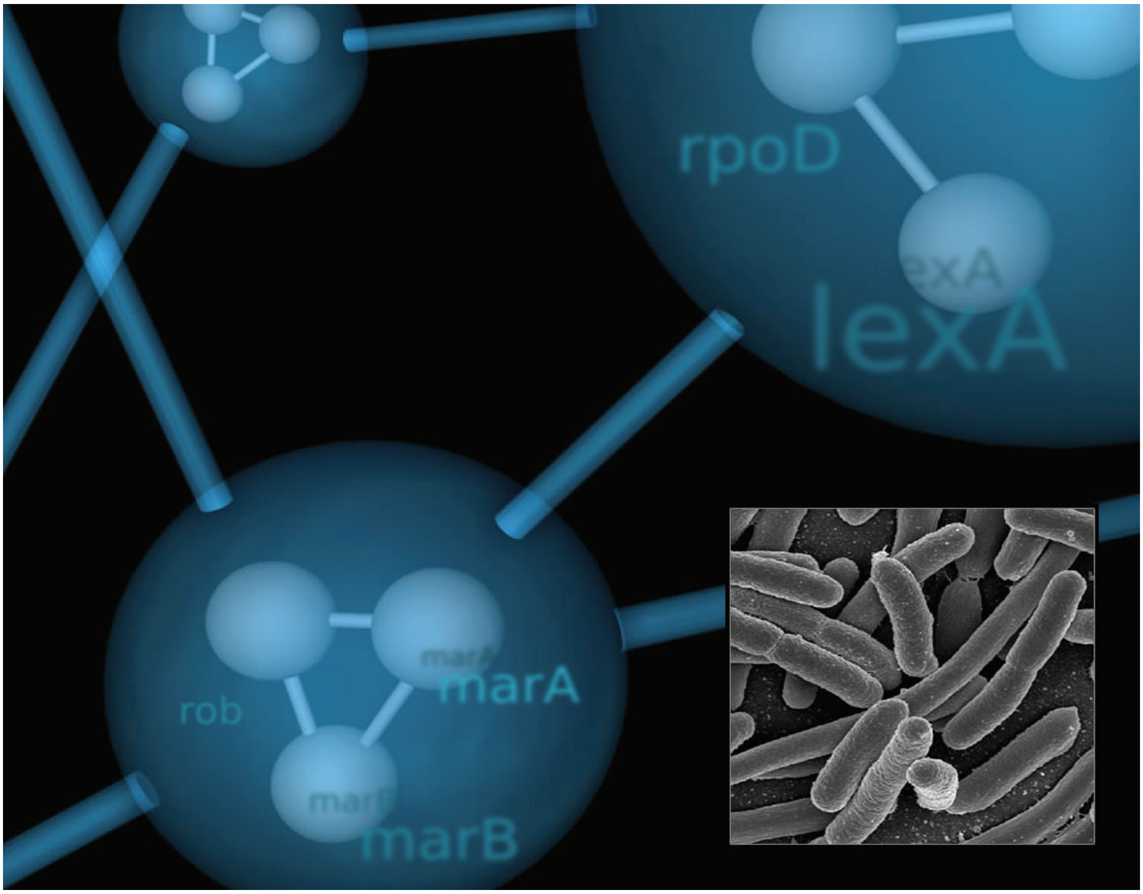
Molecular BioSystems

This article was published as part of the

Computational and Systems Biology themed issue

Please take a look at the full table of contents to access the other papers in this issue.



Predicting essential genes based on network and sequence analysis†‡

Yih-Chii Hwang,[a] Chen-Ching Lin,[ab] Jen-Yun Chang,[a] Hirotada Mori,[cd] Hsueh-Fen Juan*[b] and Hsuan-Cheng Huang*[a]

Received 13th January 2009, Accepted 23rd March 2009
First published as an Advance Article on the web 6th May 2009
DOI: 10.1039/b900611g

Essential genes are indispensable to the viability of an organism. Identification and analysis of essential genes is key to understanding the systems level organization of living cells. On the other hand, the ability to predict these genes in pathogens is of great importance for directed drug development. Global analysis of protein interaction networks provides an effective way to elucidate the relationships between genes. It has been found that essential genes tend to be highly connected and generally have more interactions than nonessential ones. With recent large-scale identifications of essential genes and protein–protein interactions in *Saccharomyces cerevisiae* and *Escherichia coli*, we have systematically investigated the topological properties of essential and nonessential genes in the protein–protein interaction networks. Essential genes tend to play topologically more important roles in protein interaction networks. Many topological features were found to be statistically discriminative between essential and nonessential genes. In addition, we have also examined sequence properties such as open reading frame length, strand, and phyletic retention for their association with the gene essentiality. Employing the topological features in the protein interaction network and the sequence properties, we have built a machine learning classifier capable of predicting essential genes. Computational prediction of essential genes circumvents expensive and difficult experimental screens and will help antimicrobial drug development.

