (f) Top 25% ranked proteins. This figure illustrates the comparison of the number of essential proteins predicted by TGSO and 13 competing methods on the Krogan dataset. The graph shows the number of truly essential proteins found by each method. The numbers in parentheses indicate the number of proteins ranked in each highest percentage.

Although there are some partial overlaps among TGSO and CVIM and TEGS, but as the number of candidate key protein increases to about 600, the prediction performance of TGSO will become significantly higher than both CVIM and TEGS, which indicates that TGSO is superior to both CVIM and TEGS. In addition, from Figure 7(a) and Figure 7(b), it is also easy to see that TGSO can achieve better performance than all these 13 competitive methods. Especially, comparing with those methods that combined PPI networks with multiple biological data, while the number of candidate essential proteins reaches 300, TGSO can achieve much better performance than all these competitive methods simultaneously.



(a) (b)

Figure 6: Comparison of Jackknife curves of TGSO and 13 other methods under the DIP database. (a) Comparison between TGSO and DC, IC, EC, SC, BC, CC, NC. (b) Comparison between TGSO and Pec, CoEWC, POEM, ION, CVIM, TEGS.

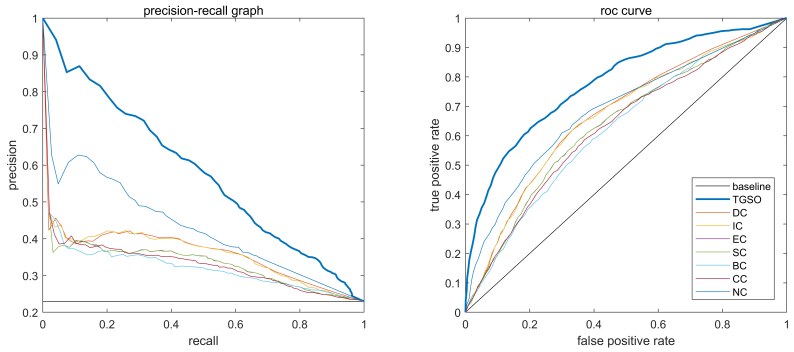


(a) (b)

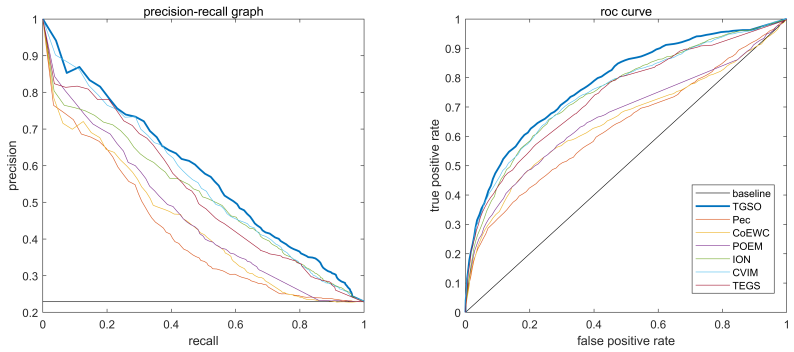
Figure 7: Comparison of Jackknife curves of TGSO and 13 other methods under the Krogan database. (a) Comparison between TGSO and DC, IC, EC, SC, BC, CC, NC. (b) Comparison between TGSO and Pec, CoEWC, POEM, ION, CVIM, TEGS.

Validation by Precision-Recall Curves and ROC Curves

In this section, we will further use the receiver operating characteristic curve (ROC curve) to evaluate the performance of TGSO. Studies show that the larger the area under the ROC curve (AUC), the better the performance of the model, and if AUC=0.5, it means a random performance [55–57]. In the three kinds of yeast cell databases including the DIP, Krogan and GAVIN databases, the proportion of key proteins is very small, and the proportion of non-essential proteins and essential proteins is about 3 to 1. Studies show that while dealing with highly skewed datasets, the precision recall (PR) curve can provide more information about the performance of an algorithm [58]. Therefore, in this section, we will further adopt the PR curves to compare TGSO with 13 competitive methods. As shown in Figure 8 and Figure 9, it is obvious that the AUCs achieved by TGSO is much higher than that of competitive methods based on both the DIP database and the Krogan database. However, from observing Figure 8(b) and Figure 9(b), we can find that the curves of TGSO and CVIM have a little overlap. Hence, in order to further evaluate TGSO and CVIM, we adopt the F1-score as well, and the comparison results are shown in Table 3.

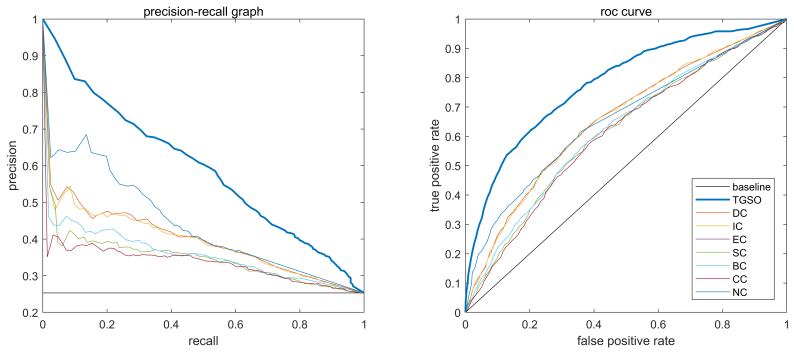


(a)

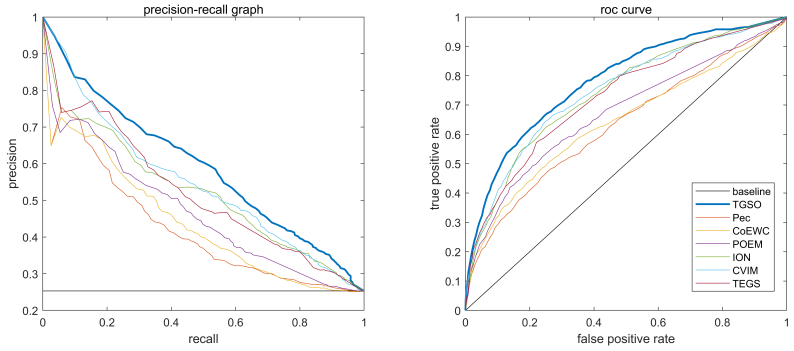


(b)

Figure 8: ROC curve and PR curve of various methods of PPI network based on the DIP database. (a) Comparison of TGSO with DC, EC, IC, SC, BC, CC and NC. (b) Comparison of TGSO with Pec, CoEWC, POEM, ION, CVIM and TEGS.



(a)



(b)

Figure 9: ROC curve and PR curve of various methods of PPI network based on the Krogan database. (a) Comparison of TGSO with DC, EC, IC, SC, BC, CC and NC. (b) Comparison of TGSO with Pec, CoEWC, POEM, ION, CVIM and TEGS.

From observing Table 3, it is obvious that not only the AUC achieved by TGSO is higher than those 13 competitive methods based on both the DIP database and the Krogan database, but also the F1-score achieved by TGSO is superior to those 13 competitive methods simultaneously. Therefore, it is reasonable to believe that TGSO has better performance than all these traditional state-of-the-art methods.

Table 3: The AUCs and F1-scores achieved by all methods based on the DIP and Krogan databases respectively

Method	AUC(DIP)	F1-score(DIP)	AUC(Krogan)	F1-score(Krogan)
TGSO	**0.7813**	**0.5466**	**0.7808**	**0.5600**
CVIM	0.7559	0.5217	0.7458	0.5411
ION	0.7522	0.5226	0.7413	0.5305
TEGS	0.7386	0.4959	0.7287	0.5148
POEM	0.6662	0.4528	0.6726	0.4704
CoEWC	0.6513	0.4528	0.6404	0.4476
Pec	0.6329	0.4062	0.6316	0.4264
NC	0.6879	0.4656	0.6584	0.4597
CC	0.6291	0.4143	0.6114	0.4282
BC	0.6250	0.4078	0.6248	0.4347
SC	0.6385	0.4233	0.6167	0.4309
IC	0.6657	0.4526	0.6573	0.4603
EC	0.6384	0.4235	0.6169	0.4308
DC	0.6705	0.4524	0.6583	0.4588

Difference analysis of TGSO and 13 competitive methods

In order to better reflect the uniqueness and differences between TGSO and these existing competitive methods, we will further compare TGSO with 13 competing prediction models based on the top 200 ranked proteins and the DIP database in this section. And the comparison results are illustrated in Table 4 and Table 5. In Table 4 and Table 5, Mi represents one of these 13 competitive models, $|TGSO \cap Mi|$ denotes the number of key proteins screened by both TGSO and Mi, while $|TGSO - Mi|$ indicates the number of critical proteins found by TGSO instead of Mi. From Table 4 and Table 5, it is not difficult to find that TGSO can screen out new key proteins that cannot discovered by any of these 13 competing methods. And in addition, From observing the fourth and fifth columns in both Table 4 and Table 5, it is easy to see that the proportion of true essential proteins screened by TGSO alone is much higher than the proportion of true essential proteins screened alone by any of these 13 competing methods, which is further demonstrated by the results illustrated in Figure 10 as well.

