SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •
Special Focus on Biomolecular Network Analysis and Application

July 2016, Vol. 59 070106:1–070106:11
doi: 10.1007/s11432-016-5583-z

Identifying essential proteins based on dynamic protein-protein interaction networks and RNA-Seq datasets

Xuequn SHANG[1], Yu WANG[1] & Bolin CHEN[1]*

School of Computer Science, Northwestern Polytechnical University, Xi'an 710072*, China*

Received April 6, 2016; accepted April 18, 2016; published online June 6, 2016

Abstract The identification of essential proteins is not only important for understanding organism structure on the molecular level, but also beneficial to drug-target detection and genetic disease prevention. Traditional methods often employ various centrality indices of static protein-protein interaction (PPI) networks and/or gene expression profiles to predict essential proteins. However, the prediction accuracy of most methods still has room to be further improved. In this study, we propose a strategy to increase the prediction accuracy of essential protein identification in three ways. Firstly, RNA-Seq datasets are employed to construct integrated dynamic PPI networks. Using a RNA-Seq dataset is expected to give more accurate predictions than using microarray gene expression profiles. Secondly, a novel integrated dynamic PPI network is constructed by considering both the co-expression pattern and the co-expression level of the RNA-Seq data. Thirdly, a novel two-step strategy is proposed to identify essential proteins from two known centrality indices. Numerical experiments have shown that the proposed strategy can increase the prediction accuracy dramatically, which can be generalized to many existing methods and centrality indices.

Keywords essential protein, dynamic protein network, RNA-Seq data, gene co-expression pattern, M2 measure

Citation Shang X Q, Wang Y, Chen B L. Identifying essential proteins based on dynamic protein-protein interaction networks and RNA-Seq datasets. Sci China Inf Sci, 2016, 59(7): 070106, doi: 10.1007/s11432-016-5583-z

1 Introduction

Essential proteins are indispensable for organisms. Recent research has revealed that essential proteins are related to many significant areas, such as disease prediction, drug design, biological research on cellular levels, and activity research on lives.

There are two kinds of traditional approaches to the detection of essential proteins. One kind is based on experimental procedures, including knockout of a single gene [1] and RNA interference [2], while the other kind is based on the topology of protein-protein interaction (PPI) networks, including using both static PPI networks and dynamic PPI networks [3]. However, the former experimental procedures are

* Corresponding author (email: blchen@nwpu.edu.cn)

often difficult and the resulting observations are accompanied with many noise data [4]. Hence, the latter approaches that prioritize associations between vertex centralities and essentiality have been proposed in many studies [5–7]. Most of those approaches focus on the topology of static networks and protein complexes, where proteins with more interactions tend to be regarded as essential proteins [8–11].

Designing efficient vertex centrality contributes one aspect to identifying essential proteins [12]. Centrality-lethality rules consist of three well-known measures, namely degree centrality (DC), closeness centrality (CC), and betweenness centrality (BC) [13, 14]. The existing studies have declared that high-degree vertices in a PPI network tend to be responsible for essential proteins [15, 16]. Several measures based on these three basic centralities have been proposed, such as eigenvector centrality (EC) [17], local average connectivity (LAC) [18], information centrality (IC) [13], sub-graph centrality (SC) [19], etc. They can all be used to identify essential proteins from various static PPI networks.

The construction of a dynamic PPI network contributes another aspect to identifying essential proteins [20]. A classical dynamic PPI network construction method uses time-course microarray gene expression profiles to generate a set of time-course temporary sub-networks [21]. Each sub-network consists of a set of temporary active vertices whose gene expression levels are larger than a threshold [22, 23]. Some forms of novel centrality, such as DLAC [24], have also been proposed to identify essential proteins from such a time-course dynamic network. However, we argue that the expression of essential proteins should not tend to vary with time very often. Hence, the method of selecting only active proteins could be further improved by considering an integrated dynamic PPI network.

Moreover, for the essential protein identification methods, although those using dynamic PPI networks can capture the characteristics of gene expression profiles with temporal information, many existing studies do not take full advantage of dynamic information [25], such as the degree of individual vertices and their connection to neighbors [24]. They often simply add the averages of topology indices of individual temporary sub-networks, instead of using the change information based on the temporal processes and the links between them.

Last but not least, with the rapid development of high-throughput sequencing technology, time-course RNA-Seq datasets have provided us with a new way to study gene expression profiles [26] and to construct more accurate dynamic PPI networks. Hence, the prediction accuracy of the associated identification of essential proteins can also benefit from using this new kind of data source.

