Li et al. BMC Bioinformatics 2016, 17(Suppl 8):279
DOI 10.1186/s12859-016-1115-5

BMC Bioinformatics

RESEARCH Open Access

Predicting essential proteins based on subcellular localization, orthology and PPI networks

Gaoshi Li[1,2], Min Li[1*], Jianxin Wang[1*], Jingli Wu[2], Fang-Xiang Wu[1,3] and Yi Pan[1,4]

From 11th International Symposium on Bioinformatics Research and Applications (ISBRA '15)
Norfolk, VA, USA. 7-10 June 2015

Abstract

Background: Essential proteins play an indispensable role in the cellular survival and development. There have been a series of biological experimental methods for finding essential proteins; however they are time-consuming, expensive and inefficient. In order to overcome the shortcomings of biological experimental methods, many computational methods have been proposed to predict essential proteins. The computational methods can be roughly divided into two categories, the topology-based methods and the sequence-based ones. The former use the topological features of protein-protein interaction (PPI) networks while the latter use the sequence features of proteins to predict essential proteins. Nevertheless, it is still challenging to improve the prediction accuracy of the computational methods.

Results: Comparing with nonessential proteins, essential proteins appear more frequently in certain subcellular locations and their evolution more conservative. By integrating the information of subcellular localization, orthologous proteins and PPI networks, we propose a novel essential protein prediction method, named SON, in this study. The experimental results on S.cerevisiae data show that the prediction accuracy of SON clearly exceeds that of nine competing methods: DC, BC, IC, CC, SC, EC, NC, PeC and ION.

Conclusions: We demonstrate that, by integrating the information of subcellular localization, orthologous proteins with PPI networks, the accuracy of predicting essential proteins can be improved. Our proposed method SON is effective for predicting essential proteins.

Keywords: Essential proteins, Protein-protein interaction network, Subcellular localization, Orthology

Background

Essential proteins are indispensable in cellular life because even if only one of these proteins is missing, organisms cannot survive or develop. The identification of essential proteins has great significance in the following facts: 1) it helps understand the minimum requirements of the survival and development of a cell. By knowing the minimum requirements of the survival and development of the cell, researchers are able to create a new cell with a minimal genome [1], which is an important content in the emerging synthetic biology. 2) It helps identify disease genes and find novel treatments for diseases [2–4]. Hence, the discovery of essential proteins facilitates to study disease genes. Because essential proteins are indispensable in bacterial cells, they are also the candidates for new antibiotics drug targets.

There are several representative biological methods to identify essential proteins, such as single gene knockout [5], conditional knockout [6] and RNA interference [7]. Because these biological experiment discovery methods are time-consuming, expensive and inefficient, it is

* Correspondence: limin@mail.csu.edu.cn; jxwang@mail.csu.edu.cn
[1]School of Information Science and Engineering, Central South University, Changsha 410083, Hunan, People's Republic of China
Full list of author information is available at the end of the article

BioMed Central

© 2016 The Author(s). **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

appealing to develop novel computational methods to improve the effectiveness of the identification.

Currently, a number of computational identification methods have been proposed. According to the features of essential proteins, these methods can be roughly divided into topology-based methods and sequence-based methods. The topology-based methods are designed based on associations between the essentiality and the topological features of essential proteins in bio-molecular networks. Degree Centrality (DC) [8], Betweenness Centrality (BC) [9], Closeness Centrality (CC) [10], Subgragh Centrality (SC) [11], Eigenvector Centrality (EC) [12], Information Centrality (IC) [13] and Neighborhood Centrality (NC) [14] are the representatives of topology-based methods. CytoNCA [15] is a cytoscape plugin for centrality analysis and evaluation of biological networks, and ClusterViz [16] is a cytoscape APP for cluster analysis of biological network. Additionally, LAC [17], TP and TP-NC [18] are also common topology-based methods.

The topology-based methods are consist of several steps as follows: Firstly, constructing a PPI network G (V, E) based on the pairs of PPI, where V denotes a set of nodes (proteins), and E denotes a set of edges of PPI network. Secondly, constructing an adjacency matrix A of PPI network G, whose element $A_{u,v}$ is 1 if there is an edge between nodes u and v, and 0 otherwise. Then, each protein in PPI network G is scored by using different centrality methods. Finally, essential proteins are determined by their scores.

The key advantage of the topology-based methods is able directly to predict essential proteins without knowing additional information. However, these methods have three main disadvantages as follows: 1) due to a lot of false positives and false negative data in PPI networks, their identification accuracies are affected. 2) These methods have difficulty in predicting essential proteins with low connectivity. 3) These methods ignore the intrinsic biological significance of essential proteins.

The sequence-based methods are another kind of computational methods to predict essential proteins. The sequence-based features are intrinsic features of an individual protein that are determined by genomic sequences. These features have been used by some methods, such as subcellular localization [19], evolutionary conservation [20–22], gene expression [23, 24].

Subcellular localization is an important feature of essential proteins. It represents a concrete location in cells that a certain protein appears. Statistical results show that essential proteins appear more frequently in certain subcellular location than nonessential proteins. Hence, we designed and used protein subcellular localization score based on the features of subcellular localization of proteins.

Evolutionary conservation is also an important characteristic of essential proteins. Because basic life process of a cell is more relevant with essential proteins. The effect of essential proteins in a negative selection is stricter than nonessential proteins [21]. Experimental results have proved that essential proteins evolve more conservative than nonessential proteins.

Gene expression is another important feature of essential proteins. The expression level of mRNA is closely associated with its essentiality. In bacteria, the higher expression level, the slower evolution of protein sequence is [23, 24]. Some studies have shown that protein sequence diversity and protein essentiality are relevant to expression level [25] in eukaryotes. So we draw a conclusion that the expression level of essential genes is higher than that of nonessential ones.

