Tsinghua Science and Technology

Volume 20 | Issue 5 Article 7

2015

A Feature Selection Method for Prediction Essential Protein

Jiancheng Zhong
the School of Information Science and Engineering, Central South University, Changsha 410083, China.
the College of Polytechnic, Hunan Normal University, Changsha 410083, China.

Jianxin Wang
the School of Information Science and Engineering, Central South University, Changsha 410083, China.

Wei Peng
the Computer Center, Kunming University of Science and Technology, Kunming 650093, China.

Zhen Zhang
the School of Information Science and Engineering, Central South University, Changsha 410083, China.

Min Li
the School of Information Science and Engineering, Central South University, Changsha 410083, China.

Recommended Citation

Jiancheng Zhong, Jianxin Wang, Wei Peng et al. A Feature Selection Method for Prediction Essential Protein. Tsinghua Science and Technology 2015, 20(5): 491-499.

TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 07/10 pp491-499
Volume 20, Number 5, October 2015

A Feature Selection Method for Prediction Essential Protein

Jiancheng Zhong, Jianxin Wang*, Wei Peng, Zhen Zhang, and Min Li

Abstract: Essential proteins are vital to the survival of a cell. There are various features related to the essentiality of proteins, such as biological and topological features. Many computational methods have been developed to identify essential proteins by using these features. However, it is still a big challenge to design an effective method that is able to select suitable features and integrate them to predict essential proteins. In this work, we first collect 26 features, and use SVM-RFE to select some of them to create a feature space for predicting essential proteins, and then remove the features that share the biological meaning with other features in the feature space according to their Pearson Correlation Coefficients (PCC). The experiments are carried out on S. cerevisiae data. Six features are determined as the best subset of features. To assess the prediction performance of our method, we further compare it with some machine learning methods, such as SVM, Naive Bayes, Bayes Network, and NBTree when inputting the different number of features. The results show that those methods using the 6 features outperform that using other features, which confirms the effectiveness of our feature selection method for essential protein prediction.

Key words: essential protein; feature selection; Protein-Protein Interaction (PPI); machine learning; centrality algorithm

1 Introduction

Essential proteins exert vital functions on cellular processes and are indispensable for each organism. The organism cannot survive and reproduce without them[1,2]. Essential proteins are composed of a set of minimal genome, which can support the cell survival with basic requirements. Recently, identification of essential proteins and their functions has attracted many pharmaceutical researchers' attention, because the essential proteins of some bacterial are lethal to bacterial, which are candidate of drug-target[3].

Many experimental and computational methods have been designed to find and predict essential proteins. The experimental methods identify essential proteins through single gene knockouts[4], conditional knockouts[5], and RNA interference[6], which are very expensive and time-consuming. Furthermore, the experimental methods are not suitable for all organisms, e.g., human. Meanwhile, with the development of high throughput experimental technologies, a variety of genome-related data, such as Protein-Protein Interaction (PPI) data, cellular localization data, protein sequence data, and gene expressing data, are available. Many features of essential proteins have been discovered through analyzing the biological information. Therefore a large number of computational methods make use of these features to predict essential proteins. Generally, these computational methods can

• Jiancheng Zhong, Jianxing Wang, Zhen Zhang, and Min Li are with the School of Information Science and Engineering, Central South University, Changsha 410083, China. E-mail: jxwang@csu.edu.cn.

• Jiancheng Zhong is also with the College of Polytechnic, Hunan Normal University, Changsha 410083, China. E-mail: jczhong@csu.edu.cn.

• Wei Peng is with the Computer Center, Kunming University of Science and Technology, Kunming 650093, China. E-mail: weipeng@kmust.edu.cn.

∗ To whom correspondence should be addressed.

Manuscript received: 2015-05-21; accepted: 2015-08-06

be divided into two categories: unsupervised and supervised machine learning-based methods.

Unsupervised methods mainly adopt the topology-based features of essential proteins in PPI. These methods are based on centrality-lethality rule, which means essential proteins tend to be the hubs of the PPI network, and removing them causes the PPI network to break down. Many centrality methods have been proposed to predict essential proteins, such as Betweenness Centrality (BC)[7,8], Closeness Centrality (CC)[9], Degree Centrality (DC)[10], Eigenvector Centrality (EC)[11], Information Centrality (IC)[12], Edge Clustering Coefficient Centrality (NC)[13], and Subgraph Centrality (SC)[14]. However, these methods only take the topological features of proteins in the PPI network into consideration. With the advent of some useful genome-related information, many researchers combine the topological features with the biological-related information to predict essential proteins. For example, Koschützki et al.[15] proposed a network-motif-based centrality to predict essential proteins by using both the functional substructures and the network centrality. Both Li et al.[16] and Tang et al.[17] weighted PPI network by gene expression profiles and proposed a novel centrality to identify essential proteins. The methods mentioned above rely heavily on the accuracy of PPI. Although modern technologies, including Yeast two-Hybrid (Y2H)[18], tandem Affinity Purification (AP)[19], and Mass Spectrometry (MS)[20], can identify PPI datasets, it is still a big challenge to obtain exact PPI datasets, due to high false positive on experiments and unstable interactions between proteins[21], which limits the effectiveness of methods for predicting essential proteins.

