BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 5 2005, pages 575–581
doi:10.1093/bioinformatics/bti058

Genome analysis

Understanding protein dispensability through machine-learning analysis of high-throughput data

Yu Chen[1,2,†] and Dong Xu[1,2,*]

[1]UT-ORNL Graduate School of Genome Science and Technology, Oak Ridge, TN, USA and
[2]Digital Biology Laboratory, Computer Science Department, 201 Engineering Building West, University of Missouri-Columbia, Columbia, MO, USA

Received on March 15, 2004; revised on September 19, 2004; accepted on September 24, 2004
Advance Access publication October 12, 2004

ABSTRACT

Motivation: Protein dispensability is fundamental to the understanding of gene function and evolution. Recent advances in generating high-throughput data such as genomic sequence data, protein–protein interaction data, gene-expression data and growth-rate data of mutants allow us to investigate protein dispensability systematically at the genome scale.

Results: In our studies, protein dispensability is represented as a fitness score that is measured by the growth rate of gene-deletion mutants. By the analyses of high-throughput data in yeast *Saccharomyces cerevisiae*, we found that a protein's dispensability had significant correlations with its evolutionary rate and duplication rate, as well as its connectivity in protein–protein interaction network and gene-expression correlation network. Neural network and support vector machine were applied to predict protein dispensability through high-throughput data. Our studies shed some lights on global characteristics of protein dispensability and evolution.

Availability: The original datasets for protein dispensability analysis and prediction, together with related scripts, are available at http://digbio.missouri.edu/~ychen/ProDispen/

Contact: xudong@missouri.edu

1 INTRODUCTION

Understanding the importance of an individual gene to the viability of an organism is critical in studying gene function and designing mutant species with the help of bioengineering techniques. In gene 'knockout' experiments, 'essential' and 'non-essential' are the two classical molecular genetics designations referring to the significance of a gene with respect to its effect on fitness in an organism (Hurst and Smith, 1999). A gene is considered to be essential if upon deletion results in lethality. On the other hand, non-essential genes are those for which knockouts do not kill the organism. Essential genes are less functionally dispensable or less redundant than non-essential genes. The deletions of different non-essential genes have different effects on the evolution and population (growth) of the organism carriers. The deletion of a non-essential gene might give the carrier a selective disadvantage, and thus, this carrier is likely to be removed from the population over time. Such selection is called purifying selection (Li, 1997). Given the role of purifying selection in determining evolution, non-essential genes that are subject to weaker purifying selection were suggested to have a higher rate of evolution (Ohta, 1973). Earlier studies showed that protein dispensability and evolutionary rate were correlated (Hirsh and Fraser, 2001; Krylov *et al*., 2003). However, the relationship between protein's dispensability and evolution could be complex due to a variety of factors involved (Pal *et al*., 2003). Recently, the balance hypothesis was proposed to study protein dispensability. It assumes that the dominance is from physiology and metabolism rather than from selection (Papp *et al*., 2003). To further investigate the mechanism of protein dispensability, yeast metabolic network was used to predict fitness effects of enzymes, which indicated that environmental specificity dominates enzyme dispensability (Papp *et al*., 2004).

A protein's dispensability is also constrained by the protein connectivity in a protein–protein interaction network. The topological structure of a large-scale protein–protein interaction network can be characterized by a scale-free network, in which only a small number of proteins are highly connected, while for a vast majority of proteins, each has only limited interactions (Jeong *et al*., 2001). It has been known that highly connected proteins are more likely to be essential and evolve slowly (Jordan *et al*., 2003). A protein's dispensability may also be related to its gene expression, an aspect that has not been examined before. Co-expressed gene are often involved in the same pathway or similar cellular function, and interacting proteins are frequently co-expressed (Jansen *et al*., 2002). Thus, the linkage of coexpression might put some constraints in protein evolution and dispensability, in a similar way as physical protein–protein interaction. Another factor that may have an impact on protein dispensability is gene duplication. Gene duplication, originated from region-specific duplication or genome-wide polyploidization, is an important feature in genome evolution (Lawton-Rauh, 2003). One can speculate that a gene with more duplicates is less likely to be essential, as the duplicates may serve as a backup if the gene is deleted. However, the relationship between the role of duplicate genes (Gu *et al*., 2003) and protein dispensability remains unknown.

While protein dispensability has been studied at the individual gene phenotype level, advances in generating high-throughput data, such as genomic sequence data, protein–protein interaction data, gene-expression data and gene fitness data, enable researchers to carry out studies at the genome scale. In this study, we

*To whom correspondence should be addressed.

†Present address: BioMarker Development, Novartis Pharmaceuticals Corporation, One Health Plaza, East Hanover, NJ 07936, USA.

conducted integrated analyses to understand the dependence of protein dispensability on protein evolutionary rate, protein-interaction connectivity, gene-expression cooperativity and gene-duplication rate at a system level by using high-throughput data from multiple resources in yeast *Saccharomyces cerevisiae*, which is a good model system for our study given that comprehensive genome-scale high-throughput data are available. Based on the dependences, we applied machine-learning methods, i.e. neural network and support vector machine (SVM), to predict protein dispensability from high-throughput data and to understand the relationship between protein dispensability and different factors involved.