In this study, we propose a strategy to increase the accuracy of the prediction of essential proteins in the following three aspects. Firstly, we use RNA-Seq data to construct dynamic PPI networks. Compared with general gene expression datasets, using RNA-Seq data to induce gene expression profiles has three advantages, which are (1) a larger dynamic range, (2) lower background noise, and (3) a better ability to detect and quantify unknown transcripts and subtypes [26]. Secondly, we propose a new method to construct an integrated dynamic PPI network by using gene co-expression patterns (CPs) and gene co-expression levels. Classic time-course dynamic PPI networks consist of many temporal sub-networks, where active proteins are generally selected from static PPI networks according to the gene expression levels. In this study, we argue that essential proteins should not exhibit significantly different expression levels across the whole time period, but rather tend to have CPs and co-expression levels. Dynamic information about individual protein pairs is then employed to construct an integrated dynamic PPI network in two stages. In the first stage, protein pairs with similar CPs are employed to select protein interactions, since they tend to have strong functional relationships [27]. In the second stage, the Pearson correlation coefficient (PCC) values of individual protein pairs are used to filter protein interactions further by setting a global cut-off threshold. By doing this, we can take advantage of both biological sense and mathematical models of gene expressions. The CP only makes sure two genes have a similar expression pattern, and it does not consider the amplitude information of the gene expressions. The PCC value, on the other hand, mathematically only indicates that two genes are co-expressed, and their expression patterns are not necessarily the same. Thirdly, based on the obtained PPI network, we propose a two-step centrality improvement strategy to increase the accuracy of prediction of essential proteins. Numerical experiments have shown that the proposed strategy can improve the prediction accuracy for well-known DC and LAC centralities, which can be generalized to many existing methods.

2 Methods

In this section, we introduce our method from four aspects: (1) the preprocessing of RNA-Seq data, (2) the construction of the integrated dynamic PPI networks, (3) the motivation for the two-step centrality improvement strategy, and (4) the evaluation methods for verifying the essentiality of the predicted proteins.

2.1 Preprocessing of RNA-Seq data

The RNA-Seq dataset used in this study is derived from *Saccharomyces cerevisiae*, which includes four time points, namely 0, 30, 60 and 180 min. Each of these includes around 2 million reads where the length of each read is around 50 bp.

Data preprocessing consists of three steps. First of all, Bowtie2 [28] is used as the alignment engine. It exploits an extremely economical data structure called the FM index [29] to store the reference genome sequences. The FM index makes the searching of sequences very fast. After that, Tophat [30], a fast splice junction mapper, is applied to align RNA-Seq reads to mammalian-sized genomes using the ultra-high-throughput short-read aligner Bowtie2. It can also be used to analyze the mapping results for identifying splice junctions between exons. Finally, the mapping results are used by Cufflinks [31] against the whole genome for assembling the reads into transcriptions. The output files contain information that illustrates the time-course expression profiles of individual genes. In this study, we use the gene expression profiles obtained from these steps to construct dynamic PPI networks.

2.2 Construction of the integrated dynamic PPI network

Tang et al. [22] propose a classical method to construct a dynamic PPI network using time-course microarray datasets and static PPI networks. They focus on searching with a suitable threshold to select the temporary active genes. Once the threshold is set, a gene with an expression level larger than the threshold is regarded as an active gene and it is then used to construct a temporal network of that specific time point. We argue that the essential proteins may not be active at all times. They may be considered to have more interactions than others. Thus, in this paper, we pay more attention to the gene CPs and gene co-expression levels between individual protein pairs [27]. Hence, both the gene CP and the PCC values are employed to construct the integrated dynamic PPI networks as follows.

2.2.1 *The CP model*

The protein interactions of static PPI networks are first filtered by using a CP model. Let g_1 and g_2 be two genes, and $g_1 = (g_{1,1}, g_{1,2}, \ldots, g_{1,k})$, $g_2 = (g_{2,1}, g_{2,2}, \ldots, g_{2,k})$ be the gene expression profiles induced from the RNA-Seq dataset with k time points. The gene CP of g_1 and g_2, is defined as follows:

$$\mathrm{CP}(g_1, g_2) = \sum_{i-1}^{k-1} |f_i(g_1) - f_i(g_2)| = \sum_{i-1}^{k-1} |\mathrm{sgn}(g_{1,i+1} - g_{1,i}) - \mathrm{sgn}(g_{2,i+1} - g_{2,i})|\,, \tag{1}$$

where $(g_{(1,i+1)} - g_{(1,i)})$ is the change of gene expression level of g_1 from time i to time $i+1$, and $\mathrm{sgn}(*)$ is the sign function. Let $f_i(g_n)$ denote the gene expression pattern of g_n changing from time i to time $i+1$, where $f_i(g_n) = 1$ means the expression level rises during this time interval, $f_i(g_n) = 0$ means the expression level is unchanged during this time interval, and $f_i(g_n) = -1$ means the expression level declines during this time interval. Mathematically, these three patterns are described as follows:

$$f_i(g_n) = \mathrm{sgn}(g_{n,i+1} - g_{n,i}) = \begin{cases} 1, & g_{n,i+1} > g_{n,i}, \\ 0, & g_{n,i+1} = g_{n,i}, \\ -1, & g_{n,i+1} < g_{n,i}. \end{cases} \tag{2}$$

If gene g_1 and g_2 have the same expression pattern, then we have $f_i(X) = f_i(Y)$ for every $i = 1, 0, -1$, so that $\mathrm{CP}(g_1, g_2) = 0$.