In contrast to unsupervised methods, supervised methods use machine learning methods to combine topological features of proteins with their biological features. Those methods first train the classifiers with the known samples and then employ the classifiers to predict the unknown samples. Seringhaus et al.[22] identified 14 sequence features that are potentially associated with essentiality, such as localization signals, Codon Adaptation Index (CAI), GC content, and overall hydrophobicity. Gustafson et al.[23] combined a lot of features including the Open Reading Frame (ORF) length, PHYletic retention (PHY), paralogs, CAI, DC, etc., by using a Naive Bayes classifier to predict the essential proteins. Hwang et al.[24] integrated the ORF length, PHY, BC, CC, and DC by using SVM classifier to infer essential proteins. Recently, some researchers have used ensemble learning methods that create an ensemble classifier to predict essential proteins. Acencio and Lemke[25] bagged the decision trees to predict essential proteins with combination of topological features (BC, CC, DC, and so on) with biological features (cellular localization and biological processes information). Deng et al.[26] trained four classifiers (C4.5 decision tree, CN2 rule, Naive Bayes classifier, and logistical regression model) to calculate scores of essential genes, respectively, and then combine the four scores to get the final predictions. Kim[27] proposed a method to combine various machine learning techniques by using a CENT-ING-GO feature space which includes GO terms and various centrality measures.

Since various types of features have been proposed for detecting essential proteins, researchers try to find some powerful prediction features from a feature spaces. For example, del Rio et al.[28] constructed a feature space with 16 different centralities measured in 18 metabolic networks for identifying essential proteins. Their results show that the prediction performance is reliable when at least 2 centrality measures are selected, and there is no improvement when 3 or 4 centrality measures are selected. Therefore, it is a challenge to select a group of features that play a key role in predicting. Few methods can select features from the feature space automatically. Researchers[24,29] usually use the statistical methods such as Pearson Correlation Coefficients (PCC) to analyze the relationship between features, and then decide which features are selected to train classifiers.

In this paper, we collect 26 features to construct a feature space which consists of topological features, such as BC, CC, DC, EC, IC, NC, and SC, biological features, such as subcellular location, and other composed features, such as ION[30], PeC[16], and WDC[17]. Then we predict essential proteins by using SVM-RFE and PCC methods using the feature spaces, which can automatically select a subset of feature from feature space with powerful prediction ability and minimal biological meaning overlap.

2 Method

In this section, we first collect 26 features to construct a feature space for predicting essential proteins, and then select suitable subsets of features from the feature space by using SVM-RFE and Pearson correlation coefficients methods.

2.1 Feature spaces construction

In order to construct the feature space for identifying essential proteins, we utilize the results obtained from previous methods of predicting essential proteins. Twenty six features are introduced as follows.

In the first step, we analyze the topological features of proteins in PPI network including BC, CC, DC, EC, IC, NC, and SC.

F1: BC[7,8].

The BC of vertex k means the relative stress centrality that can quantify the extend to which vertex k monitors the communication between other vertexes. It can be defined as the following equation.

$$\delta_{uv}(k) = \frac{p(u,k,v)}{p(u,v)}, \quad u \neq k \neq v,$$

$$\mathrm{BC}(k) = \sum_{u \in V}\sum_{v \in V} \delta_{uv}(k),$$

where $\delta_{uv}(k)$ denotes the fraction of the shortest paths that pass though the vertex k.

F2: CC[9].

The CC of a vertex u is defined as the reciprocal of the total distance between it and other vertexes in graph G. N is the number of vertices in V. It can be defined as the following equation.

$$\mathrm{CC}(u) = \frac{N-1}{\sum\limits_{v \in V} \mathrm{dis}(u,v)},$$

where $\mathrm{dis}(u,v)$ is the distance between u and v, v denotes other vertexes in the graph G.

F3: DC[10].

The DC is the most simple centrality which is defined as degree of the vertex v.

$$\mathrm{DC}(v) = \sum_{u} \mathrm{edge}(u,v).$$

F4: EC[11].

The EC assumes that the centrality value of vertex depends on the values of each adjacent vertex, which is defined as the following equation.

$$\mathrm{EC}(u) = e_{\max}(u),$$

where $\boldsymbol{e}_{\max}$ denotes the principal eigenvector of the adjacency matrix $\boldsymbol{A}$, $e_{\max}(u)$ denotes the u-th component of the principal eigenvector.

F5: IC[12].

