Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 154594, 8 pages
http://dx.doi.org/10.1155/2014/154594

Research Article

An Unsupervised Approach to Predict Functional Relations between Genes Based on Expression Data

Md. Altaf-Ul-Amin, Tetsuo Katsuragi, Tetsuo Sato, Naoaki Ono, and Shigehiko Kanaya

Computational Systems Biology Lab, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan

Correspondence should be addressed to Shigehiko Kanaya; skanaya@gtc.naist.jp

Received 1 November 2013; Revised 31 January 2014; Accepted 3 February 2014; Published 31 March 2014

Academic Editor: Farit Mochamad Afendi

This work presents a novel approach to predict functional relations between genes using gene expression data. Genes may have various types of relations between them, for example, regulatory relations, or they may be concerned with the same protein complex or metabolic/signaling pathways and obviously gene expression data should contain some clues to such relations. The present approach first digitizes the log-ratio type gene expression data of *S. cerevisiae* to a matrix consisting of 1, 0, and −1 indicating highly expressed, no major change, and highly suppressed conditions for genes, respectively. For each gene pair, a probability density mass function table is constructed indicating nine joint probabilities. Then gene pairs were selected based on linear and probabilistic relation between their profiles indicated by the sum of probability density masses in selected points. The selected gene pairs share many Gene Ontology terms. Furthermore a network is constructed by selecting a large number of gene pairs based on FDR analysis and the clustering of the network generates many modules rich with similar function genes. Also, the promoters of the gene sets in many modules are rich with binding sites of known transcription factors indicating the effectiveness of the proposed approach in predicting regulatory relations.

1. Introduction

The cell works as a system governed by integrated action of the genes indicating that genes are functionally related; for example, they may have regulatory relations between each other or they may be concerned with the same protein complex or metabolic/signaling pathways and so on. Determining functional relations between genes enables development of a genetic network which leads to the prediction of the complex rolls of the genes in different systems in the cell. Nucleotide and/or amino acid sequence similarities have been extensively used to predict functional relation between genes [1, 2]. Affinity purification [3, 4] and yeast two-hybrid assays [5, 6] are employed to determine physical association between proteins which are gene products. Synthetic lethal screens [7] measure the tendency for genes to compensate the loss of other genes. Scientists have performed numerous studies in an attempt to better understand and classify digenic epistatic relationships [8]. In [9] a probabilistic functional network of yeast genes was constructed by integrating diverse genomic data. In [10] an algorithm was proposed for regulatory networks of gene modules that combines information from genome wide location and expression data sets. Constraint-based Bayesian Structure Learning (BSL) techniques, namely, (a) PC Algorithm, (b) Grow-shrink (GS) algorithm, and (c) Incremental Association Markov Blanket (IAMB), were used to model the functional relationships between genes associated with differentiation potential of aged myogenic progenitors in the form of acyclic networks from the clonal expression profiles [11]. Attempts have been made not only to determine functional relationship between individual genes but also to measure functional relationship between gene sets [12]. Many more similar studies can be cited. Microarray gene expression data incorporating with other information have been extensively used for predicting regulatory relation between genes [13–15]. However it is logical to assume that expression data contains information about various types of functional relations between genes. In the present work we propose an approach for estimating integrated linear and probabilistic relations between expression profiles of genes and applied the concept to determine functional relations between yeast genes solely based on gene expression data. The proposed method successfully detected functionally related gene pairs that share many GO terms. The method also shows promise to be utilized in the process of detecting regulatory relations between genes.

2. Materials and Methods

2.1. Data Used in This Work. The data used in this work was previously used in other works [16–19]. The data is a 2467 × 79 matrix containing some missing values. Each data point produced by a DNA microarray hybridization experiment represents the log of the ratio of expression levels of a particular gene under two different experimental conditions. The result, from an experiment with n genes on a single chip, is a series of n log-transformed expression-level ratios. Typically, the numerator of each ratio is the expression level of the gene in the varying condition of interest, whereas the denominator is the expression level of the gene in some reference condition. The expression measurement is positive if the gene is induced (turned up) with respect to the reference state and negative if it is repressed (turned down). The data were collected at various time points during the diauxic shift, the mitotic cell division cycle, sporulation, and temperature and reducing shocks.