In order to achieve higher identification accuracy, more and more researchers are combining above-mentioned two kinds of methods. By integrating the information of GO annotations with proteins, Li et al. [26] built a weighted PPI network. In addition, by integrating the information of network topology with gene expression, they proposed a centrality method PeC [27]. Based on prior knowledge, network topology and gene expression, they also proposed two new essential protein discovery methods CPPK and CEPPK [28]. Besides the above methods, some researchers proposed to construct dynamic PPI network to reduce the impact of false positives in PPI data [29–31]. Xiao et al. [31] constructed an active PPI network and applied six typical centrality measures to identify essential proteins from the constructed active PPI network. By using PPI network and protein complexes information, Ren et al. [32] proposed an essential protein discovery method named HC. Li et al. [33] proposed a united complex centrality named UC and a parameter controlled method UC-P by using predicting protein complexes [34]. Peng et al. [35] proposed an essential protein discovery method by integrating protein domains and PPI networks. Tang et al. [36] proposed a novel method based on weighted degree centrality by integrating gene expression profiles.

There is other biological information which also was integrated with PPI network to predict essential proteins. Based on random walk model, ION [37] integrates the information of orthologous proteins with PPI networks. Zhao et al. [38] proposed their new method by using overlapping essential modules [39]. Zhong et al. [40] proposed a feature selection method by considering 26 topological or biological features for predicting essential proteins.

In this study, we propose a novel method to predict essential proteins by integrating *s*ubcellular localization, *o*rthology with PPI *n*etwork, named SON.

Experimental data

This experiment uses multiple datasets, including PPI network dataset, essential protein dataset, subcellular localization dataset and orthologous protein dataset. In

order to unify the serial number of proteins in above-mentioned databases, we use the UNIPROT [41] data files to convert protein number in each database.

PPI network dataset of S.cerevisiae is downloaded from DIP database [42] updated to Oct.10, 2010. There are 5093 proteins and 24,743 interactions without self-interactions and repeated interactions in this dataset. We select S.cerevisiae because its PPI data and gene essentiality data are most complete and reliable among various species.

Essential protein dataset is selected from MIPS [43],SGD [44],DEG [45] and SGDP [46]. There are 1285 essential proteins in this dataset, out of which 1167 are in PPI network. We take the 1167 proteins as essential proteins while other 3926(=5093−1167) proteins as non-essential ones.

Subcellular localization dataset of yeast is downloaded from knowledge channel of COMPARTMENTS database [47] on August 30, 2014. It integrates several source databases (UniProtKB [48], MGD [49], SGD [50], FlyBase [51] and WormBase [52]). As a result, it contains 5095 yeast proteins and 206,831 subcellular localization records. We select this database because both its data volume is large and it is updated in a timely manner. After preprocessing, there are still 3923 proteins in PPI network which have subcellular localization information.

Orthologous proteins dataset is taken from Version 7 of InParanoid [53]. It contains a set of pairwise comparisons among 100 whole genomes (1 prokaryote and 99 eukaryotes) that are constructed by INPARANIOD program. We only select the proteins in seed orthologous sequence pairs of each cluster generated by INPARANIOD as orthologous proteins, as it has the best match between two organisms and stands for the high homology.

Correlation analyses of subcellular localization, orthology and essentiality of proteins

To understand associations between subcellular localization and essentiality of proteins, we first count the number of essential and nonessential proteins in each subcellular location, respectively. Next, their ratios are calculated. The results are shown in Table 1. According to Table 1, the ratios of essential proteins are higher than that of nonessential proteins in Cytoskeleton, Golgi apparatus, Cytosol, Nucleus and Endoplasmic reticulum. Hence, the five subcellular locations above mentioned are positively correlated with essential proteins while the others are negatively correlated.

The associations between orthology and essentiality of proteins have been verified by Peng et al. [37]. The ratio of essential proteins is 51 % if the proteins have orthologs for at least 80 species. But if the proteins have no orthologs for reference organisms, the ratio of essential proteins is about 22 %, near to random probability [54].

Methods

Our novel method, SON, predicts essential proteins based on the information integration of subcellular localization, Orthology and PPI network. In the following subsections, we will introduce how to use these information and integrate them to calculate a protein's essentiality.

Network Centrality based on edge clustering coefficient (NC)

In the previous studies, it has been shown that network centrality is an important measure for predicting essential proteins and the network centrality based on edge

Table 1 Number and ratio of essential and nonessential proteins in each subcellular location

Subcellular location	Essential proteins number	Essential proteins ratio	Nonessential proteins number	Nonessential proteins ratio
Cytoskeleton	95	0.081	133	0.033
Golgi apparatus	61	0.052	184	0.046
Cytosol	138	0.118	289	0.073
Endosome	22	0.019	109	0.027
Mitochondrion	173	0.148	753	0.189
Plasma membrane	53	0.045	354	0.089
Nucleus	809	0.693	1407	0.353
Extracellular space	1	0.001	70	0.018
Vacuole	19	0.016	238	0.060
Endoplasmic reticulum	137	0.117	292	0.073
Peroxisome	4	0.003	61	0.015

To understand the association between subcellular localization and essentiality of proteins, we first count the number of essential and nonessential proteins in each subcellular location, respectively. Next, their ratios are calculated. According to Table 1, the ratios of essential proteins are higher than that of nonessential proteins in Cytoskeleton, Golgi apparatus, Cytosol, Nucleus and Endoplasmic reticulum. Hence, the five subcellular locations above mentioned are positive correlation with essential proteins while the others are negative correlation

clustering coefficient [14] is one of the most effective measures for the identification of essential proteins. Given a PPI network $G = (V, E)$ and a protein i, its network centrality based on edge clustering coefficient $NC(i)$ is defined as the sum of edge clustering coefficients of all edges directly connected with protein i in the graph G.

$$NC(i) = \sum_{j \in N_i} ECC(i,j) = \sum\nolimits_{j \in N_i} \frac{Z_{i,j}}{min(k_i - 1, k_j - 1)} \quad (1)$$

where N_i denotes the set of all neighbors of protein i, $Z_{i,j}$ is the number of triangles built on edge(i,j), k_i and k_j are the degrees of nodes i and j, respectively. $min(k_i-1, k_j-1)$ represents the maximal possible number of triangles that might potentially include the edge(i,j).

The edge clustering coefficient (ECC) is used to measure the degree of closeness between two nodes in a graph which has been widely applied in identifying network modules [55, 56]. Those edges which have higher ECC value are more likely to be in a module. It has been shown that essential proteins and disease genes tend to appear in the same cluster [57–59]. Therefore, if an edge with high ECC value, it is more likely to be a connection of two essential proteins. Obviously, a protein which has more neighbors and gets higher ECC values with its neighbors will have a relatively higher NC value and will tend to be an essential protein. In order to match with orthologous score and subcellular localization score whose value ranges are [0,1], here we use the normalized NC value for each protein, denoted as NNC. For a protein i, its normalized NC value $NNC(i)$ is defined as:

$$NNC(i) = NC(i)/Max\_NC \quad (2)$$

where Max_NC denotes the maximum NC value of all the proteins in the graph G.

