BMB
Reports

Partial AUC maximization for essential gene prediction using genetic algorithms

Kyu-Baek Hwang[1], Beom-Yong Ha[1], Sanghun Ju[1, †] & Sangsoo Kim[2,]*

[1]School of Computer Science and Engineering, Soongsil University, [2]School of Systems Biomedical Science, Soongsil University, Seoul 156-743, Korea

Identifying genes indispensable for an organism's life and their characteristics is one of the central questions in current biological research, and hence it would be helpful to develop computational approaches towards the prediction of essential genes. The performance of a predictor is usually measured by the area under the receiver operating characteristic curve (AUC). We propose a novel method by implementing genetic algorithms to maximize the partial AUC that is restricted to a specific interval of lower false positive rate (FPR), the region relevant to follow-up experimental validation. Our predictor uses various features based on sequence information, protein-protein interaction network topology, and gene expression profiles. A feature selection wrapper was developed to alleviate the over-fitting problem and to weigh each feature's relevance to prediction. We evaluated our method using the proteome of budding yeast. Our implementation of genetic algorithms maximizing the partial AUC below 0.05 or 0.10 of FPR outperformed other popular classification methods. [BMB Reports 2013; 46(1): 41-46]

INTRODUCTION

Identifying genes indispensable for an organism's life, as well as identifying the characteristics of those genes, is one of the central questions in biology. The essential genes of a number of organisms have been identified and catalogued through genome-wide knock-out experiments (1). However, identification of essential genes by such wet experiments is time-consuming and labor-intensive. Hence, it would be helpful to develop computational approaches to essential gene prediction for candidate screening. For example, computational prediction of essential genes in pathogenic microbes has an implication in prioritizing antimicrobial drug targets (2). In addition to the candidate list, computational predictions can identify sequence features and related properties that are relevant to essentiality.

Computationally, predicting essential genes amounts to a problem of classifying all the genes into binary classes of essential and nonessential ones based on a number of features that can be calculated for each gene. These features must have some relevance to essentiality. Previous studies have used various types of features, such as comparative genomic properties, sequence-based features, and functional genomic properties. It is well established that essential genes are more conserved than nonessential ones (3). Phyletic retention, the number of organisms in which an ortholog is conserved, has been used widely for the prediction of essential genes (4). Besides comparative genomic features, a number of sequence-based features that can be directly calculated within the sequenced genome have been employed in predicting essential genes (4, 5). For a newly sequenced genome, only the sequence-based features are typically available, and thus, essentiality prediction solely based on these features has important applications.

Proliferation of high-throughput functional genomics data provided opportunities to explore their properties in relation to essentiality. For example, a protein-protein interaction network is crucial for most biological processes. Essential genes tend to play special roles, like hubs in the network, and thus, their topological properties have been used to predict gene essentiality (6-8). In fact, Hwang *et al*. combined the protein-protein interaction network properties with the sequence features in predicting essential genes, and demonstrated that the essentiality prediction improved by adding the network properties (9). Gene expression is also a fundamental process tightly regulated at a system level, and is related to gene essentiality. It has been shown that gene expression variation is related to gene essentiality (7, 10, 11).

We integrated all the features, based on sequence information or protein-protein interaction network topology, which have been used by other methods. In addition, we included gene expression properties such as transcriptional activity and variation, as well as phyletic retention. While the inclusion of more features would improve the prediction power, some irrelevant or

*Corresponding author. Tel: +82-2-820-0457; FAX: +82-2-820-0816, E-mail: sskimb@ssu.ac.kr
†Present address: Platform R&D Team, Technical Group, Daum Commnucations, Seoul 140-894, Korea.

http://dx.doi.org/10.5483/BMBRep.2013.46.1.159

Received 2 August 2012, Revised 8 August 2012, Accepted 8 August 2012

Keywords: AUC, Classification, Essential genes, Genetic algorithms, Partial AUC

ISSN: 1976-6696 (print edition) / 1976-670X (electronic edition)

distracting features might confuse the machine learning system or cause over-fitting. Pruning such features would improve the performance of the predictor, and would help to assess the relative importance of each feature in essentiality prediction. For this, various techniques have been used. For example, the correlation coefficient of each feature with essentiality, and the näive Bayes technique were used to assess the relative importance of each feature as an essentiality predictor (4, 12). Wilcoxon rank sum tests were used to evaluate the statistical difference of essential and nonessential genes in each feature (9). The receiver operating characteristic (ROC) curves were also evaluated for individual features to assess their performances (9). We employed a more rigorous and robust method that avoids over-fitting. Using backward greedy search elimination, features were selected based on the ROC evaluation, followed by 10-fold cross-validation.