Introduction

Essential genes are the genes that are required for sustaining cellular life. Experimental identification of essential genes is usually applied by the gene-deletion technique.[1,2] Since not all functions of the deleted genes can be compensated for by their existing counterparts, gene deletion may result in the demise of the organism. Those that are directly involved in organism's lethality are called essential genes and the protein products of essential genes are called essential proteins. Several earlier single-gene-deletion experiments have identified certain essential genes in *Saccharomyces cerevisiae*[3] and *Escherichia coli*.[4] Elucidating their essentiality is a key to understand the system-level organization of living cells. On the other hand, since the absence of essential genes confers lethality, they can be considered as potential drug targets for pathogens. The ability to rapidly identify essential genes has been described to be the most important task of genomics-based drug target validation.[5] Protein–protein interactions are crucial for all biological processes. Therefore, analyzing protein–protein interaction networks provides many new insights into protein function. In addition, protein interaction networks provide a global view of the biological systems, which may help to discover the principles of cellular functions in an organism.

For a large area of the protein–protein interactions, there may be regular rules or other information that are yet to be analyzed or uncovered.[6] Recently, plenty of *E. coli* protein–protein interactions have been identified by large-scale pull-down experiments.[7,8] However, the protein–protein interaction network is too complex for us to understand easily. We overcame this problem by applying a topological analysis approach to analyze the huge amount of information. This approach has been commonly used to study social network problems in social sciences,[9,10] and can be applied to analyze protein interaction networks as well. Furthermore, topological properties are used to evaluate the properties of a network. For example, we can use the average number of interactions of a protein to know how important the protein is, as measured by the amount of protein–protein interactions. Sequence information provides the basic properties of genes, such as the size of a gene and the ortholog information, which can be found by counting the open reading frame length and performing alignment algorithms. A previous study[11] has shown that essential genes tend to be evolutionarily more conserved than nonessential genes in bacterial species. Essential genes are thought to be less functionally removable than

[a] *Institute of Biomedical Informatics, Center for Systems and Synthetic Biology, National Yang-Ming University, Taipei, 112, Taiwan. E-mail: hsuancheng@ym.edu.tw*
[b] *Department of Life Science, Institute of Molecular and Cellular Biology, Graduate Institute of Biomedical Electronics and Bioinformatics, Center for Systems Biology and Bioinformatics, National Taiwan University, Taipei, 106, Taiwan. E-mail: yukijuan@ntu.edu.tw*
[c] *Graduate School of Biological Sciences, Nara Institute of Science and Technology, Ikoma, Nara, 630-0101, Japan*
[d] *Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, 997-0035, Japan*

† This article is part of a *Molecular BioSystems* themed issue on Computational and Systems Biology.
‡ Electronic supplementary information (ESI) available: Additional data. See DOI: 10.1039/b900611g

nonessential genes, and should be conserved for the viability of cells. Predicting essential genes by a computational classifier had been studied using sequence variables of the genome of yeast as attributes.[12]

We observed that there exists strong correlation between the interaction network topology, as well as the sequence properties, and the essentiality of genes. There might be certain rules between the protein interaction network and the gene essentiality. In this study, we developed a machine learning approach, combining the protein–protein interaction network and the sequence information to predict essential genes in two well-studied model organisms, *S. cerevisiae* and *E. coli*.

Methods

Model organism data sets

We have performed the study on one eukaryotic model organism, the budding yeast *S. cerevisiae*, and one prokaryotic, the bacterium *E. coli*. The protein–protein interaction (PPI) data of *S. cerevisiae* were obtained from the DIP database.[13] The core protein–protein interaction network (PIN) version ScereCR20070107 was used in this study, which consisted of 4873 proteins and 17 166 interactions. The essentiality information for each gene was retrieved from the Yeast Deletion Database.[2] The protein product of an essential gene was regarded as an essential protein. There were 981 essential proteins and 3892 nonessential proteins in the yeast PIN. All the sequence information for yeast was obtained from the NCBI RefSeq database.[14] The experimental data from a recent large-scale PPI measurements for *E. coli* K-12[7] were used to construct the *E. coli* PIN. There were 3039 proteins and 11 477 interactions among them in this network. The essential genes of *E. coli* K-12 have been identified by genome-wide single-gene knock-out experiments.[4] There were 260 essential and 2779 nonessential genes producing proteins in the largest connected component of *E. coli* PIN. The annotation and corresponding sequence information for each gene were collected from Riley *et al.*[15] and the GeneBank database.

Topological properties

A protein interaction network may be represented as an undirected graph $G(V, E)$ that consists of a set of nodes V and a set of edges E. Each node $i \in V$ represents a unique protein, while each edge $(i,j) \in E$ represents an observed interaction between two proteins i and j. The degree K of a node (protein) i is defined as the number of edges between i and adjacent nodes, representing the observed interacting partners of a protein. Fig. 1 shows an example network to illustrate the topological properties defined here.

The clustering coefficient (CCo) of a node i is defined as the ratio between the number of edges connecting the adjacent nodes of i and the maximum number of possible edges among them.

$$\mathrm{CCo}(i) = \frac{2e_i}{K_i(K_i - 1)}, \qquad (1)$$

where e_i is the number of edges connecting the adjacent nodes of node i, and K_i is the degree of node i. Neighbors' intra-degree, NID(i), of a node i is defined as e_i, the number of edges linking adjacent nodes. NID indicates the clustering level around a node without normalization by its degree.

