Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 364 (2006) 557–564
PHYSICA A
www.elsevier.com/locate/physa

Identification of lethal cluster of genes in the yeast transcription network

K. Rho[a], H. Jeong[b], B. Kahng[a,*]

[a]School of Physics and Center for Theoretical Physics, Seoul National University, Seoul 151-747, Republic of Korea
[b]Department of Physics, Korea Advanced Institute of Science and Technology, 305-701, Republic of Korea

Received 8 June 2005; received in revised form 27 July 2005
Available online 8 November 2005

Abstract

Identification of essential or lethal genes would be one of the ultimate goals in drug designs. Here we introduce an *in silico* method to select the cluster with a high population of lethal genes, called lethal cluster, through microarray assay. We construct a gene transcription network based on the microarray expression level. Links are added one by one in the descending order of the Pearson correlation coefficients between two genes. As the link density p increases, two meaningful link densities p_m and p_s are observed. At p_m, which is smaller than the percolation threshold, the number of disconnected clusters is maximum, and the lethal genes are highly concentrated in a certain cluster that needs to be identified. Thus the deletion of all genes in that cluster could efficiently lead to a lethal inviable mutant. This lethal cluster can be identified by an *in silico* method. As p increases further beyond the percolation threshold, the power law behavior in the degree distribution of a giant cluster appears at p_s. We measure the degree of each gene at p_s. With the information pertaining to the degrees of each gene at p_s, we return to the point p_m and calculate the mean degree of genes of each cluster. We find that the lethal cluster has the largest mean degree.

PACS: 87.10.+e; 89.75.−k; 64.60.Ak

Keywords: Transcription network; Lethal genes; Percolation

1. Introduction

Thousands of genes and their products in a given living organism are believed to function in a concerted way that creates the mystery of life [1]. Such a cooperative functionality among genes can be visualized using a graph where nodes denote genes and links represent activating or repressive effects on transcription [2,3]. Traditional methods in molecular biology are very limited to analyze such large-scale interactions among thousands of genes; therefore it is difficult to obtain a global image of the gene functions. The recent advent of the microarray assays has attracted sufficient attention from researchers, allowing them to decipher gene interactions in a more efficient manner [4]. While the data obtained using microarray assays have not been yet

*Corresponding author. Tel.: +82 2 880 1326; fax: +82 2 884 3002.
E-mail address: kahng@phya.snu.ac.kr (B. Kahng).

0378-4371/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2005.08.086

sufficiently accumulated to thoroughly understand the entire genetic network and they are also susceptible to errors in detecting the expression levels, they are potential candidates for a fundamental approach to understanding large-scale gene complexes and can be used in many applications such as drug design and toxicological research.

Since microarray technology has a significant impact on genomics study, many methods for pattern interpretation have been developed, including the K-means clustering [5], the self-organizing map [6], the hierarchical method [7], the relevance network method [8], etc. All these methods, however, contain tunable thresholds, and the results obtained by these methods can be misleading on account of the thresholds artificially chosen. While these methods are useful for clustering or classifying genes, they cannot provide any information required to identify essential or lethal genes. Essential or lethal genes are the target genes for drug designs because their deletion leads to an inviable mutant of a given organism.

In this paper, we propose a novel *in silico* method to identify the essential genes in a microarray dataset. Our method is inspired by the combination of gene clustering and the close relationship between the lethality or essentiality of genes and connectivity in a network. Once the genes are clustered by using a graph theory, and then the cluster or module containing a high population of essential genes is identified by using the relationship between the lethality and connectivity of the graph [9]. The identification of lethal genes by a cluster or module proves to be comparatively more efficient in selecting essential genes than the approaches based on individual genes. Our model does not contain any artificial parameter; therefore, the essential genes can be identified in a self-organized manner. Moreover, we find that the genes belonging to the same module share a common functionality. Thus, our method can also be used to identify the functionality of unknown genes as well.