2 METHODS

2.1 Integration of high-throughput data

2.1.1 Data preparation In our analyses, we incorporated several sources of high-throughput data in yeast *S.cerevisiae*, including genomic sequence and annotation data, protein–protein interaction data, gene-expression data and mutant growth-rate data. The dispensability of a protein can be quantified by its contribution to survival and reproduction of the carrier upon gene deletion. This contribution can be measured experimentally by the growth rate of the carrier. In our study, we used two types of data for the growth rate. The first type of data is binary, i.e. 1 (for essential genes) or 0 (for non-essential genes). This type of data was used and neural network for protein dispensability prediction of SVMs. The second type of the growth-rate data is a fitness value ranging from 0 to 1. For each mutant, we estimated the deleted gene's fitness value, f_i, as $1 - r_i/r_{\max}$, where r_i is the growth rate of the strain with gene deleted and $r_{\max}$ is the maximal growth rate. The fitness values of essential gene deletion strains are 1.

Protein–protein interaction data can be represented as a weighted non-directed graph $Gp(D) = (Vp, Ep)$ with the vertex set $Vp = \{d_i | d_i \in D\}$ and the edge set $Ep = \{(d_i, d_j)|$ for $d_i, d_j \in D$ and $i \neq j\}$. Each vertex represents one protein and each edge represents one measured interaction between the two connected proteins. The spread of node degree (number of interacting nodes, k) is characterized by a distribution function $P(k)$. From gene-expression microarray data, a gene-expression cooperativity graph was constructed as $Gg(D) = (Vg, Eg)$. The vertex set $Vg = \{d_i | d_i \in D\}$ and the edge set $Eg = \{(d_i, d_j)|$ for $d_i, d_j \in D, i \neq j$ and $|r_{ij}| \geq 0.7\}$. Each vertex represents one gene and each edge represents one gene pair whose gene expression profiles correlation coefficient $|r_{ij}| \geq 0.7$. This cutoff value of $|r_{ij}|$ is determined based on our previous study (Joshi *et al*., 2004a,b).

2.1.2 Data sources We downloaded the genomic sequences and the protein annotation data of five species including budding yeast *S.cerevisiae* (http://genome-www.stanford.edu/Saccharomyces/), fission yeast *Schizosaccharomyces pombe* (http://www.sanger.ac.uk/Projects/S_pombe), *Arabidopsis thaliana* (http://www.arabidopsis.org/), *Drosophila melanogaster* (http://flybase.bio.indiana.edu/), and *Caenorhabditis elegans* (http://www.wormbase.org/). The high-throughput protein–protein interaction data based on yeast two-hybrid experiments were from Uetz *et al*. (2000) and Ito *et al*. (2001), with 5075 interactions among 3567 proteins. We combined the yeast two-hybrid data with the protein–protein interaction data in the DIP database (http://dip.doe-mbi.ucla.edu/). In total, 7231 unique binary interactions among 4067 proteins were used in this study. The gene-expression profiles of microarray data were from Gasch *et al*. (2000). The growth rates of gene deletion mutants in yeast *S.cerevisiae* were measured at the genome-scale, where 4706 homozygous diploid deletion strains were monitored in parallel under 9 different medium conditions (Steinmetz *et al*., 2002). We used the average growth rate over nine conditions in our study, similar to other studies (Hirsh and Fraser, 2001; Papp *et al*., 2003). The fitness effects of essential genes have the same phenotype (lethality) under various conditions. Thus, the essential genes defined here do not have false positives, although we might miss some 'marginal essential genes', whose deletions do not cause lethality in the well-controlled experimental conditions but do cause lethality in the wild environment (Thatcher *et al*., 1998). The growth-rate data were obtained from the public database at http://www-deletion.stanford.edu/YDPM/YDPM_index.html.

2.2 Identification of putative orthologs using reciprocal search

An all-against-all FASTA search was conducted for all the proteins coded in the *S.cerevisiae* genome to identify the putative orthologs in *S.pombe*, *A.thaliana*, *D.melanogaster* and *C.elegans*. A subclass of putative orthologs are defined as reciprocal best hits (Tatusov *et al*., 2000; Hirsh and Fraser, 2001) with additional two strict criteria: (1) FASTA (Pearson, 2000) expectation value is $<10^{-10}$ and (2) the aligned region between two protein sequences is >80% of the protein length in yeast *S.cerevisiae*. We illustrate the reciprocal process using the following example. To identify the orthologs in fission yeast *S.pombe*, we first queried one open reading frame (ORF i) in budding yeast *S.cerevisiae* against all 4940 ORFs predicted to be protein coding genes in *S.pombe* (Wood *et al*., 2002) to yield the set of hits $\{W\}$. Then, we queried the hit with the lowest expectation value in $\{W\}$ (ORF j) against all 6217 ORFs in budding yeast (Goffeau *et al*., 1996) to yield the set of hits $\{Y\}$. Finally, the protein pair {ORF i, ORF j} was considered to be putative orthologs if the member with the lowest expectation value in $\{Y\}$ is ORF i.

2.3 Identification of paralogs in *S.cerevisiae*

Duplicated genes are often referred to as paralogous genes. An all-against-all FASTA search was conducted for the whole set of *S.cerevisiae* protein sequences to identify the paralogs with two criteria: (1) The FASTA expectation value is $<10^{-10}$ and (2) the aligned region between two protein sequences is >80% of any of the two protein sequences.