The IC of a vertex u is defined as harmonic mean length of paths ending at the vertex u. It is calculated by using following equations.

$$I_{uv} = (C_{uu} + C_{vv} - C_{uv})^{-1} \ , \ \boldsymbol{C} = (\boldsymbol{D} - \boldsymbol{A} + \boldsymbol{J})^{-1},$$

$$\mathrm{IC}(u) = \left[\frac{1}{N}\sum_{v}\frac{1}{I_{uv}}\right]^{-1},$$

where $\boldsymbol{I}$ is the information matrix, $\boldsymbol{D}$ is the diagonal matrix whose values are the degrees of each vertex, $\boldsymbol{A}$ is the adjacency matrix of a network, and $\boldsymbol{J}$ is a matrix with all of its elements equal to one.

F6: NC[13].

The NC of a vertex u is calculated by the sum of the edge-clustering coefficients. It is defined as following equations.

$$\mathrm{ECC}(u,v) = \frac{z_{u,v}}{\min\{\mathrm{DC}(u)-1, \mathrm{DC}(v)-1\}},$$

$$\mathrm{NC}(u) = \sum_{v} \mathrm{ECC}(u,v),$$

where $\mathrm{ECC}(u,v)$ is the edge-clustering coefficient of $\mathrm{edge}(u,v)$. $z_{u,v}$ is the number of triangles containing $\mathrm{edge}(u,v)$ in network. The $\mathrm{DC}(u){-}1$ means the maximal number of triangles that can include vertex u.

F7: SC[14].

The SC of a vertex u accounts for the number of subgraghs in which vertex u takes part in and gives more weight to smaller subgraphs. It is defined as the following equation.

$$\mathrm{SC}(u) = \sum_{l=0}^{\infty}\frac{u_l(u)}{l!} = \sum_{v=1}^{N}[a_v(u)]^2 \mathrm{e}^{\lambda_v},$$

where $u_l(u)$ denotes the number of closed loops of length l which starts and ends at vertex u. The $(\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_N)$ is an orthonormal basis of $\mathbf{R}^N$ which consists of eigenvectors of adjacency matrix $\boldsymbol{A}$. The $\lambda_1, \lambda_2, \cdots, \lambda_N$ are eigenvalue of $\boldsymbol{A}$ and $a_v(u)$ is the u-th component of $\boldsymbol{a}_v$.

In the second step, we adopt some composited features which consider both the topological properties and biological information. Those features are proposed by methods including PeC, WDC, and ION.

F8: PeC[16].

PeC integrates gene expression profiles and PPI data to predict the essential proteins. The PeC of vertex u is defined as the sum of the ECC score of the edge connecting vertex u multiplying the corresponding PCC in terms of expression data. It can be calculated by the following equation.

$$\mathrm{PeC}(u) = \sum_{v \in N_u} \mathrm{ECC}(u,v) \times \mathrm{PCC}(u,v).$$

F9: WDC[17].

Like PeC, WDC also simultaneously considers the gene expression profiles and PPI data to infer the proteins essentiality. The WDC can be calculated by the following equation.

$$\mathrm{WDC}(u) = \sum_{v \in N_u} [(\mathrm{ECC}(u,v) \times \lambda) + (\mathrm{PCC}(u,v) \times (1-\lambda))],$$

where λ is a constant. In this paper, we assign 0.5 to λ.

F10: ION[30].

The ION integrates the orthology with PPI networks which considers both the connections between proteins and the features of their neighbors. ION calculates the score of proteins that are the linear combination of the neighbor-induced scores and orthologous property scores $(d(u))$. $\mathrm{Ne}(u)$ denotes the neighbor of u. It can be calculated by the following equation.

$$h(u,v) = \begin{cases} \mathrm{Norm}_i(\mathrm{ECC}(u,v)) = \dfrac{\mathrm{ECC}(u,v)}{\sum\limits_{w \in \mathrm{Ne}(u)} \mathrm{ECC}(u,w)}, \\ \qquad\qquad \text{if } \sum\limits_{w \in \mathrm{Ne}(u)} \mathrm{ECC}(u,w) > 0; \\ 0, \qquad\qquad \text{otherwise.} \end{cases}$$

$$\mathrm{ION}(u) = (1-a)d(u) + a \sum_{v \in \mathrm{Ne}(u)} h(u,v)\mathrm{ION}(v).$$

In the third step, we select the subcellular localization of proteins as features to construct feature space. Some researches[25] have pointed that the subcellular localization associates with gene essentiality, because localization of proteins in cellular is usually related to the proteins' functions. For example, most essential biological processes, such as DNA replication and mRNA synthesis, often take place in nuclear. In this work, 16 different localizations including Cell wall, Cytoskeleton, Cytoplasm, Endoplasmic reticulum, Endosome, Extracellular, Golgi, Lysosome, Membrane, Mitochondrion, Nucleus, Peroxisome, Secretory pathway, Vacuole, Vesicles, and Transmembrane are considered as features of protein essentiality, which correspond to F11 to F26.

