Fan *et al. BMC Bioinformatics* 2017, **18**(Suppl 13):470
DOI 10.1186/s12859-017-1876-5

BMC Bioinformatics

RESEARCH Open Access

Prediction of essential proteins based on subcellular localization and gene expression correlation

Yetian Fan[1], Xiwei Tang[2,3*], Xiaohua Hu[4], Wei Wu[5] and Qing Ping[4]

From IEEE BIBM International Conference on Bioinformatics & Biomedicine (BIBM) 2016
Shenzhen, China. 15-18 December 2016

Abstract

Background: Essential proteins are indispensable to the survival and development process of living organisms. To understand the functional mechanisms of essential proteins, which can be applied to the analysis of disease and design of drugs, it is important to identify essential proteins from a set of proteins first. As traditional experimental methods designed to test out essential proteins are usually expensive and laborious, computational methods, which utilize biological and topological features of proteins, have attracted more attention in recent years. Protein-protein interaction networks, together with other biological data, have been explored to improve the performance of essential protein prediction.

Results: The proposed method SCP is evaluated on Saccharomyces cerevisiae datasets and compared with five other methods. The results show that our method SCP outperforms the other five methods in terms of accuracy of essential protein prediction.

Conclusions: In this paper, we propose a novel algorithm named SCP, which combines the ranking by a modified PageRank algorithm based on subcellular compartments information, with the ranking by Pearson correlation coefficient (PCC) calculated from gene expression data. Experiments show that subcellular localization information is promising in boosting essential protein prediction.

Keywords: Essential proteins, Subcellular localization information, Modified PageRank algorithm, Protein-protein interaction networks

Background

Although essential proteins are only a small fraction of all proteins, they are indispensable to maintain life for an organism [1, 2]. Without these essential proteins providing all available nutrients [3], it will lead to lethality of life. Therefore, reliable identification of essential proteins is significant for biologists, for that it not only contributes to understanding the basic requirements for subcellular survival, but also plays a key role in practical implications, such as diseases analysis [4, 5], drug design [6, 7] and medical treatments [4]. This problem has attracted enormous amount of researchers, and many experimental methods have been proposed to predict and discover essential proteins through gene knock-out [8, 9], gene knockdown [10–12] and RNA interference [13]. These methods can provide an accurate prediction of essential proteins. However, the poor efficiency and high cost of experimental methods remains a significant challenge. In addition, for identification of essential proteins in some complex organisms, especially ones from humans, these experimental methods are not suitable.

To break through these experimental constraints, some researchers proposed computational methods to predict

*Correspondence: tangxiwei2010@gmail.com
[2]Department of Information Science and Engineering, Hunan First Normal University, 410205 Changsha, China
[3]College of Computer, National University of Defense Technology, 410073 Changsha, China
Full list of author information is available at the end of the article

essential proteins based on features developed in experimental studies. Especially, due to the high-throughput techniques, abundant data of essential proteins has been collected, which served as the basis for several studies that investigate the relationship between characteristics of experimentally identified essential proteins and their topological properties in protein-protein interaction networks (PPI). With the help of computational methods, the burden to test all proteins in experiments can be greatly relieved, so that only tests of top-ranked proteins based on their score of essentiality are prioritized. Jeong et al. used centrality-lethality rule to identify essential proteins in protein-protein interaction networks, which means that proteins most highly connected in the networks tend to be essential proteins [14]. Pereira-Leal et al. reported that there is higher-level correlation among essential proteins compared to that among nonessential proteins [15]. To explain this phenomenon, He and Zhang proposed the concept of essential protein-protein interactions [16]. These studies support the view that evolution of essential PPI networks are more conservative than nonessential PPI networks. Inspired by these studies that explored topological features of PPI networks, some researchers proposed computational methods to identify essential proteins, based on metrics such as betweenness centrality (BC) [17, 18], degree centrality (DC) [19], edge clustering coefficient centrally (NC) [20] and so on. However, all these methods relying on centrality metrics share some limitations. First, PPI networks generated by high-throughput technologies are often incomplete and contain false positive interactions [21]. Second, many of these methods neglect other intrinsic properties of essential proteins. To overcome these limitations, several methods are proposed to incorporate these PPI networks with other biological data. Based on the weighted PPI networks generated by gene expression profiles, Li et al. proposed an edge-aided approach named PeC to predict essential proteins [22]. Then Tang et al. proposed a modified approach named as WDC to improve the prediction performance [23].

Moreover, recently many studies found that the subcellular localization of proteins may play an important role in identifying essential proteins. Acencio and Lemke discover that integration of information from multiple sources including subcellular localization of proteins can improve the accuracy of essential proteins prediction [24]. Peng et al. proposed a Compartment Importance Centrality (CIC) method [25] that incorporate the subcellular localization information in PPI networks. One limitation of CIC method is that it may not differentiate varieties of the interactions among proteins of a large community. To overcome this limitation, in this paper, we propose a novel method that combines information of subcellular compartments with that of Pearson Correlation coefficient (SCP), based on weighted PPI networks to predict essential proteins. Additionally, a modified PageRank method is proposed to assign weights in the PPI networks more accurately.