2.2.2 *The PCC model*

The PCC value of a gene pair g_1 and g_2 is given as follows:

$$\mathrm{PCC}(g_1, g_2) = \frac{1}{s-1} \sum_{i=1}^{s} \frac{\mathrm{cov}(g_1, g_2)}{\delta_{g_1} \delta_{g_2}}, \tag{3}$$

where the value of $\mathrm{PCC}(g_1, g_2)$ ranges from -1 to 1. If $\mathrm{PCC}(g_1, g_2)$ is a positive value, there is a positive correlation between gene g_1 and g_2. By setting a global threshold, one can select protein interactions from PPI networks if the associated PCC value of two genes is larger than the threshold.

2.3 Construction of integrated dynamic PPI networks

A step-by-step description of the method to construct an integrated dynamic PPI network is given as follows.

1. The gene expression data are prepared first from the RNA-Seq dataset by using Bowtie2, Tophat, and Cufflinks.
2. According to the above results, all gene names are transformed to protein names using the ID mapping function in UniProt.
3. The gene CP model is employed to filter out the less important interactions.
4. The PCC model is employed to filter out interactions further by considering the amplitude effect of gene expressions.
5. A static PPI network is used, where only interactions that satisfy the conditions of steps 3 and 4 are kept in the final network.

2.4 Degree connectivity

According to centrality-lethality, essential proteins are considered to have more interactions than unessential proteins. Degree centrality (DC) is the most straightforward topology index of individual vertices. DC represents the direct influence of a vertex. It is defined as

$$\mathrm{DC}(i) = \frac{k_i}{n-1}, \tag{4}$$

where k_i is the degree of protein i and n is the number of all nodes in this network.

LAC [15] is another important centrality index used to identify essential proteins, which is defined as

$$\mathrm{LAC}(u) = \frac{\sum_{w \epsilon N_u} \deg^{C_u}(w)}{|N_u|}, \tag{5}$$

where u is a given vertex, N_u is the set of neighbors of u, C_u is the induced sub-graph of N_u, and $\deg^{C_u}(w)$ is the degree of any given vertex w in C_u.

In paper [32], we can see that although DC [13] is the most basic vertex centrality in networks, it performs well compared to BC, CC [13, 14], SC [19], EC [17] and IC [13] when predicting essential proteins. Based on DC, we propose a new measure called DC2 in the following section. LAC is a very good topology measure based on DC, so we also conduct some experiments based on LAC.

In this study, we propose a two-step strategy to induce a new measure M2 from any given known measure M of a vertex u, by simply adding the measure of the vertex u together with the measure of its neighbors. By doing this, the new measure can contain topological information about a larger area, which indicates the connectivity information of the target vertex u and its neighbors. Take the LAC measure, for example. The index itself describes the local connectivity of only direct neighbors, while the new LAC2 integrates the local connectivity information of vertices within the second order of neighbors. LAC2 is defined as

$$\mathrm{LAC2}(u) = \mathrm{LAC}(u) + \sum_{w \epsilon N_u} \mathrm{LAC}^{C_u}(w), \tag{6}$$

where u is a given vertex, N_u is the set of neighbors of u, C_u is the induced sub-graph of N_u, and $\mathrm{LAC}^{C_u}(w)$ is the LAC of any given vertex w in C_u. Similarly, we can get DC2 from DC based on the same principle.

Finally, we use DC2 and LAC2 to describe the essentiality of a protein. A higher value of LAC2 indicates a higher probability of being an essential protein.

Theoretically, one can also induce a similar LAC3 from LAC2 by using the same procedure. However, as suggested in [33], a higher order of an iterative step is not always good. Besides, due to the small-world property of the majority of biological networks, an index related to third-order neighbors may involve too many vertices, which may not be efficient for distinguishing the essentiality of individual vertices. Hence, we also argue that the second order of the neighborhood is enough for a topological index.

2.5 Evaluation method

The proposed strategy is evaluated by using the F-measure, fold enrichment, and the text mining method, respectively.

2.5.1 *F-measure*

The F-measure is regularly employed to measure the performance of an algorithm. Suppose X is a set of predicted essential proteins and P is the set of the known essential proteins. The *precision* and *recall* are defined as follows:

$$\text{precision} = \mathrm{Pr} = \frac{|X \bigcap P|}{|X|}, \tag{7}$$

$$\text{recall} = \mathrm{Rc} = \frac{|X \bigcap P|}{|P|}, \tag{8}$$

where $|*|$ is the cardinality of a set.

The F-measure combines Pr and Rc, and it is defined as

$$\text{F-measure} = \frac{2 \cdot \mathrm{Pr} \cdot \mathrm{Rc}}{\mathrm{Pr} + \mathrm{Rc}}, \tag{9}$$

which is the harmonic mean of precision and recall.