Subcellular localization score

It has been shown that proteins must be localized at their appropriate subcellular compartments to perform their desired functions and thus the subcellular localization information is helpful to the identification of essential proteins [59]. Here, we analyzed the associations between the subcellular localization and the topology of PPI networks. All the proteins in the PPI network are sorted in descending order according to their NNC scores. Then we calculate the numbers of subcellular location l where the top k% proteins appear and where the bottom k% proteins appear, respectively. Considering that more counting proteins may result in more false positives, we use $k = 5$ in this paper, ie., that the top/bottom 5 % proteins are selected. Let f_l be the frequency of l where the top k% proteins appear and h_l denote the frequency of l where the bottom k% proteins appear. Subcellular Localization Correlation Coefficient $LCC(l)$ is defined as

$$LCC(l) = \begin{cases} 1 - \frac{h_l}{f_l}, & f_l < h_l \\ \frac{f_l}{h_l} - 1, & f_l \geq h_l \end{cases} \quad (3)$$

When $f_l < h_l$, more proteins with low NNC values appear in the location l and a negative relationship is thought to be between the location l and protein's essentiality. On the contrary, there is a positive correlation between the location l and protein's essentiality when $f_l \geq h_l$. When $f_l = 0$, we set $LCC(l)$ as the maximum of $1 - \frac{h_l}{f_l}$ with $f_l \neq 0$. When $h_l = 0$, we set $LCC(l)$ as the maximum of $\frac{f_l}{h_l} - 1$ with $h_l \neq 0$. A protein may appear in multiple subcellular locations. For a protein i, its subcellular localization score $SL(i)$ is defined as the sum of $LCC(l)$ of all the subcellular locations it appears. Here, for each protein i we also use the normalized SL value $NSL(i)$ by using the following formula:

$$NSL(i) = \frac{SL(i) + Max\_|SL|}{Max\left(SL(i) + Max\_|SL|\right)} \quad (4)$$

Where $Max\_|SL|$ denotes the maximum value of $|SL(i)|$ for all the proteins in G and Max in the denominator takes for all the proteins in G.

Orthologous score

Orthologous score method of SON comes from ION method [37]. Given a PPI network $G = (V, E)$, let S be the set of reference species which is used to get orthologous information of V. s denotes its element and $|S|$ denotes the number of its elements. Let X_s be a subset of V whose element has orthologs in organism s. For a protein i, its orthologous score $OS(i)$ is defined as the number of reference organisms in which the protein i has orthologs, where $i \in V$ $(i = 1, \ldots, N)$. Similar to the network centrality based on edge clustering coefficient and subcellular localization score, we also use the normalized OS value $NOS(i)$ by using the following formula:

$$NOS(i) = \frac{OS(i)}{Max\_OS} \quad (5)$$

Where Max_OS denotes the maximum value of $OS(i)$ for all the proteins in G.

According to the above definition, a protein's orthologous score is 1 if its orthologs in all organisms included