The area under ROC curve (a.k.a., AUC) is a common performance measure to evaluate a classification method. When applied to the binary classification problem, the conventional full AUC maximization concerns the whole range of true positive rate and false positive rate (FPR). For the practical purpose of prioritizing candidates that would undergo experimental validation, it is important to produce a short list that is depleted of false positives. Seringhaus *et al.* (4) approached this problem by assigning relatively higher costs to false positives than to false negatives. Here, we propose partial AUC maximization that focuses on only the top scoring candidates, rather than the entire set. Partial AUC has been used in medical diagnostic screening, where a high true-positive rate to the fixed lower FPR is preferable (13). We developed a method that maximized the partial AUC directly at a given FPR of either 0.05 or 0.1 using a genetic algorithm. It is one of the heuristic search techniques that have been inspired by evolutionary biology, and has been used widely to solve complicated problems in various fields. We evaluated our method using the proteome of *Saccharomyces cerevisiae*, for which a comprehensive set of experimentally validated essential genes is available. In addition, high throughput protein-protein interaction data and gene expression profiles of this model organism are readily available from DIP (14) and GEO (15), respectively. We made comparisons with other machine learning algorithms.

RESULTS

We compiled a total of 31 features for 5825 *S. cerevisiae* genes from various sources. After removing the genes with missing values, the final dataset comprised of 3979 genes (including 940 essential genes). Ten-fold cross validation was used on this dataset for performance comparison.

Features relevant to gene essentiality

Pruning irrelevant or distracting features is an important step in a machine learning system. By maintaining only highly-relevant features, a simple and fast model can be built. We applied a feature selection wrapper to the training set of each fold. For each fold, the feature selection procedure was repeated 50 times to reduce variability (see MATERIALS and METHODS). In doing so, one would identify the features that are consistently selected. These would be the features that are the most relevant to essentiality. Table 1 summarizes the number of times each feature has been selected in each fold.

Among the 31 features, the following five have been selected 50 times in every fold: nucleus, phyletic retention, essentiality index (EI), clique level (KL), and variance of gene expression. Our findings on these highly important features for essentiality are concordant with previous studies as follows. First, essential genes tend to be localized in the nucleus, where essential biological processes such as information storage and processing take place (4, 16). Second, essential genes are known to be more evolutionarily-conserved than are nonessential genes in bacteria (3) as well as some eukaryotes (17). In this regard, Gustafson *et al.* (5) showed that phyletic retention, a surrogate for evolutionary conservation, is highly correlated with gene essentiality in yeast and *E. coli*. Third, essential genes are more likely to interact with other essential genes than nonessential genes. For example, Hwang *et al.* (9) showed that essential genes have higher values of essentiality index (the proportion of essential genes among the interacting partners) than nonessential genes. Fourth, clique level (the size of the largest cliques that a protein belongs to in a protein-protein interaction network) is related to the connectedness of a gene, which is known to be related with gene essentiality (18). It outperformed other network-based features in Hwang *et al.*'s study (9). Finally, variance in gene expression is known to be negatively correlated with gene essentiality and evolutionary conservation rate in eukaryotes, including yeast (7, 10).

Besides transcriptional variability, transcriptional activity or gene expression level has been observed to being related to essentiality, in that highly expressed genes evolve slowly from bacteria to mammals (10). In our study, mean gene expression was selected as one of the essentiality features most of the time (>90% in average). Three other sequence-based features, i.e., A3s, Gravy, and CLOSE STOP RATIO, were also chosen very often. Gravy, the general average hydropathicity, and TM_HELIX, the number of transmembrane helices, have been shown to be negatively correlated to essentiality (4). They are correlated to each other as the hydropathy is one of the important factors contributing to the prediction of transmembrane proteins. Many transmembrane proteins function as transporters and participate in metabolic or nonessential roles. In yeast, highly expressed (more likely essential than nonessential) genes prefer pyrimidine bases (C or T) at the third synonymous codon position but do not prefer codons with A+T richness, which is a characteristics of the yeast genome (38% G+C content) (19). The fact that A3s, the frequency of A at the synonymous third position of codons, was relatively frequently chosen in our feature selection experiments may then be due to its high frequency in nonessential genes. If the essential genes in yeast prefer pyrimidine bases at the third synonymous codon position, one would expect that C3s and T3s, the frequencies of C and T, respectively, at the third synonymous codon position, have similar prediction powers. Indeed, the number of times C3s was selected showed a high correlation with that of T3s (C.C. = 0.88), while that of G3s showed a negative correlation (C.C. = −0.78) in each fold of the feature selection experiments (Table 1).