2.5.2 *Fold enrichment*

Fold enrichment [34, 35] is employed to describe how those known essential proteins are enriched in the set of predicted essential proteins. Suppose X is the set of predicted essential proteins and E_x is the number of essential proteins in X. The percentage of true positives is $\mathrm{TPR} = E_x/|X|$. Let V be the set of all proteins in the static PPI network and P be the set of all known essential proteins in V. The percentage of all essential proteins in V is given as $\mathrm{EPR} = |P|/|V|$. Hence, the fold enrichment of the predicted essential proteins is defined as

$$\text{fold enrichment} = \frac{\mathrm{TPR}}{\mathrm{EPR}}. \tag{10}$$

2.5.3 *The text mining method*

The list of known essential proteins is by no means complete. Hence, we could also employ text mining technology to evaluate the predicted essential proteins by using our method. A text mining method is proposed as follows. First, we download all the articles that contain both the key words 'essential' and 'protein' (or 'gene') in the content. Then, we traverse all these articles to see whether they contain the predicted protein name. Last, we manually evaluate the predicted proteins for their essentiality in these articles. In this way, we can certify the essentiality of a protein.

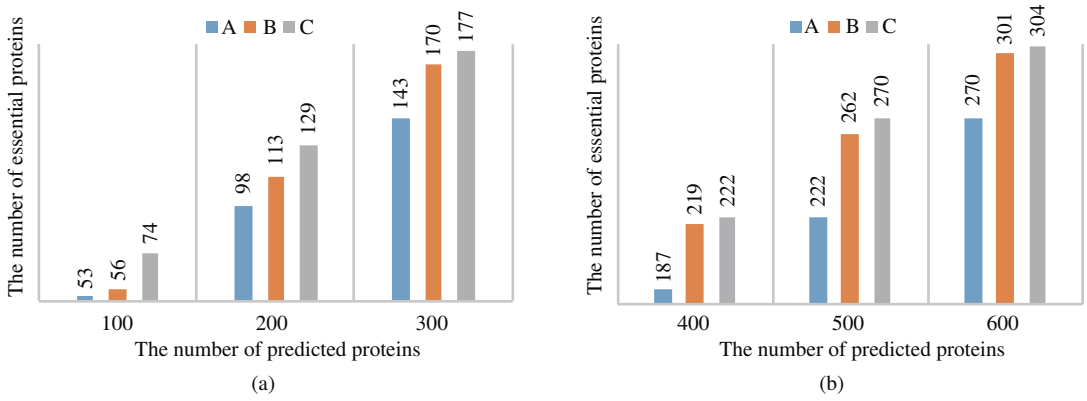

Figure 1 (Color online) The number of true positive predictions using three kinds of dataset. (a) From top 100 to top 300; (b) from top 400 to top 600.

3 Results

3.1 Experimental data sources

Three datasets are employed in this study: a static PPI dataset, an essential protein dataset, and a RNA-Seq dataset. The static PPI dataset was obtained from the DIP database (updated on April 29, 2015). It includes 5146 proteins and 22526 interactions. Redundant interaction and loops were removed from the dataset. The essential protein dataset was downloaded from http://ogeedb.embl.de/#download. Only the data for *Saccharomyces cerevisiae* were selected, which include 1136 essential proteins. The RNA-Seq dataset (SRX362640) was collected from the NCBI SRA database. We also use only the data for the yeast transcriptome during wine fermentation.

We use the ID Mapping function in UniProt to transform a gene name to a protein name. After that, 4633 proteins were left in the static PPI network and 1136 known essential proteins remained as the reference. A set of integrated dynamic PPI networks was then constructed based on the RNA-Seq gene expression profiles. The final integrated dynamic PPI network consists of 1239 proteins of which 500 are essential proteins.

3.2 Effectiveness of using the RNA-Seq dataset

Using the RNA-Seq dataset to construct dynamic PPI network is more powerful than using traditional microarray datasets in terms of predicting essential proteins. Figure 1 illustrates three groups of degree centrality predictions using the different kinds of dataset. Specifically, group A (DPPIN) was proposed by Luo et al. in [24], who selected a set of active proteins from static PPI networks. Group B (APPIN) was proposed by Xiao in [21], who filtered out noisy genes using both a time-dependent model and a time-independent model. Both methods use the time-course microarray data to describe the gene expression profiles. Group C is the network we constructed using the RNA-Seq dataset. The numbers of true positive predictions among the top 100–600 predictions are given in Figure 1. We can see from the figure that our method using the RNA-Seq data is much better than those using microarray datasets.

3.3 Construction of dynamic PPI networks

Five different kinds of protein networks are constructed to predict essential proteins. Firstly, the static PPI network obtained from DIP is employed to predict essential proteins. Secondly, the PCC values of individual protein pairs are employed to construct a dynamic PPI network, where only edges of the static PPI network with a PCC value larger than a threshold are kept. Thirdly, the gene CPs are employed to construct the dynamic PPI networks. In this study, since the RNA-Seq dataset contains only four time points, we propose two methods to describe the gene CPs at two extents. Pattern A shows when two genes are co-increased or co-decreased during at least two time intervals, while pattern B shows when

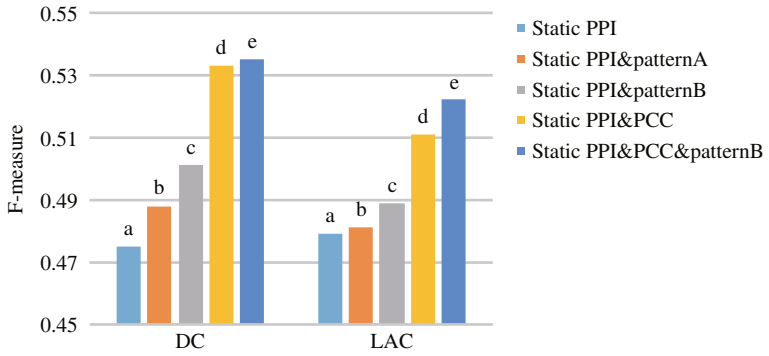

Figure 2 (Color online) The F-measure of essential proteins predicted using different PPI networks.