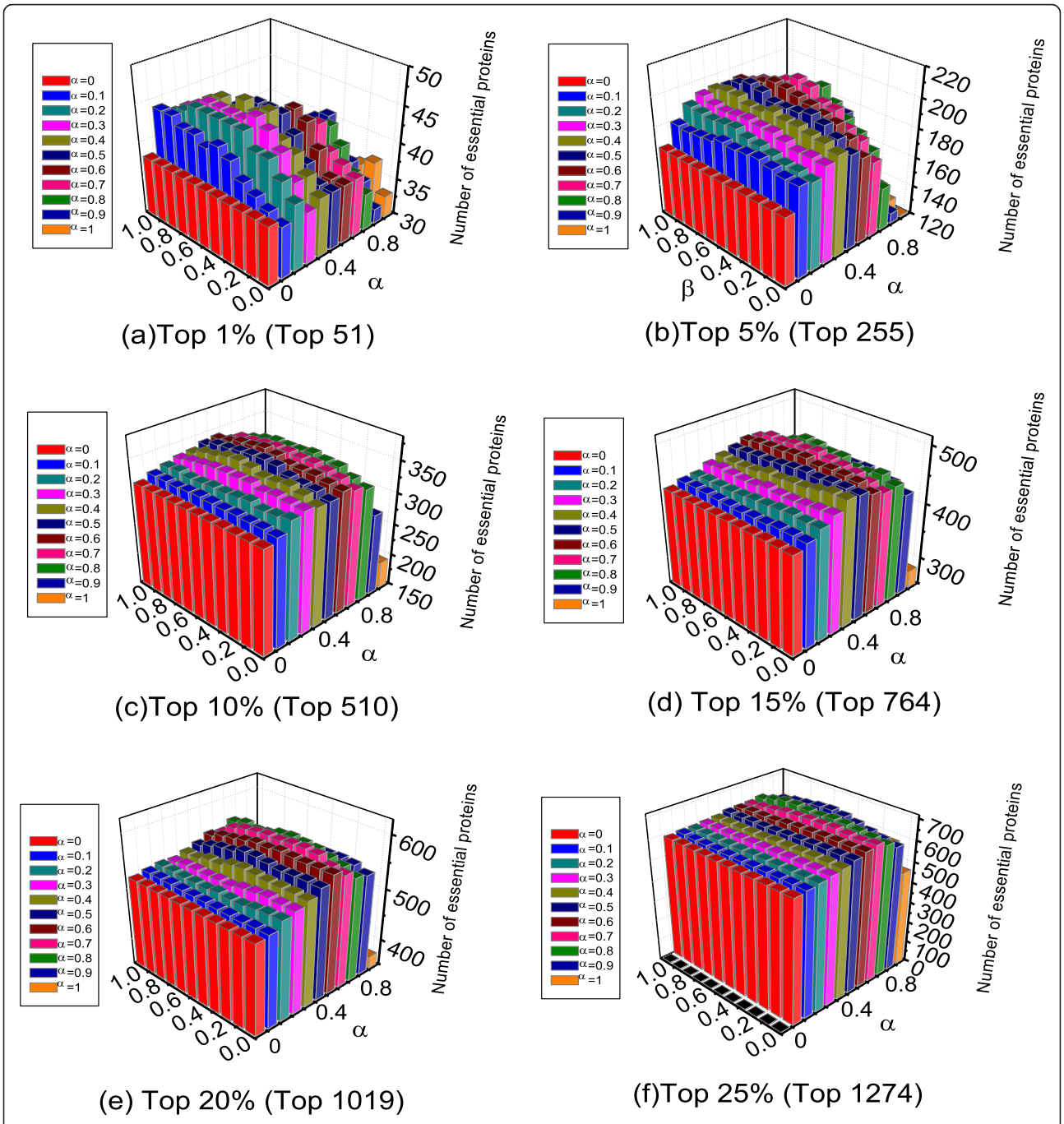

Fig. 1 Influence of parameters α and β. (**a**) Top 1 % (Top 51) (**b**) Top 5 % (Top 255) (**c**) Top 10 % (Top 510) (**d**) Top 15 % (Top 764) (**e**) Top 20 % (Top 1019) (**f**) Top 25 % (Top 1274)

in set *S*. On the contrary, its orthologous score is 0 if it does not have orthologs in any organisms in set *S*.

The sorting score and SON algorithm

The sorting score of our algorithm SON is a linear combination of the three scores: normalized network centrality based on edge clustering coefficient $NNC(i)$, normalized subcellular localization score $NSL(i)$, and normalized orthologous score $NOS(i)$. For a protein i, its sorting score is calculated as follows:

$$pr(i) = (1-\alpha) * NOS(i) + \alpha[(1-\beta) * NSL(i) + \beta * NNC(i)] \quad (6)$$

where $\alpha \in [0, 1]$ and $\beta \in [0, 1]$ are used to adjust the proportion of these three scores.