Prediction performance on gene essentiality

Genetic algorithms with different objective functions: In each

Table 1. The number of times[a] each feature has been selected in each fold

Features	Fold number										Average
	1	2	3	4	5	6	7	8	9	10	
Cytosol[1]	17	20	23	3	23	6	7	13	14	5	13.1
Extracellular[1]	0	2	0	0	0	0	1	0	0	0	0.3
Plasma membrane[1]	2	5	6	1	14	4	21	3	16	12	8.4
Mitochondria[1]	0	0	0	0	0	1	0	0	0	0	0.1
Nucleus[1]	50	50	50	50	50	50	50	50	50	50	50
TM_HELIX[2]	44	44	43	16	47	43	38	45	50	18	37.8
T3s[3]	38	21	23	28	38	10	7	19	16	25	22.5
C3s[3]	29	14	14	24	23	14	7	13	15	23	17.6
A3s[3]	50	50	50	50	50	50	50	50	50	49	49.9
G3s[3]	19	29	32	22	21	49	43	41	33	49	33.8
CAI[3]	2	0	1	6	1	0	3	2	1	2	1.8
CBI[3]	34	10	24	27	43	19	38	22	18	30	26.5
Fop[3]	12	42	18	22	10	32	11	25	29	19	22
Nc[3]	15	0	5	1	8	2	7	3	1	1	4.3
GC3s[3]	39	24	19	27	36	12	8	16	17	18	21.6
GC[3]	2	2	17	0	11	2	2	10	7	9	6.2
L_sym[3]	4	6	16	9	2	9	4	9	17	7	8.3
L_aa[3]	4	6	15	9	2	9	4	9	17	6	8.1
Gravy[3]	50	49	49	49	50	50	48	49	48	47	48.9
Aromo[3]	28	3	7	0	6	34	5	16	8	2	10.9
RAREAMINO_ACID[4]	9	4	2	0	6	6	0	5	2	0	3.4
CLOSE STOP RAIO[4]	46	44	46	50	50	48	50	45	46	47	47.2
Phyletic retention[5]	50	50	50	50	50	50	50	50	50	50	50
DegreeK[6]	2	5	0	6	3	7	2	29	8	1	6.3
CCo[6]	9	0	0	0	0	0	0	3	3	2	1.7
BC[6]	5	4	3	43	4	3	0	5	0	2	6.9
CC[6]	33	26	34	38	25	37	16	50	41	25	32.5
KL[6]	50	50	50	50	50	50	50	50	50	50	50
EI[6]	50	50	50	50	50	50	50	50	50	50	50
Mean[7]	49	48	50	47	45	41	47	50	49	47	47.3
Variance[7]	50	50	50	50	50	50	50	50	50	50	50

[a]The number of times a feature was selected. The colored background highlights the cells that were selected either every time (dark blue), more than 90% (puple), or 80% (cyan) of the 50 trials in each fold.[1]subcellular localization probabilities calculated with ConLoc. [2]Number of transmembrane helices calculated with TMHMM. [3]Condon usage freqeuncies calculated with CondonW. [4]Ratios derived from the results of CondonW. [5]Number of organisms having an ortholog. [6]Protein-protein interaction network features. [7]Gene expression features.

fold, we applied three criteria for feature selection: no filtering (Entire), ≥80% (FS_80), and ≥90% (FS_90) of the 50 trials (see Table 1 and MATERIALS and METHODS). For each selected feature subset, a linear discriminant function was trained by a genetic algorithm. As objective functions of the genetic algorithm, we employed two performance measures suitable for the task of essential gene prediction, i.e., partial AUC at 5% FPR (GA_ PAUC_0.05) and partial AUC at 10% FPR (GA_PAUC_0.10) as well as AUC (GA_ AUC). Supplementary Fig. 1 shows performance comparisons among different objective functions and feature selection criteria. Supplementary Figs. 1A and 1B compare the performance, measured by partial AUC at 5% and 10% FPR, respectively. Regardless of the feature selection criteria, GA_PAUC_ 0.05 and GA_PAUC_0.10 performed significantly better than GA_AUC in terms of partial AUC at 5% and 10% FPR, respectively. Supplementary Fig. 1C shows that GA_AUC achieved much higher AUC values than GA_ PAUC_0.05 and GA_PAUC_0.10. From these results, we can conclude that the objective function used for genetic algorithms is highly influential on their performance. Thus, the objective function should be matched to the performance measure.