Since each feature from F1 to F10 has its own value ranges, we standardize all features of numeric attributes to have zero mean and unit variance.

2.2 Feature selection

In this section, we select suitable features from the 26 features mentioned above for predicting essential proteins. The goal of feature selection is to find the suitable features that both have powerful prediction ability for protein essentiality and share minimal biological meaning between each other. To achieve the goal, we first adopt Support Vector Machines-Recursive Feature Elimination (SVM-RFE) to output a ranked list of the features. And then we use the Pearson correlation coefficients to determine the relationship between these features and rearrange the ranked list of the features. Figure 1 illustrates the workflow of SVM-RFE with PCC.

The SVM-RFE algorithm is proposed by Guyon et al.[31] which adopts a backward feature elimination

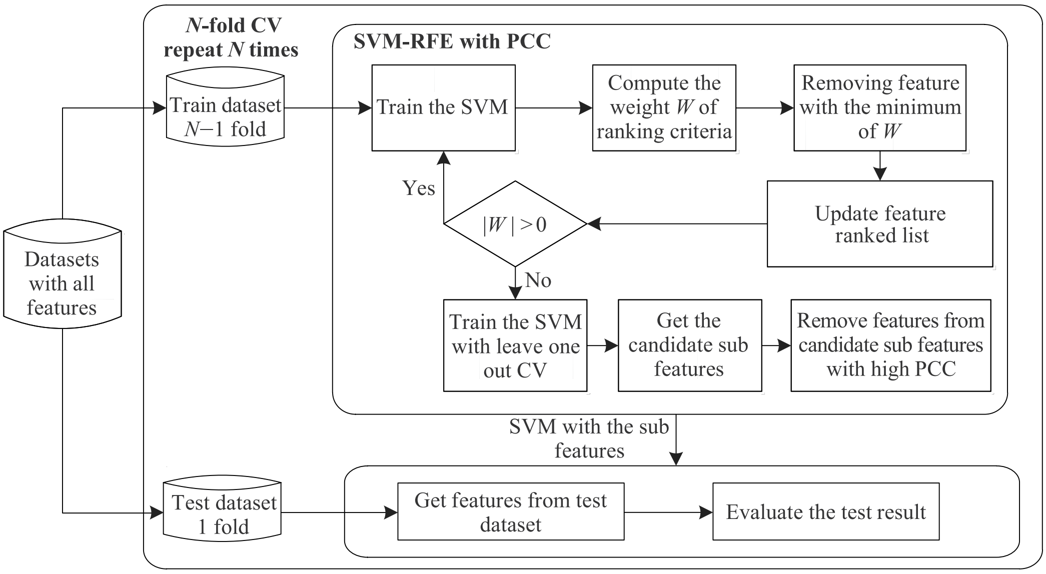

N-fold CV
repeat *N* times
SVM-RFE with PCC
Datasets with all features
Train dataset *N*−1 fold
Train the SVM
Compute the weight *W* of ranking criteria
Removing feature with the minimum of *W*
Update feature ranked list
|*W*| > 0
Yes
No
Train the SVM with leave one out CV
Get the candidate sub features
Remove features from candidate sub features with high PCC
SVM with the sub features
Test dataset 1 fold
Get features from test dataset
Evaluate the test result

Fig. 1 Proposed secure systolic Montgomery modular multiplier architecture.

strategy. It constructs sorting coefficient by weight vectors $\boldsymbol{W}$ generated by SVM, and then removes iteratively a feature with the smallest coefficient. The SVM-RFE gets the sorted list in descending order of all the features. The $\boldsymbol{W}$ is calculated by the following equation:

$$W(i)=\frac{1}{2}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{a}-\frac{1}{2}\boldsymbol{a}^{\mathrm{T}}Q(-i)\boldsymbol{a},$$

where $\boldsymbol{a}$ is the Lagrange multipliers vector, $\boldsymbol{Q}$ is a matrix defined as $Q_{ij}=K(x_i,x_j)$, K is a linear kernel function. Although SVM-RFE can output a ranked list of the features, it does not determine which one is the suitable subset features. In order to get a suitable subset, we use the 10-fold cross validation as the resampling method. Besides, SVM-RFE does not consider the redundancy of the features. To get features minimal biological meaning, we use PCC to evaluate the subset features and then remove the features that have high correlation coefficient and are ranked in front of the list. The PCC of two features is calculated by the following equation:

$$\mathrm{PCC}(f_i,f_j)=\frac{\mathrm{Cov}(f_i,f_j)}{\sqrt{\mathrm{Var}(f_i)\mathrm{Var}(f_j)}},$$

where $\mathrm{Cov}(f_i,f_j)$ denotes the covariance of f_i and f_j, $\mathrm{Var}(f_i)$ and $\mathrm{Var}(f_j)$ denote the variance of f_i and f_j, respectively. Algorithm 1 is the features selection algorithm with computational complexity of $O(mn^2)$.