This paper is organized into four sections. Our algorithm is presented in "Methods" section. Numerical experiments and results analysis are described in "Results and discussion" section. Several conclusions are drawn in "Conclusion" section.

Methods

In this section, we will present our method SCP, that can rank the importance of proteins with computed scores. The final importance scores of our SCP method is determined by two components: the results ranked by our modified PageRank algorithm (MPR) from subcellular localization information, and the results ranked by Pearson correlation coefficient (IPCC) from gene expression data:

$$\mathrm{SCP} = \lambda \cdot \mathrm{NIS}(\mathrm{MPR}) + (1-\lambda) \cdot \mathrm{NIS}(\mathrm{IPCC}), \quad \lambda \in [0,1] \tag{1}$$

where λ is an adjusting parameter for weighting the two components. In this paper the parameter λ is set as 0.5. The MPR is the importance scores computed from modified PageRank algorithm. The IPCC is the importance scores predicted by Pearson Correlation coefficient. In order to predict essential proteins, we propose a novel algorithm combining MPR with IPCC. We expect that protein with a higher SCP score would be more likely to be an essential protein. As the scores of MPR and IPCC may have different range, they should be scaled into $[0,1]$ first. We normalize the two importance scores as follows:

$$\mathrm{NIS}(Score_i) = \frac{Score_i - \min(Score)}{\max(Score) - \min(Score)}, \quad i = 1, 2, \cdots, N \tag{2}$$

MPR importance score of proteins

We first create a weighted PPI networks derived from subcellular compartments information, and then perform a modified PageRank algorithm on the network to compute importance score of proteins. For most eukaryotes, the subcellular compartments generate a specific environment that regulates the biological processes of proteins within cells. Therefore, knowing the subcellular localization of proteins may shed light on understanding the functions of these proteins. Many studies found that proteins interactions in vivo tend to co-locate in the same cellular compartment or adjacent compartments [26]. For example, 76 percent of protein-protein interactions in yeast cells are carried out in the same subcellular compartments [27]. Therefore it may be beneficial to weigh the protein-protein interactions by subcellular localization, and then

predict the importance of proteins based on the weighted protein-protein interactions.

Based on this intuition, we develop a metric to weigh the protein-protein interactions based on the information of subcellular localization. We assume that protein-protein interactions co-located in a small subcellular compartment can be more reliable in predicting essential proteins than those within a large subcellular compartment.

The importance of subcellular compartments

We model the importance of subcellular compartments based on their scales. Suppose there are K subcellular compartments $C_1, C_2, \cdots, C_K$, and the numbers of them are $N_{C_1}, N_{C_2}, \cdots, N_{C_K}$ respectively. Then the importance of subcellular compartment C_i, denoted by ISC, is defined as:

$$\mathrm{ISC}(C_i) = \frac{1}{N_{C_i}}, i = 1, 2, \ldots, K \tag{3}$$

The weight of protein-protein interactions based on subcellular compartments

The importance of protein-protein interactions can be impacted by different subcellular compartments they share. For a given protein P_i, let $\mathrm{SCL}(P_i)$ be the subcellular compartments where protein P_i located. The weight of P_i and P_j interaction is denoted by $\mathrm{W_{PPI}}(P_i, P_j)$, which is defined as:

$$\mathrm{W_{PPI}}\left(P_i, P_j\right) = \begin{cases} \max\limits_{C_i \in \mathrm{SC}(P_i, P_j)} \{\mathrm{ISC}(C_i)\}, & \mathrm{SCL}(P_i) \bigcap \mathrm{SCL}(P_j) \neq \emptyset, \\ \min\limits_{C_i \in \mathrm{SC}(P_i, P_j)} \{\mathrm{ISC}(C_i)\}, & \text{otherwise} \end{cases} \tag{4}$$

where

$$\mathrm{SC}\left(P_i, P_j\right) = \begin{cases} \mathrm{SCL}(P_i) \bigcap \mathrm{SCL}(P_j), & \mathrm{SCL}(P_i) \bigcap \mathrm{SCL}(P_j) \neq \emptyset, \\ \mathrm{SCL}(P_i) \bigcup \mathrm{SCL}(P_j), & \text{otherwise} \end{cases} \tag{5}$$

A pair of proteins may be co-located in several subcellular compartments because many proteins are annotated by multiple subcellular compartments. Here $\mathrm{SCL}(P_i) \bigcap \mathrm{SCL}(P_j)$ means the common subcellular compartments that protein P_i and P_j are co-located in. We assume that a pair of proteins in the smaller subcellular compartments is most likely to interact with each other than them in the bigger compartments. Therefore, if a pair of proteins are co-located in at least one subcellular compartment, that is $\mathrm{SCL}(P_i) \bigcap \mathrm{SCL}(P_j) \neq \emptyset$, we choose the maximum of the importance of their common subcellular compartments as the importance of the protein-protein interaction between the two proteins. Otherwise, the importance between a pair of proteins which do not share any subcellular compartments will be the minimum of all their subcellular compartments, defined as $\mathrm{SCL}(P_i) \bigcup \mathrm{SCL}(P_j)$.

