Identification of lethal cluster of genes in transcription network

K. Rho[1], H. Jeong[2] and B. Kahng[1]

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

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 a cluster with high population of lethal genes, called lethal cluster, through the microarray assay. We construct a gene transcription network based on the microarray expression level. Links are added one by one following the descending order of the Pearson correlation coefficients between two genes. As link density p increases, there are two meaningful link densities p_m and p_s. At p_m, which is smaller than the percolation threshold, the number of disconnected clusters is maximum and lethal genes are highly concentrated at a certain cluster we want to find. Thus the deletion of all genes in that cluster could lead to lethal inviable mutant efficiently. Such a lethal cluster can be identified *in silico* way. 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 of the degrees of each gene measured at p_s, we return to the point p_m and calculate the mean degree per node of each cluster. We find that the lethal cluster has the largest mean degree per node.

Key words: transcription network, lethal genes, percolation
PACS: 87.10.+e, 89.75.-k, 64.60.Ak

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 through the notion of graph where nodes denote genes and links do activating or repressing effects on transcription [2,3]. Traditional methods in molecular biology are very limited to analyze such large-scale interactions among thousands of genes, so that a global picture

of gene functions is hard to obtain. The recent advent of the microarray assay has enough attraction to researcher, allowing them to decipher gene interactions in a more efficient way [4]. While the data through the microarray assay are not sufficiently accumulated to fully understand the entire genetic network yet and they are also susceptible to errors in detecting the expression level, the microarray assay is a potential candidate for a fundamental approach to understand large-scale gene complexes and can be used in many applications such as drug design and toxicological research.

Since the microarray technology is having 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 such methods, however, contain tunable thresholds, so that the results obtained through those methods could be misled by the thresholds artificially chosen. While those methods are useful for clustering or classifying genes, they cannot give any information needed to identify essential or lethal genes. Here the essential or lethal genes mean the target genes for drug designs, because the deletion of them leads to inviable mutant of a given organism.

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

2 Formation of a giant cluster in transcription network

A network is constructed from a microarray dataset, which contains 287 single gene deletion of *S.cerevisiae* mutant strains composed of 6316 genes [10]. The deletion dataset, elucidating genetic relationships among perturbed transcriptome [11], is composed of two large, internally consistent, global mRNA expression subsets. The one provides mRNA expression levels in wile-type *S.cerevisiae* sampled 63 separate times (the 'control' set), and the other does 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 ex-

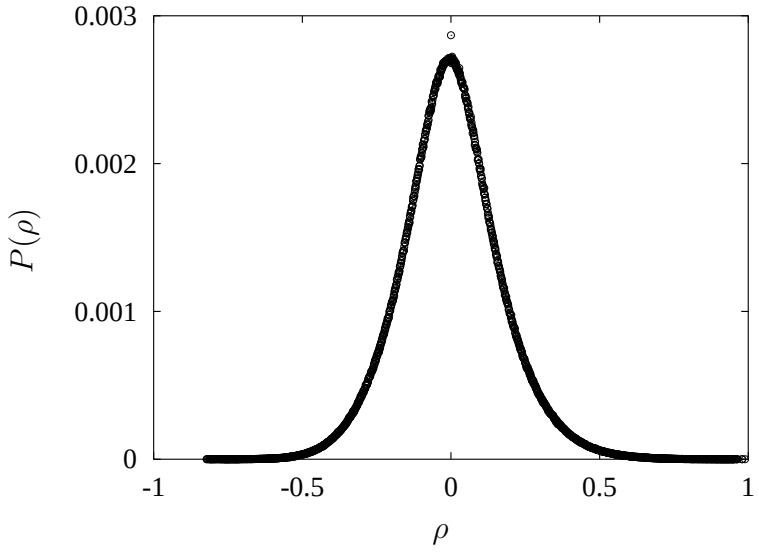

Fig. 1. The distribution of the Pearson correlation coefficients.

pression levels in the wild-type and perturbed sets for each gene. Thus the data can be written in terms of a $N \times M$ matrix with $N = 6316$ and $M = 287$, denoted as $\mathbf{C}$, representing the expression ratio of N genes for M different-deletion experiments. That is, each element $c_{i,j}$ of the matrix $\mathbf{C}$ is the logarithmic value with base 10 of the ratio of the expression levels for the i-th gene under the j-th 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, defined as

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

where $\langle\cdots\rangle$ means the average over k different deletion experiments. As shown in Fig. 1, the distribution of the correlations $\{\rho_{i,j}\}$ is of a bell shape in the range [-1,1]. We construct a network based on the set of the Pearson coefficients.