Table 4: Commonalities and differences between TGSO and 13 competing methods based on the top 200 ranked proteins and the DIP database

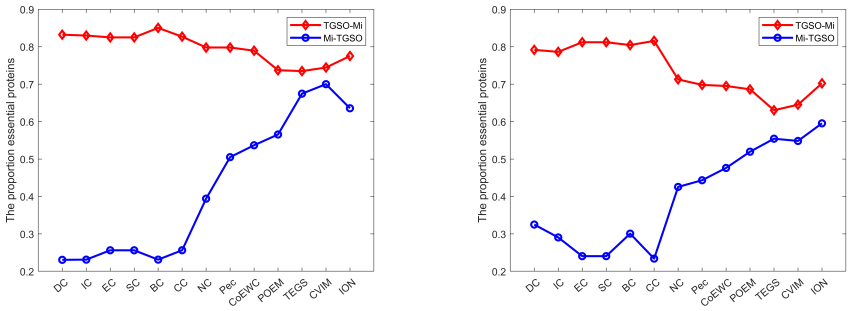
Different prediction methods (Mi)	$\|TGSO \cap Mi\|$	$\|TGSO - Mi\|$	Percentage of key proteins in $TGSO - Mi$	Percentage of key proteins in $Mi - TGSO$
DC	57	143	83.22%	23.08%
IC	53	147	82.99%	23.13%
EC	40	160	82.50%	25.63%
SC	40	160	82.59%	25.61%
BC	53	147	85.03%	23.13%
CC	44	156	82.69%	25.64%
NC	96	104	79.81%	39.42%
Pec	101	99	79.80%	50.51%
CoEWC	105	95	78.95%	53.68%
POEM	101	99	73.74%	56.57%
TEGS	117	83	73.49%	67.47%
CVIM	110	90	74.44%	70.00%
ION	71	129	77.52%	63.57%

Table 4: This table shows the commonalities and differences between TGSO and the 13 competitive methods in Table 1 based on the DIP database.

Table 5: Commonalities and differences between TGSO and 13 competing methods based on the top 200 ranked proteins and the Krogan database

Different prediction methods (Mi)	$\|TGSO \cap Mi\|$	$\|TGSO - Mi\|$	Percentage of key proteins in $TGSO - Mi$	Percentage of key proteins in $Mi - TGSO$
DC	80	120	79.17%	32.50%
IC	83	117	78.63%	29.06%
EC	67	133	81.20%	24.06%
SC	64	136	81.17%	24.05%
BC	67	133	80.45%	30.08%
CC	59	141	81.56%	23.40%
NC	106	94	71.28%	42.55%
Pec	94	106	69.81%	44.34%
CoEWC	95	105	69.52%	47.62%
POEM	98	102	68.63%	51.96%
TEGS	108	92	63.04%	55.43%
CVIM	138	62	64.52%	54.84%
ION	69	131	70.23%	59.54%

Table 5: This table shows the commonalities and differences between TGSO and the 13 competitive methods in Table 1 based on the Krogan database.



(a) DIP (b) Krogan

Figure 10: The X-axis represents 13 competing methods. The Y-axis represents the proportion of real key proteins in Mi-TGSO or TGSO-Mi.

General applicability of TGSO

In order to prove the applicability of TGSO, we will further execute some simple tests and comparisons based on the Gavin database in this section, and the experimental results are shown in the following Table 6.

Table 6: Number of essential proteins predicted by TGSO and 13 methods based on the GAVIN database

Methods	Top1%(19)	Top5%(93)	Top10%(196)	Top15%(279)	Top20%(371)	Top25%(464)
SC	0	17	87	130	190	240
EC	0	38	94	134	166	209
BC	9	40	85	122	162	201
DC	7	36	101	158	222	264
IC	16	55	119	163	213	254
CC	11	45	93	135	180	221
NC	11	51	123	170	213	259
PEC	15	69	142	193	238	285
CoEWC	16	69	136	190	237	275
POEM	17	74	148	199	249	296
ION	17	73	150	207	263	312
CVIM	16	80	160	219	271	322
TGSO	**19**	**81**	**165**	**221**	**279**	**332**

Table 6: This table shows the commonalities and differences between TGSO and the 13 competitive methods in Table 1 based on the GAVIN database.

As can be seen from Table 6, while comparing with these 13 competing methods, TGSO can achieve the best predictive performance in any range from the top 1% to 25% of potential key proteins, which demonstrates that TGSO is the best prediction model among these competitive models and has wide applicability.

Effects of parameter on performance of TGSO

In this section, we will analyze the influence of the parameter α on the performance of TGSO. In TGSO, the parameter α with value between 0 and 1 is adopted to adjust the weight of the comprehensive interaction network PCIN and the protein conservatism. During simulation, we will adjust the value of α to study its influence on the performance of TGSO. As shown in Table 7, based on the DIP database, while α is equal to 0.2, the algorithm is in the top 1% and the top 25% respectively takes the maximum value of 48 and 671. When α is 0.4, there are two maximum values of 48 and 487. When α is 0.3, the algorithm reaches the maximum value in the first 1%, the first 10%, and the first 20%. Therefore, on the DIP, 0.3 is the best parameter. In addition, from observing the Table 8, it is easy to see that based on the Krogan database, while α varying from 0.1 to 0.4, in the top 1% candidate key proteins, there are α maximum of 35 true essential proteins detected by TGSO, with the accuracy rate of 95%. While α is set to 0.2, TGSO can achieve the best accuracy rate in the top 1% and 25% candidate key proteins. When α is set to 0.3 or 0.4, TGSO achieves the best performance in the two intervals respectively. Therefore, based on the Krogan database, if α is set to 0.2 ,0.3, 0.4, TGSO can achieve the best performance. From Table 9, we can find that when α is between 0.1 and 0.4, only 0.3 occupies two maximum values. To sum up, based on these three kinds of databases, we will set α to 0.3 as the best value in experiments for comparing TGSO with these state-of-the-art competitive models in this article.

Table 7: Effects of the parameter α to TGSO based on the DIP database

α	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Top1%(51)	46	**48**	**48**	**48**	**48**	**48**	47	47	47
Top5%(255)	196	205	208	208	208	208	**209**	202	192
Top10%(510)	336	348	**368**	363	362	354	352	339	330
Top15%(764)	454	483	484	**487**	476	470	466	451	437
Top20%(1019)	558	578	**589**	584	568	556	538	528	528
Top25%(1274)	646	**671**	661	648	644	633	619	610	597

Table 7: This table shows the effects of the parameter α to TGSO based on the DIP database, and the table records the proportion of true key protein in the set of selected proteins.

Table 8: Effects of the parameter α to TGSO based on the Krogan database

α	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Top1%(37)	**35**	**35**	**35**	**35**	34	34	34	33	34
Top5%(184)	141	145	147	151	146	146	**153**	145	141
Top10%(367)	242	259	262	262	**264**	262	256	253	242
Top15%(551)	326	350	**364**	362	358	357	349	343	336
Top20%(734)	417	443	447	**449**	438	427	423	413	404
Top25%(918)	502	**524**	515	501	494	493	488	477	469

Table 8: This table shows the effects of the parameter α to TGSO based on the Krogan database, and the table records the proportion of true key protein in the set of selected proteins.

