Health Information Science and Systems

RESEARCH

Essential proteins discovery based on dominance relationship and neighborhood similarity centrality

Gaoshi Li[1,2,3], Xinlong Luo[1,2,3], Zhipeng Hu[1,2,3], Jingli Wu[1,2,3*], Wei Peng[4*], Jiafei Liu[1,2,3] and Xiaoshu Zhu[1,2,3,5]

Abstract

Essential proteins play a vital role in development and reproduction of cells. The identification of essential proteins helps to understand the basic survival of cells. Due to time-consuming, costly and inefficient with biological experimental methods for discovering essential proteins, computational methods have gained increasing attention. In the initial stage, essential proteins are mainly identified by the centralities based on protein–protein interaction (PPI) networks, which limit their identification rate due to many false positives in PPI networks. In this study, a purified PPI network is firstly introduced to reduce the impact of false positives in the PPI network. Secondly, by analyzing the similarity relationship between a protein and its neighbors in the PPI network, a new centrality called neighborhood similarity centrality (NSC) is proposed. Thirdly, based on the subcellular localization and orthologous data, the protein subcellular localization score and ortholog score are calculated, respectively. Fourthly, by analyzing a large number of methods based on multi-feature fusion, it is found that there is a special relationship among features, which is called dominance relationship, then, a novel model based on dominance relationship is proposed. Finally, NSC, subcellular localization score, and ortholog score are fused by the dominance relationship model, and a new method called NSO is proposed. In order to verify the performance of NSO, the seven representative methods (ION, NCCO, E_POC, SON, JDC, PeC, WDC) are compared on yeast datasets. The experimental results show that the NSO method has higher identification rate than other methods.

Keywords: Essential proteins, Neighborhood similarity centrality, Protein–protein interaction, Multi-feature fusion, Dominance relationship

Introduction

Essential/lethal proteins are one of the most critical macromolecules in living organisms, and their deficiency can lead to stopping growth, reproduction, and even death [1–3] of cells. Therefore, it has great significance to predict essential proteins, which helps reveal cellular molecular mechanisms [4] and discovers new biomarkers and drug targets [5].

There are roughly two types of methods for predicting essential proteins. One is biological experimental methods, and the other is computational methods. The biological experimental methods include gene knockout [6], RNA interference [7], and conditional knockouts [8], etc., while are time-consuming, expensive, and low efficient. Therefore, there is an urgent need to develop rapid, economical, high efficient essential proteins identification methods. The computational methods meet the need.

These computational strategies depend on one or many features, which can generally be divided into topology-based features and sequence-based ones. The topology-based features find the essentiality of nodes (proteins) in protein–protein interaction (PPI) networks. Degree centrality (DC) [9], betweenness centrality (BC) [10], closeness centrality (CC) [11], subgraph centrality (SC) [12],

*Correspondence: wjlhappy@mailbox.gxnu.edu.cn; weipeng1980@gmail.com
[1] Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin 541004, China
[4] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, China
Full list of author information is available at the end of the article

eigenvector centrality (EC) [13], information centrality (IC) [14] and neighborhood centrality (NC) [15],etc., are the representative topology-based methods/features. They can intuitively reflect the importance of proteins in the networks, but their recognition rate is low due to a large number of false positives in PPI networks.

The sequence-based features can be extracted based on different data, such as subcellular localization data [16], evolutionary conservation data [17–19], and gene expression data [20, 21]. However, due to the heterogeneity and incomparability of these data, the relations among extracted features from these data are less analyzed how affect the identification rate of essential proteins.

In order to improve essential proteins recognition rate, more and more researchers propose fusion methods, some are based on data fusion, others are based on multi-feature fusion. Wang et al. [22] constructed a dynamic proteins interaction network (DPIN) by combining gene expression data to PPI data, that calculated the activity threshold of each gene based on the gene expression data used the three-sigma principle. Then, many methods were proposed by fusing gene expression data to PPI data, such as NF-APIN [23], DPPN [24]. Then other data was used to purify PPI networks. Li et al. [25] constructed a spatial and temporal active proteins interaction network (ST-APIN) by integrating time-course gene expression data and subcellular localization information. From these representative methods, one can see that the effect of false positives in PPI data is reduced by fusing other data to PPI data.

The multi-feature fusion methods integrate features using fusion model, such as linear model, random walk model, Pareto Optimal Consensus model. Tang et al. [26] extracted two features, edge clustering coefficient (ECC) and Pearson correlation coefficient (PCC) from PPI network data and gene expression data, respectively. Then, a linear model was used to fuse these two features, and the WDC method was proposed. The OGN [27] method used a linear model to integrate two types of features extracted from ortholog information, gene expression profiles and PPI networks. Li et al. [28] extracted three features from PPI data, subcellular localization data and orthologous data, respectively, used a linear model to fuse them, and proposed a new method SON.

Multiplication is also used for multi-feature fusion. The PeC method [29] used multiplication to combine ECC and PCC to calculate the scores of the proteins. The JDC method [30] multiplied two classes of features, Jaccard similarity coefficient and ECC, to predict essential proteins. In the TEGS method [31], multiplication is used to combine ECC, SLC, PCC, and GO_sim.

Meanwhile, random walk model and its extended models are also used for multi-feature fusion. ION [32] used the random walk model to fuse two features extracted from ortholog information and PPI information. An extended random walk model was also adopted to integrate subcellular localization and ortholog information in the NTMEP method [33].

Li et al. found a phenomenon among multiple features, if any feature score of protein A is higher than that of protein B, then protein A is more likely to be an essential protein than protein B. This phenomenon meets Pareto Optimal Consensus (POC) theory. NCCO [34] combined orthologous feature and neighborhood closeness centrality (NCC) using an extended POC model. The E_POC method [35] also fused two kinds of features based on an extended POC model.

From the above know, there are mainly two kinds of the features used to find essential proteins, one is topology-based features, the other is sequence-based features. Due to a large number of false positives in PPI networks, the topology-based methods/features have low identification rate. The multi-feature fusion methods have higher discovery rate, but they seldom consider the relation among features. The NCCO and E_POC methods think the relation among features meet POC theory, but it is not comprehensive. After a large number of analyzing, it is found that the relation among features is more meet the dominance relation in this study. When most feature values of protein A have a large enough dominance over protein B, even if proteins A has small relatively weak feature values, protein A should prefer to be the essential protein. This phenomenon is called protein A dominates protein B.

For these cases, to reduce the reflect of false positives in PPI networks, a purified PPI network is firstly constructed in this paper. Then a new centrality with high recognition rate, neighborhood similarity centrality (NSC), is proposed based on the purified PPI network. Next, the subcellular localization score (Sub) and the ortholog score (OS) of each protein are calculated based on the subcellular localization data and the orthologous data, respectively. Finally, NSC, Sub and OS are fused based on a dominance relationship model, a new method called NSO, is proposed. In order to verify the performance of NSO, yeast datasets are used to test, and seven representative essential proteins identification methods, such as ION, NCCO, E_POC, SON, JDC, PeC and WDC, are compared. The experimental results show that the NSO method has higher identification rate than other methods.

Methods

By summarizing the existing methods, we find their common deficiencies. (1) These features have low recognition rate due to the influence of many false positives in input data; (2) these methods seldom consider relationships among features or do not consider comprehensively. To face these deficiencies, in this paper, the NSO method is proposed. Its overall flow figure is shown in Fig. 1. (1) Construction of purified PPI network. To reduce the influence of false positives in PPI network, the gene expression data is fused with the original PPI network, a purified PPI network is constructed. (2) Extraction of neighborhood similarity centrality NSC. By analyzing the similarity between proteins and their neighbors in the purified PPI network, NSC is extracted. (3) Extraction of subcellular localization score Sub. Sub is extracted from subcellular localization data. (4)Extraction of ortholog score OS. OS is extracted from orthologous proteins data. (5) Dominance relationship and NSO algorithm. The dominance relationship model is developed to fuse the three scores of proteins to obtain the final scores, and the NSO algorithm is proposed in full. The process is described in detail as follows.

Construction of purified PPI network

The construction of purified PPI networks [22] is based on the co-expression principle. The co-expression principle is that two proteins occur interaction only when they are all expression state at the same time. The interactions in PPI network are deleted when two proteins are not expression state at the same time. So, to determine the expression state of proteins is important. In this paper, the steps to determine the expression state of proteins are described as follows. Firstly, an activity threshold *Act_th* is used to determine whether a gene is expression state or not, which is defined as follows:

$$Act\_th(i) = u(i) + 3\sigma(i) \times (1 - F(i)) \tag{1}$$

$$F(i) = 1/\left(1 + \sigma(i)^2\right) \tag{2}$$

where $u(i)$ and $\sigma(i)$ denote the mean and standard deviation of gene expression values of protein i, respectively, $F(i)$ denote the volatility of gene expression values of protein i. If the expression level of a gene exceeds its *Act_th* at a certain time point, it is considered to be in an expression state at that instant. Then, the gene expression matrix is processed into the gene expression activity matrix (*GM*), which is defined as:

$$GM(i,t) = \begin{cases} 1, g(i,t) > Act\_th(i) \\ 0, otherwise \end{cases} \tag{3}$$

where $g(i,t)$ represents the gene expression value of protein i at time t. Secondly, the original PPI network is converted into an adjacency matrix (*ADM*). Then, *GM* is

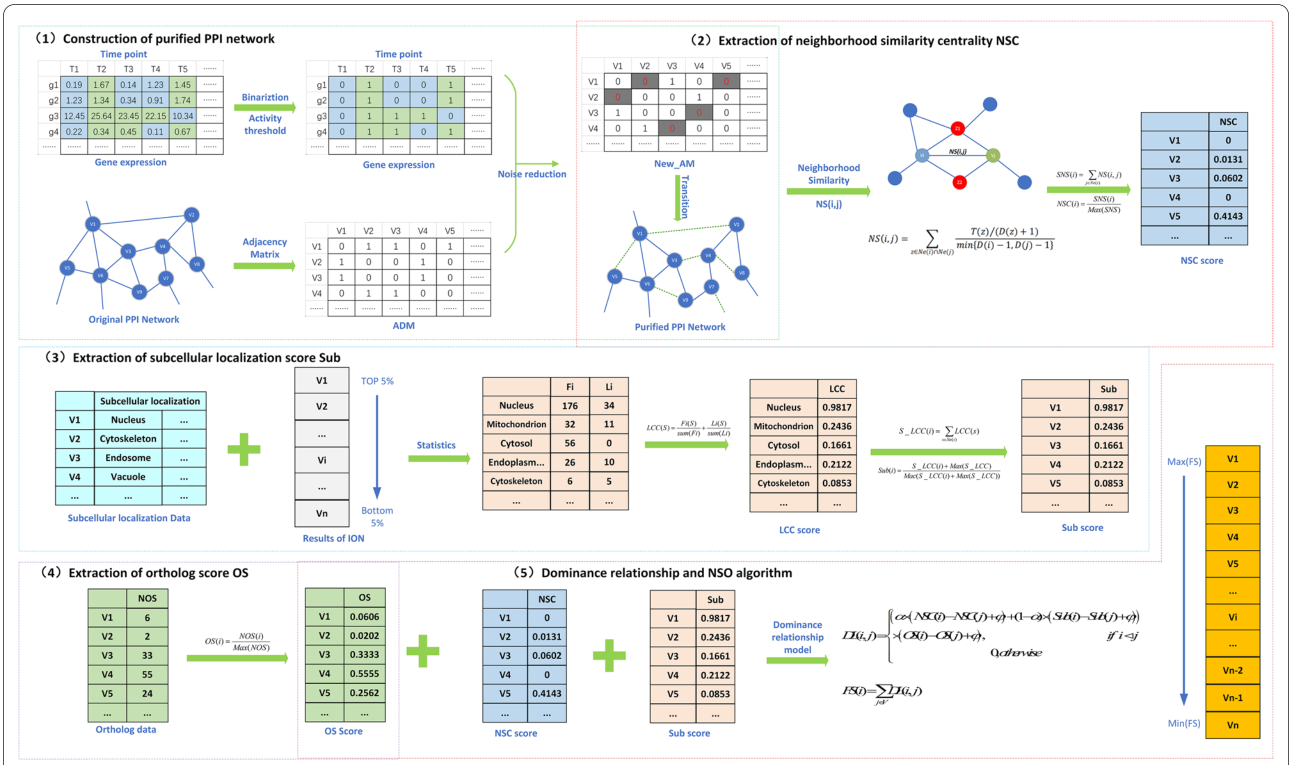

Fig.1 Theoverall flow of the NSO method

integrated into *ADM* by co-expression principle, a new purified adjacency matrix is obtained and denoted as *New_AM*, which is defined by the formula:

$$New\_AM\left(i,j\right)=\left\{\begin{array}{c}1, ifADM\left(i,j\right)=1\ and\\ \exists t, GM(i,t)=GM\left(j,t\right)=1\\ 0, otherwise\end{array}\right. \quad (4)$$

where *i,j* ∈ *V (i,j* = *1,..., N), N* is the total number of proteins in PPI network. Finally, *New_AM* is converted into a purified PPI network.

Extraction of neighborhood similarity centrality NSC

By analyzing the similar relationship between proteins and its neighbors in the purified PPI network, a phenomenon is found. If proteins *i* and *j* exist an interaction, protein *z* is their common neighbor, *T(z)* and *D(z)* are the number of triangles formed by protein *z* with its neighbors and the degree of protein *z*, respectively, the ratio of *T(z)* to *D(z)* affects the strength of interaction between proteins *i* and *j*. This phenomenon is described as the neighborhood similarity coefficient (*NSC*). Firstly, neighborhood similarity (*NS*) relation is defined as follows:

$$NS\left(i,j\right)=\sum_{z\in Ne(i)\cap Ne\left(j\right)}\frac{T(z)/(D(z)+1)}{min\left(D(i)-1,D\left(j\right)-1\right)} \quad (5)$$

where $Ne(i)\cap Ne(j)$ represents the common neighbor set between proteins *i* and *j*. *D(i)*, *D(j)* and *D(z)* denote the degrees of proteins *i*, *j* and *z*, respectively. *min()* is the minimum function. *D(z)* + *1* in the denominator is to prevent the case that *D(z)* is zero. The larger the *NS* value the interaction, the more reliable the interaction is. The S*NS(i)* value of protein *i* is the sum of the *NS* values of all neighbors of protein *i* in the purified PPI. Therefore, for protein *i*, its *SNS(i)* is defined as follows:

$$SNS(i)=\sum_{j\in Ne(i)}NS\left(i,j\right) \quad (6)$$

where *Ne(i)* represents the set of neighbors of protein *i*. In order to keep each feature value in the same range, *SNS* value is normalized to the neighborhood similarity centrality (*NSC*) value, which is defined as follows:

$$NSC(i)=\frac{SNS(i)}{max(SNS)} \quad (7)$$

where *max()* represents the maximum value function.

Extraction of subcellular localization score Sub

Studies have shown that the subcellular localization of protein is related to its function. Thus, subcellular localization information facilitates the recognition of essential proteins. The number of occurrences of the subcellular localizations of the top/bottom 5% proteins in results of ION [32] are counted and denoted as *Fi(s)* and *Li(s)*,respectively. Where *s* represents the common 11 subcellular localization. Then, subcellular localization coefficient LCC(s) is defined as:

$$LCC(s)=\frac{Fi(s)}{sum(Fi)}+\frac{Li(s)}{sum(Li)} \quad (8)$$

where, *sum(Fi)* and *sum(Li)* represent the total number of subcellular localizations in the top/bottom 5% of proteins, respectively. For a protein *i*, its subcellular localization score *S_LCC(i)* is defined as the sum of *LCC(s)* of all the subcellular localizations it appears.

$$S\_LCC(i)=\sum_{s\in Sn(i)}LCC(s) \quad (9)$$

where *Sn(i)* represents a set of subcellular localizations of protein *i*. To balance the effect of different scores, *S_LCC(i)* is normalized to obtain the subcellular localization score,*Sub(i)*, using the following formula:

$$Sub(i)=\frac{S\_LCC(i)+max(S\_LCC)}{Max(S\_LCC(i)+max(S\_LCC))} \quad (10)$$

where *max()* is the maximum value function.

Extraction of ortholog score OS

Ortholog score *OS* [32] is used to measure the conservation property of proteins. Usually, the higher the ortholog score of a protein, the more conserved it is, the more likely to be essential. The *OS(i)* of protein *i* is defined as follows:

$$OS(i)=\frac{NOS(i)}{max(NOS)} \quad (11)$$

where *NOS(i)* represents the number of ortholog reference species of protein *i* exists, and *max()* represents the maximum value function.

Dominance relationship and NSO algorithm

By analyzing a large number of essential proteins recognition methods based on multi-feature fusion, a phenomenon among features is found. When most feature values of protein *i* have a large enough dominance over protein *j*, even if proteins *i* has small relatively weak feature values, protein *i* should prefer to be the essential protein. This phenomenon is called protein *i* dominates protein *j*. The dominance relationship is defined as follows:

$$(1)\ \sum_{m=1}^{n}A_m(i)>\sum_{m=1}^{n}A_m\left(j\right)$$
$$(2)\ \forall m, A_m(i)>A_m\left(j\right)-\varphi$$

where A denotes the set of features $A(A_1,A_2,\ldots,A_n)$, A_m denote the mth feature value, φ represents the regulatory factor and is a small positive number. If protein i satisfies the characteristic dominance relation for j, it is denoted as: $i \lhd j$. Then, according to the dominance relation, a dominance relationship model is designed for multi-feature fusion. The design idea of fusion model is as follows: for three feature scores *NSC*, *Sub* and *OS*, *NSC* and *Sub* should be firstly fused due to interaction between proteins occurs only in the same subcellular localization. The dominance relationship model is defined as follows:

$$DL(i,j) = \begin{cases} \left(\alpha \times \left(NSC(i) - NSC(j) + \varphi\right) + (1-\alpha) \times \left(Sub(i) - Sub(j) + \varphi\right)\right) \\ \quad \times \left(OS(i) - OS(j) + \varphi\right),\ if\ i \lhd j \\ 0, \quad otherwise \end{cases} \tag{12}$$

where α is between [0,1], φ represents the regulatory factor, which is a small positive number. The final score for protein *i*, *FS(i)*, is defined as:

$$FS(i) = \sum_{j \in V} DL\left(i,j\right) \tag{13}$$

where V represents the set of all nodes in the PPI network. All proteins are then sorted in descending order according to their *FS* values. A protein with higher rank is more likely to be essential. The NSO algorithm is described as follows:

```
Algorithm 1: NSO algorithm
Input: PPI network $G=(V, E)$, subcellular localization score table $S$, orthologous feature score table $TM$, gene expression matrix $M$, parameter $\alpha,\varphi$.
1: According to the graph $G$ and matrix $M$, the purified PPI network is constructed using equations (1)-(4).
2: Based on the purified PPI network, the $NSC$ value of each protein is calculated using Equation (5)-(7).
3: The $Sub$ value of each protein is calculated using Equation (8)-(10) according to table $S$.
4: According to table $TM$, the $OS$ value of each protein is calculated using Equation (11).
5: for each protein in graph $G$ do
     for all proteins that satisfy $i \lhd j$ in $G$ do
          the $DL$ value is calculated using Equation (12).
     end for
   end for
6: The $FS$ values of all proteins are calculated according to Equation (13) and all proteins are sorted by their $FS$ values.
Output: Output the top 600 ranked proteins sorted by $FS$ in descending order as candidate essential proteins.
```

Results

Before the analysis of the performance of NSO, the parameter α and the regulatory factor φ in the dominance relationship model are firstly analyzed. To adequately analyze the performance of NSO, multiple yeast datasets are used to test, and seven representative essential proteins identification methods, such as ION, NCCO, E_POC, SON, JDC, PeC and WDC, are compared. Three mainstream validation methods are employed: histogram, Precision–Recall curve, and Jackknife curve. Furthermore, in order to verify the effectiveness of NSC, it is compared against six classical centrality methods: DC, IC, SC, CC, EC, and NC.

Experimental data

There are three PPI networks on S.cerevisiae (yeast) in this study. The first PPI network was downloaded from the DIP database [36], contained 5,093 proteins and 24,743 interactions, named Y5093. The second PPI network was constructed by Yu et al. [37] and contained 4,743 proteins and 23,294 interactions, called Y4743. The third PPI network derived from a paper [38], which contained 2,708 proteins and 7,123 interactions, named Y2708.

The gold standard dataset of essential proteins integrated MIPS [39], SGD [40], DEG [4] and SGDP [41], which contained 1,285 true essential proteins on S.cerevisiae.

The gene expression dataset of *S. cerevisiae* was constructed by Tu et al. [42], which contained 6,777 genes expression values at 36 time points for sampling.

The subcellular localization dataset of *S. cerevisiae* was obtained from COMPARTMENTS database [43], which contained 20,6831 subcellular localization records on 5,095 proteins.

The ortholog dataset of *S. cerevisiae* derived from Version 7 of the InParanoid database [44],which was a set of pairwise comparisons of 100 whole genomes (99 eukaryotes and 1 prokaryote) constructed by the INPARANIOD program.

Analysis of parameter α and regulatory factor φ

In the NSO algorithm, two parameters are included, which are the parameter α and the regulatory factor φ. The parameter α is used to adjust the contributions of NSC and Sub; its value is set as 0, 0.1,..., 1,respectively. The regulatory factor φ regulates the dominance relationship between proteins in certain feature value; its value is set as 0, 0.01,..., 0.1,respectively.

Table 1 shows the identification results by different values of α on Y5093 to analyze impact of parameter α on the performance of NSO. As shown in Table 1, when α values from 0.7 to 0.9, the result of NSO is better. In particular, when α is 0.9, the identification result is the best. Therefore, in this paper, the optimal value of α is set as 0.9.

The discovery results based on different values of φ on Y5093 is shown in Table 2 to analyze impact of the regulatory factor φ on the performance of NSO. In Table 2, when φ is 0, the identification performance of NSO is relatively poor; when it is set as other values, NSO can predict more essential proteins. In particular, when φ is 0.09, the overall identification capability of NSO is awfully excellent, so φ is set to 0.09.

Validated by histograms

The histogram is used to verify the performance of the NSO method and other methods. The top 100, 200, 300, 400, 500, and 600 proteins identified by these methods, respectively, are selected as candidate essential proteins. Then, based on the gold standard set of essential proteins, the number of true essential proteins correctly identified by these methods is counted.

Figure 2 shows the number of true essential proteins correctly identified by these methods on Y5093. As shown in Fig. 2a, NSO finds 92 true essential proteins, E_POC and NCCO find 84 true essential proteins, respectively, SON identifies 81 essential proteins, the other methods do not exceed 79. In Fig. 2b, the number of essential proteins identified by NSO is the most, reaching 176. Compared with SON(161), E_POC(157), NCCO(157), JDC (152), ION(150), WDC(132) and PeC(133), the number of essential proteins identified by NSO(176) increases 15, 19, 19, 24, 26, 44 and 43, respectively. As shown in Fig. 2c, NSO is only method that finds more than 253 essential proteins. The number of true identified essential proteins of SON, E_POC, NCCO, JDC, ION, WDC and PeC are 232, 227, 227, 220, 216, 196 and 189, respectively. In Fig. 2d, SON, E_POC, NCCO, JDC, ION, WDC and PeC discovery 293, 284, 282, 267, 280, 242 and 247 essential proteins, respectively. Compared with these methods, NSO finds 319 essential proteins, that improves by 8.87%, 12.32%, 13.12%, 19.47%, 13.92%, 31.81% and 29.15%, respectively. In Fig. 2e, NSO(371) is 25, 33, 34, 57, 45, 86 and 72 more than SON, E_POC, NCCO, JDC, ION, WDC and PeC, respectively. In Fig. 2f, 403, 394, 392, 356, 374, 324 and 341 essential proteins are identified by SON, E_POC, NCCO, JDC, ION, WDC and PeC, respectively, while 419 essential proteins are identified by NSO. Overall, NSO correctly identifies the most essential proteins among all methods, and it has the best performance on Y5093.

Figure 3 presents the prediction results of these methods on Y4743. Based on Fig. 3, it is evident that NSO is significantly ahead of the other methods in correctly

Table 1 Impact of parameter α on the performance of NSO on Y5093

α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
TOP 100	77	82	85	86	91	91	91	91	**93**	92	92
TOP 200	167	166	168	169	170	171	168	**172**	171	171	171
TOP 300	239	239	240	244	241	243	243	**247**	246	240	237
TOP 400	295	296	298	298	300	301	302	**305**	304	302	291
TOP 500	346	346	345	345	346	347	347	348	**349**	**349**	346
TOP 600	384	386	387	387	388	387	389	391	393	**395**	389

Bold values indicate the best performing results in each TOP dimension

Table 2 Impact of the regulatory factor φ on the performance of NSO on Y5093

φ	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
TOP 100	92	92	92	92	**93**	92	91	91	91	92	92
TOP 200	171	172	172	173	174	173	174	174	174	**176**	176
TOP 300	240	245	248	249	251	248	250	251	**253**	**253**	251
TOP 400	302	309	307	311	315	315	314	317	**319**	**319**	315
TOP 500	349	366	367	366	368	368	369	370	370	**371**	371
TOP 600	395	416	420	419	418	418	417	420	**421**	419	419

Bold values indicate the best performing results in each TOP dimension

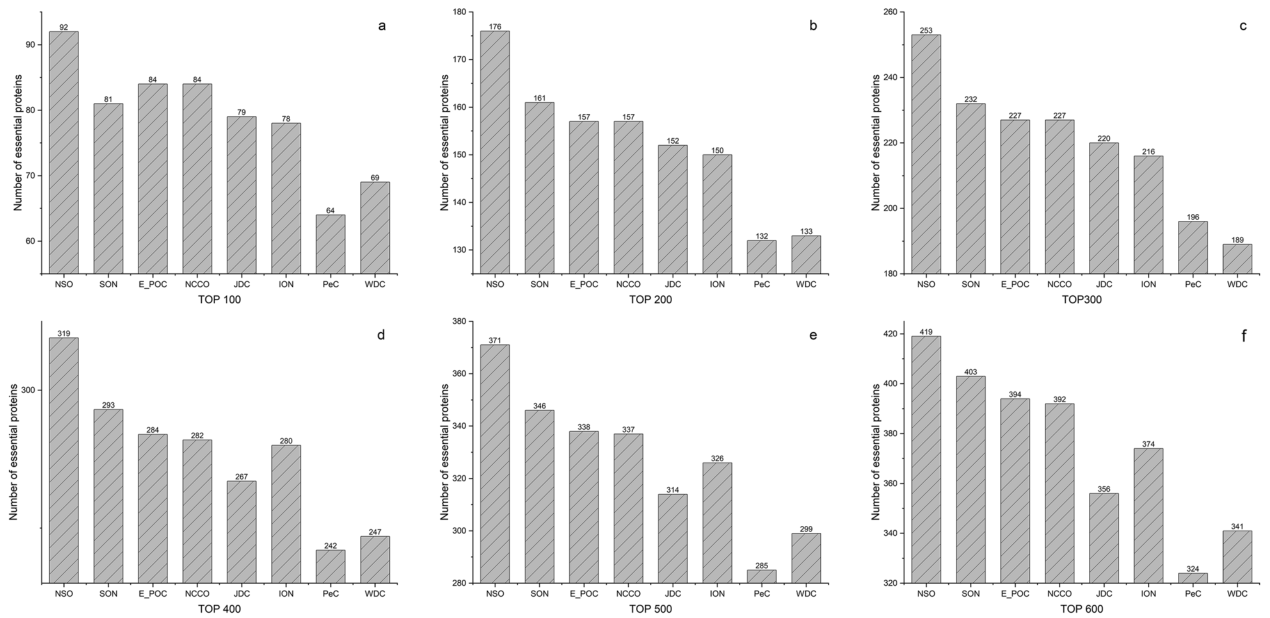

Fig.2 NSO and other methods predict the number of true essential proteins on Y5093

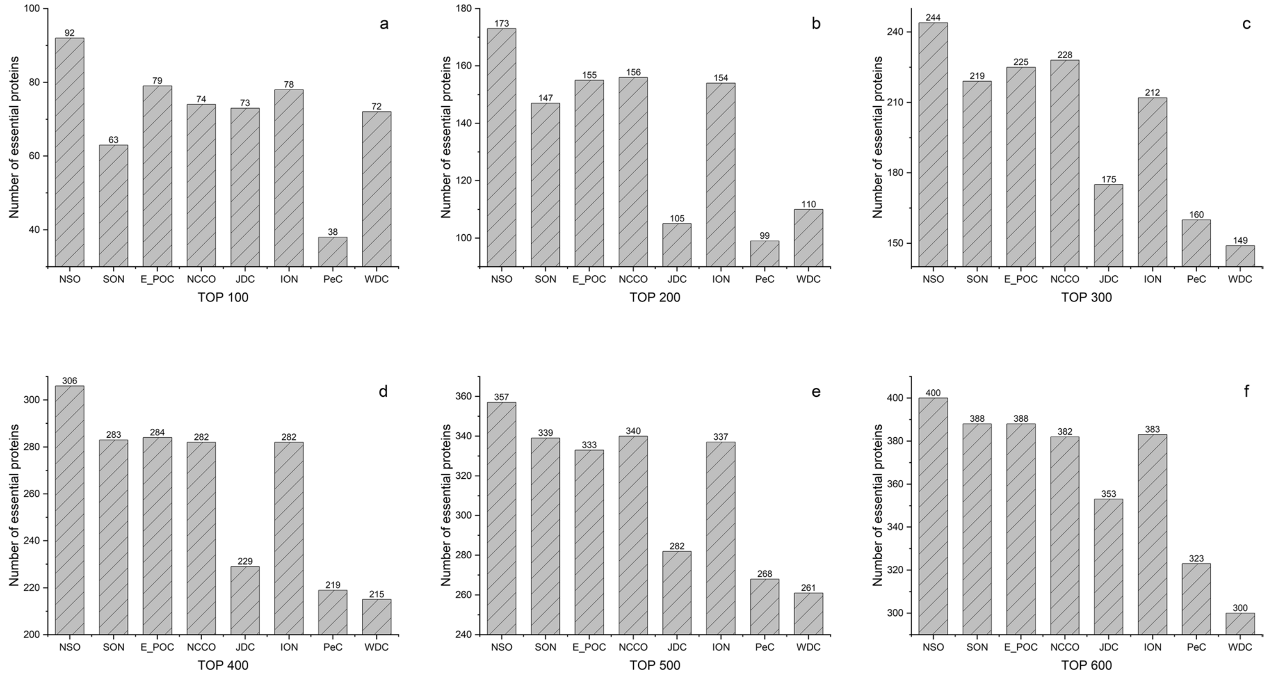

Fig.3 NSO and other methods predict the number of true essential proteins on Y4743

identifying essential proteins. In Fig. 3a, NSO identifies 92 true essential proteins, SON, E_POC, NCCO, JDC, ION, WDC and PeC identify 63, 79, 74, 73, 78, 38 and 72 respectively. In Fig. 3b, there are only four methods that find more than 150 essential proteins, and they are NSO(173), E_POC(155), NCCO(156) and ION(154). NSO is 18, 17 and 19 higher than E_POC, NCCO and ION, respectively. As shown in Fig. 3c, compared with NCCO(228), E_POC(225), SON(219), ION(212), JDC(175), PeC(160) and WDC(149), NSO(244) increases 16, 19, 25, 32, 69, 84 and 95, respectively. In Fig. 3d, NSO identifies 306, which are 22, 23, 24, 24, 77, 87 and 91 higher than E_POC(284), NCCO(283), SON(282), ION(282), JDC(229), PeC(219) and WDC(215),respectively. As shown in Fig. 3e, f, NSO identifies 357 and 400, respectively, which is significantly ahead of the other methods. In summary, NSO finds out the most essential proteins on Y4743.

The results of these methods based on Y2708 are shown in Fig. 4. In Fig. 4a, NSO, SON, E_POC and NCCO identify 80, 79, 90 and 83 essential proteins, respectively, the other methods do not exceed 73. As shown in Fig. 4b, compared with NCCO(148), E_POC(145), SON(143), ION(140), JDC(138), PeC(135) and WDC(141), NSO(159) increases 7.43%, 9.65%, 11.18%, 13.57%, 15.22%, 17.78% and 12.76%, respectively. In Fig. 4c, NSO(230) increases 17, 25, 33, 43, 34, 45 and 38 than SON(213), E_POC(205), NCCO(197), JDC(187), ION(196), WDC(185) and PeC(192), respectively. As shown in Fig. 4d–f, among the TOP 400–600 candidate proteins, NSO and SON are the two methods with the highest recognition rates among these methods. NSO identifies 289, 345, and 384,respectively. The number of true essential proteins discovered by SON is 267, 326 and 379,respectively. Undoubtedly, NSO has the best performance on Y2708.

Comparison based on Precision–Recall curve

The Precision–Recall curve is used to evaluate the performance of algorithm. The larger the AUC area of the Precision–Recall curve, the better the performance of algorithm is.

Figure 5 shows the Precision–Recall curves for these methods on Y5093. The AUC values of NSO, SON and E_POC are 0.4223, 0.4170 and 0.4095, respectively, which

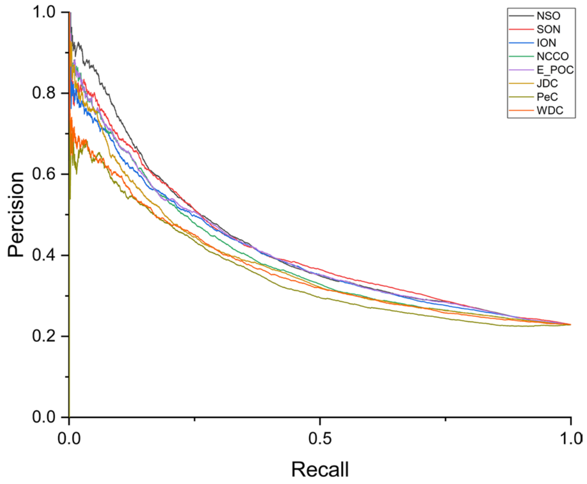

Fig.5 Comparison based on Precision–Recall curves of NSO and other methods on Y5093

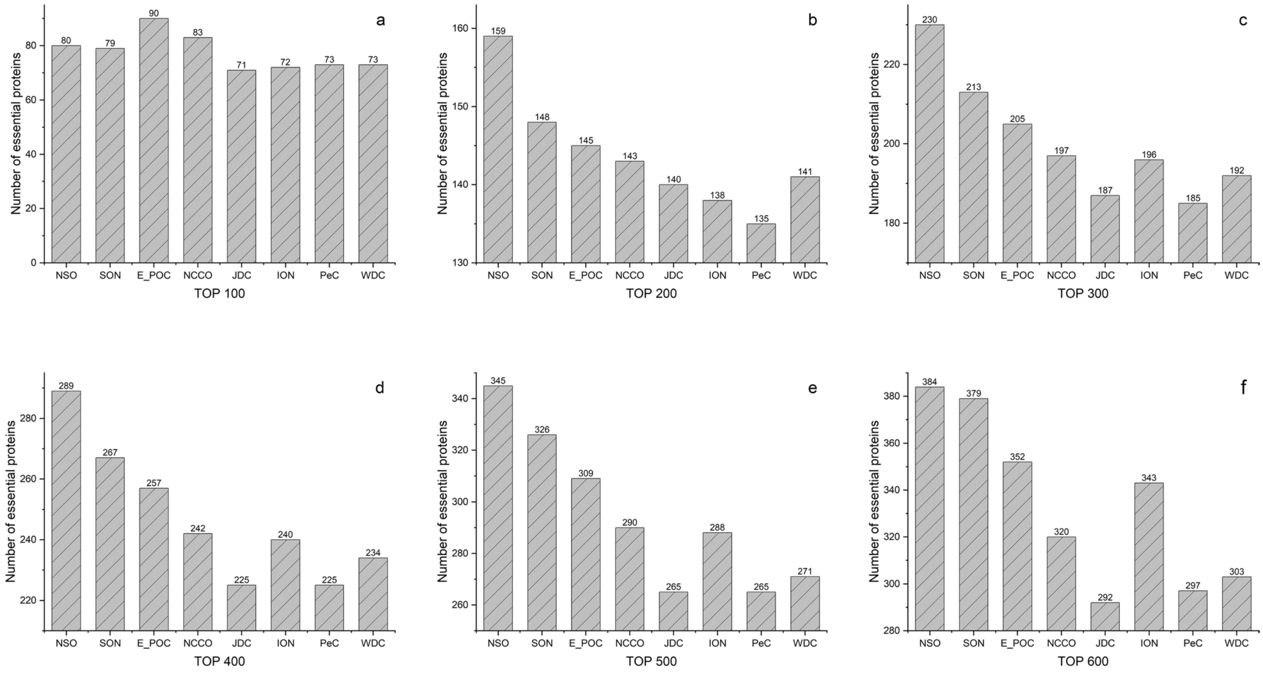

Fig.4 NSO and other methods predict the number of true essential proteins on Y2708

are significantly higher than 0.4027, 0.3946, 0.3807, 0.3512 and 0.3666 of ION, NCCO, JDC, PeC and WDC, respectively.

The Precision–Recall curve based on Y4743 is shown in Fig. 6. The Precision–Recall curve of NSO has obvious advantages over the curves of other methods, and PeC has the worst performance. Compared with ION(0.4149), E_POC(0.4088), SON(0.4002), NCCO(0.4018), JDC(0.3849), WDC(0.3587) and PeC(0.3547), the AUC values of NSO(0.4260) increases 2.67%, 4.20%, 6.44%, 6.02%, 10.67%, 18.76% and 20.10%, respectively.

The comparison results based on the Precision–Recall curve for these methods on Y2708 are shown in Fig. 7. Three curves of NSO, SON and ION are higher than the others. The AUC values of NSO, SON and ION are 0.4964, 0.4907 and 0.4886, respectively. The AUC values of NCCO, E_POC, JDC, PeC and WDC are 0.4248, 0.4752, 0.4124, 0.4081 and 0.4301, respectively.

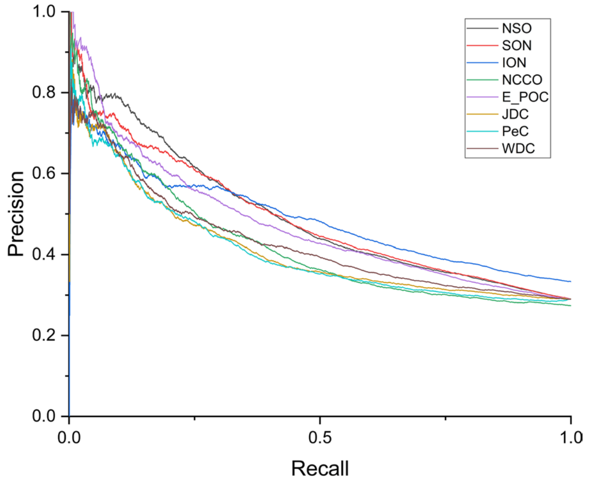

Fig.7 Comparison based on Precision–Recall curves of NSO and other methods on Y2708

Comparison based on Jackknife curves

The Jackknife curve serves as a common tool for comparing algorithm performance, assessing the strength of algorithms based on the AUC value of the Jackknife curve.

Figure 8 shows the Jackknife curves based on Y5093. As can be seen from Fig. 8, it becomes evident that NSO outshines other methods by a significant margin. The AUC value of the NSO Jackknife curve is 141,903. The AUC values of the Jackknife curves of SON, E_POC, NCCO, ION, JDC, WDC and PeC are 131,702, 128,816, 128,436, 123,974, 121,450, 110,879 and 108,080, respectively. Compared with these methods, the AUC value of NSO increases 7.74%, 10.16%, 10.48%, 14.46%, 16.84%, 27.98% and 31.29%, respectively.

On Y4743, the Jackknife curves are shown in Fig. 9. These curves can be divided into three tier. It can be seen that NSO is in the first-tier, while NCCO, E_POC, ION, and SON are in the second-tier, and JDC, PeC, and WDC are in the third-tier. The AUC value of NSO is 137,314, higher than the second-tier methods NCCO(127,340), E_POC(127,108), ION(125,568), and SON(124,668), and significantly higher than the third-tier methods JDC(104,775), WDC(95,594.5), and PeC(93,531).

Figure 10 shows the comparison results based on the Jackknife curve on Y2708. From the figure, the AUC

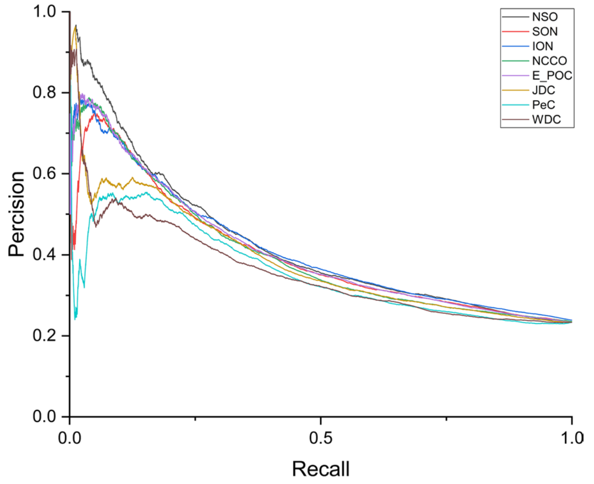

Fig.6 Comparison based on Precision–Recall curves of NSO and other methods on Y4743

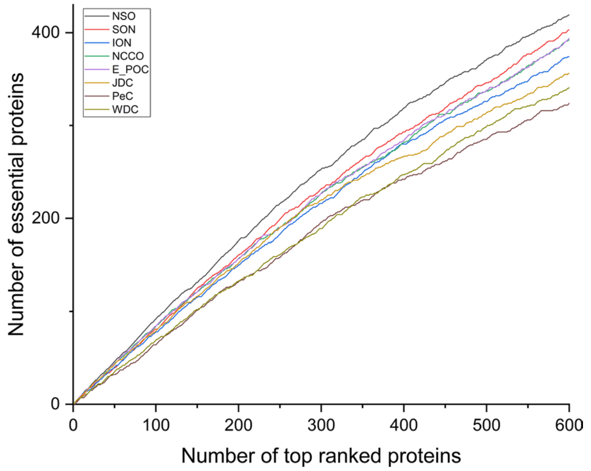

Fig.8 Comparison based on Jackknife curves between NSO and other methods on Y5093

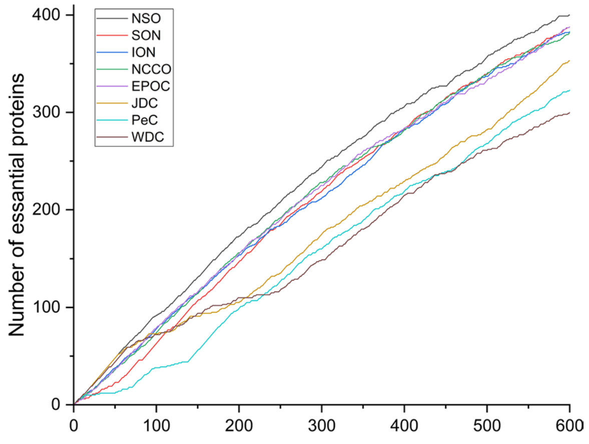

Fig.9 Comparison based on Jackknife curves between NSO and other methods on Y4743

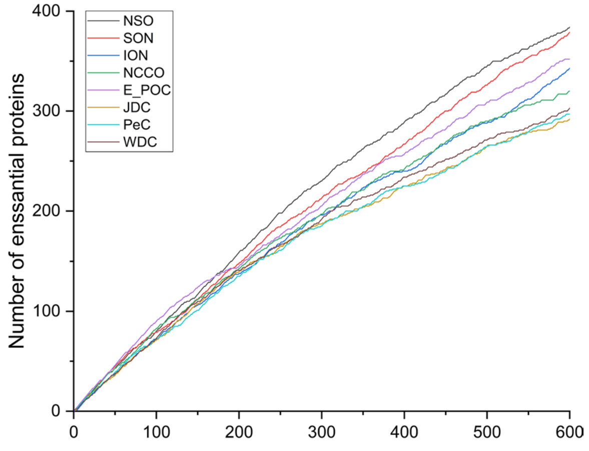

Fig.10 Comparison based on Jackknife curves between NSO and other methods on Y2708

values of NSO, SON, E_POC, NCCO, ION, WDC, JDC, and PeC are 130,026, 122,982, 119,184, 112,126, 111,020, 106,208, 103,358 and 103,092, respectively. Compared with E_POC, NCCO, ION, WDC, JDC, and PeC, the AUC value of NSO increases 7044, 10,842, 17,900, 19,006, 23,818, 26,668, and 26,934, respectively.

Comparing NSC to other 6 centralities

To evaluate the effectiveness of NSC, six classical centralities are compared on three datasets. Table 3 shows the results of these centrality methods based on the three purified PPI networks. It is evident that NSC comprehensively outperforms other centralities in terms of performance on Y5093. On Y4743 and Y2708 datasets, NSC and other centrality methods have wins and losses, but from the overall results, NSC is still able to outperform other methods.

To further evaluate the effectiveness of NSC, the average values of NSC for both essential and non-essential proteins are analyzed and listed in Table 4. As shown in Table 4, the average NSC values of essential proteins is almost more than twice that of non- essential proteins. Furthermore, based on Y5093, the top 20 proteins according to their NSC scores are listed in Table 5. In Table 5, among the top 20 ranked proteins, 85% are essential proteins. It is further confirmed the effectiveness of NSC in identifying essential proteins。

Enrichment analysis

Figure 11 shows pathways or GO terms enrichment analyses of the top 600 predicted proteins by NSO in three PPI networks. The Metascape tool is used for enrichment analysis, which is a user-friendly and powerful online gene function annotation analysis tool [45], and its website is http://metascape.org/gp/. Figure 11A shows the top 10 pathways or GO terms in the biological pathway clustering analysis of the top 600 predicted proteins by NSO on Y5093, which are mainly enriched in GO:002613 (ribonucleoprotein complex biogenesis), GO:0071826 (ribonucleoprotein complex subunit organization), GO:0016071 (mRNA metabolic process), R-SCE-983169 (Class I MHC mediated antigen processing&presentation), R-SCE-69278 (Cell Cycle, Mitotic), WP425 (Eukaryotic transcription initiation), GO:0042273 (ribosomal large subunit biogenesis),GO:0031123 (RNA 3'-end processing), sce03020 (RNA polymerase—*Saccharomyces cerevisiae*), and GO:0006281 (DNA repair). The top 10 pathways or GO terms in the biological pathway cluster analysis of the top 600 proteins predicted by NSO based on Y4743 are shown in Fig. 11B. They are mainly enriched in GO:0022613 (ribonucleoprotein complex biogenesis), GO:0071826 (ribonucleoprotein complex subunit organization), GO:0042273 (ribosomal large subunit biogenesis), sce03040 (Spliceosome—*Saccharomyces cerevisiae*), R-SCE-69278 (Cell Cycle, Mitotic), R-SCE-392499 (Metabolism of proteins), R-SCE-73894 (DNA Repair), GO:0006281 (DNA Repair), GO:0006913 (nucleocytoplasmic transport), and sce03420 (Nucleotide excision repair—*Saccharomyces cerevisiae*). The results of biological pathway clustering analysis of the top 600 proteins on Y2708 predicted by NSO are shown in Fig. 11C. Among them, the top 10 pathways or GO terms are GO:0006396 (RNA processing), GO:0032774 (RNA biosynthetic process), GO:0016071 (mRNA metabolic process), GO:0043933 (protein-containing complex organization), R-SCE-674695 (RNA Polymerase II

Table 3 Performance comparison of NSC with other centralities

	Top100	Top200	Top300	Top400	Top500	Top600
Y5093						
DC	49	101	154	205	258	298
IC	13	31	53	67	82	99
EC	67	122	185	224	265	308
SC	67	121	181	224	266	305
CC	43	83	116	155	199	235
NC	83	144	201	248	291	342
NSC	**85**	**148**	**211**	**263**	**310**	**356**
Y4743						
DC	86	145	199	264	321	359
IC	16	33	50	68	75	92
EC	**89**	**157**	**225**	281	327	355
SC	**89**	**157**	**225**	281	327	355
CC	70	138	197	237	276	307
NC	84	134	217	281	338	388
NSC	87	140	208	**285**	**341**	**389**
Y2708						
DC	**74**	**130**	182	231	264	299
IC	28	49	76	97	113	131
EC	60	117	153	193	237	273
SC	66	140	178	212	256	292
CC	54	102	140	191	236	280
NC	73	128	**184**	230	263	**312**
NSC	68	123	180	**234**	**277**	311

Bold values indicate the best performing results in each TOP dimension

Table 4 The average NSC scores of essential/non-essential proteins

	Y5093	Y4743	Y2708
T_Avg	0.0477	0.0501	0.0676
F_Avg	0.0218	0.0175	0.0395

T_Avg and F_Avg represent the average NSC values of essential and non-essential proteins, respectively

Pre-transcription Events), GO:0006974 (cellular response to DNA damage stimulus), GO:0019219 (regulation of nucleobase-containing compound metabolic process), R-SCE-69278 (Cell Cycle, Mitotic), GO:0006325 (chromatin organization), and GO:0044265 (cellular macromolecule catabolic process).

Table 5 The list of the top 20 proteins identified by NSC

Rank	Protein name	essential	NSC	Rank	Protein name	essential	NSC
1	YKR081C	1	1.00	11	YPL043W	1	0.55
2	YPR016C	1	0.96	12	YAL043C	1	0.54
3	YNL061W	1	0.95	13	YER126C	1	0.51
4	YMR049C	1	0.95	14	YLR115W	1	0.51
5	YER133W	1	0.91	15	YHR197W	1	0.50
6	YHR066W	0	0.71	16	YKL059C	1	0.48
7	YNL110C	1	0.67	17	YDR301W	1	0.47
8	YCR057C	1	0.62	18	YGL111W	1	0.45
9	YGR103W	1	0.60	19	YLR074C	0	0.45
10	YIL035C	0	0.56	20	YGR090W	1	0.44

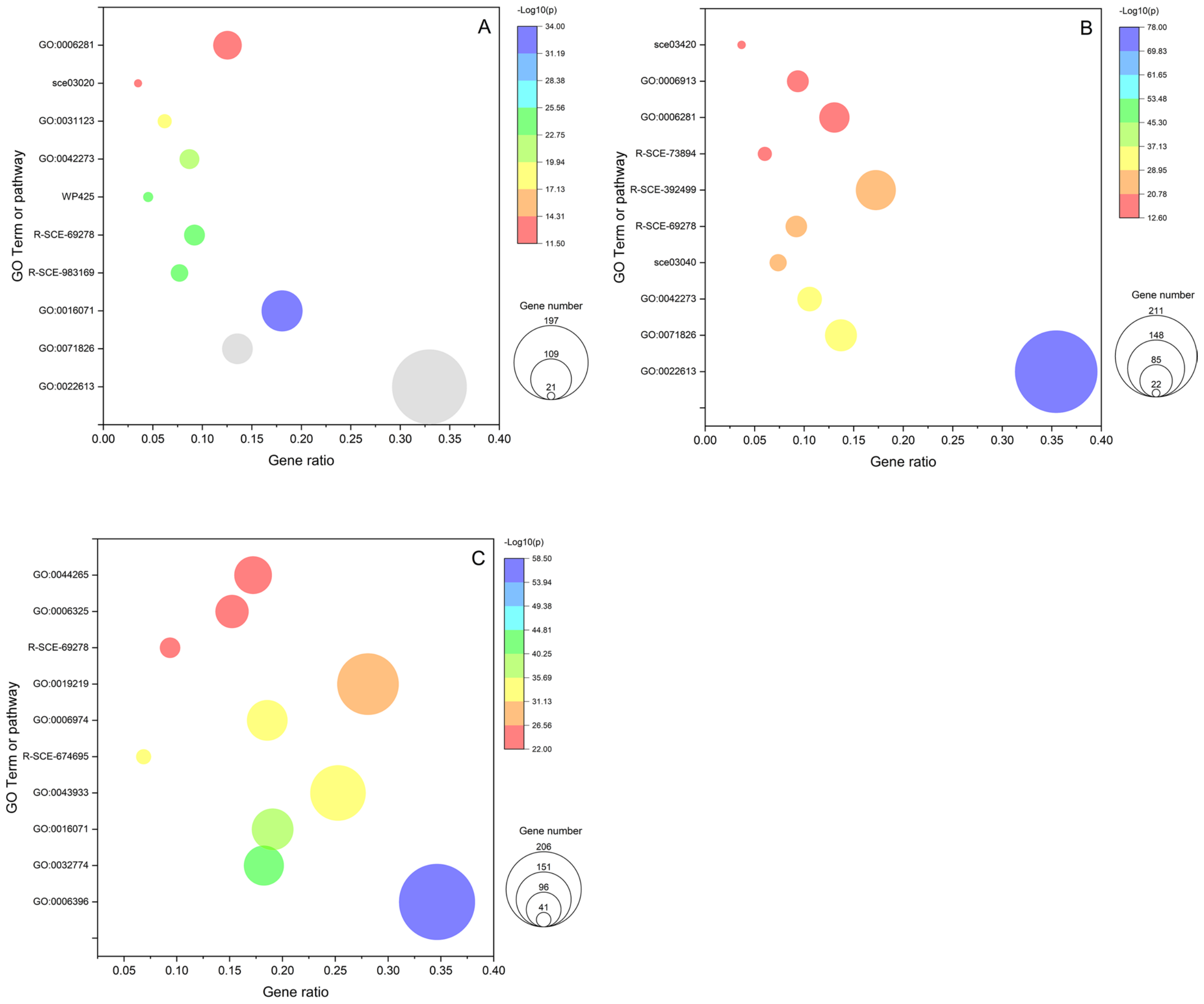

Fig. 11 Enrichment analysis of Top600 proteins predicted by NSO

Conclusion and discussion

Essential proteins play a crucial role in cellular activities. Therefore, the identification of essential proteins can help us reveal the molecular mechanism of cells and find new biomarkers and drug targets, which is of great significance. In this study, a new algorithm called NSO is proposed, which identifies essential proteins by fusing NSC, Sub, and OS using a dominance relationship model. To validate the performance of NSO, seven representative essential proteins identification algorithms are compared based on three PPI datasets of S.cerevisiae. The experimental results show that the NSO method has higher identification rate than other representative methods. Then, NSC based on purified PPI networks is compared with six representative centralities based on three PPI networks, the results show NSC can discover more essential proteins.

There are some advantages of NSO as follows: (1) NSO integrates different types of biological data, so it has strong anti-interference ability and is less affected by the quality of a single dataset; (2) NSO improves the essential proteins recognition ability; (3) The proposed feature fusion model, the dominance relation model, can be widely applied to other feature fusion methods.

The NSO method also has a disadvantage. NSO needs a lot of pre-experiments to determine parameters, which may reduce their convenience.

Funding

This research is supported by National Natural Science Foundation of China (Nos. 62141207, 62302107, 62366007, 61972185), Guangxi Natural Science Foundation (No. 2022GXNSFAA035625), Natural Science Foundation of Yunnan Province of China (No. 2019FA024), Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security (Nos. 20-A-01-03, 19-A-03-01), Guangxi Normal University Science Research Project (Natural Science) (No.2021JC008), Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing, Innovation Project of Guangxi Graduate Education (YCSW2023180).

Data availability

Data are publicly available.

Declarations

Conflict of interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval

Not applicable.

Informed consent

Not applicable.

Author details

[1]Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin 541004, China. [2]Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin 541004, Guangxi, China. [3]College of Computer Science and Engineering, Guangxi Normal University, Guilin 541004, Guangxi, China. [4]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, China. [5]School of Computer and Information Security & School of Software Engineering, Guilin University of Electronic Science and Technology, Guilin, China.

Received: 23 May 2023 Accepted: 13 October 2023
Published: 16 November 2023

References

1. Fields S, Song O-K. A novel genetic system to detect protein–protein interactions. Nature. 1989;340(6230):245–6.
2. Glass JI, Hutchison CA III, Smith HO, Venter JC. A systems biology tour de force for a near-minimal bacterium. Mol Syst Biol. 2009;5(1):330.
3. 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):901–6.
4. Zhang R, Lin Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res. 2009;37(suppl_1):D455–8.
5. Clatworthy AE, Pierson E, Hung DT. Targeting virulence: a new paradigm for antimicrobial therapy. Nat Chem Biol. 2007;3(9):541–8.
6. 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):387–91. https://doi.org/10.1038/nature00935.
7. 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):167–81.
8. Cullen LM, Arndt GM. Genome-wide screening for gene function using RNAi in mammalian cells. Immunol Cell Biol. 2005;83(3):217–23.
9. Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol. 2005;22(4):803–6. https://doi.org/10.1093/molbev/msi072.
10. Joy MP, Brock A, Ingber DE, Huang S. High-betweenness proteins in the yeast protein interaction network. J Biomed Biotechnol. 2005;2005(2):96–103. https://doi.org/10.1155/JBB.2005.96.
11. Wuchty S, Stadler PF. Centers of complex networks. J Theor Biol. 2003;223(1):45–53. https://doi.org/10.1016/s0022-5193(03)00071-7.
12. Estrada E, Rodriguez-Velazquez JA. Subgraph centrality in complex networks. Phys Rev E. 2005;71(5 Pt 2): 056103. https://doi.org/10.1103/PhysRevE.71.056103.
13. Bonacich P. Power and centrality: a family of measures. Am J Sociol. 1987;92:12.
14. Stephenson K, Zelen M. Rethinking centrality: methods and examples. Soc Netw. 1989;11(1):1–37.
15. Wang J, Li M, Wang H, Pan Y. Bioinformatics. Identification of essential proteins based on edge clustering coefficient. IEEE/ACM Trans Comput Biol. 2011;9(4):1070–80.
16. Acencio ML, Lemke N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinform. 2009;10(1):290. https://doi.org/10.1186/1471-2105-10-290.
17. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. Science. 2002;296(5568):750–2. https://doi.org/10.1126/science.1068696.
18. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 2002;12(6):962–8.
19. Batada NN, Hurst LD, Tyers M. Evolutionary and physiological importance of hub proteins. PLoS Comput Biol. 2006;2(7): e88. https://doi.org/10.1371/journal.pcbi.0020088.
20. Sharp PM. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J Mol Evol. 1991;33:23–33.
21. Rocha EP, Danchin A. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol. 2004;21(1):108–16. https://doi.org/10.1093/molbev/msh004.
22. Wang J, Peng X, Li M, Pan Y. Construction and application of dynamic protein interaction network based on time course gene expression data. Proteomics. 2013;13(2):301–12. https://doi.org/10.1002/pmic.201200277.
23. Xiao Q, Wang J, Peng X, Wu F-X, Pan Y. Identifying essential proteins from active PPI networks constructed with dynamic gene expression. BMC Genomics. 2015;16:1–7.
24. Zhang Y, Lin H, Yang Z, Wang J. Construction of dynamic probabilistic protein interaction networks for protein complex identification. BMC Bioinform. 2016;17:1–13.
25. Li M, Meng X, Zheng R, Wu FX, Li Y, Pan Y, et al. Identification of protein complexes by using a spatial and temporal active protein interaction network. IEEE/ACM Trans Comput Biol Bioinform. 2017;17:817–27.
26. Tang X, Wang J, Zhong J, Pan Y. Predicting essential proteins based on weighted degree centrality. IEEE/ACM Trans Comput Biol Bioinform. 2013;11(2):407–18.
27. Zhang X, Xiao W, Hu X. Predicting essential proteins by integrating orthology, gene expressions, and PPI networks. PLoS ONE. 2018;13(4): e0195410.
28. Li G, Li M, Wang J, Wu J, Wu F-X, Pan Y. Predicting essential proteins based on subcellular localization, orthology and PPI networks. BMC Bioinform. 2016;17(8):571–81.
29. Li M, Zhang H, Wang JX, Pan Y. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. BMC Syst Biol. 2012;6:15. https://doi.org/10.1186/1752-0509-6-15.
30. Zhong J, Tang C, Peng W, Xie M, Sun Y, Tang Q, et al. A novel essential protein identification method based on PPI networks and gene expression data. BMC Bioinform. 2021;22(1):248. https://doi.org/10.1186/s12859-021-04175-8.
31. 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 Trans Comput Biol Bioinform. 2019;17(6):2053–61.
32. 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.
33. Zhang Z, Jiang M, Wu D, Zhang W, Yan W, Qu X. A novel method for identifying essential proteins based on non-negative matrix tri-factorization. Front Genet. 2021;12: 709660.
34. Li G, Li M, Wang J, Li Y, Pan Y. United neighborhood closeness centrality and orthology for predicting essential proteins. IEEE/ACM Trans Comput

Biol Bioinform. 2020;17(4):1451–8. https://doi.org/10.1109/TCBB.2018.2889978.

35. Li G, Li M, Peng W, Li Y, Pan Y, Wang J. A novel extended Pareto optimality consensus model for predicting essential proteins. J Theor Biol. 2019;480:141–9.
36. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 2002;30(1):303–5. https://doi.org/10.1093/nar/30.1.303.
37. Yu H, Luscombe NM, Qian J, Gerstein M. Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet. 2003;19(8):422–7. https://doi.org/10.1016/S0168-9525(03)00175-6.
38. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods. 2012;9(5):471–2. https://doi.org/10.1038/nmeth.1938.
39. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, et al. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res. 2004;32:D41–4. https://doi.org/10.1093/nar/gkh092.
40. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, et al. SGD: Saccharomyces Genome Database. Nucleic Acids Res. 1998;26(1):73–9. https://doi.org/10.1093/nar/26.1.73.
41. Saccharomyces Genome Deletion Project. http://www-sequence.stanford.edu/group/.
42. Tu BP, Kudlicki A, Rowicka M, McKnight SL. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. Science. 2005;310(5751):1152–8. https://doi.org/10.1126/science.1120499.
43. COMPARTMENTS. http://compartments.jensenlab.org. Accessed 28 Dec 2014.
44. 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:D196–203.
45. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10(1):1523. https://doi.org/10.1038/s41467-019-09234-6.

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