Our feature selection method improved the prediction performance in general, and the performance of GA_PAUC_0.05 was significantly enhanced by feature selection (Supplementary Figs. 1A and 1C). GA_PAUC_0.10 showed performance enhancement by feature selection in terms of AUC (Supplementary Fig. 1C). However, its performance measured by partial AUC at 10% FPR was degraded by feature selection (Supplementary Fig. 1B). The performance of GA_AUC was improved by feature selection when measured by partial AUC at 10% FPR (Supplementary Fig. 1B). When measured by partial AUC at 5% FPR and AUC, its performance did not show significant difference by the feature selection criteria (Supplementary Figs. 1A and 1C). We also observed that the variance in performance was reduced by feature selection (Entire vs. FS_80 and FS_90; see Supplementary Tables 1, 2 and 3). By reducing the number of features, our feature selection method effectively enhances the generalization performance of the genetic algorithm in many cases.

Comparison with other classification methods: We compared our genetic algorithm-based methods (GA_PAUC_0.05, GA_ PAUC_0.10, and GA_AUC) with other popular classification methods such as multilayer perceptrons (multiLayer), logistic regression (logistic), and support vector machines with RBF (radial basis function) and polynomial kernels (smoRBF and smoPoly, respectively) (Supplementary Fig. 2). All parameters for these methods were set as default in Weka (see MATERIALS and METHODS). Our genetic algorithm-based methods outperformed all the others except only one case, i.e., performance measured by partial AUC at 5% FPR and using the entire feature set, where GA_PAUC_0.05 achieved the second highest performance (Supplementary Fig. 2A). Furthermore, our methods showed lower variance in performance than others, regardless of the performance measure and feature selection criteria (see Supplementary Tables 1, 2, and 3). Thus, we conclude that the genetic algorithm-based methods are generally more useful than other classification methods for the task where the performance measure is partial AUC. Other methods have their own objective functions. Multilayer perceptrons are usually trained by minimizing the squared error or maximizing the cross entropy function. Logistic regression is performed by maximizing the likelihood. Support vector machines are trained by maximizing the distance between the decision boundary and the closest data point. All these objective functions are calculated over all training data examples. However, partial AUC is calculated over a subset of data examples, i.e., the examples classified as positive at a specific value of FPR. Thus, the optimal classifier with respect to these objective functions can achieve suboptimal partial AUC values. On the contrary, our methods are trained by directly maximizing the performance measure, i.e., partial AUC and are likely to achieve better performance than others in terms of this measure.

Feature selection had a different impact on each classification method. Our feature selection methods greatly improved the performance of multilayer perceptrons, regardless of the performance measure and the feature selection criteria (Supplementary Fig. 2). However, feature selection did not have a strong influence on the performance of other classification methods. In our experiments, multilayer perceptrons severely overfitted the training dataset of each fold (see Supplementary Tables 4, 5, and 6). Overfitting can be reduced by eliminating less relevant features. This explains why feature selection improved the performance of multilayer perceptrons, but not the other methods.

DISCUSSION

The ultimate goal in developing computational methods for the prediction of essential genes is the predictability in under-studied organisms based on a universal set of features that are developed in well studied model organisms. The prediction system should be robust enough to be applicable to other distantly related organisms. Deng *et al.* investigated this issue in four distantly related bacteria (12). In this report, we did not address the transferability issue, but focused on two core issues, feature selection and classification algorithms.

For a newly sequenced genome, only the sequence-based features are usually available. However, with the advent of next-generation sequencing technology, gene expression profiling using RNA-seq technology can be performed concomitantly with or independent from genome sequencing. Considering that the gene expression properties are closely related to evolutionary conservation and essentiality, it is appropriate to include these features in the prediction models. Unlike gene expression profiles, high throughput protein-protein interaction data usually require lengthy experiments. As gene coexpression has been observed between permanent, not transient, protein complex partners (20), the gene coexpression network properties may compensate protein-protein interaction network properties.