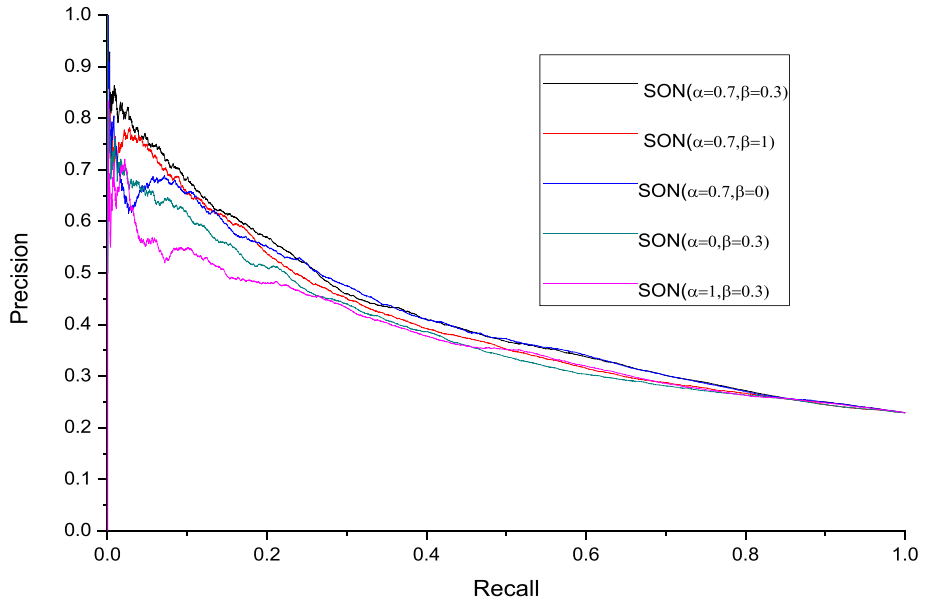

Fig. 2 Influence of parameters *α* and *β* for SON

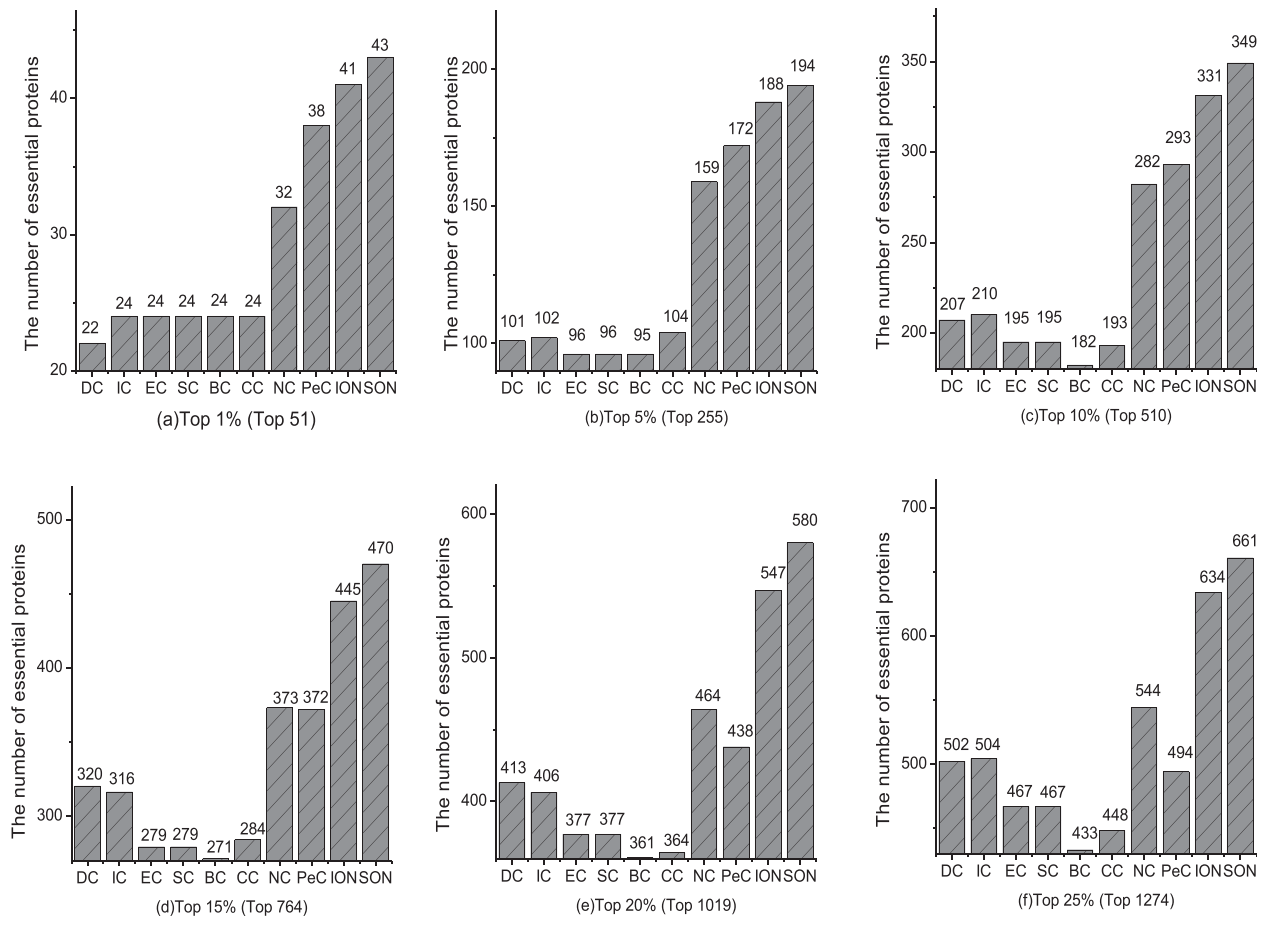

Fig. 3 SON compared with several existing methods. (**a**) Top 1 % (Top 51) (**b**) Top 5 % (Top 255) (**c**) Top 10 % (Top 510) (**d**) Top 15 % (Top 764) (**e**) Top 20 % (Top 1019) (**f**) Top 25 % (Top 1274)

SON algorithm is introduced as follows.

SON algorithm

Input: A PPI network represented as a graph $G = (V, E)$, the scoring table of subcellular localization of proteins, orthologs datasets between Yeast and 99 other organisms, parameter α, parameter β.

Output: Top K percent of proteins sorted by pr in descending order.

Step1: Calculate the value of NNC for each protein by using Equation (2).

Step2: Calculate the score of subcellular localization for each protein by using Equation (4).

Step3: Calculate orthologous score for each protein by using Equation (5).

Step4: Calculate the value of pr for each protein by using Equation (6).

Step 5: Sort proteins by the value of pr in descending order.

Step 6: Output top K percent of sorted proteins.

Results and discussion

In order to analyze and evaluate the performance of our method, SON, we perform a large number of experiments on these datasets. There are 5093 proteins and 24,743 interactions in PPI network of S.cerevisiae. Essential protein dataset is constructed by integrating MIPS,SGD,DEG and SGDP which has 1167 essential proteins in PPI network. Subcellular localization dataset includes 5095 yeast proteins and 206,831 subcellular localization records. After preprocessing, there are 3923 proteins in this dataset that have subcellular localization records. Orthologous proteins dataset is taken from Version 7 of InParanoid consisting a set of pairwise comparisons between 100 whole genomes.

In this section, we first analyze the influence of two parameters α and β towards the performance of SON algorithm. Then, SON is compared with the other existing algorithms, such as DC, BC, CC, SC, EC, IC, NC, PeC and ION. We adopted three types of popular comparison methodologies: 1) Histogram comparison methodology. Firstly, the results are sorted in descending order. Next, to select the top 1, 5, 10, 15, 20 and 25 % proteins as candidate essential proteins. Then, we compare prediction results based on the set of known essential proteins. The performance is presented in the form of histograms of the number of essential proteins predicted by each algorithm. 2) Precision-recall curves methodology. 3) Jackknife methodology. In the end, the differences of these algorithms which have high connectivity proteins and low ones are analyzed in detail.

Influence of parameter α and β

In our novel method, SON, the scoring of proteins is associated with parameters α and β. The value ranges of α and β are both from 0 to 1. When the values of α and β take 0, 0.1, 0.2, …, 0.9, 1, respectively, the number of essential proteins predicted by SON are shown in Fig. 1.

As shown in Fig. 1, when α values from 0.2 to 0.8 and β from 0.3 to 0.7, simultaneously, the result of SON is better. In particular, when $\alpha = 0$, namely only orthologous information is used, parameter β has no effect, all the results are the same.