2. Formation of a giant cluster in a transcription network

A network is constructed from a microarray dataset, containing 287 single gene deletions of *S. cerevisiae* mutant strains composed of 6316 genes [10]. The deletion dataset, which elucidates the genetic relationships among perturbed transcriptome [11], is composed of two large, internally consistent, global mRNA expression subsets. One subset provides mRNA expression levels in wild-type *S. cerevisiae* sampled separately 63 times (the 'control' set) and the other subset provides individual measurements on the genomic expression program of 287 single gene deletion mutant *S. cerevisiae* strains, which were grown under the same cell culture conditions as wide-type yeast cells (the 'perturbation' set). Individual of the microarray data is the ratio of the expression levels in the wild-type and perturbed sets for each gene. Thus, the data can be written in terms of an $N \times M$ matrix with $N = 6316$ and $M = 287$, which is denoted as $\mathbf{C}$, representing the expression ratio of N genes for M different-deletion experiments. In other words, each element $c_{i,j}$ of the matrix $\mathbf{C}$ is the logarithmic value to the base 10 of the ratio of the expression levels for the ith gene under the jth perturbation [12].

To obtain the correlations among the transcriptional genes, we consider the Pearson correlation coefficient $\rho_{i,j}$ between the expression ratio of genes i and j averaged over k different perturbations, which are defined as

$$\rho_{i,j} \equiv \frac{\langle c_{i,k}c_{j,k}\rangle - \langle c_{i,k}\rangle\langle c_{j,k}\rangle}{\sqrt{(\langle c_{i,k}^2\rangle - \langle c_{i,k}\rangle^2)\left(\langle c_{j,k}^2\rangle - \langle c_{j,k}\rangle^2\right)}}, \tag{1}$$

where $\langle\cdots\rangle$ implies the average over k different-deletion experiments. As shown in Fig. 1, the distribution of the correlations $\{\rho_{i,j}\}$ is bell shaped in the range $[-1, 1]$. We construct a network based on the obtained set of the Pearson coefficients. Links are added one by one in the descending order of the Pearson coefficients. Let p be the concentration of added links among $N(N-1)/2$ possible pairs, that is, the ratio of the number of links present in a given network to $N(N-1)/2$. When p is small, the number of links added to the graph is small, most nodes remain isolated, which form small-size components or modules. As p increases, either each cluster grows in size or the number of clusters $\mathcal{N}(p)$ containing at least two genes increases. At a certain value of p, denoted as p_m, the number of clusters becomes maximum, as shown in Fig. 2, which is estimated to be $p_m \approx 0.0002$. Beyond the value of p_m, the number of clusters decreases by the merging of two clusters; however, the mean size of the cluster increases.

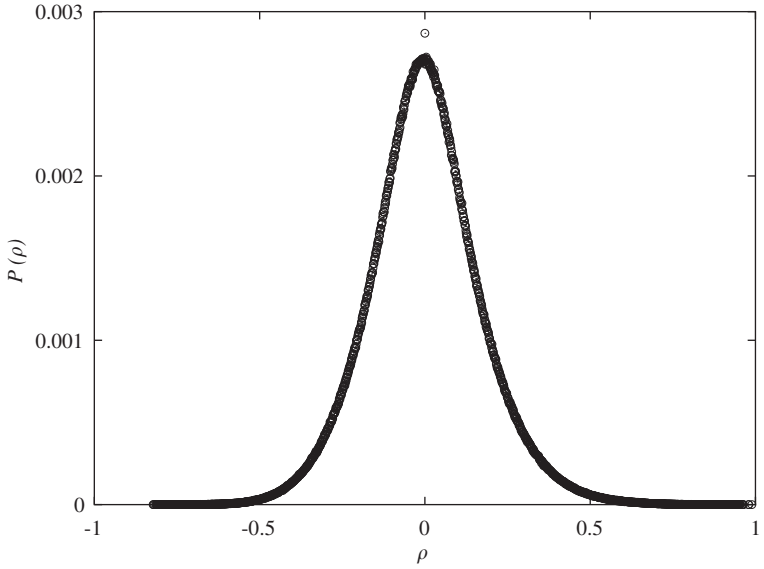

Fig. 1. The distribution of the Pearson correlation coefficients.