Algorithm 1 Selection features

```
Input: The sample of protein with 26 features $X$, and its corresponding label $Y$.
Output: The ranked list of the features.
(1) Initialize the training dataset $X$;
(2) Set rankedFeatureList= $\varnothing$ ;
(3) $i = 1$;
(4) do
      (a) Train the support vector machine with features;
      (b) Calculate the weight vector $\boldsymbol{W}$;
      (c) $f$ = the feature with smallest ranking value of $\boldsymbol{W}$;
      (d) rankedFeatureList[$i++$] $= f$;
      (e) features = features $- f$ ;
    While |features| $> 0$
(5) Reverse the rankedFeatureList;
(6) Samplesize = the size of sample;
(7) FeatureCount= |rankedFeatures| ;
(8) SelectFeatureCount = 1;
(9) For $i$ in [1.. FeatureCount]
      (a) SelectFeatures = $\varnothing$ ;
      (b) SelectFeatureCount = SelectFeatureCount$*$2;
      (c) If (SelectFeatureCount > FeatureCount) break;
      (d) SelectFeatures = SelectFeatures $\cup$
          rankedFeatures;
      (e) Calculate the accuracy vectors value by using SVM
          with SelectFeatures employed leave one out CV;
    Endfor
(10) Get the best $n$ ranked features with accuracy vectors
(11) Calculate the Pearson correlation coefficients for all
     features
(12) For $i$ in [1..$n-1$]
      (a) For $j$ in [$i$+1..$n$]
          If PCC(rankedFeatures[$i$], rankedFeatures[$j$])>
          threshold then
            i. rearrange the rankedFeatures by removing the
               feature[$j$] after feature[$n$].
            ii. $n = n - 1$;
          Endif
          Endfor
    Endfor
(13) Return the ranked list of the features;
```

3 Results and Discussion

In this section, we describe the datasets for experiments including S. cerevisiae (Bakers Yeast) datasets and its corresponding PPI datasets. Then we create the feature space of essential proteins and select the best feature subsets. Finally, we compare the performance by applying various machine learning methods on different feature subsets.

3.1 Datasets

We carried out the experiments on Bakers Yeast dataset, which is published on Oct. 10, 2010 and is available from DIP database, http://dip.doe-mbi.ucla.edu/dip/Download.cgi?SM=7[32]. The reason for selecting yeast dataset is that both its PPI and its essentialty information are complete and reliable among various species. Since some features do not use the self-interactions and repeated interactions, we filter out those interactions in the PPI network. After that, we obtain the PPI network including 5093 proteins and 24 743 interactions. To get the dataset of essential proteins, we integrate four essential protein databases: MIPS[33], SGD[34], DEG[35], and SGDP[36], which include 1285 distinct essential proteins. 1167 of them can map to the yeast PPI network. In this paper, the 1167 proteins serve as the golden dataset of essential proteins.

In order to calculate some features like ION, WDC, etc., we obtain other biological information on proteins, such as orthologous proteins, gene expression, and subcellular information. The orthologous information of proteins can be downloaded from Version 7 of the InParanoid database[37]. The gene microarray datasets can be downloaded from NCBI Gene Expression Omnibus website (http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE3431)[38], including 9335 probes at 36 different time points. The data corresponds to 6777 gene products. Among them, 4858 proteins are found in the yeast PPI network. The 16 different subcellular localizations information can be downloaded from eSLDB database[39].

3.2 The best feature subsets

In this paper, SVM-RFE and PCC are employed to select suitable features for predicting essential proteins from the feature space that integrates both topological and biological features. The topological features are calculated by some centrality methods including BC, CC, DC, EC, IC, NC, and SC, corresponding to the F1:F7 in the feature space. The biological features include both the composite features, such as PeC, WDC, and ION, and the subcellular localization information. The composite features correspond to the F8:F10 in the feature space. The subcellular localization features correspond to the F11:F26 in the feature space. There are 26 features in the feature space. First, we run the SVM-RFE with 10-fold Cross Validation (CV) to rank the feature list, listed in Table 1.

As can be seen from Table 1, ION ranked in top is a very predictive feature, which relates to the evolutionary conservation, and the essential proteins are often conserved[30]. Both WDC and PeC use the gene expression profile to determine co-expression and co-clustering of essential proteins, which achieve the good performance of prediction. Acencio ands Lemke[25] found that the proteins located in nucleus tend to perform indispensable functions, which means that the feature of nucleus is effective in predicting essentiality. Given the feature ranking list, we adopt the 10-fold CV to decide the suitable subsets of features. The Area Under the Curve (AUC) of ROC is computed for each feature subset by using SVM. Table 2 shows all features and those ranked in top 4,8,16 and their corresponding AUC of ROC.