Links are added one by one following the descending order of the Pearson coefficients. For example, if the Pearson coefficients are in order as $\rho_{1,2} > \rho_{3,4} > \cdots$, then links are added one by one according to that order as node pairs (1,2),(3,4), $\cdots$. Let p is the concentration of added links among $N(N-1)/2$ possible pairs. When p is small, i.e., the number of links added is small, most nodes remain isolated. As p increases, the size of each cluster grows or the number of clusters $\mathcal{N}(p)$ increases, where cluster means the group of nodes containing at least two nodes. At a certain value of p, denoted as p_m, the number of clusters becomes maximum shown in Fig. 2, which is estimated to be $p_m \approx 0.0002$. Beyond p_m, the number of clusters decreases by merging two clusters, however, the mean size of cluster increases.

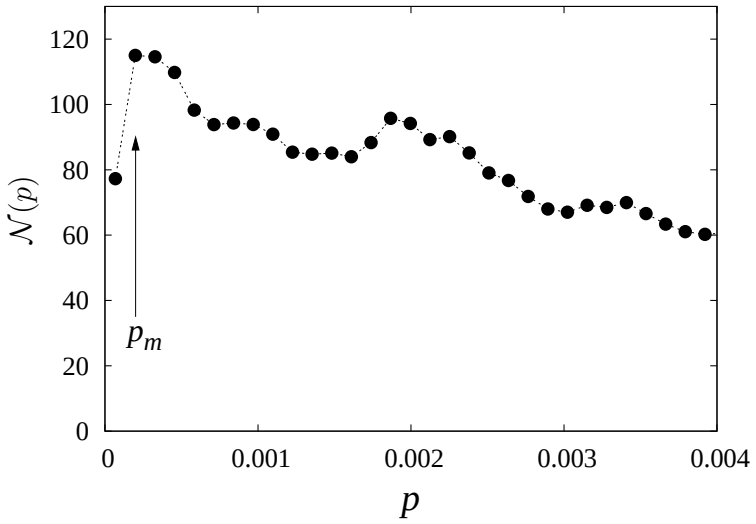

Fig. 2. Plot of the number of clusters $\mathcal{N}(p)$ as a function of the link density p.

As p increases further, the mean cluster size increases by either joining an isolated node or merging finite size clusters. At the percolation threshold p_c, a giant cluster emerges. The SF network appears when link density increases further at $p_s \approx 0.0063$ 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 when the degree exponent $\gamma < 2$. The degree exponent value $\gamma \approx 0.9$ is close to the ones obtained by others in different systems [13,14], but smaller than typical values occurring in many real world networks in the range of $2 < \gamma \leq 3$. As link density increases further away from p_s, the degree distribution no longer follows the power law.

To understand biological implication of the scale-free network at p_s, we investigate if the degree in the scale-free network is useful for detecting lethal genes. In Fig. 4, we plot the fraction of essential genes (nodes) among the genes (nodes) with degree larger than a certain degree k_0. The fraction shows an increasing behavior up to $k_0 \approx 250$, implying that the genes with larger degree are more likely to be lethal for $k_0 < 250$. However, the fraction drops rapidly beyond the degree $k_0 \approx 250$. Even for the case of $k_0 \approx 250$, the fraction of the essential genes is about 40%, which is less efficient than the case obtained from the protein interaction network where the ratio of finding essential genes is as high as 62% for highly connected proteins [9]. Thus, the identification of essential genes through the degree distribution in the scale-free transcription network is not as much efficient as the case through the protein interaction network.