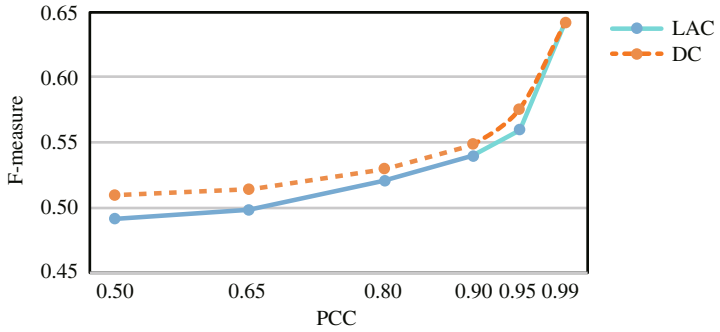

Figure 3 (Color online) The effect of changing the PCC threshold in predicting essential proteins.

two genes are co-increased or co-decreased during all three time intervals. The final kind of dynamic PPI network is constructed by considering both the PCC values and the co-expression pattern B when selecting edges from the static PPI network of DIP.

The DC and LAC methods are employed to identify essential proteins from the above five kinds of PPI networks. The F-measure is employed to evaluate the accuracy of prediction for the essential proteins.

The results are shown in Figure 2. We can see from the figure that no matter whether we use the PCC value or the CP, the performance of the constructed dynamic protein networks is always better than the static PPI network in terms of prediction of essential proteins (a vs the other values in Figure 2). To be more specific, the PCC values are better than CPs (d vs b and c); pattern B is better than pattern A (c vs b in Figure 2); while using both PCC and co-expression pattern B achieves a better result than using either of them alone (e vs c and d in Figure 2). Hence, we conclude that the quality of a dynamic PPI network is better than a static one, and using more time intervals to describe the CP is better than using fewer. Moreover, the final dynamic PPI network we construct by using both the PCC values and the CPs is relatively better than all the others.

3.4 Effect of varying PCC threshold in predicting essential proteins

In the previous section, the PCC threshold is selected as 0.9 empirically, since this value is often employed to describe a strong positive correlation between two objects. However, the value of the PCC threshold apparently influences the accuracy of prediction of essential proteins. To select a better threshold, we use the fifth kind of dynamic PPI network to predict essential proteins by changing the value of the PCC threshold from 0.5 to 0.99 when selecting the active interactions. Both the DC and the LAC methods are employed to calculate the F-measures.

Figure 3 shows the change in F-measure when varying PCC threshold values, while Figure 4 illustrates the change in the numbers of proteins, interactions and known essential proteins of the PPI networks under the same conditions. We can see from the figures that with an increase in the value of the PCC threshold, the value of the F-measure and the percentage of essential proteins rise gradually (Figure 3),

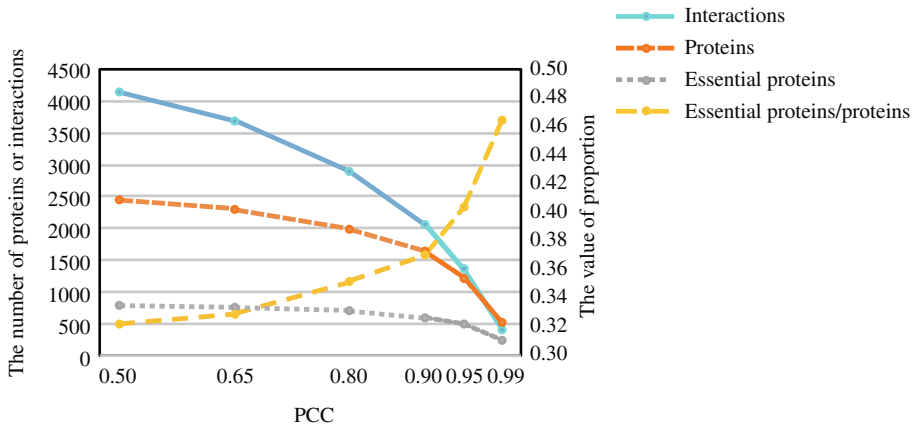

Figure 4 (Color online) Variation in number of interactions, proteins and essential proteins for different PCC threshold.

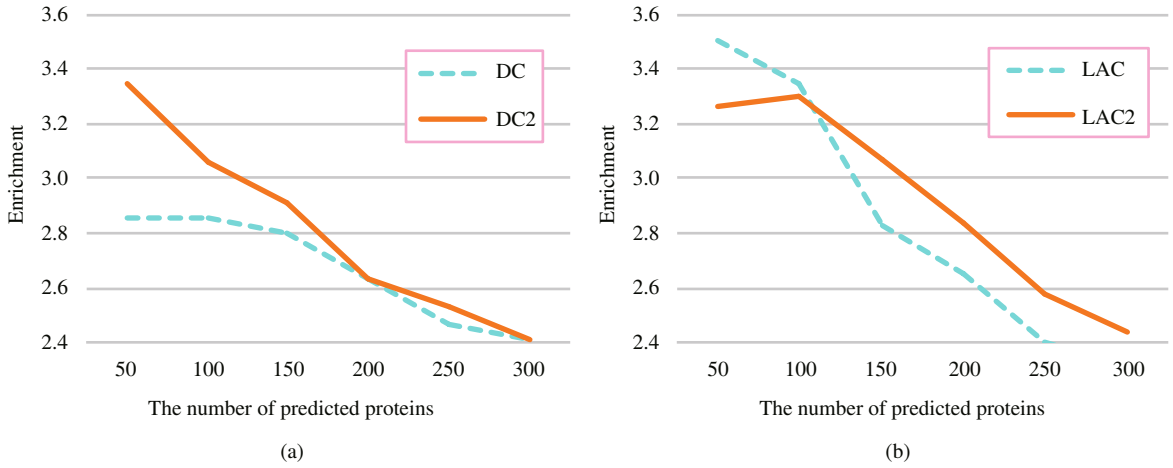

Figure 5 (Color online) The fold enrichment of essential proteins when using M and M2. (a) Compare of DC and DC2; (b) compare of LAC and LAC2.

whereas the proportions of proteins, interactions and essential proteins decrease gradually (Figure 4). Although the F-measure shows a positive correlation with the PCC threshold, we cannot use the PPI network with the PCC threshold equal to 0.99 to identify essential proteins, since the number of proteins is only 523, which is even less than the number of all known essential proteins (1136). Therefore, we select 0.95 as the best PCC threshold to construct the dynamic PPI networks.

3.5 Effect of topology in predicting essential proteins

We propose a two-step strategy to increase the prediction accuracy of initial topological indices in this study. Take DC and LAC, for example. We can see from Figure 5 that both DC2 and LAC2 are better than the initial DC and LAC, respectively. This indicates that adding the neighbors' connection is useful in increasing the prediction accuracy in terms of predicting essential proteins.

However, we suggest that the two-step method is generally a safe choice to derive a new index, and it is similar to the method of [33]. The prediction results of DC3 and LAC3 also support our conjecture. The results are shown in Figure 6. We rank the predicted proteins according to the value of the double set and triple set of DC and LAC from high to low by selecting the top-ranked predictions from 100 to 600, respectively. We find that the result for the triple set is not always better than for the double set. We guess the result is due to the small-world property of most biological networks, where the triple set of an index will involve the connectivity of the third-order neighbors of individual vertices, which would be associated with the connectivity of the majority of vertices in a network. An index associated with the third-order neighbors may not distinguish vertices in terms of their essentialities.

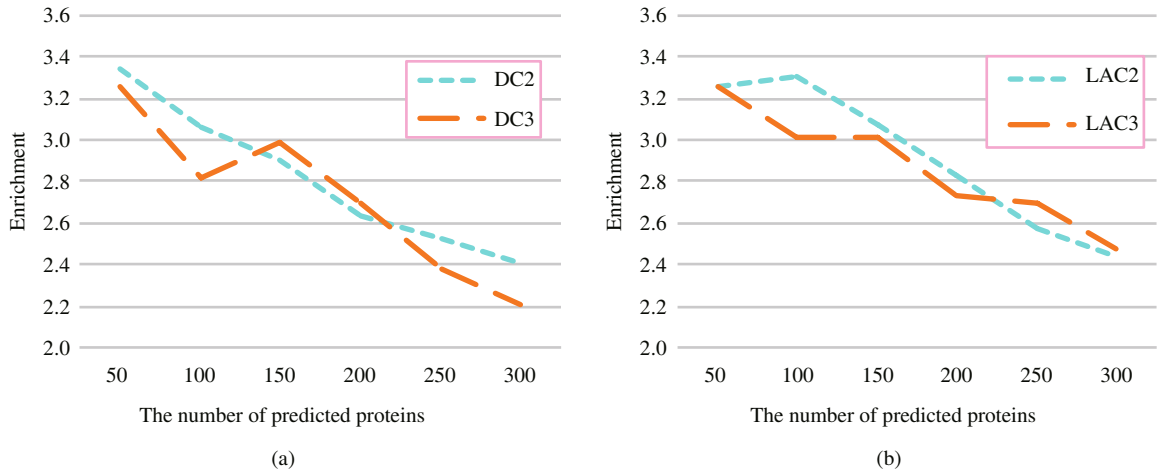

Figure 6 (Color online) The fold enrichment of essential proteins when using M2 and M3. (a) Compare of DC2 and DC3; (b) compare LAC2 and LAC3.

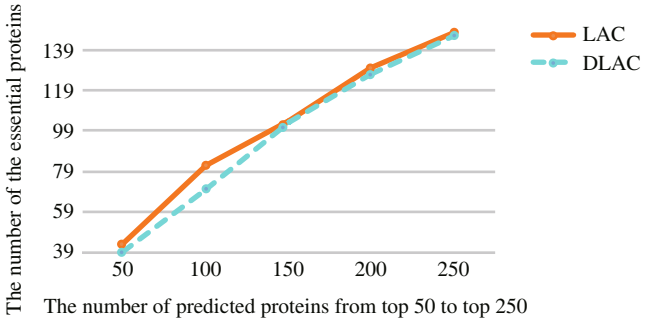

Figure 7 (Color online) The number of essential proteins when using LAC and DLAC.

3.6 The time-course dynamic network and the integrated dynamic network

Generally, a dynamic PPI network represents a set of time-course PPI networks, where the elements of vertices and/or edges change with time. However, in this study, we observe that using such a time-course dynamic PPI network does not perform well compared with our proposed integrated dynamic network in terms of the identification of essential proteins. The results are shown in Figure 7, where DLAC is calculated from a time-course dynamic PPI network constructed using the method of Tang [22], while LAC is calculated from our integrated dynamic PPI network.

Therefore, we argue that the essential proteins themselves may not be active at all times. They may be considered to have more interactions than others. Thus, the dynamic PPI network constructed by our method is relatively good in predicting essential proteins.

As well, we compare the performance of DLAC and DLAC2 in Figure 8 using the time-course dynamic network. In this figure, we can see that DLAC2 is better than DLAC in the majority of cases. This result also supports the generality of our proposed two-step strategy, which could be extended to many other topological indices and methods.

3.7 Verification of essential proteins from text mining

The list of known essential proteins is by no means of complete. Hence, we also evaluated our predictions from the public literature by using a text mining method. We predict YGR262c, YJL125c, and YBR128c as essential proteins, although they do not appear in the known essential protein list. Ref. [36] shows that disruption of the *Saccharomyces cerevisiae* YGR262c gene causes severely defective growth. Ref. [37] explains that disruption and basic functional analysis of five chromosome X novel open reading frames(ORFs) of *Saccharomyces cerevisiae* reveals YJL125c as an essential gene for vegetative growth.

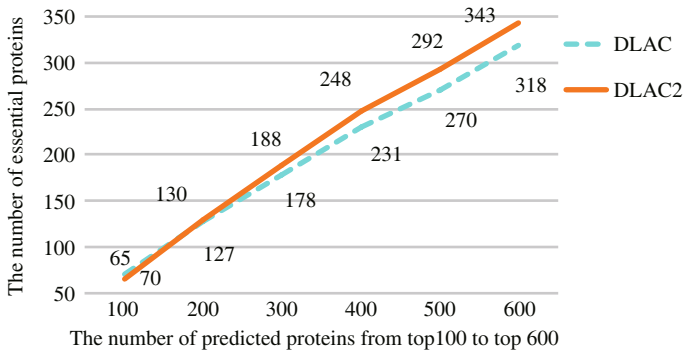

Figure 8 (Color online) The number of essential proteins when using DLAC and DLAC2.

Ref. [38] shows that disruption of YBR128c causes the infertility of the yeast. The support of these papers suggests that our proposed method is useful for predicting novel essential proteins.

4 Conclusion

In this paper, three types of improvement have been proposed to optimize the prediction of essential proteins: (1) the data source, (2) the construction of a dynamic PPI network, and (3) the novel two-step M2 measure for almost all known topological indices. Using the RNA-Seq dataset has proved its efficiency in constructing dynamic PPI networks, and the proposed two-step strategy for known topological indices has also been shown to increase the predictive accuracy. Moreover, the proposed strategy is quite flexible. It can be easily extended to many other topological indices and methods for identifying essential proteins.

Although the proposed strategies have achieved some good results, this study can also be extended in at least the following three ways. Firstly, the gene expression profiles obtained from RNA-Seq data can be further improved, since RNA-Seq data normally contain isoform information, which can be used to formulate more accurate gene expression profiles. Secondly, the RNA-Seq data we obtained has only four time points. It may not cover the gene expression profiles in detail. Therefore, if one can get RNA-Seq data with more time points, a more accurate dynamic PPI network can be constructed with it. Thirdly, for the two-step measure used in this study, we employ only the widely tested DC and LAC, as examples. These two measures consider the connective information of a node and its neighbors separately. But the combination of these two kinds of connection should be investigated further.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61272121, 61332014) and Fundamental Research Funds for the Central Universities (Grant Nos. 3102015JSJ0011, 3102015QD029).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Giaever G, Chu A M, Ni L, et al. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature, 2002, 418: 387–391
2 Cullen L M, Arndt G M. Genome-wide screening for gene function using RNAi in mammalian cells. Immun Cell Biol, 2005, 83: 217–223
3 Wang J X, Peng W, Wu F X. Computational approaches to predicting essential proteins: a survey. Proteom-Clin Appl, 2013, 7: 181–192
4 Gerdes S Y, Scholle M D, Campbell J W, et al. Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J Bacteriol, 2003, 185: 5673–5684
5 Batada N N, Hurst L D, Tyers M. Evolutionary and physiological importance of hub proteins. PLoS Comput Biol, 2006 2: e88
6 Hahn M W, Kern A D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol, 2005, 22: 803–806