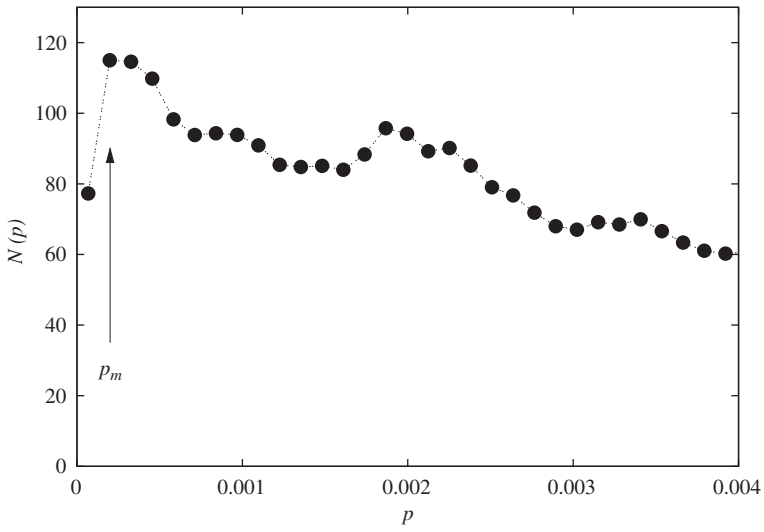

Fig. 2. Plot of the number of clusters $\mathcal{N}(p)$ as a function of the link density p.

As p increases further, the mean size of each cluster increases by either attaching to an isolated node or by merging with finite size clusters. At the percolation threshold p_c, a giant cluster emerges. The scale-free (SF) network appears when the link density increases further at $p_s \approx 0.0063$, as shown in Fig. 3. The degree distribution follows a power law $P_d(k) \sim k^{-0.9}$ with an exponential cutoff, which is a generic feature of the SF network with the degree exponent $\gamma < 2$. The degree exponent value $\gamma \approx 0.9$ is close to the those obtained in different systems [13,14], but smaller than the typical values occurring in many real world networks in the range of $2 < \gamma \leqslant 3$. As the link density increases further beyond p_s, the degree distribution ceases to follow the power law.

To understand the biological implication of the SF network at p_s, we investigate whether the degree in the SF network is useful in detecting the lethal genes. In Fig. 4, we plot a fraction of the essential genes among the genes with a degree larger than a certain degree k_0. The fraction shows an increasing trend up to $k_0 \approx 250$,

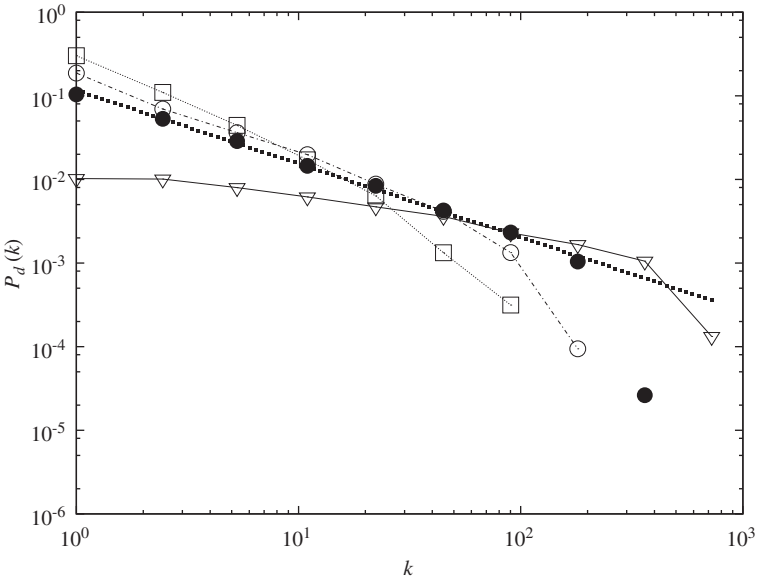

Fig. 3. Plot of the degree distribution of the gene transcription network at various link densities, $p = 0.0003$ (□), $p = 0.0016$ (◯), $p = 0.0063 \approx p_s$ (•), and $p = 0.0322$ (▽). At p_s, the degree distribution follows a power law with an exponential cutoff. The dotted line having a slope of -0.9 is drawn for guidance.

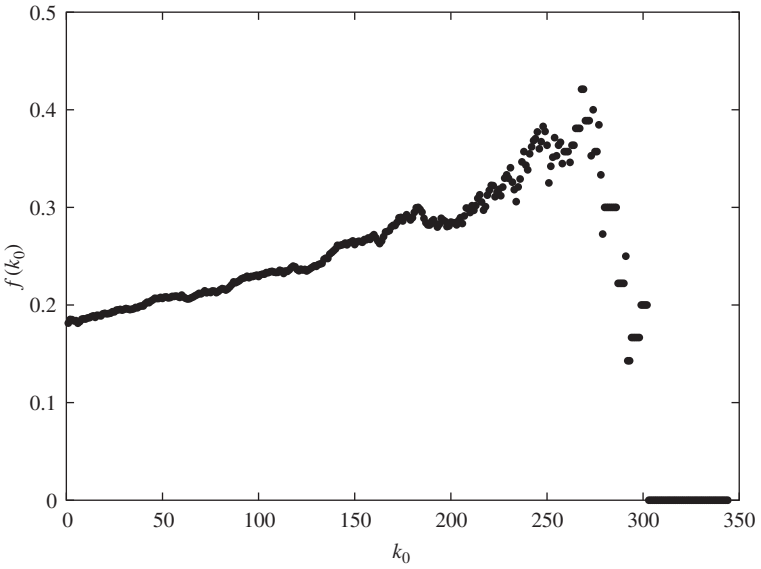

Fig. 4. Plot of a fraction of the essential genes with a degree larger than k_0 to the total number of genes as a function of k_0.

implying that the genes with larger degrees are more likely to be lethal for $k_0 < 250$. However, the fraction falls rapidly beyond the degree $k_0 \approx 250$. Even in the case of $k_0 \approx 250$, the fraction of the essential genes is approximately 40%, which is less efficient than that in the case obtained from the yeast protein interaction network, where the ratio of finding essential genes is as high as 62% for highly connected proteins. Thus, the identification of the essential genes through degree distribution in the SF transcription network alone is comparatively less efficient than that in the case obtained through degree distribution in the protein interaction network.

3. Identification of essential gene cluster

Here we introduce a new method to identify the essential genes from the microarray data, which is based on the idea that the unit of selection is a group of genes with similar functionality instead of individual genes. The selection method is as follows. Initially, $N = 6316$ genes are present and they are not connected to each other, as shown in Fig. 5A. At each time step, links are added one by one in the descending order of the Pearson coefficient $\rho_{i,j}$. Simultaneously, the number of clusters $\mathcal{N}(p)$ is measured, where the isolated nodes are not counted as individual clusters. The link density p is defined as a fraction of the number of links added to all possible pairs of nodes, $N(N-1)/2$. As p increases, we identify p_m where the number of clusters becomes maximum, as defined before in Fig. 2. At this point, we identify each cluster and their members, as shown in Fig. 5B. We also record the network configuration for further discussion. Following this, more links are added until the link density reaches the density value of p_s, where the network is scale-free in the degree distribution. At p_s, we measure the degree of each node, as indicated in Fig. 5C. Maintaining the degree of each node at p_s, we return to the network configuration recorded earlier at p_m. We then calculate the average degree per node

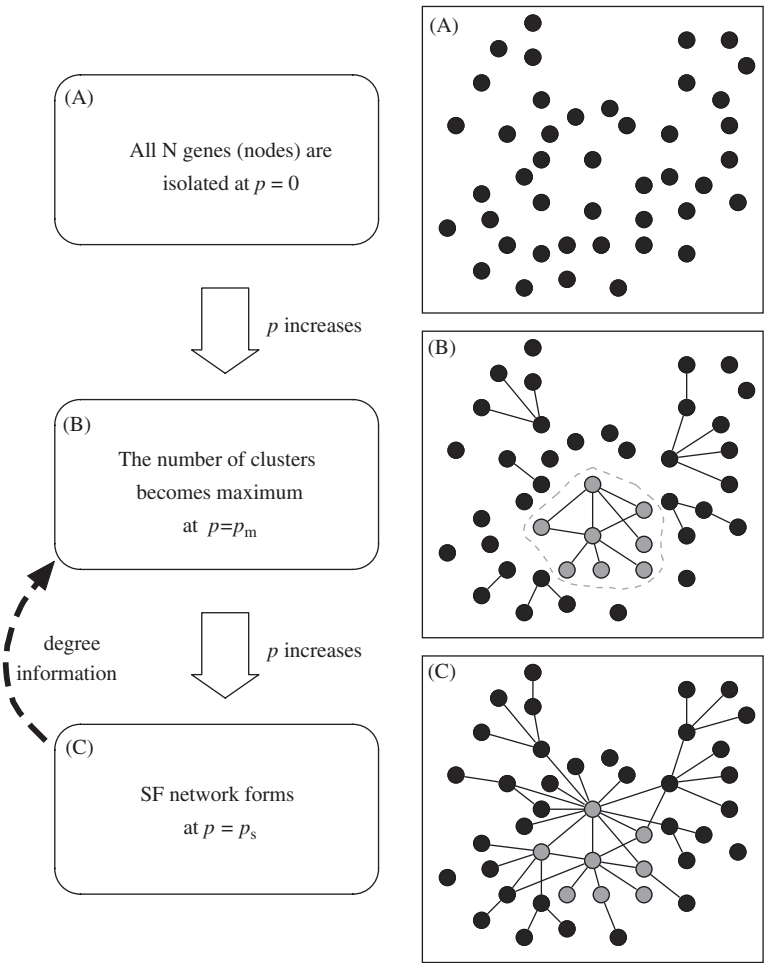

Fig. 5. Schematic diagram of how to identify the lethal cluster of genes. (A) From the initial state with N isolated vertices at $p = 0$, links are added one by one in the descending order of the Pearson coefficients. (B) At $p = p_m$ where the number of clusters becomes maximum, each node recognizes the cluster to which it belongs. (C) At $p = p_s$ where the network is scale-free in the degree distribution, the degree of each node is measured. Maintaining the degree of each node measured in (C), we return to the network configuration in (B). We calculate the mean degree per node in each cluster of (B) based on the degrees measured in (C). For example, the mean degree per gray-colored node belonging to the cluster denoted by the dashed line is $\frac{37}{8}$, which is the largest among those of the other clusters. We propose that this cluster is lethal.

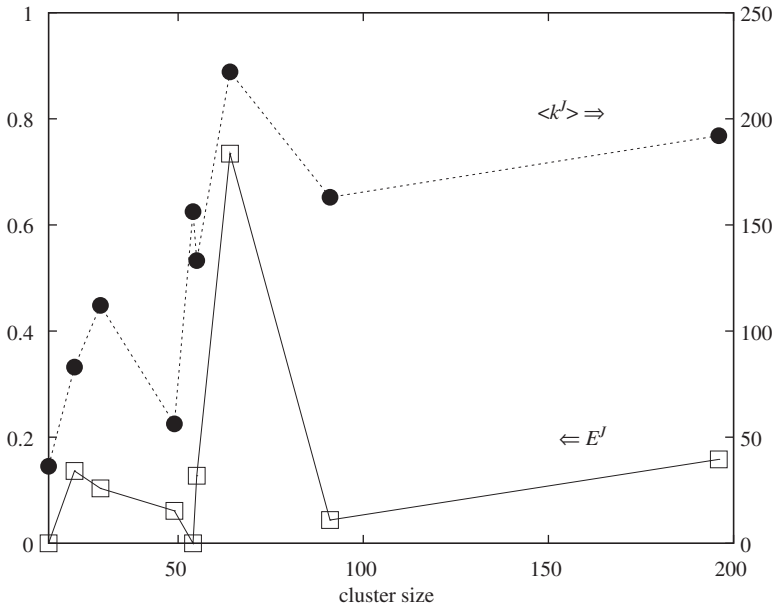

Fig. 6. The comparison between $\langle k^J \rangle$ (•) and $\mathscr{E}^J$ (□) for each cluster indexed by a cluster size of p_m.