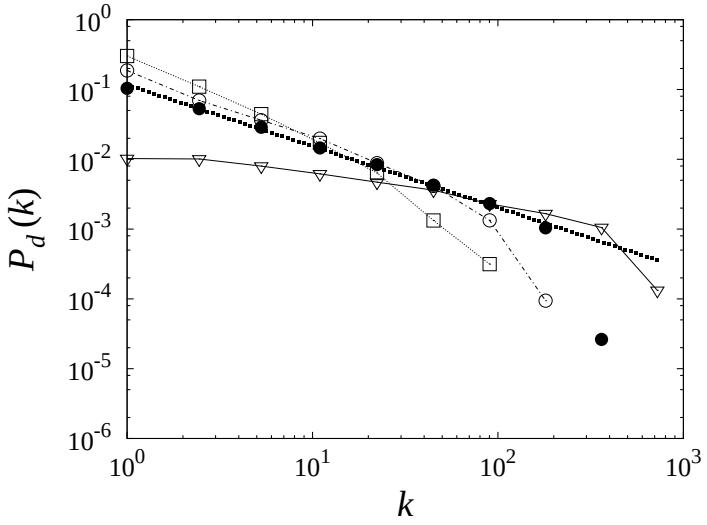

Fig. 3. Plot of the degree distribution of the gene transcription network at various link density, $p = 0.0003$ (□), $p = 0.0016$ (○), $p = 0.0063 \approx p_s$ (●), and $p = 0.0322$ (▽). At p_s, the degree distribution follows a power law with an exponential cutoff. Dotted line with slope -0.9 is drawn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

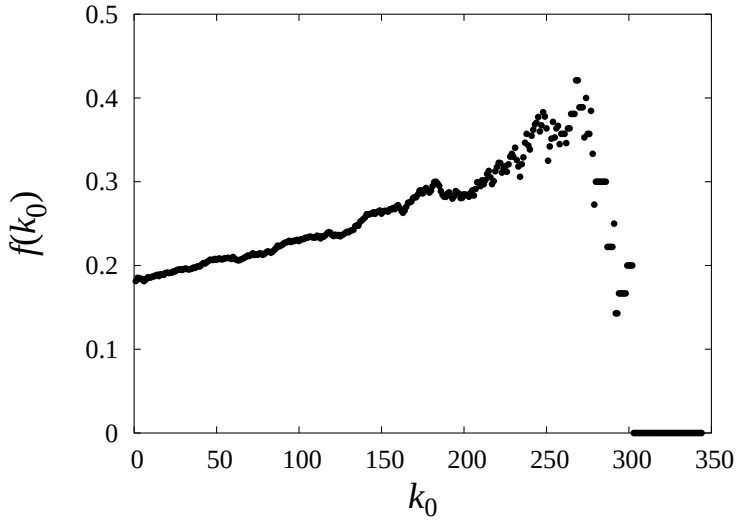

Fig. 4. Plot of the fraction of the essential genes with degree larger than k_0 to the total number of genes as a function of k_0.

3 Identification of essential gene cluster

Here we introduce a new method to identify 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: At initial, $N = 6316$ genes are present and they are not connected each other as shown in Fig.5A. At each time step, links are added one by one by following the descending order of the Pearson coefficient $\rho_{i,j}$. At the same time, the number