The importance of proteins

By analyzing the weighted protein-protein interaction network, we can achieve prior estimate on the importance of each protein. The proteins which have stronger interactions with others to be more important proteins (essential proteins). Guided by this idea, we sum up all the weights of protein-protein interactions related to a protein P_i as its prior importance (denoted by $\mathrm{IPSC}(P_i)$):

$$\mathrm{IPSC}(P_i) = \sum_{P_j \in \mathrm{SCL}(P_i)} W_{\mathrm{PPI}}\left(P_i, P_j\right) \tag{6}$$

Modified PageRank algorithm

PageRank is one of the most famous methods that rank the importance of nodes in networks based on link structures of nodes. The basic idea of PageRank algorithm is that the importance of a node is determined by the importance of their parents nodes and the number of their parents nodes. Therefore, by analyzing the quantity and quality of their parents nodes, PageRank algorithm can give a rough importance estimates for all nodes in networks.

In the classic PageRank algorithm, the importance of nodes can be defined as follows:

$$PR(P_i) = \alpha \sum_{P_j \in \mathrm{SCL}(P_i)} \frac{1}{\mathrm{L}(P_j)} PR(P_j) + (1 - \alpha)\frac{1}{N} \tag{7}$$

where N is the number of the nodes, and $\mathrm{L}(P_j)$ is the number of outbound links for node P_j, which belongs to the set of nodes that link to P_i, also denoted by $\mathrm{SCL}(P_i)$. α is a dampening factor set to 0.85 in this paper.

Equation 7 can be re-written in a matrix form as:

$$PR = M \times PR \tag{8}$$

where

$$M = \alpha M_1 + (1 - \alpha) M_2, \quad \alpha \in [0, 1] \tag{9}$$

and

$$M_1(i, j) = \begin{cases} \frac{1}{\mathrm{L}(P_j)}, & if\ P_j \in \mathrm{SCL}(P_i), \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

$$M_2 = \frac{1}{N} \mathbf{1}_{N \times N} \tag{11}$$

We propose a modified PageRank algorithm to calculate the importance of nodes MPR, defined as follows:

$$\tilde{MPR}^{k+1} = \hat{M} \times MPR^k \tag{12}$$

Here the modified iterator matrix $\hat{M}$ is divided into two matrices:

$$\hat{M} = \alpha \hat{M}_1 + (1 - \alpha)\hat{M}_2, \quad \alpha \in [0, 1] \tag{13}$$

where sparse hyperlink matrix $\hat{M}_1$ are generated from the weighted protein-protein interaction networks:

$$\hat{M}_1(i,j) = \begin{cases} \frac{W_{PPI}(P_i,P_j)}{\sum_{P_k \in SCL(P_i)} W_{PPI}(P_i,P_k)}, & if\ P_j \in SCL(P_i), \\ 0, & otherwise \end{cases} \quad (14)$$

and the reset probability matrix M2 comes from the prior importance of proteins:

$$\hat{M}_2(i,j) = \frac{IPSC(P_i)}{\sum_{k=1}^{N} IPSC(P_k)} \quad (15)$$

Finally, the importance of nodes is normalized as follows:

$$MPR^{k+1} = \frac{\tilde{MPR}^{k+1}}{\left\| \tilde{MPR}^{k+1} \right\|} \quad (16)$$

Pearson correlation coefficient

Pearson correlation coefficient (PCC) is a popular method to measure linear correlation between two variables. Here we utilize PCC, derived from gene expression data, to calculate the importance of protein-protein interactions. Given gene expression data of two proteins, denoted by $X = (x_1, \cdots, x_m)$ and $Y = (y_1, \cdots, y_m)$, the importance of protein-protein interactions between the two proteins can be calculated as follows:

$$PCC(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{m} (y_i - \bar{y})^2}} \quad (17)$$

Finally, the importance of each protein Pi, denoted as IPCC(Pi), is computed by summing up all weights of protein-protein interaction importance of protein P_i:

$$IPCC(P_i) = \sum_{P_j \in SCL(P_i)} PCC(P_i, P_j) \quad (18)$$

Results and discussion

In this section, experiments are carried out to evaluate the effectiveness of our algorithm. We take advantage of three types of datasets, namely protein-protein interactions data, gene expression data and subcellular localization data, to predict essential proteins for Saccharomyces cerevisiae. We compare the performance of our algorithm SCP against other five methods (CIC, DC, NC, PeC, WDC) on real dataset of essential proteins. The results show that our method SCP outperforms the other five methods.

Experimental data

Protein-protein interactions data

We downloaded protein-protein interaction networks from the Biogrid database (BIOGRID-3.2.111), which is a freely accessible database to provide physical and genetic interactions [28]. The network consists of 6304 proteins and 81,614 interactions between them.

Gene expression data

The gene expression data of yeast was obtained from the NCBI Gene Expression Omnibus website. This dataset was collected at 36 different times from 9335 probes (uploaded on April 14, 2011), since there is evidence that the expression of gene is periodic during metabolic cycle of Saccharomyces cerevisiae [29]. In total 6777 genes are present in the dataset, some of which have more than one expression profiles. For genes that have multiple expression profiles, we select the profile whose average is maximum.

Subcellular localization data

The COMPARTMENTS database [30] contains subcellular localization information from several data sources, such as literature, high-throughput microscopy-based screens, prediction from primary sequence and text mining. The dataset includes 819 subcellular compartments in total, which was downloaded on April 20, 2014.