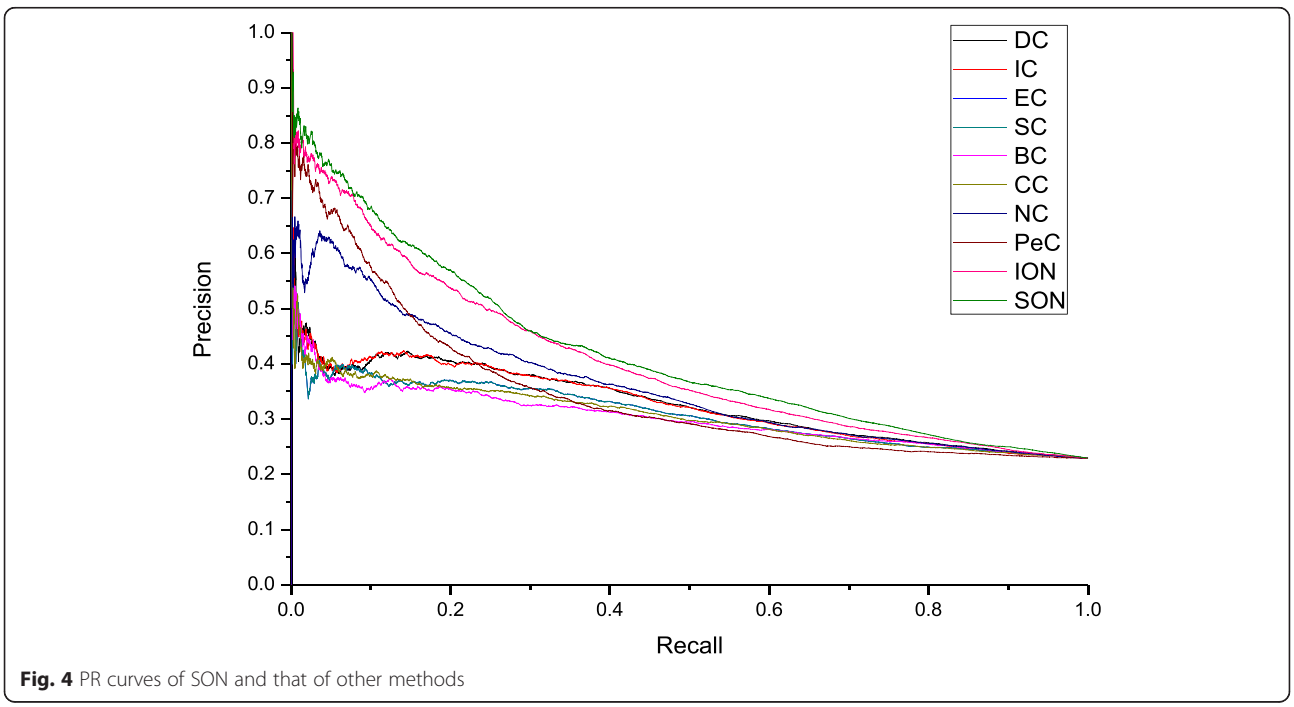

Fig. 4 PR curves of SON and that of other methods

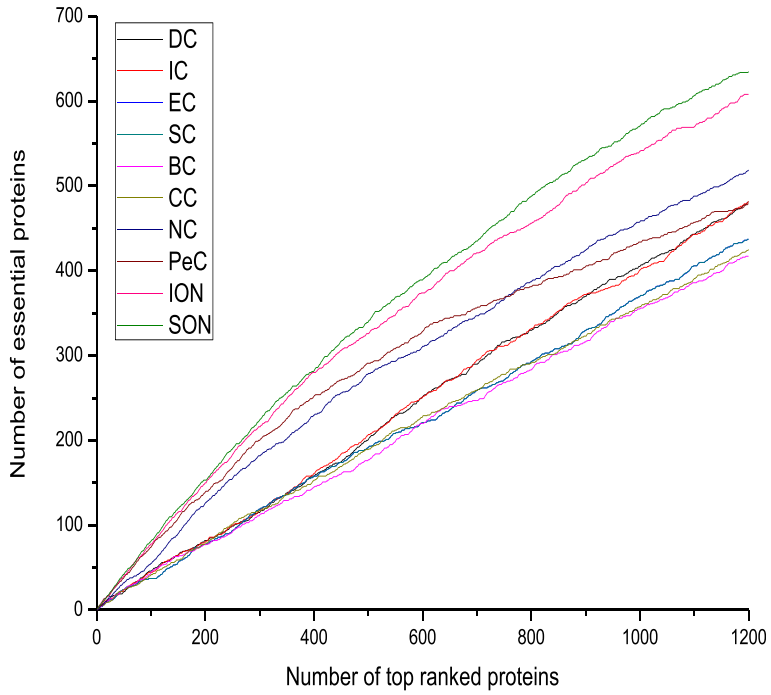

Fig. 5 Jackknife curves of SON and other nine methods

In order to further analyze the influence of the parameters α and β, we utilize the precision-recall curves methodology with five sets of parameters α and β, such as $\alpha = 0.7$and $\beta = 0.3$,$\alpha = 0.7$ and $\beta = 1$,$\alpha = 0.7$ and $\beta = 0$,$\alpha = 0$ and $\beta = 0.3$, $\alpha = 1$ and $\beta = 0.3$. The results are shown in Fig. 2. According to Fig. 2, when $\alpha = 0.7$and $\beta = 0.3$, namely, the proportions of orthologous information, NC, and subcellular localization information are 30, 21, and 49 %, respectively, the result is the best. In this paper, we consider the optimal values to be $\alpha = 0.7$ and $\beta = 0.3$.

Comparison with nine existing methods

In this section, the performance of SON is compared with nine existing methods. We select the top 1, 5, 10, 15, 20 and 25 % proteins predicted by DC, BC, CC, SC, EC, IC, NC, PeC, ION and SON as candidate essential proteins to compare, respectively. The results are shown

Table 2 Number of predicting high and low connectivity essential proteins by using SON and other nine existing methods

	K	DC	IC	EC	SC	BC	CC	NC	PeC	ION	SON
degree < =10	1 %	0	0	0	0	0	0	0	0	17	14
	5 %	0	0	0	0	0	0	3	40	66	64
	10 %	0	0	0	0	1	0	27	84	108	116
	15 %	0	0	8	8	18	7	66	117	146	156
	20 %	0	0	28	28	41	20	101	153	188	193
	25 %	11	20	73	73	76	55	156	192	253	220
degree > 10	1 %	22	24	24	24	24	24	32	39	24	29
	5 %	101	102	96	96	95	104	156	133	122	129
	10 %	207	210	195	195	181	193	255	209	223	233
	15 %	320	316	271	271	253	277	307	255	299	313
	20 %	413	406	349	349	320	344	363	285	359	387
	25 %	491	484	394	394	357	393	388	302	381	441

As shown in the top part of Table 2 (degree < = 10), it is weak for eight centrality methods to predict low connectivity essential proteins. When taking the top 20 % proteins ranked in descending order according to their ranking scores computed by DC and IC, the numbers of predicting essential proteins are 0. The performance of SON overall is better than that of eight centrality methods (DC, IC, EC, SC, BC, CC, NC and PeC). When *K* is 10, 15, 20 %, respectively, the performance of SON is also better than that of ION

in Fig. 3. From Fig. 3, it is easy to see that the result of SON is clearly the best.

Comparison the experimental results based on precision-recall curve

Precision-recall (PR) curve is another common methodology to validate algorithm performance. In terms of the corresponding area under the PR curve (AUC) value, the overall performance of each method is evaluated. At the beginning, according to their scores computed for each method, all proteins are sorted in descending order. Then the top K proteins are selected as candidate essential proteins while the remaining proteins in PPI networks as candidate nonessential ones. The values of K range from 1to 5093. The results are shown in Fig. 4. As shown in Fig. 4, PR curve of SON is obviously higher than that of other methods. Note that the curves of EC and SC are almost identical.