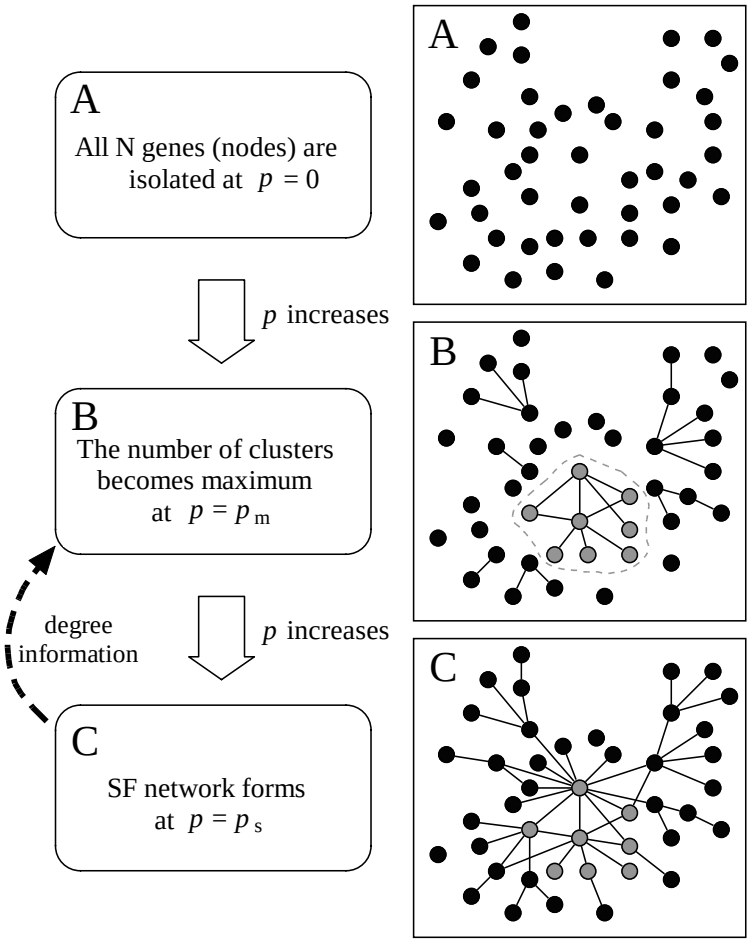

Fig. 5. The schematic diagram how to identify the lethal cluster of genes. (A) From the initial state with N isolated vertices at $p = 0$, links are added one by one in the descending order of the Pearson coefficients. (B) At $p = p_m$ where the number of cluster becomes maximum, each node recognizes to which cluster it belongs. (C) At $p = p_s$ where the network is scale-free in the degree distribution, the degree of each node is measured. Keeping the degree of each node vertex measured in (C), we return to the network configuation 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 grey-color node belonging to the cluster denoted by the dashed line is $37/8$, which is the largest among those of other clusters. That cluster is lethal, we propose.

of clusters $\mathcal{N}(p)$ is measured, where isolated nodes are not counted as individual clusters. The link density p is defined as the 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. After that, links are added more until the link density reaches the density p_s, where the network is scale-free in the degree distribution. At p_s, we measure the degree of each node as depicted in Fig.5C. Keeping the degree of each node at p_s, we return to the network configura-

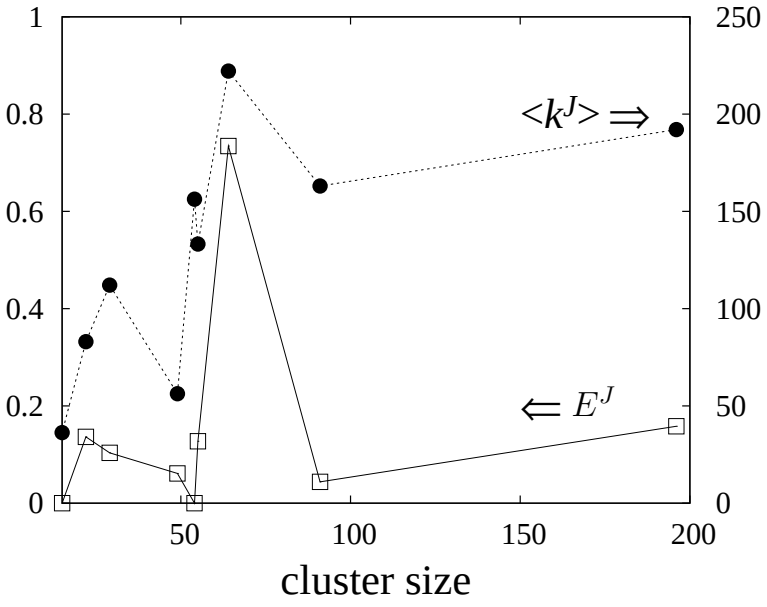

Fig. 6. The comparison between $\langle k^J \rangle$ (●) and $\mathcal{E}^J$ (□) for each cluster indexed by cluster size at p_m.

tion recorded before at p_m. We then calculate the average degree per node 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 the node i belongs to, which was assigned at p_m. $N^J(p_m)$ is the number of nodes belonging to the cluster J at p_m. Then we propose that the cluster with the largest value of $\langle k^J \rangle$ contains high density of essential genes, which is based on the fact that genes with larger degree are more likely to be essential in the protein interaction networks [9].

To check this proposal, we directly measure the essentiality $\mathcal{E}^J$, 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 $\mathcal{E}^J$, behave in the same manner. Thus we can confirm that the cluster containing the largest fraction of essential genes can be found in *in silico* way through $\langle k^J \rangle$. For the yeast dataset, we identify the third largest cluster with 64 genes turns out to be the most essential cluster containing 47 essential genes, 17 nonessential genes, and 1 unidentified gene (Fig. 7). Thus the certainty of selecting essential genes is remarkably improved as high 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 network as we studied in the previous section.