Table 9: Effects of the parameter α to TGSO based on the Gavin database

α	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Top1%(19)	17	18	**19**	18	18	18	18	18	18
Top5%(93)	80	82	81	83	83	83	**86**	**86**	79
Top10%(196)	159	163	165	167	167	**169**	167	162	158
Top15%(279)	204	218	221	218	223	**225**	222	216	204
Top20%(371)	247	266	279	**281**	280	280	273	261	255
Top25%(464)	294	304	**332**	326	324	316	311	308	303

Table 9: This table shows the effects of the parameter α to TGSO based on the Gavin database, and the table records the proportion of true key protein in the set of selected proteins.

DISCUSSION

Essential proteins are indispensable materials to sustain life activities.In recent years, the development of computational methods for essential protein recognition has become a research hotspot, and many researchers have successively developed

various algorithms based on PPI networks. With the gradual improvement of high-throughput biodata, more efficient prediction models have been proposed by combining PPI networks with biodata including the subcellular information and lineal homology information to screen essential proteins. Inspired by this, in this paper, a novel detection method called TGSO is designed to identify essential proteins based on multiple data fusion.And experimental results show that the method can achieve excellent prediction results, which provides a good reference for the future researches.

Conclusions

In this paper, we propose a new prediction model:TGSO. In TGSO, DBN is introduced to construct the node aggregation degree interactive network (ADN), PCC is adopted to construct the protein co-expression interactive network (CEN), and the subcellular localization information is adopted to construct the protein co-localization interactive network (CLN) firstly. And then, by integrating these three kinds of interactive networks, a comprehensive protein interaction network (PCIN) is obtained. Next, through combining protein conservatism scores with the PCIN, an iterative algorithm is proposed to calculate the essentiality score for each protein, which can be used to screen essential proteins efficiently. Finally, intensive experiments have been conducted to estimate the performance of TGSO based on the DIP, Krogan and Gavin databases separately, and experimental results show that TGSO can achieve more satisfactory performance than traditional state-of-the-art methods. In future work, we will introduce more biological information such as the protein-domain interactions and the gene ontology information to further improve the prediction performance of TGSO.

Declarations

Abbreviations

BC: Betweenness Centrality; CC: Closeness Centrality; CoEWC: Co-Expression Weighted by Clustering coefficient; DC: Degree Centrality; EC: Eigenvector Centrality; IC: Information Centrality; NC: Neighbor Centrality; PPI: Protein-Protein Interaction; SC: Subgraph Centrality; RWHN: Randomly Walking in the Heterogeneous Network; DBN:Density between nodes; ADC:Aggregation degree interaction network; CEN:Co-expression interaction network; Co-localization interaction netwrok; PCIN: Protein comprehensive interaction network

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the first author or corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

SYL and ZPC conceived the study. XYL, YHT and ZZ improved the study based on the original model. ZPC, YHT and LW supervised the study. SYL, ZPC and YHT wrote the manuscript of the study. All authors reviewed and improved the manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 61873221 and Grant 61672447, by the Natural Science Foundation of Hunan Province under Grant 2018JJ4058 and Grant 2019JJ70010, in part by the College Students' Research Learning and Innovative Experiment Plan Project of Hunan Province(S201911077006).

Acknowledgements

Not applicable.

Author details

[1]College of Computer Engineering and Applied Mathematics, Changsha University, 410022 Changsha, China. [2]Hunan Province Key Laboratory of Industrial Internet Technology and Security, Changsha University, 410022 Changsha, China. [3]College of Electronic Information and Electrical Engineering, Changsha University, 410022 Changsha, China.

References

1. Roemer, T., Jiang, B., Davison, J., Ketela, T., Bussey, H.: Large-scale essential gene identification in candida albicans and applications to antifungal drug discovery. Molecular Microbiology **50**(1), 167–181 (2010)
2. Zhang, Z., Wu, F.X., Wang, J., Qi, L., Zheng, R., Min, L.: Prioritizing disease genes by using search engine algorithm. Current Bioinformatics **11**(2), (2016)
3. Glass, J.I., Hutchison Iii, C.A., Smith, H.O., Venter, J.C.: A systems biology tour de force for a near-minimal bacterium. Molecular Systems Biology **5** (2014)
4. Perocchi, F., Mancera, E., Steinmetz, L.M.: Systematic screens for human disease genes, from yeast to human and back. Molecular Biosystems **4**(1), 18–29 (2007)
5. Cullen, L.M., Arndt, G.M.: Genome-wide screening for gene function using rnai in mammalian cells. Immunology & Cell Biology **83**(3), 217 (2005)
6. Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B.: Functional profiling of the saccharomyces cerevisiae genome. Nature **418**(6896), 387–391 (2002)
7. Yu, H., Kim, P.M., Sprecher, E., Trifonov, V., Gerstein, M.: The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. PLOS Computational Biology **3**(4), 1–8 (2007). doi:10.1371/journal.pcbi.0030059
8. Li, M., Wang, J., XiangChen, Wang, H., Pan, Y.: A local average connectivity-based method for identifying essential proteins from the network level. Computational Biology & Chemistry **35**(3), 143–150 (2011)
9. Li, M., Lu, Y., Wang, J., Wu, F.X., Pan, Y.: A topology potential-based method for identifying essential proteins from ppi networks. IEEE/ACM Transactions on Computational Biology & Bioinformatics **12**(2), 372 (2015)
10. Jeong, H.M., Mason, S.P., Barabási, A.L., Oltvai, Z.N.: Lethality and centrality in protein networks. Nature **411**(6833), 41–2 (2001)
11. Hahn, M.W., Kern, A.D.: Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks. Molecular Biology and Evolution **22**(4), 803–806 (2004). doi:10.1093/molbev/msi072. https://academic.oup.com/mbe/article-pdf/22/4/803/13433323/msi072.pdf
12. Stephenson, K., Zelen, M.: Rethinking centrality: Methods and examples. Social Networks **11**(1), 1–37 (1989). doi:10.1016/0378-8733(89)90016-6
13. Bonacich, Phillip: Power and centrality: A family of measures. American Journal of Sociology **92**(5), 1170–1182 (1987)
14. Estrada, E., Rodríguez-Velázquez, J.A.: Subgraph centrality in complex networks. Physical Review E Statistical Nonlinear & Soft Matter Physics **71**(5), 056103 (2005)
15. Joy, M.P., Brock, A., Ingber, D.E., Huang, S.: High-betweenness proteins in the yeast protein interaction network. Journal of Biomedicine & Biotechnology **2005**(2), 96 (2014)
16. Wuchty, S., Stadler, P.F.: Centers of complex networks. Journal of Theoretical Biology **223**(1), 45–53 (2003)
17. Wang, J., Li, M., Wang, H., Pan, Y.: Identification of essential proteins based on edge clustering coefficient. IEEE/ACM Transactions on Computational Biology & Bioinformatics **9**(4), 1070–1080 (2012)
18. Kuchaiev, O., Rašajski, M., Higham, D.J., Pržulj, N.: Geometric de-noising of protein-protein interaction networks. PLOS Computational Biology **5**(8), 1–10 (2009). doi:10.1371/journal.pcbi.1000454
19. Sprinzak, E., Sattath, S., Margalit, H.: How reliable are experimental protein-protein interaction data? Journal of Molecular Biology **327**(5), 919–923 (2003)
20. 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 Systems Biology **6** (2012)
21. Zhang, X., Xu, J., Xiao, W.X.: A new method for the discovery of essential proteins. Plos One **8** (2013)
22. Zhao, B., Wang, J., Li, M., Wu, F.X., Pan, Y.: Prediction of essential proteins based on overlapping essential modules. IEEE Transactions on Nanobioscience **13**(4), 415–424 (2014)
23. Zhao, B., Zhao, Y., Zhang, X., Zhang, Z., Wang, L.: An iteration method for identifying yeast essential proteins from heterogeneous network. BMC Bioinformatics **20**(1) (2019)
24. Ashburner, Michael, Ball, Catherine, A., Blake, Judith, A., Botstein, David: Gene ontology: tool for the unification of biology. Nature Genetics (2000)
25. Kim, W.: Prediction of essential proteins using topological properties in go-pruned ppi network based on machine learning methods. Tsinghua Ence & Technology **17**(006), 645–658 (2012)
26. Lei, X., Yang, X., Wu, F.: Artificial fish swarm optimization based method to identify essential proteins. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 1–1 (2018)
27. Zhang, W., Xu, J., Li, Y., Zou, X.: Detecting essential proteins based on network topology, gene expression data, and gene ontology information. IEEE/ACM Transactions on Computational Biology & Bioinformatics **15**(1), 109–116 (2016)
28. Zhang, W., Xu, J., Zou, X.: Predicting essential proteins by integrating network topology, subcellular localization information, gene expression profile and go annotation data. IEEE/ACM Transactions on Computational Biology and Bioinformatics **PP**(99), 1–1
29. 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 Systems Biology **6**(1), 1–17 (2012)