The power to predict essential genes varies among various types of features. Unlike most of the previous studies that measured each feature's power independently, using statistical measures, we employed the feature selection wrapper technique, which evaluated each feature based on classification models involving all the other features. As the measure of relevance to essentiality, we used the number of times a feature was selected. The majority of the most consistently selected features in our study had been previously reported as the most powerful predictors by the other studies. It was possible to rationalize their selection in terms of biological knowledge. As such, feature selection in essential gene prediction is not only one of the essential steps in machine learning but also has a strong biological foundation.

While the computational prediction of biological effects can shed insight into what features play major roles, its results can have a bigger impact through experimental validations. It would then be better to suggest a candidate list that is short and depleted with false positives in order to minimize the labor-intensive and time-consuming wet experiments. In such cases, partial AUC, which is measured at fixed lower FPR, is then more important than full AUC. We implemented genetic algorithms to maximize directly the partial AUC. While other classification methods have their own objective functions, genetic algorithms allow the flexibility of formulating one's own objective functions. To our knowledge, this is the first implementation of genetic algorithms in maximizing partial AUC in biomedical applications. We believe that its flexibility and robustness may stimulate wide acceptance.

MATERIALS AND METHODS

Data preparation

Sequence-based features: We downloaded 5885 ORF sequences of *S. cerevisiae* from ftp://genome-ftp.stanford.edu. CodonW (http://codonw.sourceforge.net/) was used to calculate 14 features related to the codon usage (see Supplementary Table 7). ConLoc (http://sbi.postech.ac.kr/conloc/) was used to predict the subcellular localizations: cytosol, extracellular matrix, plasma membrane, mitochondria, and nucleus. The probabilities for these five locations were estimated and used as features, respectively. TMHMM Server (http://www.cbs.dtu.dk/ services/ TMHMM/) predicted the number of transmembrane helices of each of the 5825 amino acid sequences. Finally, ratios of rare amino acids and amino acids close to stop codons were calculated as suggested in (4).

Protein-protein interaction network features: We followed the protocol established by Hwang *et al.* (9). Briefly, a protein-protein interaction data of S. cerevisiae was downloaded from DIP (version Scere20091230), which contained 5033 proteins and 22,118 interactions. From this yeast protein-protein interaction network, we calculated the following topological properties as suggested in (9): degree (degreeK), clustering coefficient (CCo), betweenness centrality (BC), closeness centrality (CC), clique level (KL), and essentiality index (EI). A list of essential genes from DEG (16) was used for calculating the EI of 4662 genes.

Gene expression features: We downloaded gene expression profiles of yeast obtained by Affymetrix Yeast Genome S98 Array (9335 probes) from GEO (http://www.ncbi.nlm.nih.gov/geo/) (15). Total number of samples was 2015. From the raw CEL files, the expression level of each probe was calculated by RMA (21). Then, we calculated mean and variance of expression of 5885 genes as in (10). When multiple probes were mapped to one gene, we used average values.

Phyletic retention features: We obtained phyletic retention data for 4728 genes of *S. cerevisiae* from (5), where the phyletic retention was defined as the number of organisms having an ortholog.

Final dataset: Finally, we obtained a dataset consisting of 3979 genes and 31 features by excluding genes having a missing value. There were 940 essential genes. In our experiments, 10-fold cross validation was used.

Feature selection

We employed a backward search-based wrapper for feature selection (22). Feature selection wrappers are known to produce better results than filter methods, such as information gain, because they consider classification models involved. AUC (Area under the ROC Curve) with 3-fold cross validation was adopted as a performance measure for the wrapper. As a classification model for the wrapper, we used logistic regression. We applied the wrapper 50 times and selected features based on the number of times they have been selected. According to the result of the 50 trials, we made three different feature groups: no filtering, more than 80% selected, and more than 90% selected.

Partial AUC maximization by genetic algorithms

The partial AUC (area under the ROC curve) is a sub-area of AUC measured over a range of false positive rates (see Supplementary Fig. 1). In order to maximize partial AUC on a training dataset, we used genetic algorithms. A linear discriminant function was used for scoring each essential gene candidate, c, as follows.

$$\mathrm{score}(c) = \mathbf{w} \cdot \mathbf{f},$$

where **w** is a weight vector and **f** denotes a set of feature values of *c*. Each feature value f_i was normalized into [0, 1] as follows:

$$f_i = (f_i - \min(f_i)) / (\max(f_i) - \min(f_i)).$$

A genetic algorithm was used for finding **w** which maximizes partial AUC. See Supplementary Table 8 for parameter setting of the genetic algorithm. In our experiments, we used partial AUCs at FPRs (false positive rates) of 5% and 10% as examples of acceptable FPRs in subsequent validation.