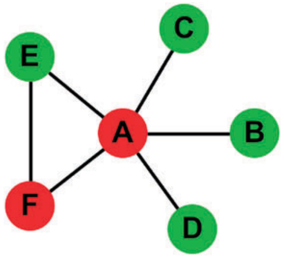


Fig. 1 An example network showing the calculation of network properties. Each node represents a protein and each edge represents a protein–protein interaction. Red nodes represent essential proteins, while green nodes represent non-essential proteins. We take node *A* as an example. The degree of node *A* is 5 because it has 5 edges, connecting with nodes *B*, *C*, *D*, *E* and *F*. The clustering coefficient of node *A* is calculated by 1/(5!/2!/3!) as there is one edge (*E*, *F*) out of 5!/2!/3! possible edges among the 5 neighbouring nodes of node *A*. The betweenness centrality is 2 × 3/(5!/2!/3!) since the 2 × 3 shortest paths between either of the 2 nodes *E*, *F* and any of the 3 nodes *B*, *C*, *D* must pass through node *A*. The closeness centrality, the reciprocal of the average distance from node *A* to other nodes, is 1. In this example, node *A* joins only one clique, which consists of nodes *A*, *E*, and *F*; thus, the KL of node *A* is 3. The essentiality index, the proportion of essential proteins interacting with node *A*, is 1/5.

The betweenness centrality (BC) of a node i measures the ratio of the shortest paths going through a node i.[16,17]

$$\mathrm{BC}(i) = \sum_{s \neq t \neq i \in V} \frac{\sigma_{st}(i)}{\sigma_{st}}, \qquad (2)$$

where σ_{st} is the number of shortest paths from node s to node t, and $\sigma_{st}(i)$ is the number of shortest paths from s to t that pass through node i.

Closeness centrality (CC) of a node i is the reciprocal of the sum of average shortest distances from all other nodes in the network to node i.

$$\mathrm{CC}(i) = \frac{N-1}{\sum_j d(i,j)}, \qquad (3)$$

where N is the total number of nodes in the network, and $d(i,j)$ is the length of the shortest path from node i to node j.

A clique is a complete subgraph in the PIN. The size of a clique is defined as the number of nodes consisted of the clique. We identify all the cliques in the protein interaction networks. In *E. coli*, the maximal clique is of size 7; and in yeast, it is 10. KL(i) is defined as the largest size of cliques that a node i can join, which ranges either from 3 to the size of the largest clique in the whole PIN, or is set to 0 if there exists no clique containing protein i.

$$\mathrm{KL}(i) = \begin{cases} 0 & \text{if there exists no clique with node } i \\ k & \end{cases}, \qquad (4)$$

where k represents the size of largest clique containing node i.

To understand whether essential proteins tend to cluster together, we defined the essentiality index (EI) as the proportion of essential proteins interacting with a node (protein) i.

$$\mathrm{EI}(i) = n_e(i)/K_i, \qquad (5)$$

Table 1 Examples of weight (w_{ij}) calculation for obtaining common-function degree (CFK)

Bait	Prey	Component	Process	Function	Weight
fkpB	rpoC	6	4	—	11
caiF	tatE	5	—	—	6
caiF	sdhA	6	5	1	13
caiE	caiF	—	7	—	8
caiE	yncD	—	—	—	1

where $n_e(i)$ represents the number of essential proteins among the adjacent nodes (proteins) and K_i is the degree of node i (the number of adjacent nodes).

We defined a score called, common-function degree (CFK), to measure the amount of common-function adjacent nodes. Based on the Gene Ontology annotation, we assigned weight to each edge according to the ontology depth of shared functions between two connected nodes. The Gene Ontology annotations for *E. coli* were collected from Baba *et al.*[4] and yeast from Saccharomyces Genome Database.[18]

$$\mathrm{CFK}(i) = \sum w_{ij} = \sum_j 1 + d_{ij}^{(\mathrm{B})} + d_{ij}^{(\mathrm{C})} + d_{ij}^{(\mathrm{M})}, \qquad (6)$$

where j represents any node adjacent to node i and $d_{ij}{}^{(\mathrm{B,C,M})}$ the ontology depth of common function shared by i and j in the category of biological process (B), cellular component (C), and molecular function (M), respectively. Some examples of w_{ij} are listed in Table 1.

Sequence-related properties

It has been found that the genes with protein products of longer amino acid chains, that is, with longer sequences of open reading frames (ORF), tend to be essential.[19] Based on this we defined the ORF length of a gene as a sequence feature for essentiality prediction.

$$\mathrm{ORFL}(i) \ = \ \text{length (bp) of the ORF of gene } i \qquad (7)$$

Since the genes located on a leading strand are more efficiently transcribed into proteins, an essential gene prefers to be located in the leading strand rather than in the lagging one.[20] We defined another sequence feature according to the location of genes.

$$\mathrm{ST}(i) = \begin{cases} 1 & \text{if gene } i \text{ is on leading strand} \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

It has been reported that essential genes tend to be more conserved than nonessential ones.[19] We defined phyletic retention, PR(i), as the number of organisms in which an ortholog is present. For *E. coli*, we compared it with 374 complete bacteria genomes available in the NCBI Genome database. An ortholog of gene i in genome M must satisfy the following criteria: (1) bidirectional best hit, which means that not only the ortholog must be the best match of gene i to genome M but gene i must also be the best hit of the ortholog, as identified by BLAST searches; (2) minimal sequence conservation of 40% between them; and (3) the difference in length between the two sequences should not exceed 30%. For yeast, the ortholog of each gene was obtained from Gustafson *et al.*[19]