7 Yu H, Greenbaum D, Lu H X, et al. Genomic analysis of essentiality within protein networks. Trends Genet, 2004, 20: 227–231
8 Estrada E. Virtual identification of essential proteins within the protein interaction network of yeast. Proteomics, 2006, 6: 35–40
9 Li M, Lu Y, Wang J X, et al. A topology potential-based method for identifying essential proteins from PPI networks. IEEE/ACM Trans Comput Biol Bioinform, 2015, 12: 372–383
10 Ren J, Wang J X, Li M, et al. Discovering essential proteins based on PPI network and protein complex. Int J Data Min Bioinform, 2015, 12: 24–43
11 Li M, Zheng R Q, Zhang H H, et al. Effective identification of essential proteins based on priori knowledge, network topology and gene expressions. Methods, 2014, 67: 325–333
12 Tang Y, Li M, Wang J X, et al. CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks. Biosystems, 2015, 127: 67–72
13 Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press, 1994
14 Freeman L C. Centrality in social networks conceptual clarification. Soc Netw, 1979, 1: 215–239
15 Zotenko E, Mestre J, O'leary D P, et al. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. PLoS Comput Biol, 2008, 4: e1000140
16 Jeong H, Mason S P, Barabási A L, et al. Lethality and centrality in protein networks. Nature, 2001, 411: 41–42
17 Bonacich P. Power and centrality: a family of measures. Amer J Sociol, 1987, 92: 1170–1182
18 Li M, Wang J X, Chen X, et al. A local average connectivity-based method for identifying essential proteins from the network level. Comput Biol Chem, 2011, 35: 143–150
19 Estrada E, Rodriguez-Velazquez J A. Subgraph centrality in complex networks. Phys Rev E, 2005, 71: 056103
20 Wang J X, Peng X Q, Peng W, et al. Dynamic protein interaction network construction and applications. Proteomics, 2014, 14: 338–352
21 Xiao Q H, Wang J X, Peng X Q, et al. Identifying essential proteins from active PPI networks constructed with dynamic gene expression. BMC Genomics, 2015, 16: S1
22 Tang X W, Wang J X, Liu B B, et al. A comparison of the functional modules identified from time course and static PPI network data. BMC Bioinform, 2011, 12: 339
23 Jin R M, Mccallen S, Liu C C, et al. Identifying dynamic network modules with temporal and spatial constraints. In: Proceedings of Pacific Symposium on Biocomputing, Big Island of Hawaii, 2009. 203–214
24 Luo J W, Kuang L. A new method for predicting essential proteins based on dynamic network topology and complex information. Computl Biol Chem, 2014, 52: 34–42
25 Chen B L, Fan W W, Liu J, et al. Identifying protein complexes and functional modules from static PPI networks to dynamic PPI networks. Brief Bioinform, 2014, 15: 177–194
26 Oh S, Song S, Grabowski G, et al. Time series expression analyses using RNA-Seq: a statistical approach. BioMed Res Int, 2013, 203681
27 Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol, 2005, 4: 17
28 Langmead B, Salzberg S L. Fast gapped-read alignment with Bowtie 2. Nat Methods, 2012, 9: 357–359
29 Ferragina P, Manzini G. Opportunistic data structures with applications. In: Proceedings of IEEE 41st Annual Symposium on Foundations of Computer Science, Redondo Beach, 2000. 390–398
30 Trapnell C, Pachter L, Salzberg S L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009, 25: 1105–1111
31 Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. Nat Protoc, 2012, 7: 562–578
32 Wang J X, Li M, Wang H, et al. Identification of essential proteins based on edge clustering coefficient. IEEE/ACM Trans Comput Biol Bioinform, 2012, 9: 1070–1080
33 Liu G M, Wong L, Chua H N. Complex discovery from weighted PPI networks. Bioinformatics, 2009, 25: 1891–1897
34 Lage K, Karlberg E O, Størling Z M, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol, 2007, 25: 309–316
35 Chen Y X, Wang W H, Zhou Y Y, et al. In silico gene prioritization by integrating multiple data sources. PLoS ONE, 2011, 6: e21137
36 Stocchetto S, Marin O, Carignani G, et al. Biochemical evidence that *Saccharomyces cerevisiae* YGR262c gene, required for normal growth, encodes a novel Ser/Thr-specific protein kinase. FEBS Lett, 1997, 414: 171–175
37 Jaquet L, Jauniaux J C. Disruption and basic functional analysis of five chromosome X novel ORFs of *Saccharomyces cerevisiae* reveals YJL125c as an essential gene for vegetative growth. Yeast, 1999, 15: 51–61
38 Huang M E, Cadieu E, Souciet J L, et al. Disruption of six novel yeast genes reveals three genes essential for vegetative growth and one required for growth at low temperature. Yeast, 1997, 13: 1181–1194