Essential protein set

This set of essential proteins were downloaded from DEG [3], MIPS [31], SGD [32] and SGDP. It contains 1204 essential proteins in all.

ROC curves

The proteins of Saccharomyces cerevisiae are classified into essential and nonessential proteins, so the prediction

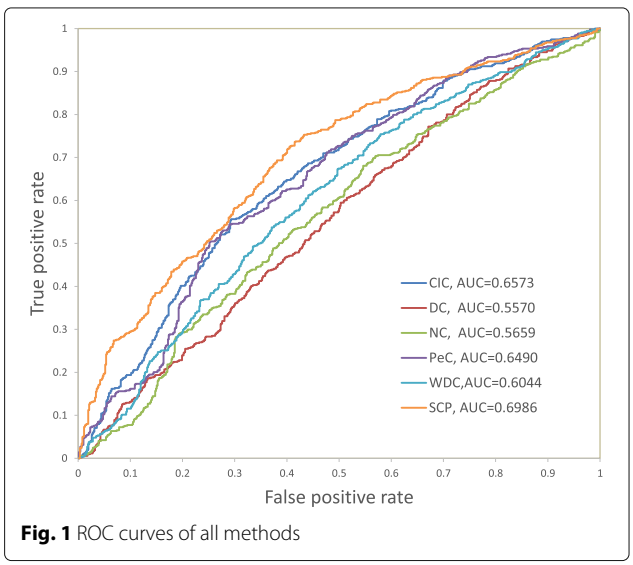

Fig. 1 ROC curves of all methods

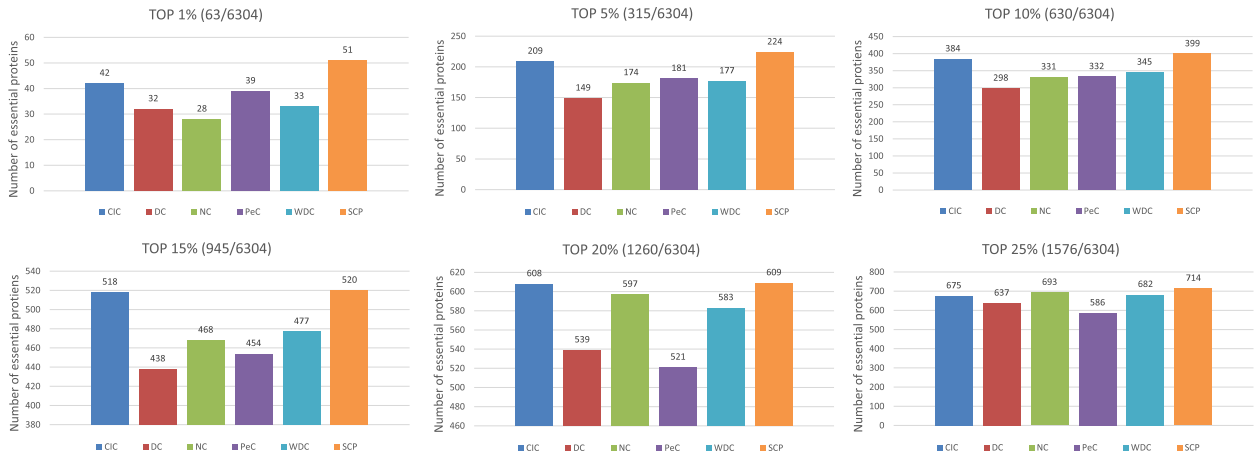

Fig. 2 Number of essential proteins in ranked proteins

of essential proteins is actually a two-class classification problem. Hence, ROC curve is a proper metric to evaluate the performance of a binary classifier, plotted at different thresholds. In an ROC curve, the horizontal axis represents the values of false positive rate (FPR) and vertical axis represents the values of the true positive rate (TPR). The false positive rate is also known as specificity and the true positive rate is also known as sensitivity or recall. They are defined as follows:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{19}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{20}$$

where FP is the number of false positive, which means a prediction is positive and the actual value is negative. Conversely, FN is the number of false negative, which means the prediction is negative while the actual value is positive. Then TP is the number of true positive when both the prediction and actual value are positive. TN is the number of true negative when both the prediction and true value are negative.

Furthermore, the size of the area under the curve, named AUC, is used to evaluate the performance of a binary classifier. Therefore, the larger the AUC value is, the better classifier is. In Fig. 1, ROC curves are plotted to analyze the top 1204 proteins ranked by all six algorithms, because our dataset contains 1204 essential proteins in total. As DC is a simple topological centrality algorithm, the AUC for DC is only 0.5570. Then NC