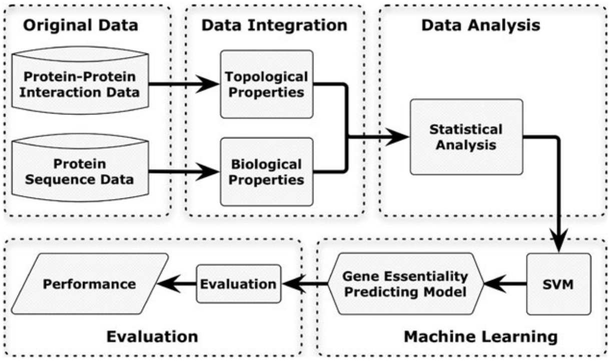


Fig. 2 The workflow for prediction of essential genes integrating network and sequence information using SVM approach.

Classification model

We constructed a support vector machine[21] (SVM) classification to predict the essential genes using network and sequence features. The workflow of the classification model is show in Fig. 2. The classifier was performed using WEKA (Waikato environment for knowledge analysis).[22] WEKA is a JAVA software package, providing a collection of machine learning algorithms for data mining tasks. We used the Sequential Minimal Optimization algorithm[23] implemented in WEKA to train the SVM with the default polynomial kernel function for the analysis with default parameters.

The selection, training, and evaluation of the machine learning algorithms were performed using the two data sets, yeast and *E. coli*, as described above. Since we only considered the largest connected components of the PIN for network analysis, there were 4815 genes in the data set of yeast (975 essential and 3840 nonessential) and 2974 of *E. coli* (256 essential and 2718 nonessential). All the topological properties and sequence characteristics described in above were encoded as input data (feature vectors) for SVM. We used Wilcoxon rank sum test[24] to estimate the statistical significance of difference between topological properties, as well as sequence characteristics, of essential and nonessential genes. Wilcoxon rank sum test (also called the Mann–Whitney U test) is a non-parametric test for assessing whether two samples are from the same distribution or not.

The amounts of essential genes and nonessential genes, in yeast, have apparent disproportion (ess : noness = 975 : 3840), and so do those in *E. coli* (ess : noness = 256 : 2718). To deal with the amount imbalance, we randomly selected a balanced subsample of nonessential genes with the same amount of essential genes and repeated the procedure 10 times to evaluate the performance in average.

Performance measures

The performance of the SVM classifier was measured as an average value in a 10-fold cross-validation analysis, where each dataset was randomly divided into 10 parts—9 parts for model learning (training) and the remaining part for validation (testing). In a classification problem, a prediction can either be a true or false positive, or true or false negative. From the counts of each case (TP, FP, TN, and FN), four performance measures were used: precision ($\mathscr{P}$), recall ($\mathscr{R}$), F-measure ($\mathscr{F}$), and Matthews Correlation Coefficent (MCC)[25] , as defined below.

$$\mathscr{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

$$\mathscr{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

$$\mathscr{F} = \frac{2 \cdot \mathscr{P} \cdot \mathscr{R}}{\mathscr{P} + \mathscr{R}} \tag{11}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{12}$$

MCC is generally regarded as a balanced measure which can be used for comparison of the results with different positive-to-negative ratios, whereas precision and recall (and hence F-measure) are sensitive to dataset imbalance.

Results and discussion

Network analysis of essential genes

Topological analysis of the yeast protein interaction network shows that the means of degree (11.55), clustering coefficient (0.16), and betweenness centrality (0.001) of essential genes in yeast are about twice as large as those in nonessential genes (5.92, 0.08, and 5.2 × 10^{-4}, respectively, as shown in Table 2). The results are consistent with previous studies.[26] The genes with higher degrees are hubs in PIN and may control many regulatory pathways. The genes with larger clustering coeficients mean that many of their interacting partners interact with each other. The genes with larger betweeness centrality represent the bottlenecks on the interaction network or pathways they participate in. Other topological properties we analyzed were also statistically discernable between essential and nonessential genes, with significant *P*-values < 10^{-15}, indicating that the genes playing important roles in PIN tend to play essential roles in the mechanism of the cell survival. To investigate the capability of each topological property for identification of essential genes and compare their performance, we ranked each topological property and obtained the proportions of essential genes in the top 1% to 30% of each property. The clique level (KL, the size of the largest cliques a gene belongs to) we proposed outperformed other topological properties as shown in Fig. 3. The other two properties NID and CFK also show better performance than degree. The averaged essentiality index (EI), as well as clustering coefficient (CCo), of essential genes are significantly larger than nonessential (EI: 0.46 *vs.* 0.27, in Table 2), indicating that essential genes tend to interact with each other closely. However, they perform worse than expected for the top 10% ranked genes although both of them perform well in other regions (Fig. 3). The reason why EI and CCo perform poorly only for the top 10% ranked genes is due to a large number of nonessential genes with very low degree (say, K = 1 or 2; correlation between EI and K is shown in the ESI‡). If a gene with K = 1, its EI can only be either 1 or 0, depending on whether it interacts with an essential or nonessential one. Since there are an abundance of genes with K = 1 (a characteristic of protein–protein interaction networks) and the essential genes tend to have larger values of K, there exist many nonessential genes with K = 1 and EI = 1. Similarly, there are also a lot nonessential genes with K = 2 and EI = 1 (or 0.5). Note that the averaged EI of essential genes is 0.46. As a result, while ranked by EI, genes with EI = 1 are listed at the top but consist of more nonessential than essential ones. For clustering coefficients a similar singularity happens for genes with very low degree. For K = 2, the CCo value can only be either 1 or 0, depending on whether the two interacting partners of the gene interact with each other or not.