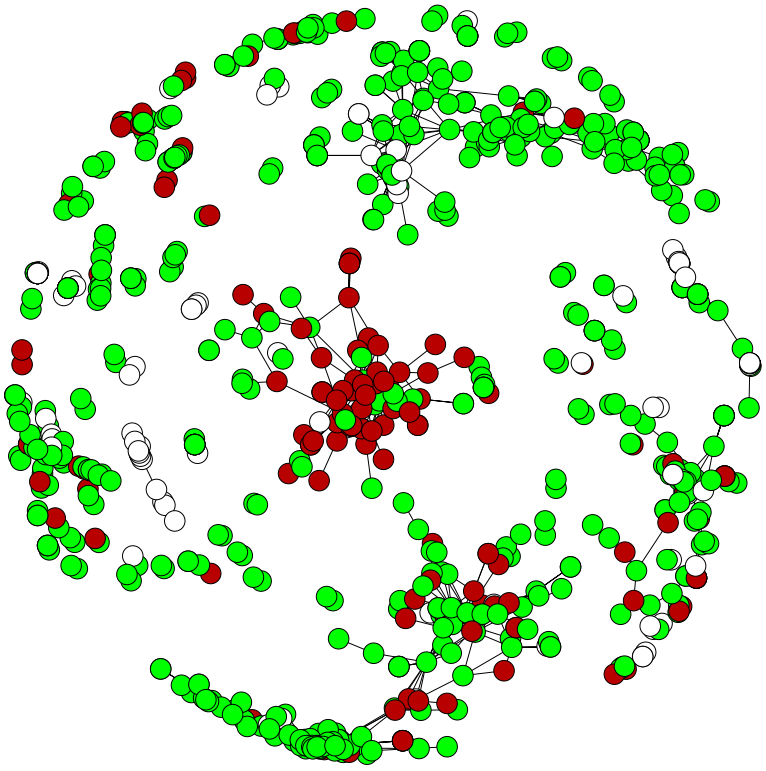

Fig. 7. The gene transcription network of the yeast *S. cerevisiae* at p_m. Red (●), green (●) and white (○) nodes represent essential, nonessential, and unknown genes, respectively (color online).

of each cluster at p_m, that is,

$$\langle k^J \rangle = \frac{\sum_{i \in J} k_i^J(p_s)}{N^J(p_m)}, \tag{2}$$

where $k_i^J(p_s)$ is the degree of node i measured at p_s and J is the cluster index to which node i belongs, which was assigned at p_m. $N^J(p_m)$ is the number of nodes belonging to cluster J at p_m. We then propose that the cluster with the largest value of $\langle k^J \rangle$ contains a high density of essential genes, which is based on the fact that genes with a larger degree are more likely to be essential in the protein interaction networks [9].

To check this proposal, we directly measure essentiality $\mathscr{E}^J$, which is defined as the fraction of known essential genes to the total number of genes belonging to a given cluster J. Indeed, as shown in Fig. 6, the two quantities, $\langle k^J \rangle$ and $\mathscr{E}^J$, behave in a similar manner. Thus, we can confirm that the cluster containing the largest fraction of essential genes can be found in the *in silico* method through $\langle k^J \rangle$. With regard to the yeast dataset, we identify the third largest cluster with 64 genes that proves to be the most essential cluster containing 47 essential genes, 17 nonessential genes, and 1 unidentified gene (Fig. 7). Thus, the certainty of