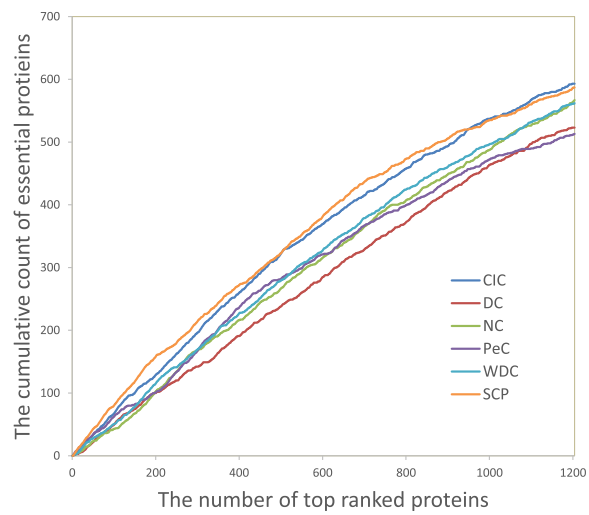

Fig. 3 Jackknife curves of all methods

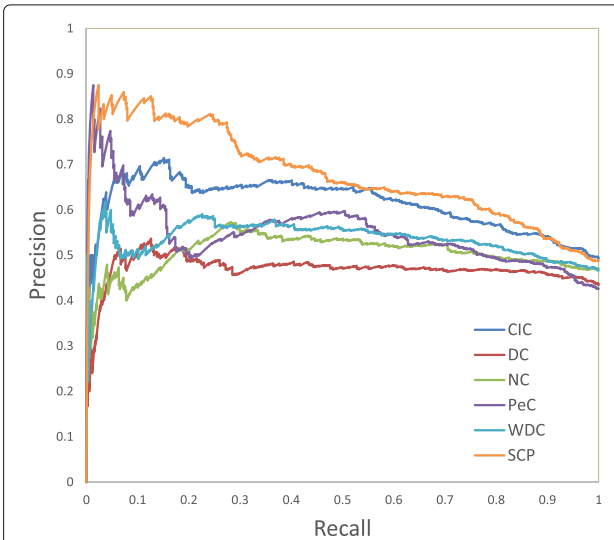

Fig. 4 Precision-recall curves of all methods

is a method applying the edge-clustering coefficient to predict essential proteins, which achieves a litter better performance than DC. PeC and WDC have higher AUC values than DC and NC since they both incorporate gene expression data with PPI data to boost classification performance. CIC performs better than PeC, WDC, NC and DC, since it combines the subcellular localization information with other types of data. Lastly, our method SCP outperforms all the other five methods with a considerable margin. This shows the effectiveness of our fusion method.

Analysis of essential proteins of top ranked proteins

In this section, we attempt to visualize the proportion of essential proteins in top ranked proteins by all methods, including our method SCP and other five methods. First, we rank proteins by their importance scores in descending order computed by all six methods. Second, we select the top $1, 5, \cdots, 25$ percent of all 6304 proteins in their ranked order as essential protein candidates. Then we count the number of real essential proteins in these essential protein candidates according to the golden standard dataset of real essential proteins. The comparative results are shown in Fig. 2. From this figure, we can observe that the SCP outperforms all the other five algorithms on all six proportions of essential proteins.

In the Fig. 2, let us take the top 1% ranked proteins as an example: our method achieves considerable margin compared to other five methods (51 true essential proteins versus 42,32,28,39 and 33 for CIC, DC, NC, PeC and WDC respectively). In addition, Fig. 2 shows that DC and PeC performs better at top 1% and 5% than NC and WDC. However, from top 15 to 25%, the performances of NC and WDC are better than those of DC and PeC. The performance of CIC is good except at the top 25% ranked proteins, when it ranks fourth, and is only better than DC and PeC. In summary, our method achieves the best performances consistently at various percentage of top ranked proteins.

Jackknife curves

In this section, we compare our method with five other methods by the jackknife curves, which is proposed by Holman et al. [33] to show the ability to recover known