Table 1 Essential proteins feature rankings for Yeast datasets.

Rank No.	Feature name	Rank No.	Feature name
1	ION	14	Endoplasmic reticulum
2	WDC	15	BC
3	Nucleus	16	Mitochondrion
4	PeC	17	Membrane
5	DC	18	Transmembrane
6	NC	19	Secretory pathway
7	Cytoplasm	20	Cell wall
8	IC	21	Cytoskeleton
9	Vacuole	22	CC
10	EC	23	Vesicles
11	Endosome	24	Golgi
12	SC	25	Extracellular
13	Peroxisome	26	Lysosome

The highest AUC value 0.609 is obtained under the feature subset of size 8. We use the PCC to evaluate the correlations between the top 8 features, which is shown in Table 3.

In Table 3, the values of PCC between WDC and PeC, WDC and NC features are greater than 0.8, which indicate that those features are closely related to each other. We analyse the methods including WDC, PeC, and NC, those methods are all based on the edge-clustering coefficient, and PeC and WDC use the similar biological information, such as gene expression information. We remove the PeC and NC from the feature space, and then run SVM with 10-fold CV with the rest of features. The AUC of ROC for SVM with six features has the highest value 0.610.

We compare the performance of a list of machine learning methods when using 6 features, 8 features or all features. Some measures, such as True Positive rate (TP rate), False Positive rate (FP rate), precision, F-Measure, Matthews Correlation Coefficient (MCC), ROC area, PRC area, were used to evaluate the prediction performance. The results are shown in Table 4. The feature space with 6 features performs better than other feature spaces in SVM, Naive Bayes, Bayes

Table 2 The AUC of ROC by using SVM with different feature subsets.

Number of features	Features name	AUC of ROC
4	ION,WDC,Nucleus,PeC	0.608
8	ION, WDC, Nucleus, PeC, DC, NC, Cytoplasm, IC	0.609
16	ION, WDC, Nucleus, PeC, DC, NC, Cytoplasm, IC, Vacuole, EC, Endosome, SC, Peroxisome, Endoplasmic reticulum, BC, Mitochondrion	0.607
26	ALL features	0.577

Table 3 The correlation of top 8 features.

	ION	WDC	Nucleus	PeC	DC	NC	Cytoplasm	IC
ION	1							
WDC	0.425 232	1						
Nucleus	0.240 546	0.195 593	1					
PeC	0.349 71	0.801 061	0.156 85	1				
DC	0.388 783	0.579 09	0.119 282	0.348 574	1			
NC	0.440 695	0.810 345	0.209 794	0.555 108	0.724 148	1		
Cytoplasm	0.160 436	0.052 004	0.202 167	0.034 506	0.069 622	0.063 784	1	
IC	0.581 611	0.464 07	0.229 772	0.293 526	0.616 673	0.544 119	0.092 517	1

Table 4 Comparison of different methods when using 6 features, 8 features or all features.

Method	Number of features	TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC area	PRC area
SVM	6	0.805	0.586	0.791	0.805	0.767	0.341	0.61	0.706
	8	0.805	0.587	0.791	0.805	0.766	0.34	0.609	0.706
	All	0.801	0.646	0.808	0.801	0.745	0.313	0.577	0.691
Naive Bayes	6	0.79	0.477	0.773	0.79	0.778	0.352	0.748	0.795
	8	0.79	0.52	0.768	0.79	0.771	0.328	0.745	0.796
	ALL	0.782	0.526	0.758	0.782	0.764	0.304	0.744	0.792
Bayes Network	6	0.76	0.405	0.769	0.76	0.764	0.344	0.75	0.812
	8	0.755	0.397	0.768	0.755	0.76	0.341	0.747	0.809
	ALL	0.71	0.394	0.75	0.71	0.725	0.285	0.73	0.797
NBTree	6	0.811	0.509	0.793	0.811	0.789	0.387	0.755	0.81
	8	0.806	0.533	0.787	0.806	0.78	0.364	0.755	0.806
	ALL	0.806	0.534	0.787	0.806	0.78	0.363	0.755	0.809

Network, and NBTree.

4 Conclusions

Although different feature-based methods for predicting essential proteins have been proposed, integrating different features and selecting suitable features to predict the essential proteins is still a big challenge. In this work, we adopt SVM-RFE and PCC-based method to select suitable features related to essentiality of proteins from their topological and biological features. We conduct experiments on S. cerevisiae data. (1) The features are ranked by SVM-RFE. The features ranked in top 8 are the suitable subsets of features because they achieve the highest AUC of ROC in 10-fold CV when inputting different number of features. (2) In terms of the correlation between the 8 features, a new feature space is constructed by removing the features PeC and NC from the top 8 features because they are highly related to the WDC. (3) We evaluate the impact of different feature spaces on the prediction performance, such as 6 features, 8 features, and all features, using different machine learning methods. The results show that the feature space with 6 features performs better than other feature space. Thus, our method is able to select the features that are not only powerful to predict essential proteins but also share minimal biological meaning with each other.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 61232001, 61502166, 61502214, 61379108, and 61370024) and Scientific Research Fund of Hunan Provincial Education Department (Nos. 15CY007 and 10A076).

References

[1] R. S. Kamath, A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, and M. Sohrmann, Systematic functional analysis of the caenorhabditis elegans genome using rnai, *Nature*, vol. 421, no. 6920, pp. 231–237, 2003.

[2] J. Wang, X. Peng, W. Peng, and F. Wu, Dynamic protein interaction network construction and applications, *Proteomics*, vol. 14, nos. 4&5, pp. 338–352, 2014.

[3] N. Judson and J. J. Mekalanos, Tnaraout, a transposon-

based approach to identify and characterize essential bacterial genes, *Nature Biotechnology*, vol. 18, no. 7, pp. 740–745, 2000.

[4] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Vronneau, S. Dow, A. Lucau-Danila, K. Anderson, and B. Andr, Functional profiling of the saccharomyces cerevisiae genome, *Nature*, vol. 418, no. 6896, pp. 387–391, 2002.

[5] T. Roemer, B. Jiang, J. Davison, T. Ketela, K. Veillette, A. Breton, F. Tandia, A. Linteau, S. Sillaots, and C. Marta, Large-scale essential gene identification in candida albicans and applications to antifungal drug discovery, *Molecular Microbiology*, vol. 50, no. 1, pp. 167–181, 2003.

[6] L. M. Cullen and G. M. Arndt, Genome-wide screening for gene function using rnai in mammalian cells, *Immunology and Cell Biology*, vol. 83, no. 3, pp. 217–223, 2005.

[7] L. C. Freeman, A set of measures of centrality based on betweenness, *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[8] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, High-betweenness proteins in the yeast protein interaction network, *BioMed Research International*, vol. 2005, no. 2, pp. 96–103, 2005.

[9] S. Wuchty and P. F. Stadler, Centers of complex networks, *Journal of Theoretical Biology*, vol. 223, no. 1, pp. 45–53, 2003.

[10] R. R. Vallabhajosyula, D. Chakravarti, S. Lutfeali, A. Ray, and A. Raval, Identifying hubs in protein interaction networks, *PLoS One*, vol. 4, no. 4, p. e5344, 2009.

[11] P. Bonacich, Power and centrality: A family of measures, *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[12] K. Stephenson and M. Zelen, Rethinking centrality: Methods and examples, *Social Networks*, vol. 11, no. 1, pp. 1–37, 1989.

[13] J. Wang, M. Li, H. Wang, and Y. Pan, Identification of essential proteins based on edge clustering coefficient, *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 9, no. 4, pp. 1070–1080, 2012.

[14] E. Estrada and J. A. Rodriguez-Velazquez, Subgraph centrality in complex networks, *Physical Review E*, vol. 71, no. 5, p. 056103, 2005.

[15] D. Koschützki, H. Schwöbbermeyer, and F. Schreiber, Ranking of network elements based on functional substructures, *Journal of Theoretical Biology*, vol. 248, no. 3, pp. 471–479, 2007.

[16] M. Li, H. Zhang, J.-X. Wang, and Y. Pan, A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data, *BMC Systems Biology*, vol. 6, no. 1, p. 15, 2012.

[17] X. Tang, J. Wang, J. Zhong, and Y. Pan, Predicting essential proteins based on weighted degree centrality, *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 11, no. 2, pp. 407–418, 2014.

[18] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proceedings of the National Academy of Sciences*, vol. 98, no. 8, pp. 4569–4574, 2001.

[19] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Sèraphin, The tandem affinity purification (tap) method: A general procedure of protein complex purification, *Methods*, vol. 24, no. 3, pp. 218–229, 2001.

[20] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, and K. Boutilier, Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry, *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[21] E. Sprinzak, S. Sattath, and H. Margalit, How reliable are experimental proteincprotein interaction data? *Journal of Molecular Biology*, vol. 327, no. 5, pp. 919–923, 2003.

[22] M. Seringhaus, A. Paccanaro, A. Borneman, M. Snyder, and M. Gerstein, Predicting essential genes in fungal genomes, *Genome Research*, vol. 16, no. 9, pp. 1126–1135, 2006.

[23] A. M. Gustafson, E. S. Snitkin, S. C. Parker, C. DeLisi, and S. Kasif, Towards the identification of essential genes using targeted genome sequencing and comparative analysis, *Bmc Genomics*, vol. 7, no. 1, p. 265, 2006.