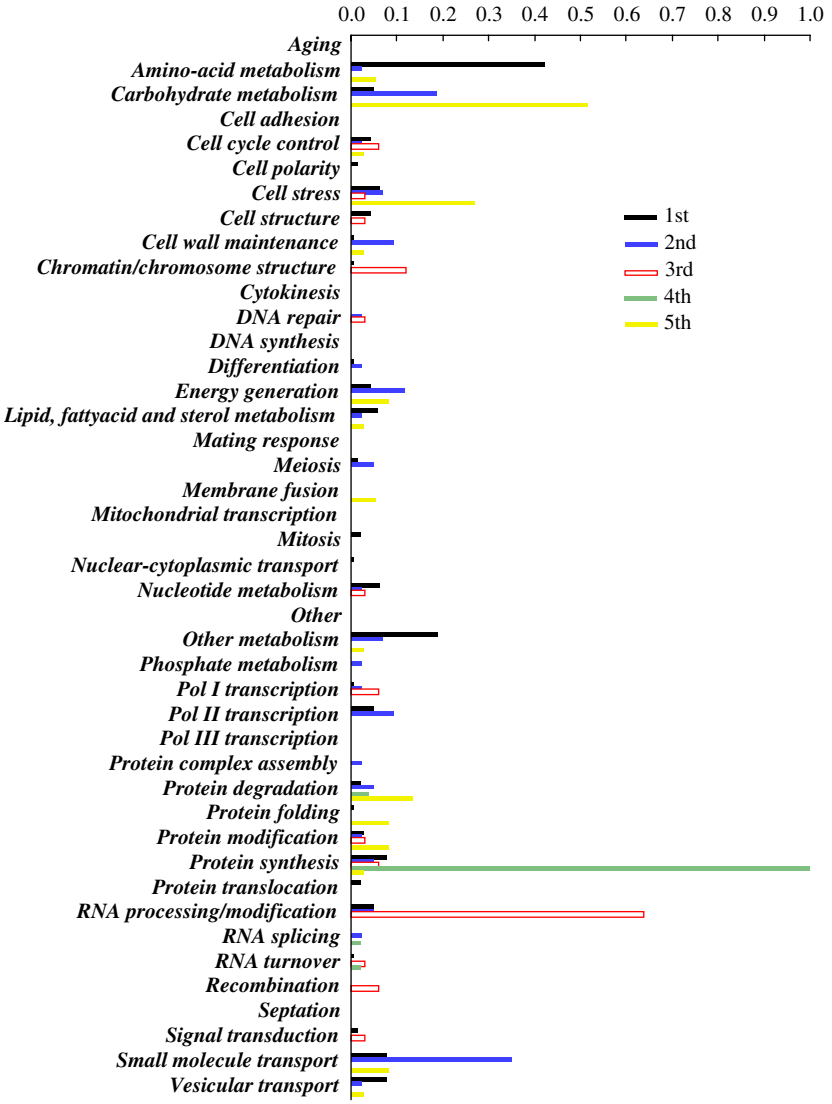

Fig. 8. Genes' ratio specific to each functional category for the genes belonging to the first five largest clusters at p_m (color online).

selecting essential genes improves remarkably by as much as 73% or even higher when the unidentified gene is excluded. This fraction is much larger than the one obtained only through the degree information in the gene transcription networks that we studied in the previous section.

4. Functional modules

It is well known that the biochemical network is composed of a modular structure based on its functionality. In the case of the yeast, 43 functional categories are known [10]. We identify 43 functional categories of genes belonging to the first five largest clusters at p_m, the ratio of which is shown in Fig. 8. This figure indicates that each cluster at p_m has a major population of genes with a specific functionality. For example, the majority of the genes in the largest cluster belong to the functional class of amino acid metabolism. The genes in the second, third and fourth largest cluster are from the class of small molecule transport, RNA processing/modification, and protein synthesis, respectively. This functional clustering within the gene transcription network is rooted in the genes of the same functional category that are likely to respond to an external perturbation in a similar manner. As a result, the Pearson correlation coefficients between them are large, making clusters at a small p_m disconnect with each other. Our result is consistent with the recent discovery of modular structures in the yeast protein interaction network [15] and in the metabolic networks [16]. Based on these properties, we may assign functionally unknown genes as functional candidates based on the major functionality in an identical cluster.

5. Conclusions and discussion

We have introduced a new method to identify the cluster containing a high population of essential genes in the transcription network by using the two known properties that genes with the same functionality are highly correlated to the expression level of the microarray assay and that the essential genes are likely to have larger degree than others in the scale-free network. The certainty of selecting essential genes is proved to be as high as 73%. Thus, such a selection method can be useful in resolving various knockout problems such as drug designs. It should be noted that our method does not include any tuning parameter, and the selection can be performed in a self-organized manner with less ambiguity compared with other existing methods.

This work is supported by the 21C Frontier Microbial Genomics and Application Center Program, MOST (Grant MG05-0203-1-0) and the Korean Ministry of Sciences and Technology through M1 03B5000-00110. The authors would like to thank A.-L. Barabási for helpful discussions and hospitality during their visit to the University of Notre Dame, where this work was initiated.

References

[1] A.-C. Gavin, et al., Nature (London) 415 (2002) 141.
[2] E.M. Marcotte, M. Pellegrini, M.J. Thompson, T.O. Yeates, D. Eisenberg, Nature (London) 402 (1999) 83.
[3] A.J. Enright, I. Iliopoulos, N.C. Kryrpides, C.A. Ouzounis, Nature (London) 402 (1999) 86.
[4] I.S. Kohane, A.J. Butte, A. Kho, Microarrays for Integrative Genomics, MIT Press, Cambridge, MA, 2002.
[5] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Proc. Nat. Acad. Sci. USA 95 (1998) 14863.
[6] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub, Proc. Nat. Acad. Sci. USA 96 (1999) 2907.
[7] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Proc. Nat. Acad. Sci. USA 96 (1999) 6745.
[8] A.J. Butte, P. Tamayo, D. Slonim, T.R. Golub, I.S. Kohane, Proc. Nat. Acad. Sci. USA 97 (2000) 12182.
[9] H. Jeong, S.P. Mason, A.-L. Barabási, Z.N. Oltvai, Nature London 411 (2001) 41.
[10] M.C. Costanzo, et al., Nucleic Acids Res. 29 (2001) 75.
[11] G. Giaever, et al., Nature (London) 418 (2002) 387.
[12] T.R. Hughes, et al., Cell 102 (2000) 109.
[13] N. Guelzim, S. Bottani, P. Bourgine, F. Képès, Nat. Genet. 31 (2002) 60.
[14] P. Provero, cond-mat/0207345.
[15] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, N. Barkai, Nat. Genet. 31 (2002) 370.
[16] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barabási, Science 297 (2002) 1551.