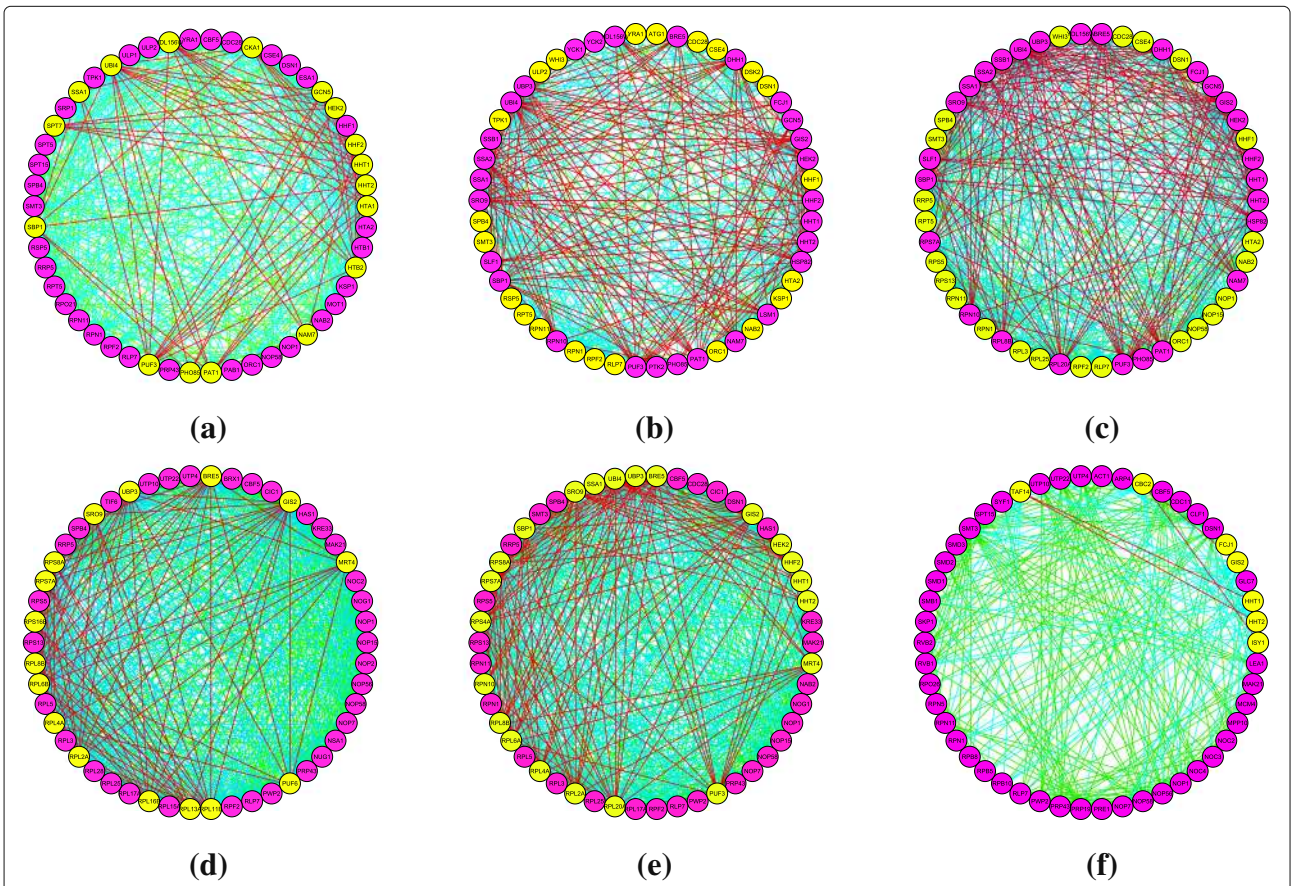

Fig. 5 The comparative results of protein-protein interaction links by six methods. The figure shows the networks of the proteins ranked in top 50 by all six methods, and the links between them. The pink nodes represent the essential proteins, and the yellow nodes represent the nonessential proteins. Red, blue and green links represent Noness-Noness, Ess-Noness and Ess-Ess interactions respectively. **a** CIC. **b** DC. **c** NC. **d** PeC. **e** WDC. **f** SCP

essential proteins. The results are shown in Fig. 3. The horizontal axis of the jackknife curves represents the proteins ranked by scores of importance in descending order from left to right. In this section, we choose the top 1204 proteins of all the six methods to analyze the performance.The vertical axis is the cumulative count of essential proteins. Compared with other five methods, the AUC of our method is the largest. The Jackknife curves also reveal that the performance of our method SCP is better than the other methods.

Precision-recall curves

In this section, we employ precision-recall (PR) curves to compare the performance of our method SCP with the other methods. The recall has been defined as the true positive rate (TPR) in "ROC curves" section. The precision is defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{21}$$

To analyze a binary classification, precision is a measure of the proportion of results that are relevant to the query, and recall is a measure of the proportion of results relevant to the query that are successfully retrieved. If AUC is high, both precision and recall are high. High score of precision suggests the classifier achieves accurate results, while high recall indicates the classifier obtains a majority of all positive results. Because there are 1204 essential proteins in our dataset, we also plot PR curves to analyze the top 1204 proteins ranked by all six algorithms. It is shown in Fig. 4 that SCP achieves the best performance among all the methods.

The analysis of links between top ranked proteins

In this section, we will do some further analysis of the links between top ranked proteins for all the methods. We construct small PPI networks based on the top 50 ranked proteins and the links depending on the whole yeast PPI networks. The results are shown in Fig. 5. Pink nodes represent essential proteins, while yellow nodes represent nonessential proteins identified by six methods. In this study, 43 essential proteins are obtained by our method SCP in the top 50 proteins, while for CIC, DC, NC, PeC, WDC, it is only 33, 22, 23, 34 and 28 respectively. Meanwhile, we analyze the links between top ranked proteins. As the number of links between top ranked proteins is different for various methods, we calculate the proportion of the links between essential proteins (Ess-Ess), between essential proteins and nonessential proteins (Ess-Noness), and between nonessential proteins (Noness-Noness). In Fig. 5, red, blue and green links represent Noness-Noness, Ess-Noness and Ess-Ess interactions respectively. From the Fig. 5, it is easy to find for SCP, the number of Noness-Noness interactions is much less than those of the other methods. For Ess-Ess and Ess-Noness interactions, it is not easy to distinguish the difference of all the methods as these kinds of links are too many. Therefore, in order to show more details of the comparison of SCP and other methods, many experiments are carried out shown in Table 1. It shows the proportions of Ess-Ess, Ess-Noness and Noness-Noness from top 100 to top 400 ranked proteins for all six methods. From the table, it shows SCP obtained the best performance of all the methods. For instance, in the top 100 ranked proteins, the proportion of Noness-Noness for our method is only 4.11%, which is much lower than other methods, while the proportion of Ess-Ess for our method is up to 63.58%, which is the highest of all the methods.

Table 1 Analysis of link proportion

Top	Link	CIC	DC	NC	PeC	WDC	SCP
100	Ess-Ess	44.64%	27.82%	18.34%	42.22%	26.43%	**63.58%**
	Ess-Noness	43.21%	45.86%	45.52%	35.91%	44.92%	32.31%
	Noness-Noness	12.15%	26.32%	36.14%	21.87%	28.64%	**4.11**%
200	Ess-Ess	45.91%	26.78%	23.86%	35.74%	34.03%	**66.05%**
	Ess-Noness	41.70%	47.80%	42.88%	35.94%	41.50%	28.21%
	Noness-Noness	12.39%	25.33%	33.27%	28.32%	24.46%	**5.74%**
300	Ess-Ess	45.74%	23.58%	30.33%	37.20%	35.02%	**53.90%**
	Ess-Noness	41.68%	47.01%	42.62%	36.18%	40.96%	35.84%
	Noness-Noness	12.58%	29.41%	27.05%	26.62%	24.02%	**10.26%**
400	Ess-Ess	46.15%	23.74%	30.89%	39.58%	35.35%	**51.23%**
	Ess-Noness	40.94%	46.22%	42.36%	36.39%	40.96%	37.20%
	Noness-Noness	12.92%	30.04%	26.75%	24.04%	23.70%	**11.56%**

(Optimal values are denoted by boldface)

Table 2 Number of essential proteins in top ranked proteins from SCP on various value of λ

λ	1%	5%	10%	15%	20%	25%
0	45	173	335	437	521	589
0.5	**51**	**224**	399	**520**	**609**	**714**
1	49	216	**403**	517	603	700

(Optimal values are denoted by boldface)

The analysis of parameter λ

In this section, we discuss the selection of parameter λ. As the prediction of essential proteins is an unsupervised learning procedure, we can't learn a best parameter λ from the data. Therefore, we only choose $\lambda \in \{0, 0.5, 1\}$ to analyze the performance of our algorithm SCP. In reality, when $\lambda = 0$, the results of SCP only come from IPCC. Conversely, the results will only be calculated by MPR when $\lambda = 1$. In this paper, we chose λ as 0.5, which means the results of SCP integrate MPR and IPCC. In order to compare the performance of the method on various λ, we calculate the number of essential proteins at different top percentages of ranked proteins (top 1%, 5%, 10%, 15%, 20%, 25%). From Table 2, it demonstrates that when $\lambda = 0.5$, SCP obtains the best performance. Therefore, in this paper the parameter λ is set as 0.5. As a result, SCP successfully integrates the results of MPR and IPCC and has achieved a great boost on the performance of essential proteins prediction.

The analysis of the performance of CIC and SCP

In this section, we will analyze the performance of CIC and SCP. Both CIC and SCP utilize the subcellular localization information to predict the essential proteins, while SCP also use the information of the gene expression data. Therefore, we will compare CIC with modified PageRank (MPR), part of our method SCP, which only uses the subcellular localization information as CIC does to predict essential proteins. The results are shown in Table 3. Although the performance of MPR is worse than SCP, MPR achieves better performance than CIC in most cases, except for top 15 and 20 percentages, where the number of essential proteins identified by MPR is a little less than those does by CIC.

Table 3 Number of essential proteins in top ranked proteins identified by CIC, MPR and SCP

Method	1%	5%	10%	15%	20%	25%
CIC	42	209	384	518	608	675
MPR	49	216	**403**	517	603	700
SCP	**51**	**224**	399	**520**	**609**	**714**

(Optimal values are denoted by boldface)

Conclusion

Essential proteins are crucial to the development and survival of life. Many computational methods are proposed to detect essential proteins based on biological and topological features of proteins. In our study, we also found that integration of information from multiple sources can boost the identification of essential proteins. Specifically, the utilization of subcellular localization information can make a remarkable contribution to the prediction of essential proteins. In this paper, a SCP method is proposed, which integrates the ranking function by a modified PageRank algorithm with weighted subcellular localization with Pearson correlation coefficient based on gene expression data. Several experiments are carried out to compare the performance of SCP with five other methods in identification of essential proteins. Experimental results show that our method SCP performs the best among all six methods.

Acknowledgements
Not applicable.

Funding
This work was supported by the National Natural Science Foundation of China (No. 61473059, 61472133), the Fundamental Research Funds for the Central Universities of China and NSFC 61532008. Publication of this article was funded by the the National Natural Science Foundation of China (No. 61472133).

Availability of data and materials
The source code and data for implementing our method are available from the corresponding author. The datasets used in this study are downloaded at https://thebiogrid.org http://moment.utmb.edu/cgi-bin/main_cc.cgi https://compartments.jensenlab.org/Downloads.

About this supplement
This article has been published as part of *BMC Bioinformatics* Volume 18 Supplement 13, 2017: Selected articles from the IEEE BIBM International Conference on Bioinformatics & Biomedicine (BIBM) 2016: bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-18-supplement-13.