Comparison the experimental results based on jackknife methodology

To further investigate the performance of SON, jackknife methodology is also employed. The results are shown in Fig. 5. The x-axis represents the number of proteins in PPI networks ranked in descending order according to their sorting scores computed from all above-mentioned methods while the y-axis represents the cumulative count of essential proteins. The areas under the curves are used to measure the performances of the above-mentioned methods. According to Fig. 5, SON is clearly better than DC, IC, EC, SC, BC, CC, NC, PeC and ION. Note that the curves for EC and SC are almost identical.

Differences between SON and nine existing methods

In order to further analyze SON, we compared its ability to identify low/high connectivity essential proteins with nine existing methods. After statistical analysis, we notice that the low connectivity (less than or equal to 10) proteins are about 76 % in the yeast PPI network and 58 % of essential proteins in known essential protein list are low connectivity in the yeast PPI network. Hence, it is very important for essential protein prediction method to identify low connectivity essential proteins. The results of predicting essential proteins with low and high connectivity for several above-mentioned methods are illustrated in Table 2.

As shown in the top part of Table 2 (degree ≤ 10), it is weak for eight centrality methods to predict low connectivity essential proteins. When taking the top 20 % proteins from DC and IC, the numbers of predicting essential proteins are 0. The performance of SON overall is better than that of eight centrality methods (DC, IC, EC, SC, BC, CC, NC and PeC). When K is 10, 15 and 20 %, respectively, the performance of SON is also better than that of ION.

As shown in the bottom part of Table 2 (degree > 10), we can see that DC and IC have good performance in predicting high connectivity essential proteins. However, SON in predicting high connectivity essential proteins outperforms EC, SC, BC, CC and ION.

Conclusions

Although identification of essential proteins is of great significance, biological experimental methods for identifying essential proteins are time-consuming, costly and inefficient. Hence it is necessary to use computational methods to identify essential proteins. In this paper, by the integration of subcellular localization, orthologous and PPI, we proposed a novel method, SON, to predict essential proteins.

First, we analyze the correlation between subcellular localization, orthologous proteins and essentiality of proteins. Then, we propose our novel method SON. By comparing with nine existing methods (DC, IC, EC, SC, BC, CC, NC, PeC and ION), we conclude that the overall performance of SON is the best among them. We further analyze the performance of SON in predicting low/high connectivity essential proteins, and discover that SON can predict a large number of low connectivity essential proteins ignored by the eight existing centrality methods.

Acknowledgment

An abstract of this paper was published by the 11th International Symposium on Bioinformatics Research and Applications (ISBRA2015) [60].

Declarations

This article has been published as part of BMC Bioinformatics Volume 17 Supplement 8, 2016. Selected articles from the 11th International Symposium on Bioinformatics Research and Applications (ISBRA '15): bioinformatics. The full contents of the supplement are available online https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-17-supplement-8.

Funding

Publication of this article has been funded by the National Natural Science Foundation of China (No. 61370024, No. 61428209, and No. 61232001).

Availability of data and materials

The program of the proposed algorithm SON and the data (the PPI network, the subcellular localization dataset, and the list of essential proteins) used in this paper are available from http://bioinformatics.csu.edu.cn/resources/softs/SON/index.html.

Authors' contributions

GSL obtained the protein-protein interaction data, essential proteins, orthologous data and subcellular localization data. GSL, ML and JLW designed the new method, SON. GSL, ML, JXW and JLW analyzed the results. GSL, JXW, YP and FXW discussed extensively about this study and drafted the manuscript together. YP and FXW participated in revising the draft. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]School of Information Science and Engineering, Central South University, Changsha 410083, Hunan, People's Republic of China. [2]Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, Guangxi, People's Republic of China. [3]Department of Mechanical Engineering and Division of Biomedical Engineering, University of Saskatchewan, Saskatoon S7N 5A9, SK, Canada. [4]Department of Computer Science, Georgia State University, Atlanta 30302-4110, GA, USA.