2.4 Prediction of protein dispensability

We predicted a protein's dispensability based on the combination of protein evolution rate, protein-interaction connectivity, gene-expression cooperativity and gene-duplication data. Neural network and SVM were used to extract features of essential genes and non-essential genes in the training process, and the trained models were used to predict protein dispensability. The data size (number of genes) is 5409. We randomly selected 70% of the data for training and used the remaining 30% of the data for independent testing. For a given yeast gene, there were nine input units for neural network and SVM as follows: (1) the number of species where the yeast protein has orthologs in the other four selected species (*S.pombe*, *A.thaliana*, *D.melanogaster* and *C.elegans*); (2–5) the four sequence identity values, each for an ortholog pair identified from reciprocal search between *S.cerevisiae* and another species (if no ortholog is found in a species, the input value is set to 0); (6) the number of interacting partners in protein–protein interaction network; (7) the number of neighbors in gene expression cooperativity network; (8) the number of parologs in yeast; and (9) the protein size. Features 2–5 provide supplemental information about protein evolutionary rate. Protein size is related with biological functionality and diversity (Ryden and Hunt, 1993). All nine features were scaled into the range of [0, 1] for training and testing.

For the expected output of SVM or neural network, each protein was labeled as 1 (for essential genes) or 0 (for non-essential genes). For SVM prediction, the initial output fitness value is a floating-point value; while for neural network prediction, the initial output fitness value is within the range between 0 and 1. After choosing a cutoff value, we can make a binary choice whether this gene is essential or non-essential as 0 or 1. For the neural network, the back-propagation learning algorithm and a logistic activation function were used. The learning rate was usually between 0.1 and 1.0. This neural network had one hidden layer with three hidden units. During each training, the iteration is set from 100 to 1000 cycles. After every 50 cycles, a model was saved and evaluated using 3-fold cross-validation of the training data. The performance of each model was ranked by the average Matthews correlation coefficient (Mathews, 1975). The model with the best performance was chosen as the predictor to predict the essential or non-essential genes in

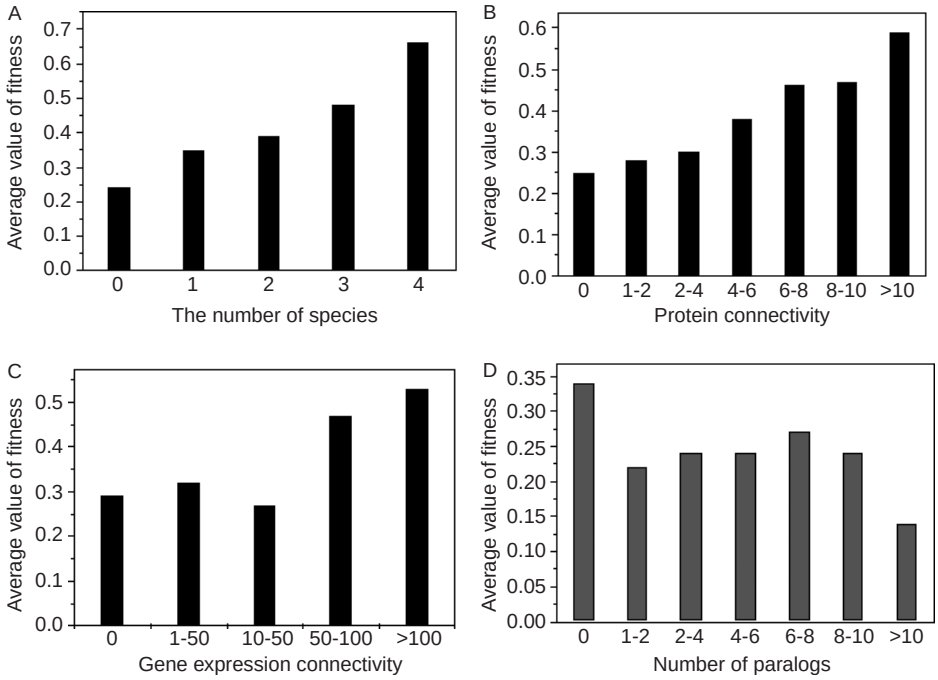

Fig. 1. The fitness (indispensability) of a protein in yeast *S.cerevisiae* versus (**A**) number of other species (among *S.pombe*, *A.thaliana*, *D.melanogaster* and *C.elegans*) having orthologs of the protein, (**B**) protein connectivity in the protein–protein interaction network, (**C**) gene cooperativity in the gene expression network and (**D**) number of paralogs in yeast *S.cerevisiae*.

the test set. For SVM, the polynomial kernel was used and two parameters C_+ and C_- were applied for the tradeoff between the generalization ability and mis-classification error of the unbalanced data. The software packages used were SNNS 4.2 (Stuttgart Neural Network Simulator) (Zell *et al*., 1993) for the neural network and LIBSVM 2.4 (a library for SVMs) (Schölkopf *et al*., 2000) for the SVM. The related parameters, other than those specified in the paper, were chosen from the default of the software packages. The scripts for data training and testing using SNNS 4.2 and LIBSVM 2.4 are available upon request.