Table 2 Summary of the network and sequence features for yeast

Feature	Essential	Noness.	*P*-value
Network			
Degree	11.55	5.92	$<10^{-15}$
Betweenness centrality	0.00114	0.00052	$<10^{-15}$
Closeness centrality	0.254	0.243	$<10^{-15}$
Clustering coefficient	0.16	0.08	$<10^{-15}$
Neighbors' intra-degree	15.06	4.30	$<10^{-15}$
Clique level	3.07	1.32	$<10^{-15}$
Essentiality index	0.46	0.27	$<10^{-15}$
Common function degree	196.57	81.51	$<10^{-15}$
Characteristic path length	3.81	4.22	$<10^{-15}$
Diameter	10	11	—
Sequence			
Phyletic retention	3.14	1.51	$<10^{-15}$
ORF length (bp)	1543.2	1257.3	$<10^{-10}$
Leading/Lagging strand	474/501	1853/1937	—

The averages of essential and nonessential genes for each feature are shown here. *P* values are the statistical significance of the difference between essential and nonessential according to Wilcoxon rank sum tests.[24] The bottom row shows the numbers of essential (nonessential) genes located in leading and lagging strands, respectively. Diametrer and characteristic path length were not used as SVM features.

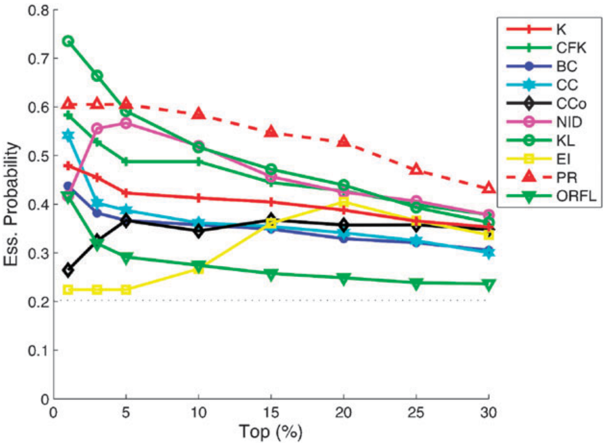


Fig. 3 Proportions of essential genes in the top percentages of yeast genes ranked by each different feature. The dotted line indicates the overall proportion of essential genes in *S. cerevisiae*.

Applying the same analysis on the protein interaction network of *E. coli*, we observed similar results as shown in Table 3. The average degree and betweeness centrality of essential genes (17.17 and 0.00238) are 3–4 times higher than those of nonessential genes (6.68 and 0.00064). The characteristic path length (3.24 in average) and diameter (8) between essential genes are also shorter than nonessential

Table 3 Summary of the network and sequence features for *E. coli*

Feature	Essential	Noness.	*P*-value
Network			
Degree	17.17	6.68	$<10^{-9}$
Betweenness centrality	0.00238	0.00064	$<10^{-4}$
Closeness centrality	0.308	0.303	0.13
Clustering coefficient	0.056	0.065	0.14
Neighbors' intra-degree	18.35	2.42	$<10^{-5}$
Clique level	1.95	1.39	$<10^{-5}$
Essentiality index	0.21	0.18	0.0020
Common function degree	99.36	24.77	$<10^{-15}$
Characteristic path length	3.24	3.38	$<10^{-175}$
Diameter	8	9	—
Sequence			
Phyletic retention	247.33	86.91	$<10^{-15}$
ORF length (bp)	1021.7	995.5	0.86
Leading/Lagging strand	175/80	1450/1264	—

See the footnote in Table 2 for detailed descriptions.

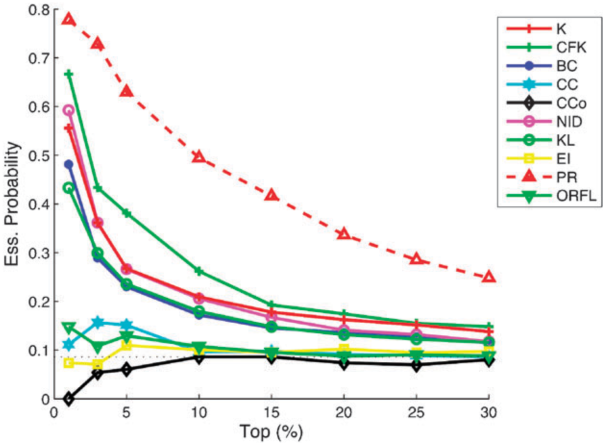


Fig. 4 Proportions of essential genes in the top percentages of *E. coli* genes ranked by each different feature. The dotted line indicates the overall proportion of essential genes in *E. coli*.

genes (3.38 and 9), confirming previous results in yeast. While the behavior of the characteristic path length indicates that essential genes are significantly closer to each other in the network, the clustering coefficient and closeness centrality are undistinguishable between essential and nonessential genes in *E. coli*. Although these two topological properties do not reflect our expectation that a node which is densely clustered around might imply a more important role in the network and hence higher biological significance, this might be due to simpler subcellular organization of prokaryotes or noisy systematic measurements of protein–protein interaction in *E. coli* (see ESI‡). To further explore the role of a protein in a denser region of the interaction network, we investigated other related topological features for essential and nonessential genes and found that the subgraph features of clique and NID are actually closely related to gene essentiality. Fig. 4 shows the evaluation and comparison of all the topological properties used in this study by their essentiality in *E. coli*.

Sequence analysis of essential genes

As shown in Table 2, the averaged phyletic retention of essential genes is more than two times of nonessential in yeast. In *E. coli*, the difference is even larger (Table 3). These confirm the previous conclusions that essential genes tend to be less dispensable than nonessential ones and conserved through natural selection. Identification of essential genes using phyletic retention performed the best in both yeast and *E. coli* (Fig. 3 and 4). More than 60% of the top 5% genes with the highest phyletic retention are essential in both organisms.

The size of an essential gene, 1543.2 base pairs on average, counted by the length of its open reading frame, tends to be larger than that of nonessential genes (1257.3) in yeast, while the difference in *E. coli* is not significant. The proportion of essential genes located in the leading strand during replication (leading: 175 *vs.* lagging: 80) was found to be higher than nonessential (1450 *vs.* 1264) in *E. coli*, but there is no such preference in yeast.

Performance of essential gene prediction

We wanted to construct a classifier that could identify an essential gene using its sequence and topological property in protein interaction networks. SVMs are known to produce classifiers that generalize well to unseen data[21] and have been used widely for sequence-based and other bioinformatics prediction. We used the SMO implementation of SVMs in WEKA, with a polynomial kernel function and default parameters to construct a classifier that separates between essential and nonessential genes. The SVM feature vectors included both network and sequence properties that were selected based on the results of our analysis.

We evaluated the predictive power of our classifier by 10-fold cross validation. For yeast, we obtained precision of 0.76 and recall of 0.71, while combining all the network and sequence features, as listed in Table 4. Receiver operating characteristic (ROC) curves of the classifier and the included individual features for *S. cerevisiae* are shown in Fig. 5A. To compare the predicting results for different methods, we also calculated the corresponding F-measure and MCC. Using network features alone, we obtained an F-measure of 0.68 and MCC of 0.35, slightly worse than using sequence features alone (F-meausure, 0.71; MCC, 0.44). Combining both kinds of features significantly improves the performance of the classifier (F-measure, 0.74; MCC, 0.49). Note that all the performance measures shown above were based on balanced data. Using imbalanced data in which the ratio of essential and nonessential genes remained intact (1 : 3.94), we reconstructed the SVM classifier and obtained the precision of 0.77, recall of 0.23, and MCC of 0.36. Compared with the previous study,[12] which used only sequence features and imbalanced data, our method performs better with near two times

Table 4 Performance of the prediction for essential genes in yeast

Attributes	Precision(%)	Recall(%)	F-measure(%)	MCC(%)
All	76.3	71.3	73.7	49.2
Seq. only	73.4	68.8	71.0	43.9
Network only	66.9	69.4	68.1	35.1
All (imbalanced)	77.2	23.0	35.4	35.9
Previous	69.0	10.0	17.5	20.1
E. coli model	75.4	64.2	69.3	43.7

Note that imbalanced data were used in previous study[12] for performance evaluation.

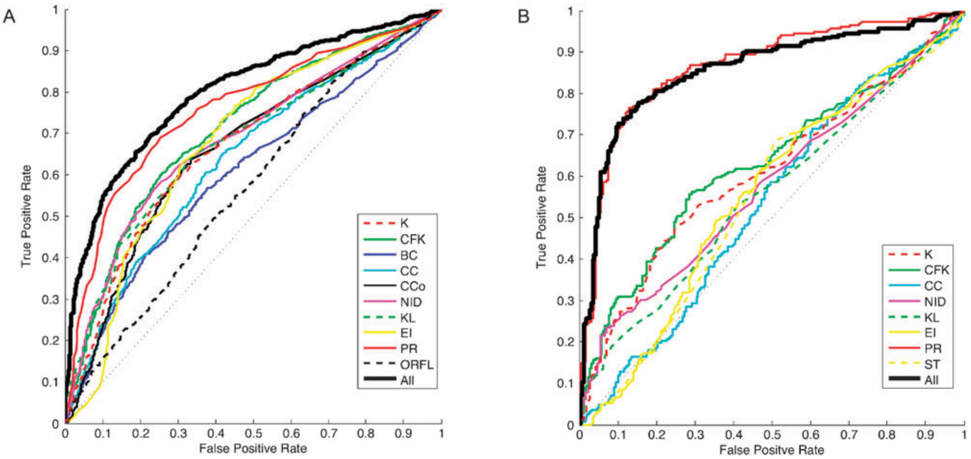


Fig. 5 Receiver operating characteristic (ROC) curves for the essential gene prediction classifiers on (A) *S. cerevisiae* and (B) *E. coli* data sets with individual and combined (ALL) input features. The ROC curves plot the true positive rates *vs.* the false positive rates for different thresholds of classifier output.

improvement in terms of MCC (0.36 *vs.* 0.20). A major reason for the improved performance is the integration of various network features into our classifier. Besides, previous studies used a special machine learning approach with a slight bias to reduce false-positives, while our SVM-based classifer can readily find the optimal solution. Hence, our classifier can perform better, even with only sequence features.