Authors' contributions
YF conceived, designed and implemented this study. XT performed the data collection and analysis. YF and QP drafted the manuscript. XT, XH and WW contributed useful discussion and suggestion to complete the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]School of Mathematics, Liaoning University, 110036 Shenyang, China. [2]Department of Information Science and Engineering, Hunan First Normal University, 410205 Changsha, China. [3]College of Computer, National University

of Defense Technology, 410073 Changsha, China. [4]College of Computing and Informatics, Drexel University, 19104 Philadelphia, USA. [5]School of Mathematical Sciences, Dalian University of Technology, 116023 Dalian, China.

Published: 1 December 2017

References

1. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K. Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. Science. 1999;285:901–6.
2. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R. Systematic functional analysis of the caenorhabditis elegans genome using rnai. Nature. 2003;421:231–7.
3. Zhang R, Lin Y. Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res. 2009;37:455–8.
4. Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D. Systematic screen for human disease genes in yeast. Nature Gene. 2002;31:400–4.
5. 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.
6. Judson N, Mekalanos JJ. Tnaraout, a transposon-based approach to identify and characterize essential bacterial genes. Nat Biotechnol. 2000;18(7):740–5.
7. Lamichhane G, Zignol M, Blades NJ, et al. A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to mycobacterium tuberculosis. Proc Natl Acad Sci. 2003;100(12):7213–8.
8. Giaever G, Chu AM, Ni L, Connelly C. Functional profiling of the saccharomyces cerevisiae genome. Nature. 2002;418(6896):387–91.
9. Chen L, Ge X, Xu P. Identifying essential streptococcus sanguinis genes using genome-wide deletion mutation. Gene Essentiality Methods Protoc. 2015;1279:15–23.
10. Roemer T, Jiang B, Davison J, et al. Large-scale essential gene identification in candida albicans and applications to antifungal drug discovery. Mol Microbiol. 2003;50(1):167–81.
11. Harborth J, Elbashir SM, Bechert K, et al. Identification of essential genes in cultured mammalian cells using small interfering rnas. J Cell Sci. 2001;114(24):4557–65.
12. Zhang B, Ji Y, Van SF, et al. Identification of critical staphylococcal genes using conditional phenotypes generated by antisense rna. Science. 2001;293:2266–9.
13. Cullen LM, Arndt GM. Genome-wide screening for gene function using rnai in mammalian cells. Immunol Cell Biol. 2005;83(3):217–23.
14. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001;411:41–2.
15. Pereira-Leal JB, Audit B, Peregrin-Alvarez JM, Ouzounis CA. An exponential core in the heart of the yeast protein intercation network. Mol Biol Evol. 2005;22(3):421–5.
16. He X, Zhang J. Why do hubs tend to be essential in protein networks? PLoS Genet. 2006;2(6):826–34.
17. Freeman LC. A set of measures of centrality based on betweenness. Sociometry. 1977;40(1):35–41.
18. Joy MP, Brock A, Ingber DE, Huang S. High-betweenness proteins in the yeast protein interaction network. BioMed Res Int. 2005;2:96–103.
19. Vallabhajosyula RR, Chakravarti D, Lutfeali S, et al. Identifying hubs in protein interaction networks. PLoS One. 2009;4(4):5344.
20. Wang J, Li M, Wang H, Pan Y. Identification of essential proteins based on edge clustering coefficient. IEEE/ACM Trans Comput Biol Bioinforma. 2012;9(4):1070–80.
21. Sprinzak E, Sattath S, Margalit H. How reliable are experimental protein-protein interaction data? J Mol Biol. 2003;327(5):919–23.
22. 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(1):15.
23. Tang X, Wang J, Zhong J, Pan Y. Predicting essential proteins based on weighted degree centrality. IEEE/ACM Trans Comput Biol Bioinforma. 2014;11(2):407–18.
24. Acencio ML, Lemke N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinformatics. 2009;10:290–307.
25. Peng XQ, Wang JX, Zhong JC, et al. An efficient method to identify essential proteins for different species by integrating protein subcellular localization information. IEEE Int Conf Bioinforma BioMed (BIBM). 2015;2015:277–80.
26. Kumar A, Agarwal S, Heyman JA, et al. Subcellular localization of the yeast proteome. Genes Dev. 2002;16:707–19.
27. Schwikowski B, Uetz P, Field S. A network of protein-protein interactions in yeast. Nat Biotechnol. 2000;18:1257–61.
28. Stark C, Breitkreutz BJ, Reguly T, et al. Biogrid: A general repository for interaction datasets. Nucleic Acids Res. 2006;34:535–9.
29. Tu B, Kudlicki A, Rowicka M, McKnight S. Logic of the yeat metabolic cycle: Temporal compartmentalization fo cellular processes. Scinence. 2005;310:1152–8.
30. Binder JX, Pletscher-Frankild S, Tsafou K, et al. Compartments: unification and visualization of protein subcellular localization evidence. Database. 2014;2014:900.
31. Mewes HW, Frishman D, Munsterkotter KFX, et al. Mips: Analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Res. 2006;34(1):169–72.
32. Cherry JM, Adler C, Ball C, et al. Sgd: Saccharomyces genome database. Nucleic Acids Res. 1998;26(1):73–9.
33. Holman A, Davis P, Foster J, et al. Computational prediction of essential genes in an unculturable endosymbiotic bacterium, wolbachia of brugia malayi. BMC Microbiol. 2009;9:243.