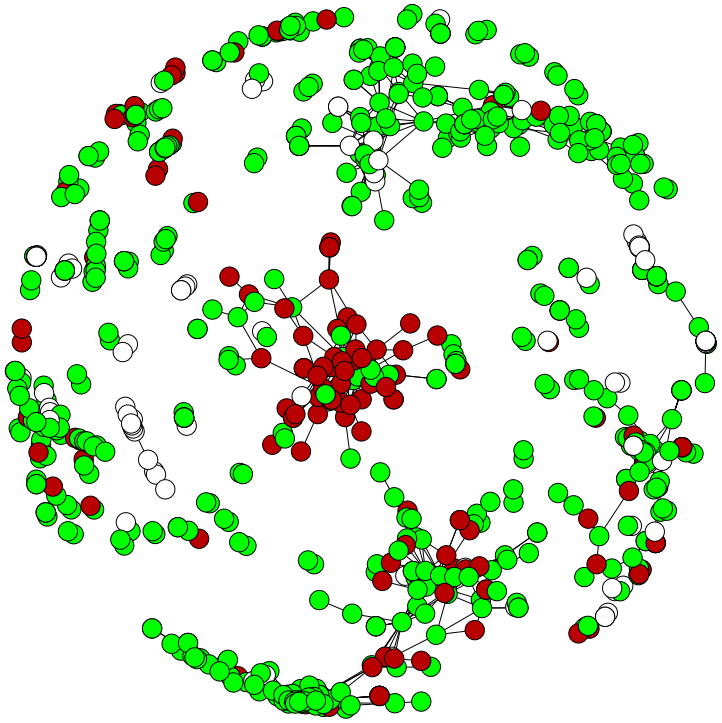

Fig. 7. The gene transcription network of the yeast *S.cerevisiae* at p_m. The red (●), green (●) and white (○) nodes represent essential, nonessential, and unknown genes, respectively.

4 Functional modules

It is well known that biochemical network is composed of modular structure based on its functionality. For 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, of which ratio is shown in Fig. 8. From this figure, one can find 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. Those of the second, third and fourth largest cluster are from the class of small molecule transport, RNA processing/modification, and protein synthesis, respectively. Such a functional clustering in the gene transcription network is rooted from that genes of the same functional category are likely to respond to an external perturbation in a similar way. As a result, the Pearson correlation coefficients between them are large, making clusters at small p_m disconnected 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 such properties, we may assign functional candidates to functionally unknown genes as the most popular functionality in the same cluster.