Published: 31 August 2016

References

1. Glass JI, Hutchison 3rd CA, Smith HO, Venter JC. A systems biology tour deforce for a near-minimal bacterium. Mol Syst Biol. 2009;5:330.
2. Furney SJ, Alba MM, Lopez-Bigas N. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics. 2006;7:165.
3. Li M, Zheng R, Li Q, Wang J, Wu F, Zhang Z. Prioritizing Disease Genes By Using Search Engine Algorithm. Curr Bioinforma. 2016;11(2):195–202.
4. Lan W, Wang J, Li M, Peng W, Wu F. Computational approaches for prioritizing candidate disease genes based on PPI networks. Tsinghua Sci Technol. 2015;20(5):500–12.
5. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, et al. Functional profiling of the Saccharomyces cerevisiae genome. Nature. 2002;418:387–91.
6. Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, Tandia F, Linteau A, Sillaots S, Marta C, et al. Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. Mol Microbiol. 2003;50:167–81.
7. Cullen LM, Arndt GM. Genome-wide screening for gene function using RNAi in mammalian cells. Immunol Cell Biol. 2005;83:217–23.
8. Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol. 2005;22:803–6.
9. Joy MP, Brock A, Ingber DE, Huang S. High-betweenness proteins in the yeast protein interaction network. J Biomed Biotechnol. 2005;2:96–103.
10. Wuchty S, Stadler PF. Centers of complex networks. J Theor Biol. 2003;223:45–53.
11. Estrada E, Rodriguez-Velazquez JA. Subgraph centrality in complex networks. Phys Rev E. 2005;71:056103.
12. Bonacich P. Power and centrality: A family of measures. Am J Sociol. 1987;92:12.
13. Karen S, Zelen M. Rethinking centrality: Methods and examples. Soc Networks. 2002;11:37.
14. Wang JX, Li M, Wang H, Pan Y. Identification of Essential Proteins Based on Edge Clustering Coefficient. IEEE/ACM trans comput biol bioinforma/IEEE, ACM. 2012;9:1070–80.
15. Tang Y, Li M, Wang JX, Pan Y, Wu FX. CytoNCA: a cytoscape plugin for centrality analysis and evaluation of biological networks. BioSysts. 2015;127: 67–72. doi:10.1016/j.biosystems.2014.11.005.
16. Wang J, Zhong J, Chen G, Li M, Wu F-X, Pan Y. ClusterViz: A Cytoscape APP for Cluster Analysis of Biological Network. IEEE/ACM Trans Comput Biology Bioinform. 2015;12(4):815–22.
17. Li M, Wang JX, et al. A local average connectivity-based method for identifying essential proteins from the network level. Comput Biol Chem. 2011;35:143–50.
18. Li M, Lu Y, Wang JX, Wu FX, Pan Y. A topology potential-based method for identifying essential proteins from PPI networks. IEEE/ACM Trans Comput Biol Bioinform. 2015;12(2):372–83.
19. Acencio ML, Lemke N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinform. 2009;10:290.
20. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, et al. Evolutionary rate in the protein interaction network. Science. 2002;296:750–2.
21. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 2002;12:962–8.
22. Batada NN, Hurst LD, Tyers M. Evolutionary and physiological importance of hub proteins. PLoS Comput Biol. 2006;2, e88.
23. 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.
24. Rocha EPC. Danchin A, An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. Mol Biol Evol. 2004;21:108–16.
25. Krylov DM, Wolf YI, Rogozin IB, Koonin EV. Gene Loss: Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution. Genome Res. 2003;13:2229–35.
26. Li M, Wang JX, Wang H, Pan Y. Identification of Essential Proteins from Weighted Protein Interaction Networks. J Bioinform Comput Biol. 2013;11(3):1341002.
27. Li M, Zhang H, Wang JX, et al. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. BMC Syst Biol. 2012;6:15.
28. Li M, Zheng RQ, Zheng HH, Wang JX, Pan Y. Effective identification of essential proteins based on prior knowledge, network topology and gene expressions. Methods. 2014;67(3):325–33.
29. Li M, Wu XH, Wang JX, Pan Y. Towards the identification of protein Complexes and Functional Modules by integrating PPI network and gene expression data. BMC Bioinform. 2012;13:109.
30. Tang XW, Wang JX, Liu BB, Li M, Chen G, Pan Y. A comparison of the functional modules identified from time course and static PPI network data. BMC Bioinform. 2011;12:339.
31. Xiao QH, Wang JX, Peng XQ, Wu FX, Pan Y. Identifying essential proteins from active PPI networks constructed with dynamic gene expression. BMC Genomics. 2015;16 Suppl 3:S1.
32. Ren J, Wang JX, Li M, Wu FX. Discovering essential proteins based on PPI network and protein complex. Int J DataMing Bioinform. 2015;12(1):24–43.
33. Li M, Lu Y, Niu ZB, Wu FX: United complex centrality for identification of essential proteins from PPI networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics. DOI 10.1109/TCBB.2015.2394487
34. Li M, Chen JE, Wang JX, Hu B, Chen G. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. BMC Bioinform. 2008;9:398.
35. Peng W, Wang J, Cheng Y, et al. UDoNC: an algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks [J]. IEEE/ACM Trans Comput Biol Bioinform. 2015;12(2):276–88.
36. Tang X, Wang J, Zhong J, Pan Y. Predicting essential proteins based on weighted degree centrality. Comput Biology Bioinform, IEEE/ACM Transactions on. 2014;11(2):407–18.
37. Peng W, Wang JX, Wang WP, et al. Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. BMC Syst Biol. 2012;6:87.
38. Zhao B, Wang J, Li M, Wu F-X, Pan Y. Prediction of essential proteins based on overlapping essential modules. IEEE Trans Nanobioscience. 2014;13(4):1–10.
39. Li M, Wang JX, Chen JE, Cai Z, Chen G. Identifying the Overlapping Complexes in Protein Interaction Networks. Int J DataMing Bioinform. 2010; 4(1):91–108.
40. Zhong JC, Wang JX, Peng W, Zhang Z, Li M. A Feature Selection Method for Prediction Essential Protein. Tsinghua sci Technol. 2015;20(5):491–9.
41. Consortium TU. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. 2010;38:D142–8.
42. Xenarios I, Salwinski L, Duan XQJ, 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;30:303–5.
43. Mewes HW, Frishman D, Mayer KFX, Munsterkotter M, Noubibou O, Pagel P, et al. MIPS: analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Res. 2006;34:D169–72.
44. Cherry JM. SGD: Saccharomyces Genome Database. Nucleic Acids Res. 1998;26:9.
45. Zhang R, Lin Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res. 2009;37:D455–8.
46. Saccharomyces Genome Deletion Project [http://yeastdeletion.stanford.edu/]. Accessed 20 June 2012.
47. COMPARTMENTS [http://compartments.jensenlab.org]. Accessed 28 Dec 2014.
48. Magrane M and Consortium U: UniProt Knowledgebase: a hub of integrated protein data. Database, 2011: doi:10.1093/database/bar009.
49. Eppig JT, Blake JA, Bult CJ, et al. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res. 2012;40:D881–6.

50. Cherry JM, Hong EL, Amundsen C, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 2011;40:D700–5.
51. Mcquilton P, St Pierre SE, Thurmond J, et al. FlyBase 101—the basics of navigating FlyBase. Nucleic Acids Res. 2011;40:D706–14.
52. Harris TW, Antoshechkin I, Bieri T, et al. WormBase: a comprehensive resource for nematode research. Nucleic Acids Res. 2009;38:D463–7.
53. Ostlund G, Schmitt T, Forslund K, et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 2010;38:D196–203.
54. Estrada E. Virtual identification of essential proteins within the protein interaction network of yeast. Proteomics. 2006;6:35–40.
55. Wang JX, Li M, Chen JE, Pan Y. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. IEEE/ACM Trans Comput Biol Bioinform. 2011;8(3):607–20.
56. Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks. Proc Nat Acad Sci U S A. 2004;101:2658–632.
57. Hart GT, Lee I, Marcotte E. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinform. 2007;8:236.
58. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási AL. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015;347(6224):1257601.
59. Peng X, Wang J, Wang J, Wu F-X, Pan Y: Rechecking the Centrality-Lethality Rule in the Scope of Protein Subcellular Localization Interaction Networks. Plos ONE, DOI:10.1371/journal.pone.0130743.
60. Li G, Li M, Wang J, Wu F.X and Pan Y: A novel method for predicting essential proteins based on subcellular localization, orthology and PPI networks. Proceeding of International Symposium on Bioinformatics Research and Applications (ISBRA2015), 2015;9096 pp.427, June 2015.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit