

Prediction of Protein Dispensability through Integrated Analysis of Multiple-Source High-Throughput Data

Yu Chen^{1,2} and Dong Xu^{1,2*}

¹UT-ORNL Graduate School of Genome Science and Technology, Oak Ridge, TN, USA.

²Digital Biology Laboratory, Computer Science Department, University of Missouri-Columbia, Columbia, MO, USA.

* Correspondence author: xudong@missouri.edu

Abstract

*Protein dispensability is fundamental to understanding of gene function and evolution. It is usually studied at the individual gene phenotype level. 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. In our studies, protein dispensability was represented as a fitness score that was measured by the growth rate of gene-deletion mutants. Through analyses of high-throughput data in yeast *Saccharomyces cerevisia*, 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. Correspondence analysis also showed such significant dependencies, which imply that the integration of high-throughput data can provide substantial information on protein dispensability. Thus neural network and support vector machines were applied to predict protein dispensability. Our study provides a "proof-of-principle" for a global understanding of protein dispensability through computational analyses of high-throughput data.*

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 through bioengineering. In gene "knockout" experiments, "essential" and "nonessential" are two classical molecular genetics designations referring to the significance of a gene with respect to its effect on fitness in an organism¹. A gene is considered to be essential if deleting it results in lethality. On the other hand, nonessential genes are those for which knockouts do not kill the organism. Essential genes are less functionally dispensable or redundant than those nonessential genes. Deletions of different nonessential

genes have different effects on the evolution and population (growth) of the organism carrier. A deletion of a nonessential gene might give the carrier a selective disadvantage and this carrier is likely to be removed from the population over time. Such selection is called purifying selection². Given the role of purifying selection in determining evolution, nonessential genes that are subject to weaker purifying selection should have a higher rate of evolution³. The effect of gene deletion can be characterized using protein dispensability, i.e., to what extent the deletion of a protein's gene can affect the viability of the organism in terms of its growth. Early study showed that protein dispensability and evolutionary rate are correlated⁴. However, the relationship between protein's dispensability and its evolution has not been studied in details and could be complex since a variety of factors are involved⁵.

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. That is, a vast majority of proteins have only limited interactions⁶. Recently, Jordan *et al.*⁷ suggested that the correlation between a protein's evolutionary rate and its number of protein interactions arises only because a few, highly interactive proteins evolve more slowly than all other proteins. Gene expression level is used frequently to infer the activity and function of gene products. Co-expressed gene are usually involved in the same pathway or similar cellular function. Thus, we expected that gene expression profiles might reflect the dispensability of gene products. Gene duplication, arising from region-specific duplication or genome-wide polyploidization, is an important feature in genome evolution⁸. However, the relationship between the role of duplicate genes⁹ and protein dispensability remained unknown.

Recent 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 a scale that is extended from individual proteins to the proteome. In this study, we conducted the integrated analyses to understand the dependence of protein dispensability on protein evolutionary rate, protein connectivity, gene-expression connectivity 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 dependencies, we applied machine-learning methods, i.e., neural network and support vector machine, to predict protein dispensability from high-throughput data.

2. Method

2.1 Integration of high-throughput data

Data preparation

In our analyses we incorporated several sources of high-throughput data, including genomic sequence and annotation data, protein-protein interaction data, gene-expression data and yeast mutant growth-rate data. The dispensability of a protein can be quantified using the growth rate based on its contribution to survival and reproduction upon gene deletion. For each mutant, we estimated the deleted gene's fitness, 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 range of fitness value is between 0 and 1. The fitness values of essential genes deletion strains are 1.

A protein-protein interaction network can be viewed as a weighted non-directed graph $G_p(D) = (V_p, E_p)$. The vertex set $V_p = \{d_i | d_i \in D\}$ and the edge set $E_p = \{(d_i, d_j) | \text{for } d_i, d_j \in D \text{ and } i \neq j\}$. 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 connectivity graph was constructed as $G_g(D) = (V_g, E_g)$. The vertex set $V_g = \{d_i | d_i \in D\}$ and the edge set $E_g = \{(d_i, d_j) | \text{for } d_i, d_j \in D \text{ and } i \neq j \text{ and } |r_{ij}| = 0.7\}$. Each vertex represents one gene and each edge represents one gene pair whose gene expression profiles correlation coefficient $|r_{ij}| = 0.7$. The cutoff value of $|r_{ij}|$ is determined based on our previous study¹⁰.

Data resources

We downloaded the genomic sequence 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

protein-protein interaction data from high-throughput yeast

two-hybrid interaction experiments were from Uetz et al¹¹

and Ito et al¹². There were 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 binary

interactions among 4067 proteins were used in this study.

The gene-expression profiles of microarray data were from

Gasch et al¹³. The growth rates of yeast gene deletion

mutants were measured in the genome-scale, where 4706

homozygous diploid deletion strains were monitored in

parallel in 9 different medium conditions. The data were

obtained from the public database at

http://www-deletion.stanford.edu/YDPM/YDPM_index.html.

We used the average growth rate over 9 conditions in

our study.

2.2 Identification of putative orthologs and paralogs using reciprocal searching

An all-against-all FASTA search was conducted for the

whole set of *S. cerevisiae* protein sequences to identify the

putative orthologs in fission yeast *Schizosaccharomyces*

pombe, *Arabidopsis thaliana*, *Drosophila melanogaster*

and *Caenorhabditis elegans*. A subclass of putative

orthologs are defined as reciprocal best hits with additional

two strict criteria: (1) FASTA¹⁴ expectation value is less

than 10^{-10} and (2) The FASTA-alignable region between two

proteins is longer than 80% of the protein length in yeast *S.*

cerevisiae. The reciprocal process is done in the following

way: To identify the orthologs in fission yeast *S. pombe*, we

first queried one budding yeast ORF i against all 4940 open

reading frames (ORFs) predicted to be protein coding genes

in *S. pombe*¹⁵ to yield the set of hits $\{W\}$. Then we queried

the fission yeast ORF j against all 6217 ORFs in budding

yeast¹⁶ to yield the set of hits $\{Y\}$. Finally the protein pair

$\{ORF\ i, ORF\ j\}$ were considered to be putative orthologs

only if i is the member of $\{Y\}$ with the lowest expectation

value and j is the member of $\{W\}$ with the lowest expectation

value. Similarly we identified duplicated genes, which are

often referred to as paralogous genes through an

all-against-all FASTA search for the whole set of *S.*

cerevisiae protein sequences, with the two criteria but not

the reciprocal process.

2.3 Correspondence analysis

Correspondence analysis is an exploratory technique that uses singular value decomposition to analyze two-way contingency tables¹⁷. In a typical correspondence analysis, a cross-tabulation table of frequencies for categorical variables is calculated. Suppose the variable X is to have a range of values consisting of r categories and the variable Y is to have c categories, the cell density of cell (i, j) in a (r×c) contingency table is denoted by $o_{ij} = n_{ij}/n$, where n_{ij} denotes the sample frequency in cell(i, j). The (r×c) matrix of cell densities is denoted by O and is called correspondence matrix. The (r×1) vector of row marginals $o_{i.}$, $i = 1, 2, \dots, r$ is denoted by r and similarly the (c×1) vector of column marginals $o_{.j}$, $j = 1, 2, \dots, c$, is denoted by c. These row and column vectors can be written as $r = Oe_c$ and $c = Oe_r$, where e_c (c×1) and e_r (r×1) are vectors of unities. The vectors r and c are also referred to respectively as row and column masses. Diagonal matrices constructed from the row and column masses are denoted by D (r×r) and Dc (c×c), respectively. The matrix difference (O - rc') is a measure of the deviation from independence. For the matrix (O - rc'), the generalized singular value decomposition subject to the conditions $A'D_r^{-1}A = I$, $B'D_c^{-1}B = I$ is given by

$$(O - rc') = AD_m B' = \sum_{k=1}^K m_k a_k b_k'$$

where the columns of A (r×K) and B (c×K) are denoted by a_k and b_k respectively and $\mu_1, \mu_2, \dots, \mu_k$ are the diagonal elements of the matrix D_k (K×K) and called the singular values of (O - rc'). The total inertia can therefore be written as

$$\text{tr}[D_r^{-1}(O - rc')D_c^{-1}(O - rc')] = \sum_{k=1}^K m_k$$

The coordinates for the row and column profile deviations are given by $V(r \times K) = D_r^{-1}AD_m$ and $W(c \times K) = D_c^{-1}BD_m$ respectively. Correspondence analysis was implemented by using Matlab.

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 connectivity and gene-duplication data. Neural network and support vector machines were used to extract features of the binary classes

(essential genes and non-essential genes) in the training process and the trained models were used to predict protein dispensability. The data size is 5878. We randomly selected 70% of the data for training and used the remaining 30% of the data for testing. There were 8 input units including number of species out of four species, where yeast proteins had ortholog hits, the sequence identity of one pair of putative orthologs, the number of interacting protein interaction partners, gene expression connectivity, the number of paralogs and protein size. For neural network the back-propagation learning algorithm and the logistic activation function were used. This neural network had one hidden layer with 8 hidden units. For SVM, the polynomial kernel was used and we used two parameters C_r and C_c to tradeoff the generalization ability and miss-classification error of the unbalanced data. The software packages we used were SNNS 4.2 (Stuttgart Neural Network Simulator)¹⁸ and LIBSVM 2.4¹⁹.

3. Results

3.1 Integrated analysis of protein dispensability

To characterize the properties of protein dispensability, we investigate the relationship 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 connectivity, and 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. Protein dispensability was quantified by using the growth rates of gene deletion mutants and represented as a fitness value within the range [0-1]. The relevant variables derived from sequence data, protein-protein interaction data, and gene expression data are:

- $X_{E_o} = \{X_{O_i} : O_i \in E_o\}$, distribution of proteins with different evolutionary rates in yeast *S. cerevisiae*. $E_o = \{0, 1, 2, 3, 4\}$ represents the number of species which contain orthologs of a given protein in budding yeast *S. cerevisiae* using reciprocal searching methods against *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, *Drosophila melanogaster* and *Caenorhabditis elegans* (see Section 2.2). X_{O_i} represents 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 in yeast *S. cerevisiae*. X_{P_i} is the number of proteins with the node degree (number of interactions) P_i of a given protein in the interaction graph, and $E_p = \{0, 1, 2, 3, 4, \dots\}$.

- $X_{E_g} = \{X_{G_i} : G_i \in E_g\}$, distribution of proteins (genes) with different connectivities in the gene expression graph of the microarray data in 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 $E_g = \{0, 1, 2, 3, 4, \dots\}$.
- $X_{E_d} = \{X_{D_i} : D_i \in E_d\}$, 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 $E_d = \{0, 1, 2, 3, 4, \dots\}$.

We measured the relationships between protein dispensability and distributions of X_{E_o} , X_{E_p} , X_{E_g} and X_{E_d} (see Figure 1). Figure 1A shows that the average fitness of deleting 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 proteins involved in more interactions have higher values of fitness (i.e., more likely to be essential). Figure 1C demonstrates that genes with more strong correlated gene partners (which means more coordination 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.

3.2. Correspondence analysis of dependency of protein dispensability

It was shown in section 3.1 that protein dispensability correlated with protein evolutionary rate, protein connectivity, gene-expression connectivity, and gene-duplication rate. We further used correspondence analysis to statistically identify the major trends in the variation of data. Moreover, correspondence analysis simultaneously characterizes the relationship between the rows and columns of the data matrix. Correspondence

analysis requires contingency tables, which are the counts of the joint occurrences of rows and columns in a table. In tables 1-4 rows represented fitness data categories and columns represented the categorical variables of the number of species in which *S. cerevisiae* has ortholog hits, protein connectivity, gene expression connectivity and duplication data, respectively. We also calculated the expected frequency of each cell (i,j). If rows and columns are independent, the expected joint occurrences can be expressed as the product of the corresponding marginal densities f_i and f_j multiplied by the total number of observations N . Obviously, the rows and column are not independent. Thus, Pearson Chi-square statistics was calculated to test the significant dependency between the rows and column for each table:

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

This equation can be also expressed in the form:

$$G^2 = n \times \text{tr} \left[D_r^{-1} (O - rc') D_c^{-1} (O - rc') \right]$$

The Chi-square values for tables 1-4 are 542.8, 405.4, 204 and 176.8 respectively with p-values < 0.001, indicating the significant dependency between rows and columns of each table. Moreover, the relationships among row variables and column variables can be measured as the distances when such variables were coordinated in the principle axes (Fig. 2). Each factorial axis represents a fraction of the whole information contained in the analyzed matrix. Figure 2 shows the scatter plots of correspondence analysis of fitness data and number of species with orthologs, protein connectivity, gene-expression connectivity and the number of paralogs, respectively. We can find that different effects of fitness have obviously different profiles, whose relationships with column variables can be easily recognized. The more a protein is essential, the slower evolution rate and the higher protein connectivity, gene expression connectivity, and lower duplication rate.

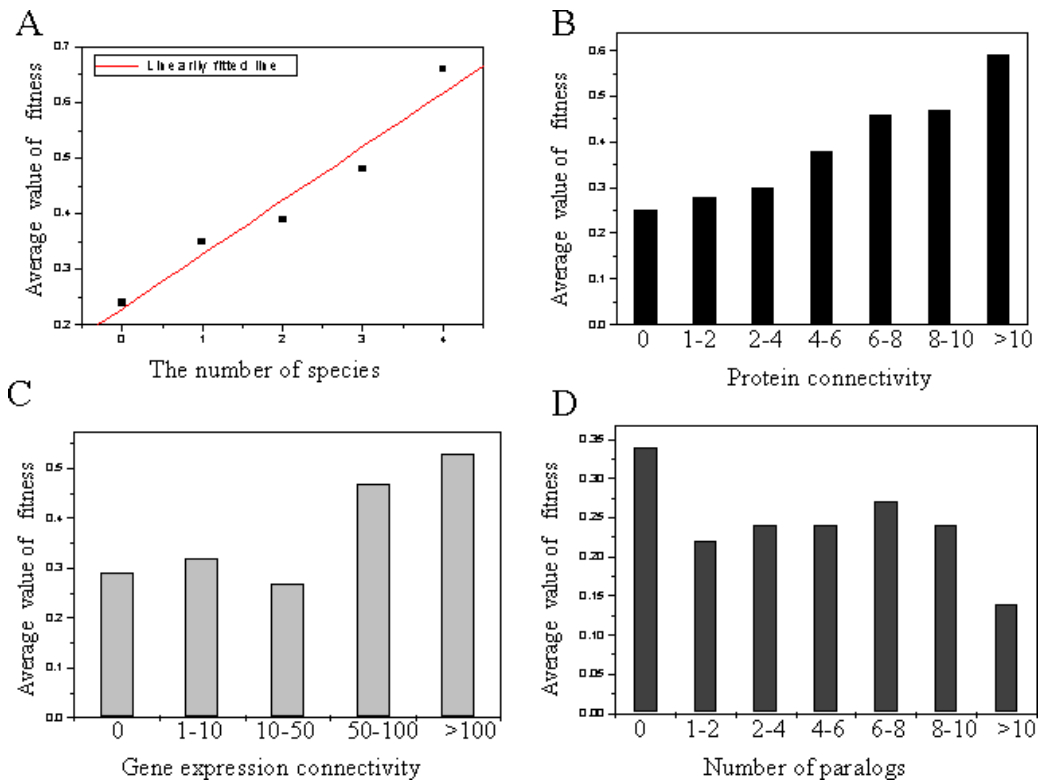


Figure 1: The fitness (indispensability) of a protein in yeast versus (A) number of other species having homologs of the protein, (B) protein connectivity in the protein-protein interaction network, (C) protein connectivity in the gene expression network, and (D) number of paralogs in yeast.

Table 1: Contingency table of occurrence distribution for fitness versus the number of species in which a yeast protein has ortholog hits. The number in the parenthesis is the expected value.

Fitness	Number of species out of four species				
	0	1	2	3	4
Weak effect (0 =< fitness < 0.1)	2106 (1859)	427 (471)	168 (214)	117 (166)	60 (168)
Moderate effect (0.1 =< fitness < 0.5)	1135 (1117)	294 (283)	142 (129)	81 (99)	77 (101)
Moderate effect (0.5 =< 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)
----------------------------------	-----------	-----------	----------	----------	----------

Table 2: Contingency table of occurrence distribution for fitness versus protein connectivity in protein-protein interaction network. The number in the parenthesis is the expected value.

Fitness	Protein connectivity (number of interacting partners)						
	0	1-2	2-4	4-6	6-8	8-10	=10
Weak effect (0 =< fitness < 0.1)	1092 (940)	796 (725)	635 (640)	203 (235)	70 (107)	35 (63)	51 (172)
Moderate effect (0.1 =< fitness < 0.5)	538 (565)	418 (436)	408 (384)	132 (141)	65 (64)	44 (38)	125 (103)
Moderate effect (0.5 =< 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)

Table 3: Contingency table of occurrence distribution for fitness versus gene expression connectivity from microarray data. The number in the parenthesis is the expected value.

Fitness	Number of connection in gene-expression profile				
	0	1-10	10-50	50-100	=100
Weak effect (0 =< fitness < 0.1)	2333 (2230)	304 (299)	175 (160)	42 (62)	52 (155)
Moderate effect (0.1 =< fitness < 0.5)	1323 (1332)	160 (178)	92 (95)	34 (37)	126 (93)
Moderate effect (0.5 =< 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)

Table 4: Contingency table of occurrence distribution fitness versus gene-duplication rate. The number in the parenthesis is the expected value.

	Number of paralogs						
	0	1-2	2-4	4-6	6-8	8-10	= 10
Weak effect ($0 \leq \text{fitness} < 0.1$)	1884 (2064)	581 (503)	264 (195)	51 (46)	41 (32)	17 (14)	44 (28)
Moderate effect ($0.1 \leq \text{fitness} < 0.5$)	1278 (1239)	324 (302)	73 (117)	28 (27)	11 (19)	7 (14)	9 (17)
Moderate effect ($0.5 \leq \text{fitness} < 1$)	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)

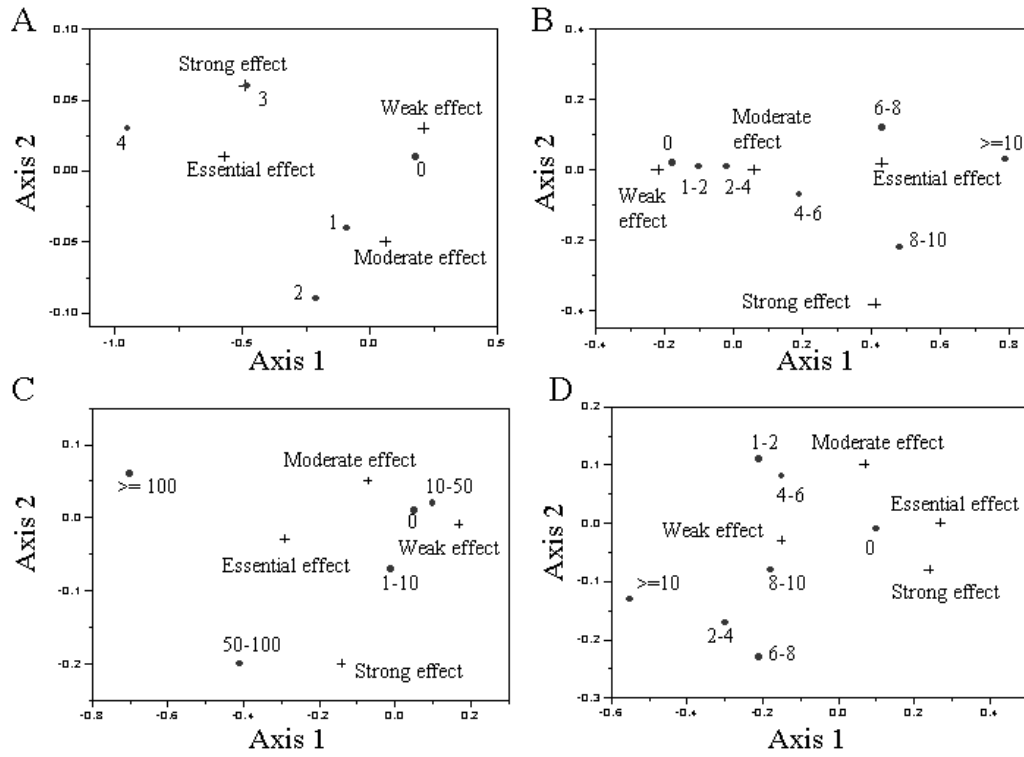


Figure 2: Correspondence analysis scatter plot of data in the first two principal dimensions. A. The data for fitness effect and the number of species with orthologs. For axes 1 and 2 the percentages of total inertia are 98.60% and 1.38%. 0, 1, 2, 3 and 4 mean the number of species in which a yeast protein has ortholog hit. B. The data for fitness effect and protein-protein interaction connectivity. For axes 1 and 2 the percentages of total inertia are 95.60% and 2.32%. C. The data for fitness effect and the number of gene connection in microarray data. For axes 1 and 2 the percentages of total inertia are 95.10% and 4.82%. D. The data for fitness effect and the number of paralogs. For axes 1 and 2 the percentages of total inertia are 81.7% and 16.5%.

3.3 Prediction of protein dispensability

The dependence of protein fitness effect on protein evolution, protein interaction connectivity, gene expression connectivity and gene duplication suggests that it may be possible to predict protein dispensability based on high-throughput data. We applied supervised machine learning methods including neural network and support vector machine (SVM) for the prediction. Each gene deletion was labeled as “essential effect” and “non-essential effect” and supervised classifiers can be trained to capture the ability to distinguish “essential effect” from “non-essential effect” examples. The input units are integrated features on protein evolution rate,

protein connectivity, gene expression connectivity and gene duplication rate. The output of SVM is a binary value while the output of a neural network is a score whose range is from 0 to 1. Each trained model was evaluated using threefold cross-validation and ranked by the Matthews correlation coefficient²⁰. Then the best model was tested by test dataset. The testing results were evaluated in terms of the performance of sensitivity and specificity:

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}); \quad \text{Specificity} = \text{TN}/(\text{FP}+\text{TN})$$

where TP: true positives; FN: false negatives; FP: false positives; TN: true negatives. Figure 3A is the performance of neural network and SVM. The overall accuracy of prediction can reach 80%.

We also conducted the three-class prediction of protein dispensability using neural network. The fitness data were grouped into three classes: essential effect (fitness value = 1); weak effect (fitness value < 0.1); and intermediate effect (0.1 = fitness value < 1). We partition the k class problem into a series of two-class problem: one class contains data in one 'true' class, and the 'others' class combines all other classes. This procedure is repeated for each of the k classes, leading to k two-way trained predictors. Figure 3B is the performance of the three classes. An independent dataset (612 fitness data) was used to test the three-class prediction result. The standard Q percentage accuracy was calculated. Suppose we have $N =$

$n_1 + n_2 + \dots + n_k$ data points (n_1 are observed to belong to class F_1 etc). Out of n_1 observations, c_1 are correctly and uniquely recognized as belonging to F_1 etc., so that total $C = c_1 + c_2 + \dots + c_k$ are correctly recognized. The overall accuracy is $Q = C/N$. For three-class prediction, Q is 56%. The prediction accuracy of intermediate effect class was less than 10%. We found that most of observations in intermediate effect class were predicted to being into weak effect class. This suggested the existence of some unknown factors affecting protein dispensability, which will be further investigated to fully characterize protein dispensability.

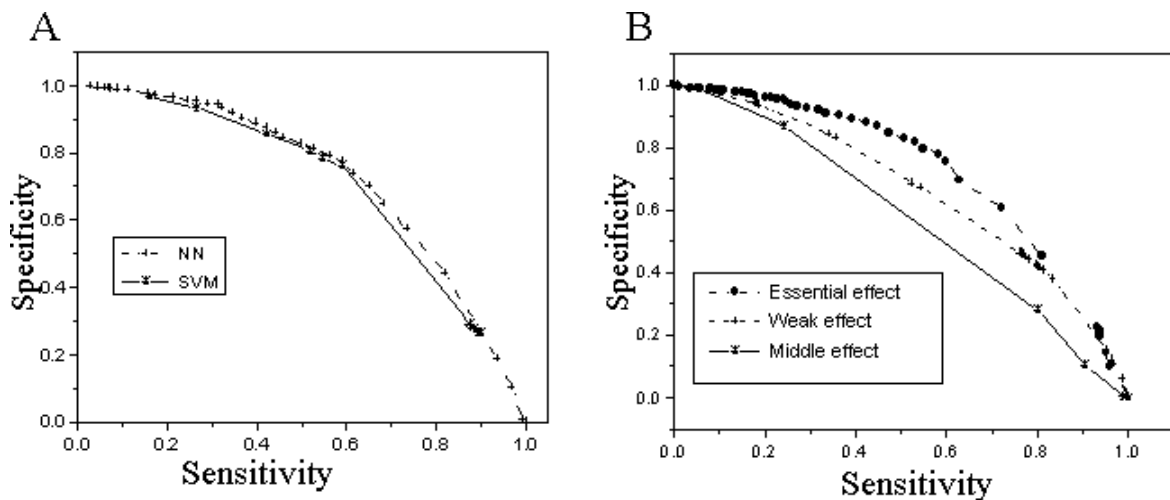


Figure 3. Performance of protein dispensability prediction. A. Sensitivity and specificity of neural network and support vector machine for one-class prediction. B. Sensitivity and specificity of neural network for three-class prediction.