30. Zhang, X., Xiao, W., Hu, X., Irene, S.N.: Predicting essential proteins by integrating orthology, gene expressions, and ppi networks. Plos One **13**(4), 0195410 (2018)
31. Lei, X., Yang, X., Schreiber, G.: A new method for predicting essential proteins based on participation degree in protein complex and subgraph density. Plos One **13**(6) (2018)
32. Li, G., Li, M., Wang, J., Li, Y., Pan, Y.: United neighborhood closeness centrality and orthology for predicting essential proteins. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1–1 (2018)
33. Zhang, F., Peng, W., Yang, Y., Dai, W., Song, J.: A novel method for identifying essential genes by fusing dynamic protein–protein interactive networks. Genes **10**(1) (2019)
34. Li, S., Chen, Z., He, X., Zhang, Z., Wang, L.: An iteration method for identifying yeast essential proteins from weighted ppi network based on topological and functional features of proteins. IEEE Access **PP**(99), 1–1 (2020)
35. Pereira-Leal, J.B., Benjamin, A., Peregrin-Alvarez, J.M., Ouzounis, C.A.: An exponential core in the heart of the yeast protein interaction network. Molecular Biology & Evolution (3), 421 (2005)
36. Horyu, D., Hayashi, T.: Comparison between pearson correlation coefficient and mutual information as a similarity measure of gene expression profiles. Japanese Journal of Biometrics **33**(2), 125–143 (2013)
37. Hart, G.T., Lee, I., Marcotte, E.M.: A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics **8**(1), 236–236 (2007)
38. Dezső, Z., Oltvai, Z.N., Barabási, A.: Bioinformatics analysis of experimentally determined protein complexes in the yeast saccharomyces cerevisiae. Genome Research **13**(11), 2450 (2003)
39. 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 Computational Biology **4**(8), 1–16 (2008). doi:10.1371/journal.pcbi.1000140
40. Kumar, A.: Subcellular localization of the yeast proteome. Genes Dev **16**(6), 707–719 (2002)
41. Lei, X., Zhao, J., Fujita, H., Zhang, A.: Predicting essential proteins based on rna-seq, subcellular localization and go annotation datasets. Knowledge-Based Systems, 095070511830159 (2018)
42. Schwikowski, B., Uetz, P., Fields, S.: A network of protein-protein interactions in yeast. Nature Biotechnology **18**(12), 1257–1261 (2001)
43. Ioannis, X., Lukasz, S., Duan, X.J., Patrick, H., Sul-Min, K., David, E.: Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nuclc Acids Research (1), 303 (2002)
44. Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Krogan, N.J., Cagney, G., Yu, H.: Global landscape of protein complexes in the yeast saccharomyces cerevisiae. Nature **440**(7084), 637–43 (2006)
45. Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B.: Proteome survey reveals modularity of the yeast cell machinery. Nature **440**(7084), 631–6 (2006)
46. W, M.H., D, F., X, M.K.F., M, M., O, N., P, P., T, R., M, O., A, R., V, S.: Mips: analysis and annotation of proteins from whole genomes in 2005. Nuclc Acids Research (suppl_1), 169–72
47. Michael, C.J., Caroline, A., Catherine, B., A, C.S., S, D.S., T, H.E., Yankai, J., Gail, J., TaiYun, R., and, S.M.: Sgd: Saccharomyces genome database. Nuclc Acids Research (1), 1 (1998)
48. Saccharomyces Genome Deletion Project. http://yeastdeletion.stanford.edu/
49. Zhang, R., Lin, Y.: Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nuclc Acids Research **37**(Database issue), 455–8 (2008)
50. Tu, B., P.: Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. Ence **310**(5751), 1152 (2005)
51. Östlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S., Frings, O., Sonnhammer, E.L.L.: InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Research **38**(suppl_1), 196–203 (2009). doi:10.1093/nar/gkp931. https://academic.oup.com/nar/article-pdf/38/suppl_1/D196/16772408/gkp931.pdf
52. Peng, X., Wang, J., Zhong, J., Luo, J., Yi, P.: An efficient method to identify essential proteins for different species by integrating protein subcellular localization information. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2015)
53. Binder, J.X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S.I., Schneider, R., Jensen, L.J.: COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database **2014** (2014). doi:10.1093/database/bau012. bau012. https://academic.oup.com/database/article-pdf/doi/10.1093/database/bau012/8244417/bau012.pdf
54. Holman, A.G., Davis, P.J., Foster, J.M., Carlow, C.K., Kumar, S.: Computational prediction of essential genes in an unculturable endosymbiotic bacterium, wolbachia of brugia malayi (2009)
55. Ping, P., Wang, L., Kuang, L., Ye, S., Iqbal, M.F.B., Pei, T.: A novel method for lncrna-disease association prediction based on an lncrna-disease association network. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 1–1 (2018)
56. Li, J., Li, X., Feng, X., Wang, B., Wang, L.: A novel target convergence set based random walk with restart for prediction of potential lncrna-disease associations. BMC Bioinformatics **20**(1) (2019)
57. Chen, Z., Meng, Z., Liu, C., Wang, X., Wang, L.: A novel model for predicting essential proteins based on heterogeneous protein-domain network. IEEE Access **PP**(99), 1–1 (2020)
58. DAVIS, J.: The relationship between precision-recall and roc curves. In: Proceedings of the 23th International Conference on Machine Learning, 2006 (2006)

Figures

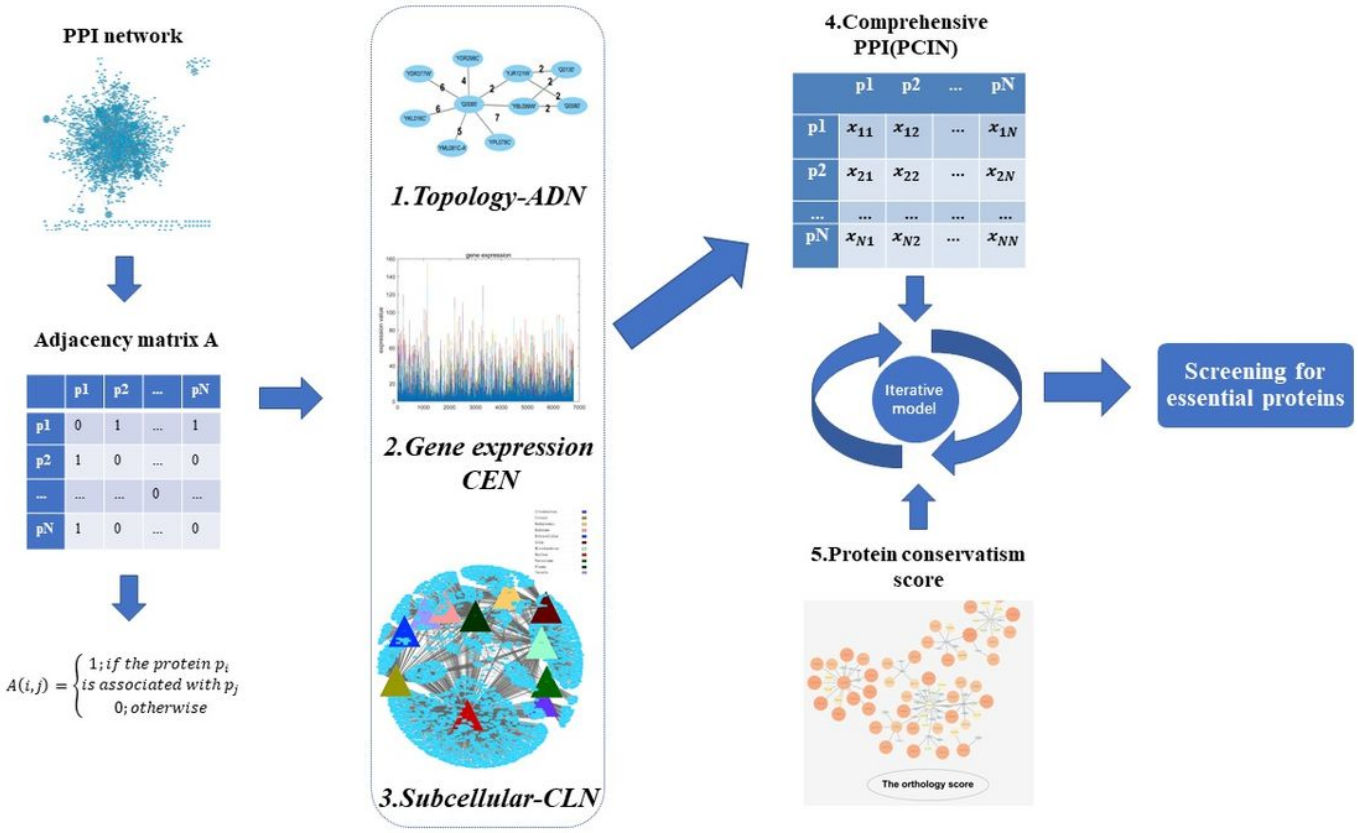


Figure 1

Flow chart of TGSO.(All the subgraphs in this gure were created by the rst author using the open source software Cytoscape in combination with existing experimental data, without any borrow- ing.) The initial PPI, combined with subcellular localization and gene expression data as well as network topology information, was integrated to obtain the com- prehensive protein interaction network, and the network and protein conservative score were put into the iterative model to obtain the nal required protein score.

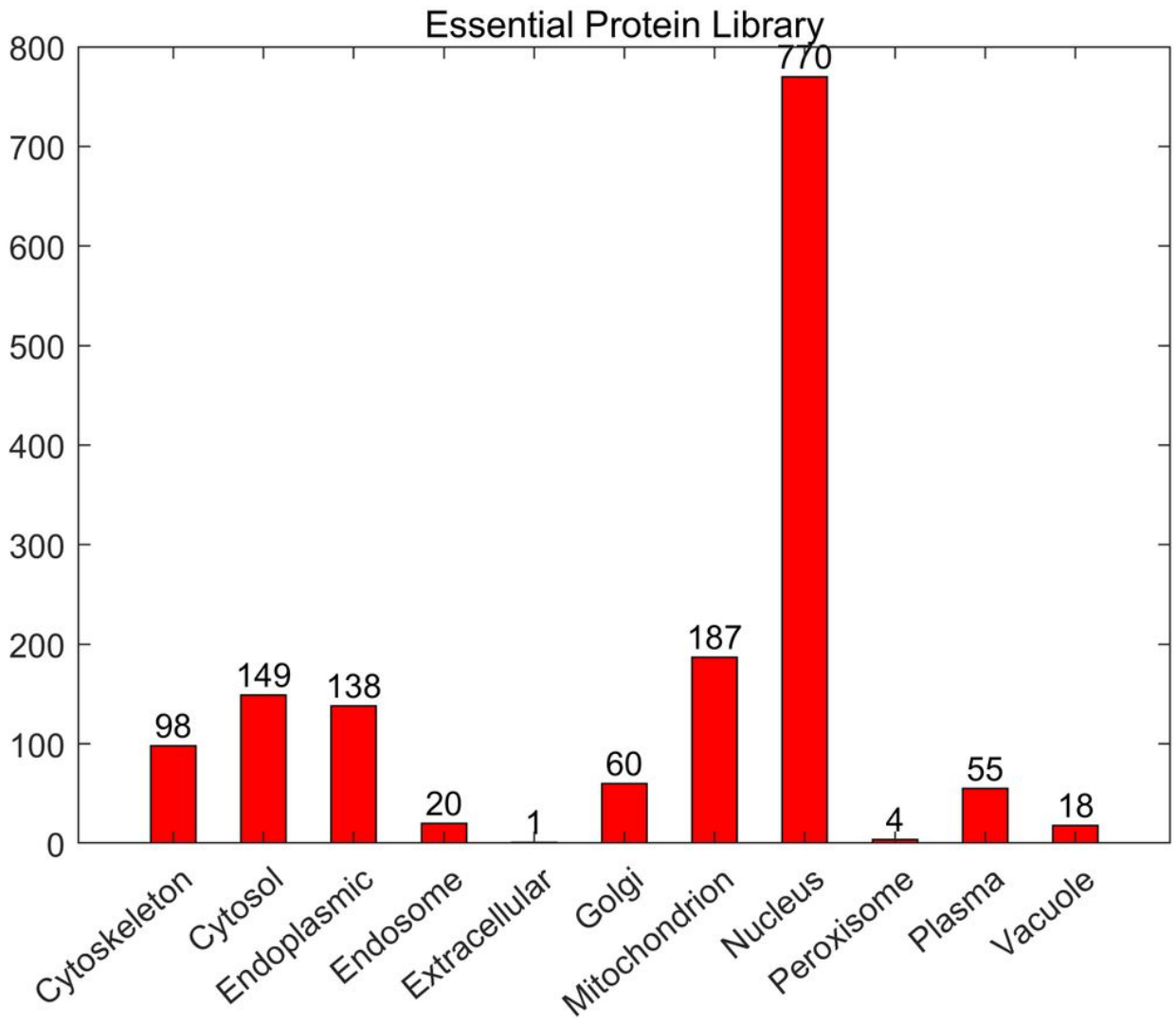


Figure 2

The number of essential proteins about eleven subcellular locations in the Essential Protein Library.

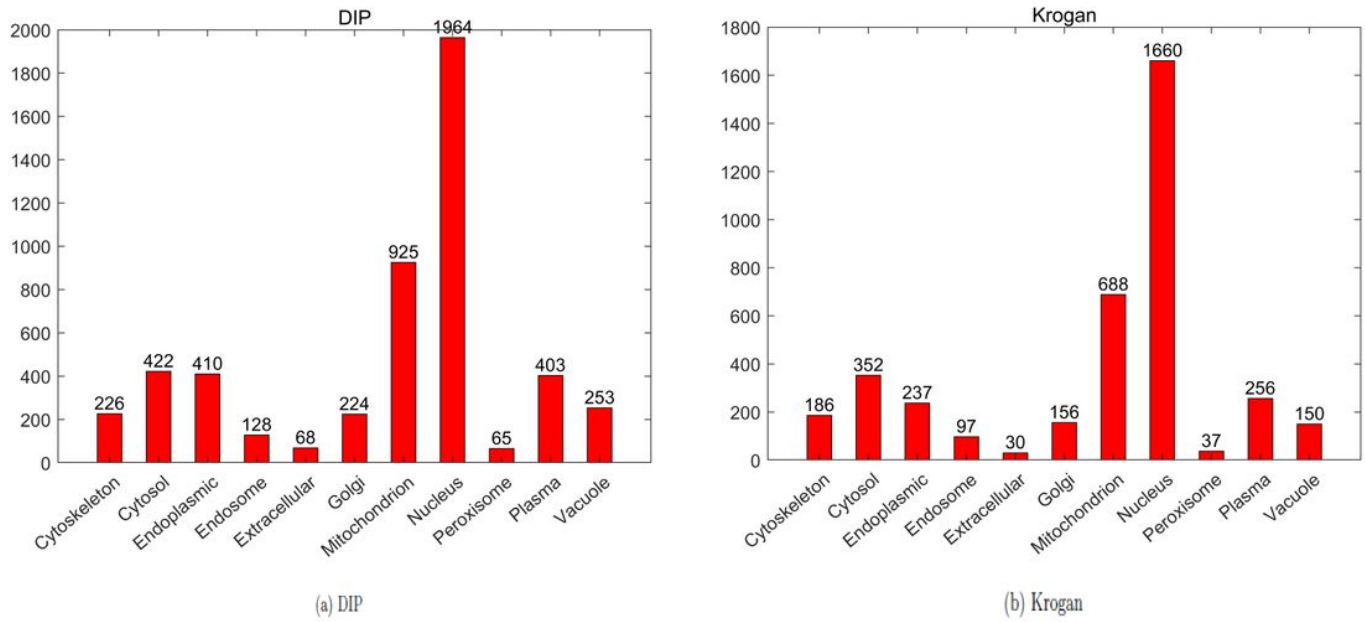


Figure 3

The number of proteins about eleven subcellular locations in the DIP and Krogan protein databases.

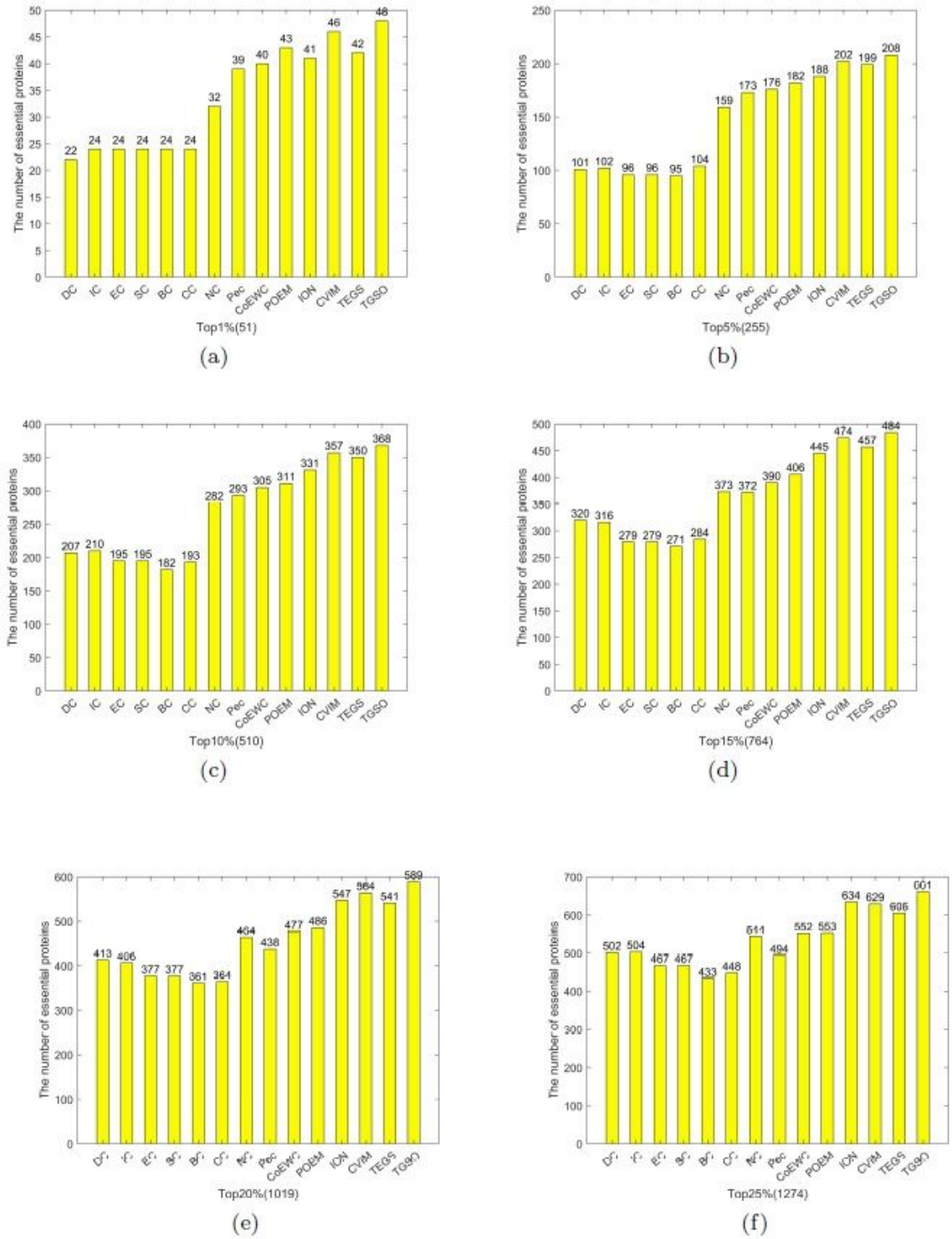


Figure 4

(a) Top 1% ranked proteins. (b) Top 5% ranked proteins. (c) Top 10% ranked proteins. (d) Top 15% ranked proteins. (e) Top 20% ranked proteins. (f) Top 25% ranked proteins. This gure illustrates the comparison of the number of essential proteins predicted by TGSO and 13 competing methods on the DIP dataset. The graph shows the number of truly essential proteins found by each method. The numbers in parentheses indicate the number of proteins ranked in each highest percentage.

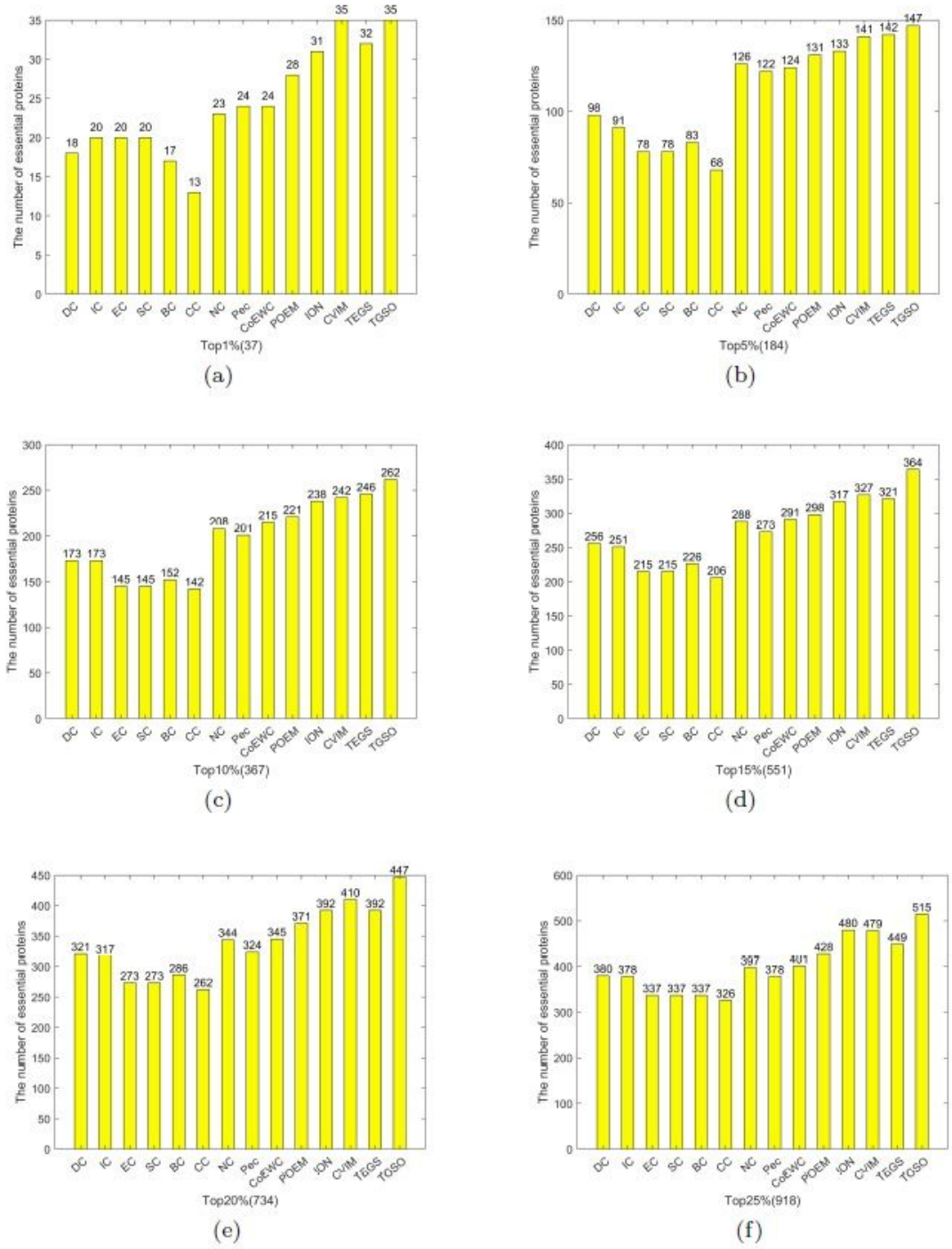


Figure 5

(a) Top 1% ranked proteins. (b) Top 5% ranked proteins. (c) Top 10% ranked proteins. (d) Top 15% ranked proteins. (e) Top 20% ranked proteins. (f) Top 25% ranked proteins. This gure illustrates the comparison of the number of essential proteins predicted by TGSO and 13 competing methods on the Krogan dataset. The graph shows the number of truly essential proteins found by each method. The numbers in parentheses indicate the number of proteins ranked in each highest percentage.

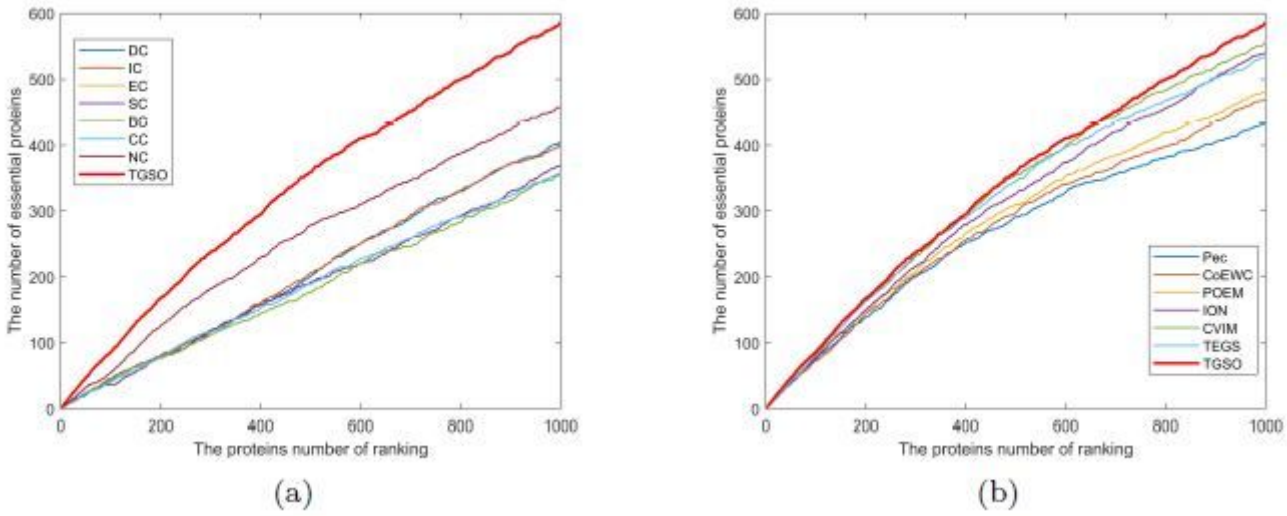


Figure 6

Comparison of Jackknife curves of TGSO and 13 other methods under the DIP database. (a) Comparison between TGSO and DC, IC, EC, SC, BC, CC, NC. (b) Comparison between TGSO and Pec, CoEWC, POEM, ION, CVIM, TEGS.

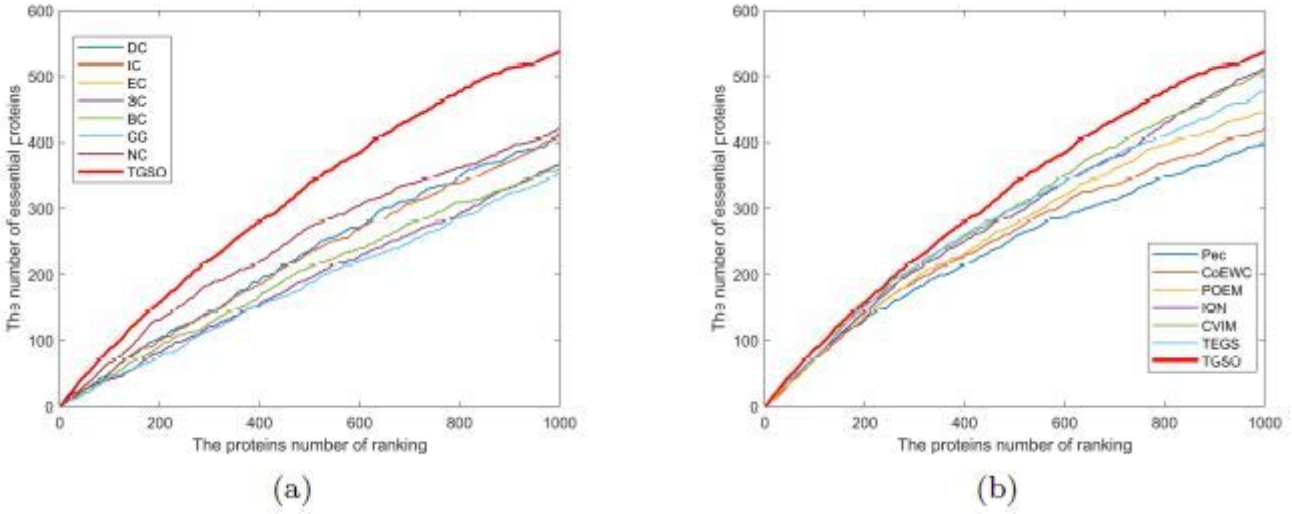
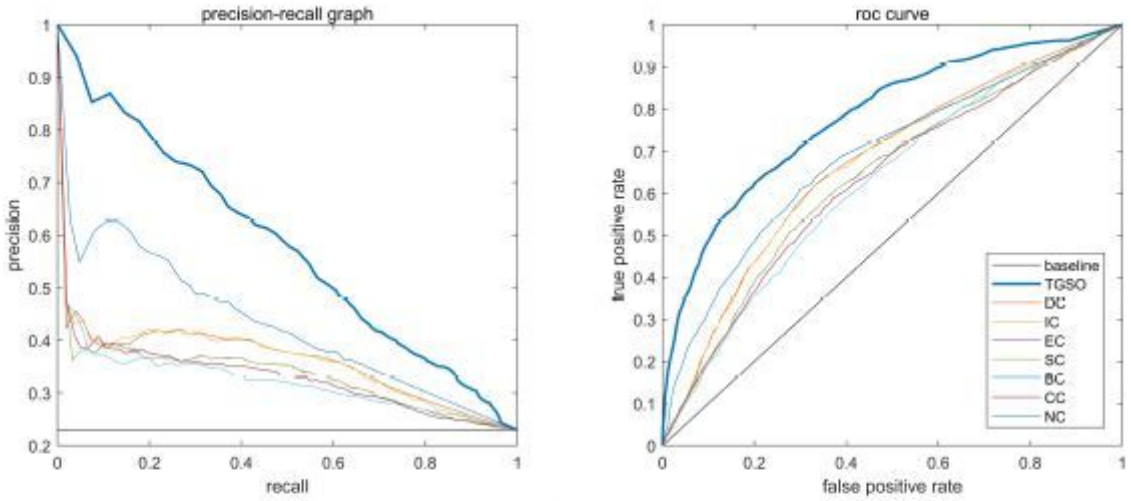
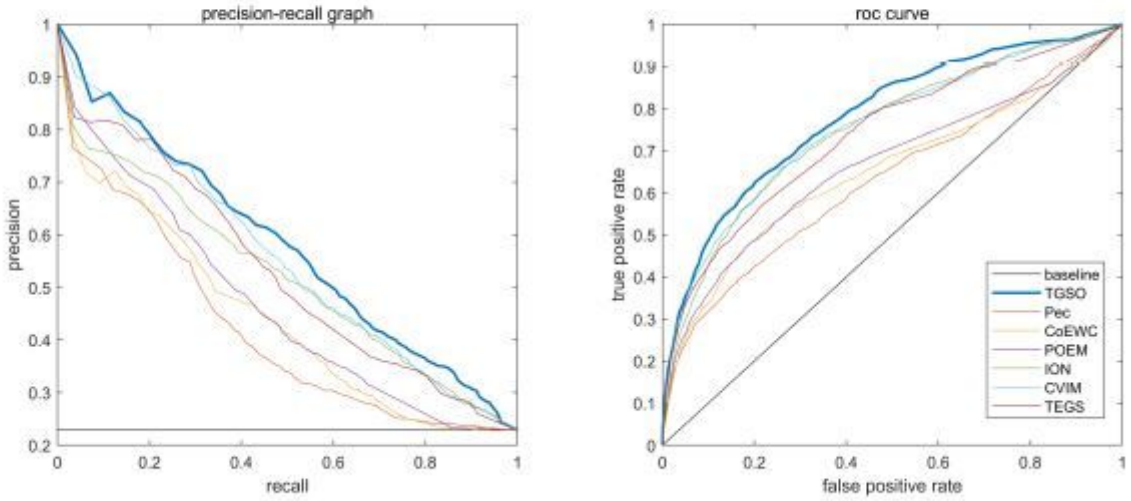