3 RESULTS

In this section, we will first carry out statistical analyses to identify the global relationship between the fitness of yeast protein and each individual factor derived from the high-throughput data. Then, we will use the statistical information to select inputs of the neural network and SVM for predicting the fitness of individual protein from the related factors.

3.1 Integrated analysis of protein dispensability

To characterize the properties of protein dispensability, we investigated the relationships between the fitness of a protein in budding yeast *S.cerevisiae* and protein evolutionary rate, protein connectivity in the protein–protein interaction network, gene-expression cooperativity or gene-duplication rate. We incorporated four types of high-throughput data into our analysis, including growth rates of mutants, protein sequence data, protein–protein interaction data and gene-expression data. The relevant variables derived from sequence data, protein–protein interaction data and gene-expression data are as follows:

- $X_{E_o} = \{X_{O_i} : O_i \in E_o\}$, distribution of proteins with different evolutionary rates in yeast *S.cerevisiae*. $E_o = \{0, 1, 2, 3, 4\}$ represents the number of species that contain orthologs of a given protein in budding yeast *S.cerevisiae* using the reciprocal search against *S.pombe*, *A.thaliana*, *D.melanogaster* and *C.elegans* (see Section 2.2). X_{O_i} represents the number of proteins that contain orthologs in O_i species.
- $X_{E_p} = \{X_{P_i} : P_i \in E_p\}$, distribution of proteins with different connectivities in the protein–protein interaction network of yeast *S.cerevisiae*. X_{P_i} is the number of proteins with the node degree (number of interactions) P_i, and $E_p = \{0, 1, 2, 3, 4, \ldots\}$.
- $X_{E_g} = \{X_{G_i} : G_i \in E_g\}$, distribution of proteins (genes) with different connectivities in the gene expression cooperativity graph of the microarray data yeast *S.cerevisiae*. G_i is the node degree, i.e. number of genes whose expression profiles have correlation coefficient ≥ 0.7 with a given gene (protein). X_{G_i} is the number of proteins with the node degree G_i, and $E_g = \{0, 1, 2, 3, 4, \ldots\}$.
- $X_{E_d} = \{X_{D_i} : D_i \in E_d\}$, distribution of proteins (genes) with different gene duplication rates in yeast *S.cerevisiae*. X_{D_i} is the number of proteins with D_i paralogs in the yeast *S.cerevisiae* obtained from the FASTA search, and $E_d = \{0, 1, 2, 3, 4, \ldots\}$.

We carried out statistical analysis for the relationships between protein dispensability and distributions of X_{E_o}, X_{E_p}, X_{E_g} and X_{E_d} (Fig. 1). Figure 1A shows that the average fitness of a gene in yeast *S.cerevisiae* has a positive correlation with the number of species in which the gene product (protein) has ortholog hits. This implies that highly conserved proteins across species or slowly evolved proteins are less dispensable. Figure 1B shows the relationship between protein connectivity in a protein–protein interaction network and fitness. We can see that the proteins involved in more interactions have

Table 1. Contingency table of fitness distribution versus number of species in which a protein in yeast *S.cerevisiae* has ortholog hits in the other four selected species (*S.pombe*, *A.thaliana*, *D.melanogaster* and *C.elegans*)

Fitness	Number of species out of four species				
	0	1	2	3	4
Weak effect ($0 \leq$ fitness < 0.1)	2106 (1859)	427 (471)	168 (214)	117 (166)	60 (168)
Moderate effect ($0.1 \leq$ fitness < 0.5)	1135 (1117)	294 (283)	142 (129)	81 (99)	77 (101)
Moderate effect ($0.5 \leq$ fitness < 1)	37 (50)	13 (13)	7 (6)	8 (4)	12 (4)
Essential genes (fitness $= 1$)	519 (771)	227 (195)	121 (89)	132 (69)	195 (70)

The numbers in parentheses denote the expected values.

Table 2. Contingency table of fitness distribution versus protein connectivity in a protein–protein interaction network

Fitness	Protein connectivity (number of interacting partners)						
	0	1–2	2–4	4–6	6–8	8–10	$\geq$10
Weak effect ($0 \leq$ fitness < 0.1)	1092 (940)	796 (725)	635 (640)	203 (235)	70 (107)	35 (63)	51 (172)
Moderate effect ($0.1 \leq$ fitness < 0.5)	538 (565)	418 (436)	408 (384)	132 (141)	65 (64)	44 (38)	125 (103)
Moderate effect ($0.5 \leq$ fitness < 1)	18 (25)	15 (19)	15 (17)	12 (6)	2 (3)	6 (2)	9 (5)
Essential genes (fitness $= 1$)	272 (390)	252 (301)	248 (301)	131 (97)	81 (44)	43 (26)	167 (71)

The numbers in parentheses denote the expected values.

Table 3. Contingency table of fitness distribution versus gene expression cooperativity from microarray data

Fitness	Number of connection in gene-expression profile				
	0	1–10	10–50	50–100	$\geq$100
Weak effect ($0 \leq$ fitness < 0.1)	2333 (2230)	304 (299)	175 (160)	42 (62)	52 (155)
Moderate effect ($0.1 \leq$ fitness < 0.5)	1323 (1332)	160 (178)	92 (95)	34 (37)	126 (93)
Moderate effect ($0.5 \leq$ fitness < 1)	55 (59)	10 (8)	3 (3)	4 (2)	5 (4)
Essential genes (fitness $= 1$)	831 (921)	134 (123)	55 (66)	46 (26)	134 (64)

The numbers in parentheses denote the expected values.

Table 4. Contingency table of fitness distribution versus number of paralogs that a protein in yeast *S.cerevisiae* has

Fitness	Number of paralogs						
	0	1–2	2–4	4–6	6–8	8–10	$\geq$10
Weak effect ($0 \leq$ fitness < 0.1)	1884 (2064)	581 (503)	264 (195)	51 (46)	41 (32)	17 (14)	44 (28)
Moderate effect ($0.1 \leq$ fitness < 0.5)	1278 (1239)	324 (302)	73 (117)	28 (27)	11 (19)	7 (14)	9 (17)
Moderate effect ($0.5 \leq$ fitness < 1)	65 (55)	7 (13)	2 (5)	2 (1)	0 (1)	0 (0)	1 (1)
Essential genes (fitness $= 1$)	986 (855)	115 (208)	60 (81)	12 (19)	13 (13)	5 (6)	3 (12)

The numbers in parentheses denote the expected values.

higher values of fitness (i.e. more likely to be essential). Figure 1C demonstrates that genes with more correlated gene partners (which means more cooperativity with other genes in biological pathways) tend to be less dispensable. Figure 1D shows that a protein having more paralogs in *S.cerevisiae* (or more gene duplications) is more likely to be dispensable. Tables 1–4 are contingency tables of fitness distribution versus number of species in which a protein in yeast *S.cerevisiae* has ortholog hits in the other four species, connectivity in protein–protein interaction network, gene-expression cooperativity from microarray data and the number of paralogs that a protein in yeast *S.cerevisiae* has. Each distribution frequency is compared with the expected value. The expected value at position (i, j) in the table is calculated based on a random distribution according to the following formula:

$$\text{Expect}(i, j) = \text{Sum}(i) * \text{Sum}(j)/\text{Total},$$

where Sum(i) is the sum of all the observed values in the i-th row in the table, Sum(j) is the sum of all the observed values in the j-th column and Total is the sum of all the values in the table. This assumes that there is no correlation between column and row. The observations in Figure 1 are in line with the distributions in Tables 1–4. For example, in Table 2, we can find that the number of essential proteins is less than expected when the number of interacting partners is small and larger than expected when the number of interacting partners is large. Obviously, the hypothesis that the row variable and column variable are independent is not true. Thus, Pearson χ^2-statistics was calculated to test the significant dependence between the rows and columns for each table:

$$G^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - n_i n_j/n)^2}{n_i n_j/n},$$

where n_{ij} denotes the observations in cell (i, j), $n_{i.}$ denotes the sum of observed values of row $i, i = 1, 2, \ldots, r, n_j$ denotes the sum of observed values of column j, $j = 1, 2, \ldots, c$, and n denotes the total of all the values in the table. The χ^2-values for Tables 1–4 are 543, 405, 204 and 177, respectively with P-values <0.001, indicating significant dependence between rows and columns in each table.

The correlation between protein size and protein dispensability is very weak (data not shown). We can conclude that protein size does not have a straightforward trend in terms of determining protein dispensability like other four factors as shown in Figure 1. Nevertheless, protein size might have implicit effect on protein dispensability especially when this feature is combined with other features such as protein evolution rate. Hence, we still included protein size as one input feature in the neural network and SVM. Although the effect of protein size is small, the neural network and SVM can automatically give the protein size a small weight.

3.2 Prediction of protein dispensability

The dependence of protein fitness effects on protein evolution, protein-interaction connectivity, gene-expression cooperativity and gene duplication suggests that it may be possible to predict protein dispensability based on high-throughput data. We found that the dependence is not strong enough to be expressed in an explicit function through techniques such as linear regression. Thus, we applied supervised machine learning methods including neural network and SVM for the prediction. The neural network and SVM were trained to distinguish 'essential effect' from 'non-essential effect' based on the high-throughput data. The test results were evaluated in terms of the performance of sensitivity and specificity:

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}) \quad \text{and}$$

$$\text{Specificity} = \text{TN}/(\text{FP}+\text{TN}),$$

where TP is for true positives, FN is for false negatives, FP is for false positives and TN is for true negatives. For different cutoffs for converting the floating-point values of initial neural network/SVM outputs into the final binary outputs between 'essential effect' and 'non-essential effect', different sensitivity/specificity values can be obtained. To assess the tradeoffs between sensitivity and specificity (Hastie *et al.*, 2001), Figure 2A shows the receiver operating characteristic (ROC) curves of the performance of neural network and SVM for the testing dataset. The neural network has a slightly better performance than the SVM.

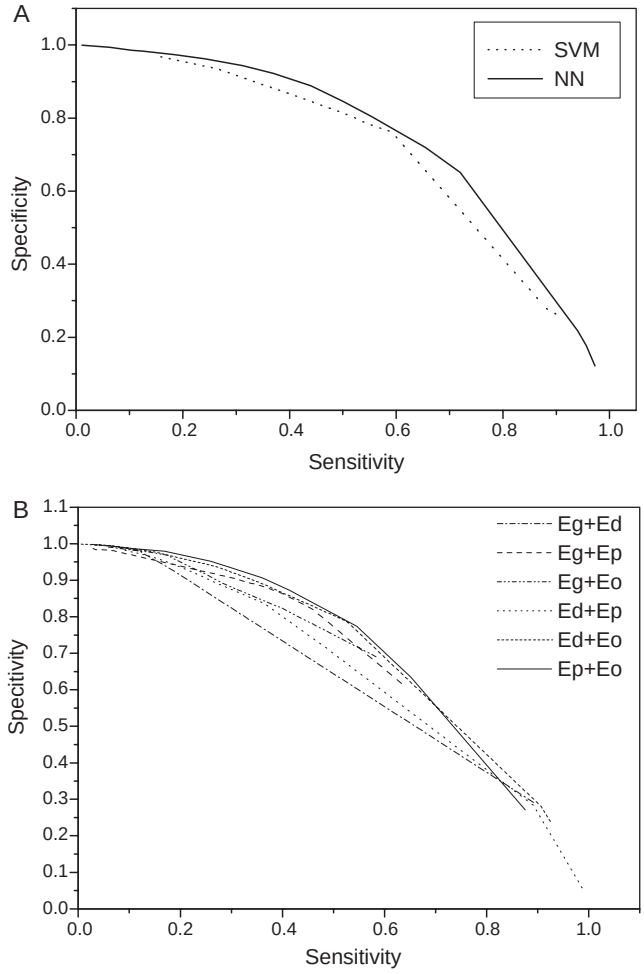

Fig. 2. ROC curves of protein dispensability predictions. (**A**) Sensitivity and specificity of neural network (NN) and SVM. (**B**) Sensitivity and specificity of neural network for different feature combinations (e.g. connectivity of gene-expression data; E_d, gene duplication data; E_p, protein connectivity in protein-interaction network; and E_o, gene conservation rate).

To further understand how different factors contribute to the protein dispensability, we investigated the effect of different combinations of factors on protein dispensability prediction. For every pair of the factors (E_o, E_p, E_g and E_d), i.e. protein evolutionary rate, protein-interaction connectivity, gene-expression cooperativity and gene-duplication rate, respectively, a neural network was trained and tested to get the optimal performance. The data with detail descriptions are available at http://digbio.missouri.edu/~ychen/ProDispen/. As shown in Figure 2B for the ROC curves, the combination that has the highest impact for protein dispensability prediction is $E_o + E_p$, i.e. protein evolutionary rate and protein-interaction connectivity. $E_o + E_d$ has a significant better performance than $E_p + E_d$, while $E_o + E_g$ and $E_g + E_p$ have similar performance. This suggests that E_o is likely to be more important than E_p for protein dispensability. Based on such arguments, we can rank the relative importance for protein dispensability, from high to low, as E_o, E_p, E_g, and E_d. This result is consistent with the statistical analysis of contingency Tables 1–4. The χ^2-values for Tables 1–4, which correspond to as E_o,

Table 5. Protein dispensability, connectivities in protein-interaction network and gene conservation rates of components of the Anaphase-Promoting Complex/Cyclosome (APC/C)

ID	Gene	Species where ortholog found	Protein connectivity	Fitness value
YDL008W	*APC11*		10	1
YNL172W	*APC1*		10	1
YOR249C	*APC5*		10	1
YDR118W	*APC4*		10	1
YKL022C	*CDC16*	*S.pombe*; *D.melanogaster*	10	1
YBL084C	*CDC27*		10	1
YLR127C	*APC2*	*A.thaliana*; *S.pombe*	12	1
YLR102C	*APC9*		13	0.1
YFR036W	*CDC26*		12	0.05
YHR166C	*CDC23*	*A.thaliana*; *D.melanogaster*; *S.pombe*; *C.elegans*	15	1

E_p, E_g and E_d are 543, 405, 204, and 177, respectively as showed in Section 3.1. The larger χ^2-value, the stronger dependence underlying the data. Hence, the χ^2-analysis gives the same order for E_o, E_p, E_g and E_d as the ROC analysis in terms of their relative importance for protein dispensability.

Our studies showed that gene conservation rate is the most important factor to determine protein dispensability. The more ancient a gene is, the more likely it is essential. This is also supported by a pilot gene deletion project of *S.pombe* (Decottignies *et al*., 2003). On the other hand, gene dispensability is also closely correlated with functionality. Some organism-specific genes, which may play important roles in cell life, can be essential genes too. This aspect may be characterized by protein–protein interaction network. Protein–protein interaction network is tolerant to error but highly vulnerable to attack on the highly connected proteins (Albert *et al*., 2000; Barabasi and Albert, 1999). Highly connected proteins constitute functional units essential for the life of cell and they have high likelihood to be essential genes. This is particularly the case for some molecular machines that consist of multiple protein units. Missing one unit can cause the instability of the entire molecular machine structure and thus the loss of function. Table 5 shows an example of the protein dispensability, conservation rate and protein connectivity for the components of the Anaphase-Promoting Complex/Cyclosome, which is an ubiquitin ligase complex that degrades mitotic cyclins and anaphase inhibitory protein, thereby triggering sister chromatid separation and exiting from mitosis. In this complex, the components are highly connected as shown in Table 5. Most of the components are essential proteins although they are not conserved in evolution. In other words, gene conservation rate and protein functionality capture the different aspects of protein dispensability. Such effect may not be completely captured by our current study, and is subjected to future investigation.