Other classification methods

We used the Weka package (http://www.cs.waikato.ac.nz/ml/

weka/) for comparing the performance of our GA-based methods with the following popular classification methods: multilayer perceptrons, logistic regression, and support vector machines. Default settings for these models in Weka were used. For multilayer perceptrons, the number of hidden layers was one, and the number of hidden nodes was 16. The backpropagation algorithm was used for training multilayer perceptrons. A ridge estimator was used for the logistic regression model, and RBF (radial basis function) and polynomial kernels were used for support vector machines.

Acknowledgements
This work was supported by the Soongsil University Research Fund (2009).

REFERENCES

1. Zhang, R. and Lin, Y. (2009) DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.* **37**, D455-D458.
2. Plaimas, K., Eils, R. and König, R. (2010) Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst. Biol.* **4**, 56.
3. Jordan, K., Rogozin, I. B., Wolf, Y. I. and Koonin, E. V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12**, 962-968.
4. Seringhaus, M., Paccanaro, A., Borneman, A., Snyder, M. and Gerstein, M. (2006) Predicting essential genes in fungal genomes. *Genome Res.* **16**, 1126-1135.
5. Gustafson, A. M., Snitkin, E. S., Parker, S. C., DeLisi, C. and Kasif, S. (2006) Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics.* **7**, 265.
6. Dezso, Z., Oltvai, Z. N. and Barabasi, A. L. (2003) Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccharomyces cerevisiae. *Genome Res.* **13**, 2450-2454.
7. Jeong, H., Oltvai, Z. N. and Barabasi, A. -L. (2003) Prediction of protein essentiality based on genomic data. *ComPlexUs* **1**, 19-28.
8. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X. and Gerstein, M. (2004) Genomic analysis of essentiality within protein networks. *Trends Genet.* **20**, 227-231.
9. Hwang, Y. C., Lin, C. C., Chang, J. Y., Mori, H., Juan, H. F. and Huang, H. C. (2009) Predicting essential genes based on network and sequence analysis. *Mol. Biosyst.* **5**, 1672-1678.
10. Choi, J. K., Kim, S. C., Seo, J., Kim, S. and Bhak, J. (2007) Impact of transcriptional properties on essentiality and evolutionary rate. *Genetics.* **175**, 199-206.
11. Zhou, L., Ma, X. and Sun, F. (2008) The effects of protein interactions, gene essentiality and regulatory regions on expression variation. *BMC Syst. Biol.* **2**, 54.
12. Deng, J., Deng, L., Su, S., Zhang, M., Lin, X., Wei, L., Minai, A. A., Hassett, D. J. and Lu, L. J. (2011) Investigating the predictability of essential genes across distantly related organisms using an integrative approach. *Nucleic Acids Res.* **39**, 795-807.
13. Takenouchi, T., Komori, O. and Eguchi, S. (2012) An extension of the receiver operating characteristic curve and auc-optimal classification. *Neural Comput.* **24,** 2789-2824.
14. Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. and Eisenberg, D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32**, D449-451.
15. Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muertter, R. N., Holko, M., Ayanbule, O., Yefanov, A. and Soboleva, A. (2011) NCBI GEO: archive for functional genomics data sets-10 years on. *Nucleic Acids Res.* D1005-1010.
16. Zhang, C. T. and Zhang, R. (2008) Gene essentiality analysis based on DEG, a database of essential genes. *Methods Mol. Biol.* **416**, 391-400.
17. 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.
18. Jeong, H., Mason, S. P., Barabási, A. -L. and Oltvai, Z. N. (2001) Lethality and centrality in protein networks. *Nature* **411**, 41-42.
19. Sharp, P. M., Tuohy, T. M. and Mosurski, K. R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* **14**, 5125-5143.
20. Jansen, R., Greenbaum, D. and Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12**, 37-46.
21. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003) Exploration, normalization, and summarization of high density oligonucleotide array probe level data. *Biostatistics.* **4**, 249-264.
22. Kohavi, R. and John, G. H. (1997) Wrappers for feature selection. *Artif. Intell.* **97**, 249-256.