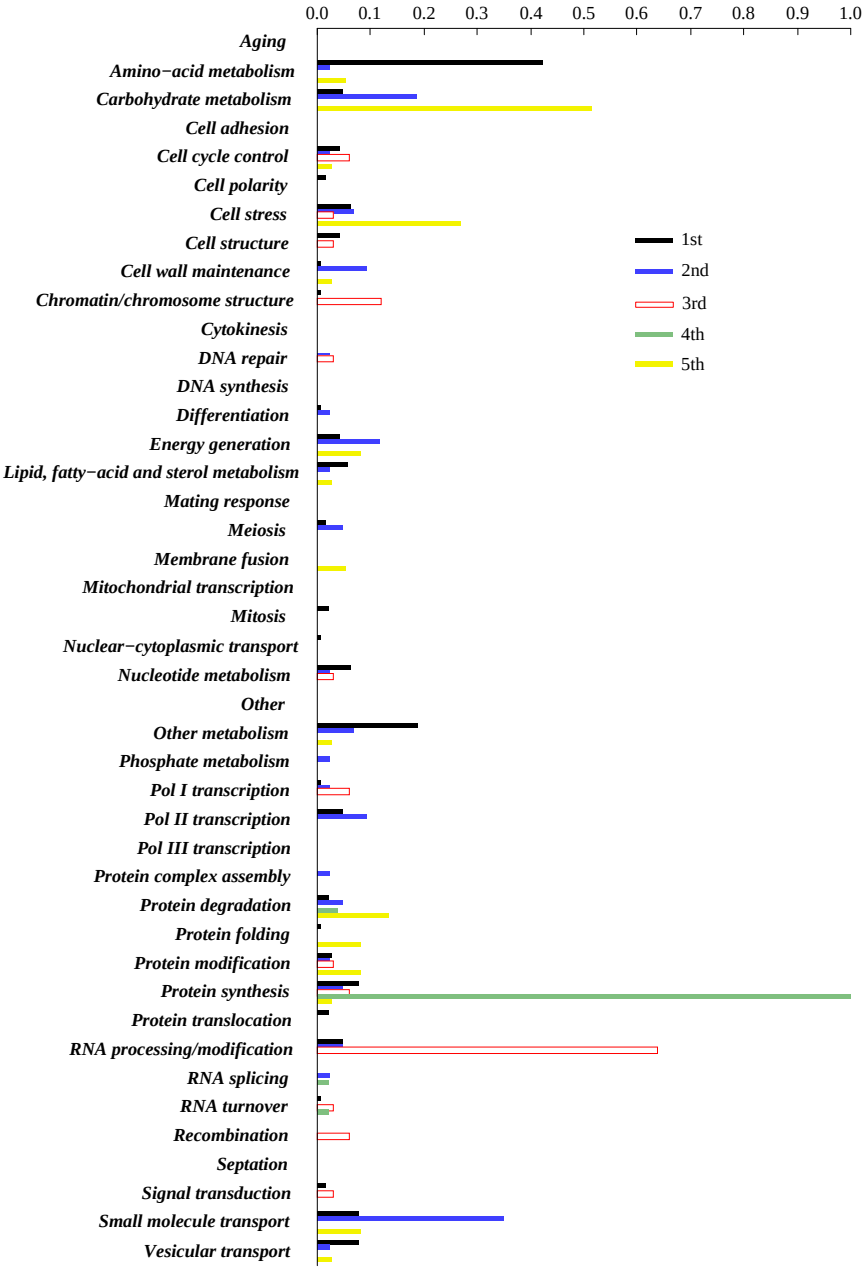

Fig. 8. The genes ratio belonging to each functional category for the genes belonging to the first five largest clusters at p_m.

5 Conclusions and discussion

We have introduced a new method to identify the cluster containing high population of essential genes in the transcription network by using the two known properties

that the genes with the same functionality are highly correlated in the expression level of the microarray assay and the essential genes are likely to have larger degree than others in scale-free network. The certainty of selecting essential genes turns out to be as high as 73%. Thus, such a selecting method could be useful in various knockout problems such as drug designs. Note that our method does not include any tuning parameter, and the selection can be performed in self-organized way with less ambiguity compared with other existing methods.

This work is supported by the KRF Grant No. R14-2002-059-01000-0 in the ABRL program and by 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 at the University of Notre Dame, where this work was initiated.

References

[1] A.-C. Gavin, *et al.*, Nature (London) **415**, 141 (2002).

[2] E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg, Nature (London) **402**, 83 (1999).

[3] A. J. Enright, I. Iliopoulos, N.C. Kryrpides and C.A. Ouzounis, Nature (London) **402**, 86 (1999).

[4] I.S. Kohane, A. J. Butte, and A. Kho, *Microarrays for integrative genomics* (MIT Press, 2002).

[5] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, Proc. Natl. Acad. Sci. USA **95**, 14863 (1998).

[6] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T. R. Golub, Proc. Natl. Acad. Sci. USA **96**, 2907 (1999).

[7] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, Proc. Natl. Acad. Sci. USA **96,** 6745 (1999).

[8] A.J. Butte, P. Tamayo, D. Slonim, T.R. Golub, and I.S. Kohane, Proc. Natl. Acad. Sci. USA **97**, 12182 (2000).

[9] H. Jeong, S.P. Mason, A.-L. Barabási, and Z.N. Oltvai, Nature (London) **411**, 41 (2001).

[10] M.C. Costanzo, *et al.*, Nucleic Acids Res. **29**, 75 (2001).

[11] G. Giaever, *et al.*, Nature (London) **418,** 387 (2002).

[12] T. R. Hughes, *et al.*, Cell **102**, 109 (2000).

[13] N. Guelzim, S. Bottani, P. Bourgine, and F. Képès, Nature Genetics **31**, 60 (2002).

[14] P. Provero, (cond-mat/0207345).

[15] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai, Nature Genetics **31**, 370 (2002).

[16] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.-L. Barabási, Science **297**, 1551 (2002).