4 DISCUSSION

We have studied the protein dispensability at the genome scale by the integrated analyses of high-throughput data. We have shown the dependences of protein dispensability on protein-evolution rate, protein-interaction connectivity, gene-duplication rate and gene-expression cooperativity, in an order of decreasing importance. Moreover, we provided a framework for predicting protein dispensability based on such dependences. The approach described in this study is most probably applicable to other organisms for which various high-throughput data are becoming available (Brown and Balling, 2001). Although gene deletion strains in yeast *S.cerevisiae* are available and phenotype assay has been performed for all genes, we expect that protein dispensability prediction will be particularly useful for organisms in which mutant strains are less available and more difficult to assay (Brown *et al*., 1996).

It has been proposed that essential genes evolve more slowly by the adaptive theory of mutation rates (Tourasse and Li, 2000). However, Hurst and Smith (1999) argued that in rat and mouse essential and non-essential genes evolve at the same rate. These observations suggested that sequence mutation rate might not be the most biologically relevant measure of the evolutionary conservation of gene. In our studies, we explored an alternative, i.e. the number of species in which budding yeast *S.cerevisiae* has orthologous hits through comparison of multiple complete genomes using the reciprocal search method. The conservation rate based on such an approach correlates well with the dispensability of a gene, and in fact, it is the most important factor among the four that we considered. Our method is similar to an previous study by Krylov *et al*. (2003), who introduced the concept of propensity for gene loss (PGL) to measure gene conservation rate to study the correlations among gene loss, protein sequence divergence and gene dispensability. However, they did not apply any machine learning technique (neural network or SVM) in predicting gene dispensability, as we did.

Although we quantified the general dependences of protein dispensability on different factors, the actual relationship between protein dispensability and these factors in individual cases can be complicated. Our studies indicated that gene conservation rate and protein-interaction connectivity are the most important factors to determine protein dispensability. This does not contradict the results of Papp *et al*. (2003, 2004) that protein dispensability is not the results of selection to favor resilience but the consequence of environmental specificity. Various interaction networks (i.e. regulation, coexpression and metabolism) are abstract representations of biological functionality and genes/proteins are organized in a dynamic fashion within these dynamic networks. It is possible that a change of environmental condition triggers certain interactions of proteins or correlated gene expression, leading to a detected fitness change. Further systems-level understanding of the organization and dynamics of protein interactions will shed lights on mechanistic basis of protein dispensability.

The dispensability is related not only to the connectivity of protein physical interactions, but also to the connectivity of genetic interactions. Genetics interactions reveal network components performing related functions or connecting pathways that converge on the same essential endpoint of functionality (Ozier *et al*., 2003). Genetic interactions can be mapped in a large scale by synthetic genetic array (SGA) analysis in which two single deletions that cause no evident phenotype individually are lethal in combination. A comprehensive identification of synthetic lethal interactions in budding yeast was conducted by crossing mutations in 132 query genes with the complete set of 4800 viable yeast gene deletion mutants (Tong *et al*., 2004). In Figure 3, we measured the relationship between protein

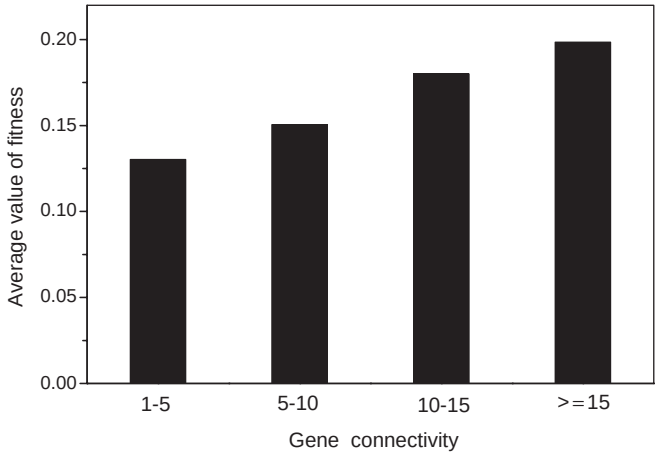

Fig. 3. The fitness (indispensability) of 132 query genes in yeast *S.cerevisiae* versus gene connectivity in the genetic interaction network determined through the SGA analysis. Gene connectivity is measured by the node degree (number of interactions) in the network.

dispensability and gene connectivity (the node degree of components in genetic interaction network) for these 132 query genes. It shows that highly connected 'hub genes' are more important for fitness, as they may participate in more biological activities, and thus they are more essential to cell life. It has been argued that increased functional connectivity of network underlies the evolution of more complex species (Pawson and Nash, 2003). The study on the relationship between dispensability and genetic interactions is ongoing, which may shed some lights on gene dispensability from a different perspective.

ACKNOWLEDGEMENTS

We would like to thank the two anonymous reviewers for helpful comments and suggestions. This research was sponsored in part by the US Department of Energy's Genomes to Life program (www.doegenomestolife.org) under project 'Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling' (www.genomes-to-life.org). It was also partially funded by Nation Science Foundation (EIA-0325386).