Figure 7

Comparison of Jackknife curves of TGSO and 13 other methods under the Krogan database. (a) Comparison between TGSO and DC, IC, EC, SC, BC, CC, NC. (b) Comparison between TGSO and Pec, CoEWC, POEM, ION, CVIM, TEGS.



(a)



(b)

Figure 8

ROC curve and PR curve of various methods of PPI network based on the DIP database. (a) Comparison of TGSO with DC, EC, IC, SC, BC, CC and NC. (b) Comparison of TGSO with Pec, CoEWC, POEM, ION, CVIM and TEGS.

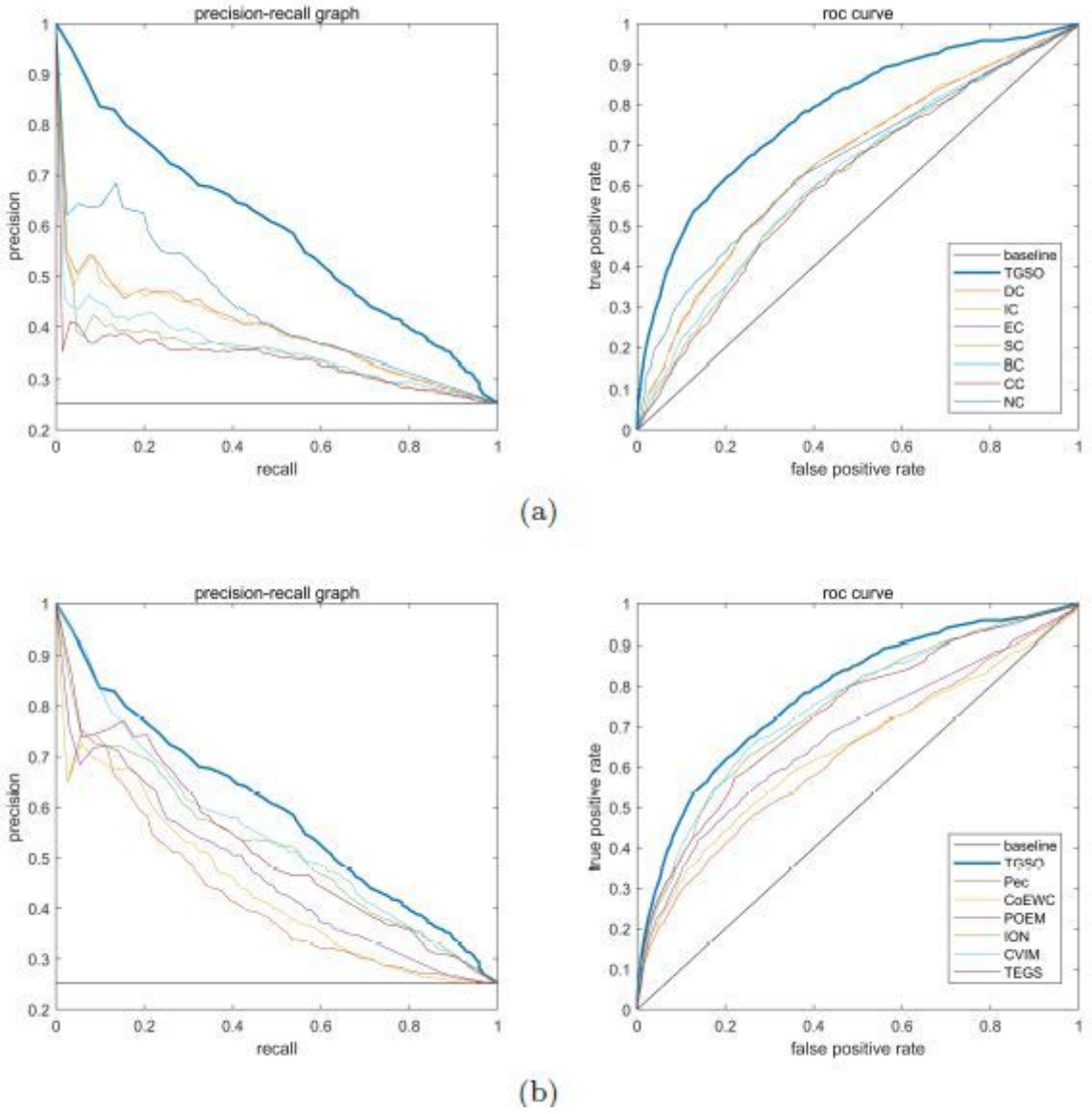


Figure 9

ROC curve and PR curve of various methods of PPI network based on the Krogan database. (a) Comparison of TGSO with DC, EC, IC, SC, BC, CC and NC. (b) Comparison of TGSO with Pec, CoEWC, POEM, ION, CVIM and TEGS.

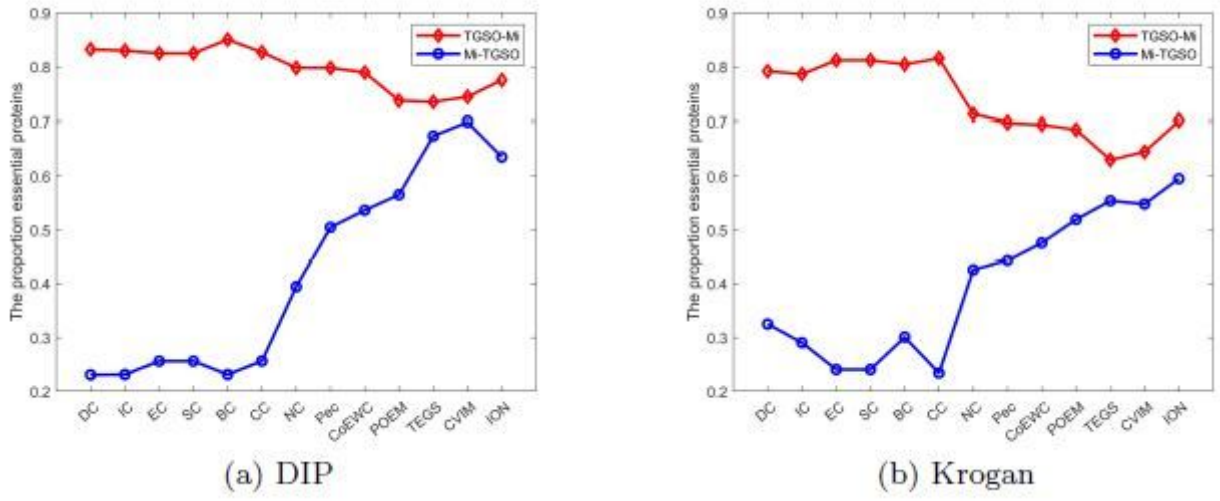


Figure 10

The X-axis represents 13 competing methods. The Y-axis represents the proportion of real key proteins in Mi-TGSO or TGSO-Mi.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- bmcartbiblio.sty
- bmcarticle.bib
- bmcart.cls
- bmcarticle1.tex
- bmcarticle1.bbl