Using the same feature vectors and approaches, we constructed an SVM classifier for *E. coli* and evaluated its performance as shown in Table 5. ROC curves of the classifier and those of individual features are shown in Fig. 5B. The resulting precision for predicting essential genes in *E. coli* is 0.83, and the recall is 0.75, better than those in yeast. The phyletic retention has much higher predictive power in *E. coli* than yeast, largely because there are many more complete bacterial genomes for comparison with *E. coli.* Using the sequence features alone, the MCC already achieves 58.9%. However, the network features alone could not classify essential genes well since the known protein interaction network of *E. coli* is still incomplete and less confident than the yeast network. We need more confident interaction network data to improve the prediction accuracy.

Considering that the developed classifier might be applied to predict essential genes in other organisms, we also evaluated the performance of classifiers by cross prediction between *S. cerevisiae* and *E. coli* (Table 4 and 5). Since the sequence and network features usually vary across different organisms, we assumed that their trends remain similar and used their rank percentages, instead of raw feature values, as the input feature vectors for SVM. Using the classification model constructed on *S. cerevisiae*, we obtained its predictive performance on *E. coli* with precision of 0.72, recall of 0.72, and an MCC of 0.44. *Vice versa*, using the classification model constructed by *E. coli* data, we could predict the essential genes in *S. cerevisiae* with precision of 0.75, recall of 0.64, and an MCC of 0.44. The results indicated that the proposed method could work to predict essential genes in a new species based on the information from another one.

Table 5 Performance of the prediction for essential genes in *E. coli*

Attributes	Precision(%)	Recall(%)	F-measure(%)	MCC(%)
All	82.8	74.5	78.4	59.3
Seq. only	82.6	74.3	78.2	58.9
Network only	58.4	46.4	51.7	13.6
All (imbalanced)	31.1	74.5	43.9	40.9
Yeast model	72.2	72.1	72.2	44.3

Topological analysis of protein–protein interaction networks provide global characteristics of proteins (genes) in a biological system. Proteins are the system components and the protein interaction network describes connectivity between the components. The drawback of such a network diagram is that it is descriptive only, and cannot predict the quantitative and dynamic behavior of a given system.[27] However, the network diagram serves as a tremendous visual aid in understanding biological processes and allows us to build qualitative models to represent a biological system. Although topological analysis reveals only statistical information of a given system, we have shown that it is feasible for machine learning algorithms to infer a predictor model of essential genes from topological features alone. Integrating additional informations like protein (gene) expression, localization, and temporal profiles, we may extend the network diagram to more precise system models, towards quantitative and dynamic understanding of biological systems.

Conclusions

We have identified the topological properties and the sequence characteristics to discriminate essential and nonessential genes with statistical significance. Similar to yeast, we found that essential genes in *E. coli* also tend to play topologically more important roles in protein–protein interaction networks. Our newly proposed network properties of clique level, neighbors' intra-degree, and common-function degree were strongly correlated with the gene essentiality both in yeast and *E. coli.* Based on these features, we have developed a computational method for predicting essential genes using machine learning approach. While sequence information alone was known to effectively classify essential and nonessential genes, our results indicated that combining network information substantially improves the prediction accuracy. As satisfactory results can be obtained for both yeast and *E. coli* using the proposed approaches, we may apply the method to predict essential genes of other organisms, once their protein interaction network data are available.

The ability to rapidly identify essential genes of pathogens has been described as the most important task of genomics-based drug target validation. Experimental identification of essential genes, achieved through whole-genome knockout techniques, usually requires great time and expense. The proposed computational methodology allows for predicting essential genes as potential drug targets of novel and/or newly sequenced pathogens. While large-scale experimental screening data for protein–protein interactions were not available for the target organism, computational prediction of protein–protein interactions can be applied to construct a plausible protein interaction network. Thereafter, the proposed methodology may still work with possibly declined accuracy. Furthermore, previous studies have shown that human disease genes often play more "central" role than others in protein interaction network or other biological pathways. Our method might also be applied to help potential drug targets for human disease genes, as well as pathogens.

Acknowledgements

This research has been supported by grants from the National Science Council, Taiwan; National Science and Technology Program for Agricultural Biotechnology, Taiwan; National Health Research Institutes, Taiwan; and NTU Frontier and Innovative Research Projects.

References

1 K. M. Cadigan, U. Grossniklaus and W. J. Gehring, Functional redundancy: the respective roles of the two sloppy paired genes in drosophila segmentation, *Proc. Natl. Acad. Sci. U. S. A.*, 1994, **91**(14), 6324–8.

2 E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. El Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M'Rabet, P. Menard, M. Mittmann, C. Pai, C. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens, M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, S. Veronneau, M. Voet, G. Volckaert, T. R. Ward, R. Wysocki, G. S. Yen, K. Yu, K. Zimmermann, P. Philippsen, M. Johnston and R. W. Davis, Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis, *Science*, 1999, **285**(5429), 901–6.