REFERENCES

Albert,R., Jeong,H. and Barabasi,A. (2000) Error and attack tolerance of complex networks. *Nature*, **406**, 378–381.

Barabasi,A.L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Brown,S.D. and Balling,R. (2001) Systematic approaches to mouse mutagenesis. *Curr. Opin. Genet. Dev.*, **11**, 268–273.

Brown,J.R., Ye,H., Bronson,R.T., Dikkes,P. and Greenberg,M.E. (1996) A defect in nurturing in mice lacking the immediate early gene *fosB*. *Cell*, **86**, 297–309.

Decottignies,A., Sanchez-Perez,I. and Nurse,P. (2003) *Schizosaccharomyces pombe* essential genes: a pilot study. *Genome Res.*, **13**, 399–406.

Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Cell. Biol.*, **11**, 4241–4257.

Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al*. (1996) Life with 6000 genes. *Science*, **274**, 563–567.

Gu,Z., Steinmetz,L.M., Gu,X., Scharfe,C., Davis,R.W. and Li,W.H. (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature*, **421**, 63–66.

Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, NY.

Hirsh,A.E. and Fraser,H.B. (2001) Protein dispensability and rate of evolution. *Nature*, **411**, 1046–1049.

Hurst,L.D. and Smith,N.G. (1999) Do essential genes evolve slowly. *Curr. Biol.*, **9**, 747–750.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jansen,R., Greenbaum,D. and Gerstein,M. (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res.*, **12**, 37–46.

Jeong,H., Mason,S.P., Barabasi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Jordan,I.K., Wolf,Y.I. and Koonin,E.V. (2003) No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol. Biol.*, **3**, 1–6.

Joshi,T., Chen,Y., Becker,J.M., Alexandrov,N. and Xu,D. (2004a) Cellular function prediction for hypothetical proteins in yeast *Saccharomyces cerevisiae* using multiple sources of high-throughput data, *Proceedings of the World Multi-Conference on Systemics, Cybernetics and Informatics*, Vol. IX, pp. 17–20.

Joshi,T., Chen,Y., Becker,J.M., Alexandrov,N. and Xu,D. (2004b) Genome-scale gene function prediction using multiple sources of high-throughput data in yeast *Saccharomyces cerevisiae*. *OMICS* (in press).

Krylov,D.M., Wolf,Y.I., Rogozin,I.B. and Koonin,E.V. (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.*, **13**, 2229–2235.

Lawton-Rauh,A. (2003) Evolutionary dynamics of duplicated genes in plants. *Mol. Phylogenet. Evol.*, **29**, 396–409.

Li,W.H. (1997) *Molecular Evolution*, 1st edn. Sinauer Associates Inc., Sunderland, MA.

Mathews,B. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–455.

Ohta,T. (1973) Slightly deleterious mutant substitutions in evolution. *Nature*, **246**, 96–98.

Ozier,O., Amin,N. and Ideker,T. (2003) Global architecture of genetic interactions on the protein network. *Nat. Biotechnol.*, **21**, 490–491.

Pal,C., Papp,B. and Hurst,L.D. (2003) Genomic function: rate of evolution and gene dispensability. *Nature*, **421**, 496–497.

Papp,B., Pal,C. and Hurst,L.D. (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature*, **424**, 194–197.

Papp,B., Pal,C. and Hurst,L.D. (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, **429**, 661–664.

Pawson,T. and Nash,P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.

Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.

Ryden,L.G. and Hunt,L.T. (1993) Evolution of protein complexity: the blue copper-containing oxidases and related proteins. *J. Mol. Evol.*, **36**, 41–66.

Schölkopf,B., Smola,A., Williamson,R. and Bartlett,P.L. (2000) New support vector algorithms. *Neural Comput.*, **12**, 1207–1245.

Steinmetz,L.M., Scharfe,C., Deutschbauer,A.M., Mokranjac,D., Herman,Z.S., Jones,T., Chu,A., Giaever,G., Prokisch,H., Oefner,P.J. and Davis,R.W. (2002) Systematic screen for human disease genes in yeast. *Nat. Genet.*, **31**, 400–404.

Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.

Thatcher,J.W., Shaw,J.M. and Dickinson,W.J. (1998) Marginal fitness contributions of nonessential genes in yeast. *Proc. Natl Acad. Sci. USA*, **95**, 253–257.

Tong,A.H., Lesage,G., Bader,G.D., Ding,H., Xu,H., Xin,X., Young,J., Berriz,G.F., Brost,R.L., Chang,M. *et al*. (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.

Tourasse,N.J. and Li,W.H. (2000) Selective constraints, amino acid composition, and the rate of protein evolution. *Mol. Biol. Evol.*, **17**, 656–664.

Uetz,P., Giot,I., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Wood,V., Gwilliam,R., Rajandream,M.A., Lyne,M., Lyne,R., Stewart m,A., Sgouros,J., Peat,N., Hayles,J., Baker,S. *et al*. (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.

Zell,A., Mache,N., Sommer,T. and Korb,T. (1993) The SNNS Neural Network Simulator, GWAI-91, 15. Fachtagung für Künstliche Intelligenz, Bonn, Informatik-Fachberichte. Springer-Verlag, NY, Vol. 285, pp. 254–263.