4 Discussion

Our studies have provided some global insights into protein dispensability by the integrated analyses of high-throughput data. We have shown the dependencies of protein dispensability on protein-evolution rate, protein-interaction connectivity, gene-expression connectivity and gene-duplication rate. Moreover, we provided proof of principle for that protein dispensability can be predicted based on such dependencies. The approach described in this study will most likely be applicable to other organisms for which various high-throughput data are becoming available²¹. Although yeast gene deletion strains are available and phenotype assay has been performed for all genes, we expected that

protein dispensability prediction will be particularly useful for organisms in which mutant strains are less available and more difficult to assay²².

Acknowledgment

This research was sponsored 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).

Reference:

- [1] Hurst, LD and Smith, NG, "Do essential genes evolve slowly", *Curr. Biol.*, **9**: 747-750, 1999.
- [2] Li, WH, *Molecular Evolution* 1st edition, Sinauer Associates Inc., 1997.
- [3] Ohta, T, "Slightly deleterious mutant substitutions in evolution", *Nature*, **246**:96-98, 1973.
- [4] Hirsh AE and Fraser HB, "Protein dispensability and rate of evolution", *Nature*. **411**:1046-1049, 2001.
- [5] Pal C, Papp B and Hurst LD, "Genomic function: Rate of evolution and gene dispensability", *Nature*. **421**:496-497, 2003.
- [6] Jeong H, Mason SP, Barabasi AL, Oltvai ZN, "Lethality and centrality in protein networks", *Nature*. **411**:41-42, 2001.
- [7] Jordan IK, Wolf YI, Koonin EV, "No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly", *BMC Evolutionary Biology* **3**:1-6, 2003
- [8] Lawton-Rauh A, "Evolutionary dynamics of duplicated genes in plants", *Molecular Phylogenetics and Evolution*, **29**: 396-409, 2003.
- [9] Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW and Li WH, "Role of duplicate genes in genetic robustness against null mutations", *Nature*, **421**:63-66, 2003.
- [10] Joshi T, Chen Y, Becker JM, Alexandrov N and Xu D, "Cellular Function Prediction for Hypothetical Proteins in Yeast *Saccharomyces cerevisiae* Using Multiple Sources of High-Throughput Data", submitted.
- [11] Uetz P, Giot I, Cagney G, et al, "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*", *Nature*, **403**: 623-627, 2000.
- [12] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y, "A comprehensive two-hybrid analysis to explore the yeast protein interactome", *PNAS*, **98**:4569-4574, 2001.
- [13] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D and Brown PO, "Genomic expression programs in the response of yeast cells to environmental changes", *Mol Cell Biol*, **11**: 4241-4257, 2000.
- [14] Pearson WR, "Flexible sequence similarity searching with the FASTA3 program package", *Methods Mol. Biol*. **132**:185-219, 2000.
- [15] Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, et al, "The genome sequence of *Schizosaccharomyces pombe*", *Nature*, **415**:871-880, 2002.
- [16] Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H et al., "Life with 6000 genes", *Science*, **274**: 563-567, 1996.
- [17] Greenacre MJ, "Theory and applications of correspondence analysis", Academic Press, London, 1984.
- [18] Zell A, Mache N, Sommer T, and Korb T, "The SNNs Neural Network Simulator", *GWAF91*, 15. Fachtagung für Künstliche Intelligenz, Bonn, Informatik-Fachberichte 285, Springer, 254- 263, 1993.
- [19] Schölkopf B, Smola A, Williamson R, and Bartlett PL, "New support vector algorithms", *Neural Computation*, **2**:1207-1245, 2000.
- [20] Mathews B, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme", *Biochim. Biophys. Acta*, **405**: 442-455, 1975.
- [21] Brown SD and Balling R, "Systematic approaches to mouse mutagenesis", *Curr Opin Genet Dev*. **11**:268-73, 2001.
- [22] Brown JR, Ye H, Bronson RT, Dikkes P, Greenberg ME, "A defect in nurturing in mice lacking the immediate early gene *fosB*", *Cell*, **86**:297- 309, 1996.