3 L. M. Steinmetz, C. Scharfe, A. M. Deutschbauer, D. Mokranjac, Z. S. Herman, T. Jones, A. M. Chu, G. Giaever, H. Prokisch, P. J. Oefner and R. W. Davis, Systematic screen for human disease genes in yeast, *Nat. Genet.*, 2002, **31**(4), 400–4.

4 T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner and H. Mori, Construction of *escherichia coli* k-12 in- frame, single-gene knockout mutants: the keio collection, *Mol. Syst. Biol.*, 2006, **2**, 2006–2008.

5 A. F. Chalker and R. D. Lunsford, Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach, *Pharmacol. Ther.*, 2002, **95**(1), 1–20.

6 A. L. Barabasi and Z. N. Oltvai, Network biology: understanding the cell's functional organization, *Nat. Rev. Genet.*, 2004, **5**(2), 101–13.

7 M. Arifuzzaman, M. Maeda, A. Itoh, K. Nishikata, C. Takita, R. Saito, T. Ara, K. Nakahigashi, H. C. Huang, A. Hirai, K. Tsuzuki, S. Nakamura, M. Altaf-Ul-Amin, T. Oshima, T. Baba, N. Yamamoto, T. Kawamura, T. Ioka-Nakamichi, M. Kitagawa, M. Tomita, S. Kanaya, C. Wada and H. Mori, Large-scale identification of protein–protein interaction of escherichia coli k-12, *Genome Res.*, 2006, **16**(5), 686–91.

8 G. Butland, J. M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt and A. Emili, Interaction network containing conserved and essential protein complexes in *escherichia coli*, *Nature*, 2005, **433**(7025), 531–7.

9 S. Wasserman and K. Faust, *Social NetworkAnalysis: Methods and Applications*, Cambridge University Press, 1994.

10 R. Albert and A. L. Barabasi, Statistical mechanics of complex networks, *Rev. Mod. Phys.*, 2002, **74**, 47–97.

11 I. K. Jordan, I. B. Rogozin, Y. I. Wolf and E. V. Koonin, Essential genes are more evolutionarily conserved than are nonessential genes in bacteria, *Genome Res.*, 2002, **12**(6), 962–8.

12 M. Seringhaus, A. Paccanaro, A. Borneman, M. Snyder and M. Gerstein, Predicting essential genes in fungal genomes, *Genome Res.*, 2006, **16**(9), 1126–35.

13 L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie and D. Eisenberg, The database of interacting proteins: 2004 update, *Nucleic Acids Res.*, 2004, **32**(90001), 449D–51.

14 K. D. Pruitt, T. Tatusova and D. R. Maglott, Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res.*, 2005, **33**(database issue), D501–4.

15 M. Riley, T. Abe, M. B. Arnaud, M. K. Berlyn, F. R. Blattner, R. R. Chaudhuri, J. D. Glasner, T. Horiuchi, I. M. Keseler, T. Kosuge, H. Mori, N. T. Perna, r. Plunkett, G. K. E. Rudd, M. H. Serres, G. H. Thomas, N. R. Thomson, D. Wishart and B. L. Wanner, *Escherichia coli* k-12: a cooperatively developed annotation snapshot—2005, *Nucleic Acids Res.*, 2006, **34**(1), 1–9.

16 M. Joy, A. Brock, D. Ingber and S. Huang, High-betweenness proteins in the yeast protein interaction network, *J. Biomed. Biotechnol.*, 2005, **2005**(2), 96–103.

17 H. Yu, P. Kim, E. Sprecher, V. Trifonov and M. Gerstein, The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics, *PLoS Comput. Biol.*, 2007, **3**(4), e59.

18 E. L. Hong, R. Balakrishnan, Q. Dong, K. R. Christie, J. Park, G. Binkley, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, C. J. Krieger, M. S. Livstone, S. R. Miyasato, R. S. Nash, R. Oughtred, M. S. Skrzypek, S. Weng, E. D. Wong, K. K. Zhu, K. Dolinski, D. Botstein and J. M. Cherry, Gene ontology annotations at sgd: new data sources and annotation methods, *Nucleic Acids Res.*, 2008, **36**(Database issue), D577–81.

19 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*, 2006, **7**, 265.

20 E. P. Rocha and A. Danchin, Essentiality, not expressiveness, drives gene-strand bias in bacteria, *Nat. Genet.*, 2003, **34**(4), 377–8.

21 B. Schölkopf, *Support Vector Learning*, Oldenbourg Verlag, Munich, 1997.

22 I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, 2nd edn, 2005.

23 J. Platt, *Fast training of support vector machines using sequential minimal optimization*, in *Advances in Kernel Methods-Support Vector Learning*, ed. B. Schölkopf, C. Burges and A. Smola, MIT Press, 1998.

24 F. Wilcoxon, Individual comparisons by ranking methods, *Biometrics Bull.*, 1945, **1**(6), 80–83.

25 B. Mathews, Comparison of the predicted and observed secondary structure of t4 phage lysozyme, *Biochem. Biophys. Acta*, 1975, **405**(2), 442–451.

26 H. Yu, D. Greenbaum, H. Xin Lu, X. Zhu and M. Gerstein, Genomic analysis of essentiality within protein networks, *Trends Genet.*, 2004, **20**(6), 227–31.

27 L. You, Toward computational systems biology, *Cell Biochem. Biophys.*, 2004, **40**(2), 167–84.

This journal is © The Royal Society of Chemistry 2009