[24] Y.-C. Hwang, C.-C. Lin, J.-Y. Chang, H. Mori, H.-F. Juan, and H.-C. Huang, Predicting essential genes based on network and sequence analysis, *Molecular BioSystems*, vol. 5, no. 12, pp. 1672–1678, 2009.

[25] M. L. Acencio and N. Lemke, Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information, *BMC Bioinformatics*, vol. 10, no. 1, p. 290, 2009.

[26] J. Deng, L. Deng, S. Su, M. Zhang, X. Lin, L. Wei, A. A. Minai, D. J. Hassett, and L. J. Lu, Investigating the predictability of essential genes across distantly related organisms using an integrative approach, *Nucleic Acids Research*, vol. 39, no. 3, pp. 795–807, 2011.

[27] W. Kim, Prediction of essential proteins using topological properties in go-pruned ppi network based on machine learning methods, *Tsinghua Science and Technology*, vol. 17, no. 6, pp. 645–658, 2012.

[28] G. del Rio, D. Koschtzki, and G. Coello, How to identify essential genes from molecular networks? *BMC Systems Biology*, vol. 3, no. 1, p. 102, 2009.

[29] K. Plaimas, R. Eils, and R. König, Identifying essential genes in bacterial metabolic networks with machine learning methods, *BMC Systems Biology*, vol. 4, no. 1, p. 56, 2010.

[30] W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, and Y. Pan, Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks, *BMC Systems Biology*, vol. 6, no. 1, p. 87, 2012.

[31] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Gene selection for cancer classification using support vector machines, *Machine Learning*, vol. 46, nos. 1–3, pp. 389–422, 2002.

[32] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, Dip, the database of interacting proteins: A research tool for studying cellular networks of protein interactions, *Nucleic Acids Research*, vol. 30, no. 1, pp. 303–305, 2002.

[33] H.-W. Mewes, D. Frishman, K. F. Mayer, M. Münsterkötter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, and V. Stümpflen, Mips: Analysis and annotation of proteins from whole genomes in 2005, *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D169–D172, 2006.

[34] J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, et al., Sgd: Saccharomyces genome database, *Nucleic Acids Research*, vol. 26, no. 1, pp. 73–79, 1998.

[35] R. Zhang and Y. Lin, Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes, *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D455–D458, 2009.

[36] M. Antoniotti, G. D. Bader, G. Caravagna, S. Crippa, A. Graudenzi, and G. Mauri, Gestodifferent: A cytoscape plugin for the generation and the identification of gene regulatory networks describing a stochastic cell differentiation process, *Bioinformatics*, vol. 29, no. 4, pp. 513–514, 2013.

[37] G. Östlund, T. Schmitt, K. Forslund, T. Köstler, D. N. Messina, S. Roopra, O. Frings, and E. L. Sonnhammer, Inparanoid 7: New algorithms and tools for eukaryotic orthology analysis, *Nucleic Acids Research*, vol. 38, no. suppl 1, pp. D196–D203, 2010.

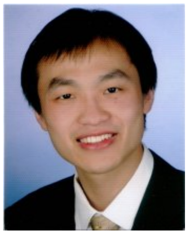
[38] B. P. Tu, A. Kudlicki, M. Rowicka, and S. L. McKnight, Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes, *Science*, vol. 310, no. 5751, pp. 1152–1158, 2005.

[39] A. Pierleoni, P. L. Martelli, P. Fariselli, and R. Casadio, esldb: Eukaryotic subcellular localization database, *Nucleic Acids Research*, vol. 35, no. suppl 1, pp. D208–D212, 2007.

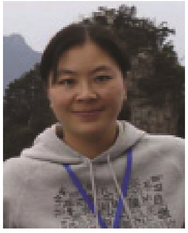
Jiancheng Zhong received the bachelor and master degrees in computer science from Hunan Normal University, China, in 2004 and 2007, respectively. Currently, he is a PhD student in School of Information Science and Engineering, Central South University, China. His current research interests include bioinformatics, data mining, and machine learning.

Jianxing Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in 2001. He is the chair of and a professor in Department of Computer Science, Central South University, China. His current research interests include algorithm analysis and optimization, parameraized algorithm, bioinformatics and computer network. He is a member of the IEEE.

Wei Peng received the PhD degree in computer science from Central South University, China, in 2013. Currently, she is an associate professor of Kunming University of Science and Technology, China. Her current research interests include molecular systems biology, biological system identification, and data mining.



Zhen Zhang received his BS degree in environmental engineering from Hunan University in 1996 and his MS degree in computer science from Technical University of Munich in 2007. He is currently a PhD candidate in bioinformatics at Central South University. His research interests include structural variant discovery and genome assembly.



Min Li received the PhD degree in computer science from Central South University, China, in 2008. She is currently an associate professor at the School of Information Science and Engineering, Central South University, China. Her main research interests include bioinformatics and systems biology.