2.2. Missing Value Imputation. In microarray gene expression data missing values often occur due to various reasons, such as insufficient resolution, image corruption, dust, or scratches on the slide. Usually, microarray datasets are estimated to have more than 5% missing values and up to 90% of genes are affected [20, 21]. The gene expression data considered in this work contains 3760 missing values. The missing values were filled based on principal component analysis (PCA) by using the *R* package pcaMethods [22]. Using PCA we can model a matrix M by defining two parameter matrices, the scores, T, and the loadings, P, such that when multiplied with each other they well reconstruct the original matrix as follows:

$$M = 1 \times \overline{m} + TP^t + E, \tag{1}$$

where E is the error matrix and $1 \times \overline{m}$ denotes the original variable averages. Now if M contains missing values but P and T can be completely estimated, then we can use

$$\widehat{M} = 1 \times \overline{m} + TP^t \tag{2}$$

as an estimate for M_{ij} if M_{ij} is missing.

2.3. Digitization of Gene Expression Matrix. After missing value imputation, let us denote the gene expression data matrix as M. For each row of M we calculate the average and standard deviation. Let for the ith row the average and standard deviations be denoted as avg_i and sd_i. Now, the digitized matrix D is created as follows:

$$\begin{aligned} D_{ij} &= 1 \quad \text{if } M_{ij} \geq \text{avg}_i + \text{th} \times \text{sd}_i \\ D_{ij} &= -1 \quad \text{if } M_{ij} \leq \text{avg}_i - \text{th} \times \text{sd}_i \\ D_{ij} &= 0 \quad \text{otherwise.} \end{aligned} \tag{3}$$

In the above equations "th" is a threshold which should be a real number and in most practical cases it is within 0 to 2. We digitized the data using the values of threshold "th" as 0.5, 1, and 1.5. For each case the distribution of the genes with respect to the count of 1 s in their profiles is shown in Figure 1. In case of th = 0.5, the distribution approaches roughly normal and we observed similar trend in case of −1. Hence in this work we considered th = 0.5 for the digitization of the gene expression data.

Table 1: Nine joint probabilities calculated for each gene pair.

a/b	1	0	−1
1	$P(1,1)$	$P(1,0)$	$P(1,-1)$
0	$P(0,1)$	$P(0,0)$	$P(0,-1)$
−1	$P(-1,1)$	$P(-1,0)$	$P(-1,-1)$

2.4. Probability Density Mass Function Table. Based on a digitized matrix containing only 1, 0, and −1 a probability density mass function table can be constructed corresponding to each gene pair indicating nine joint probabilities as shown in Table 1.

Any element of the above table $P(k, k')$ (corresponding to two genes say, gene a and gene b) where $k, k' \in \{1, 0, -1\}$ can be calculated by assuming TRUE = 1 and FALSE = 0 in (4) as follows:

$$P\left(k, k'\right) = \frac{\sum_{i=1}^{N} D_{ai} == k \text{ AND } D_{bi} == k'}{N}. \tag{4}$$

Here N is the width of matrix D.

We assume that the joint probabilities of Table 1 and corresponding conditional probabilities contain important clues to estimate functional relations between genes.

2.5. Hypothesis. In this work we hypothesize that when gene a is positively functionally related to gene b, then $P(b = 1 \mid a = 1)$ should be statistically high. Using Bayes rule we can write $P(b = 1 \mid a = 1) = P(a = 1, b = 1)/P(a = 1)$. Now if $P(a = 1)$ is very small, then $P(b = 1 \mid a = 1)$ can be very high and that can sometimes happen because of noisy data. To avoid this problem we can consider $P(b = 1, a = 1)$ as an indicator that gene a is positively functionally related to gene b. To further strengthen the case we consider that when both $P(b = 1, a = 1)$ and $P(b = 1, a = 1) + P(b = 0, a = 0) + P(b = -1, a = -1)$ are statistically significant then gene a and gene b are positively functionally related. Considering other joint probability masses might be useful for finding functional relations between some multi function genes. By intuition we can realize that the sum of

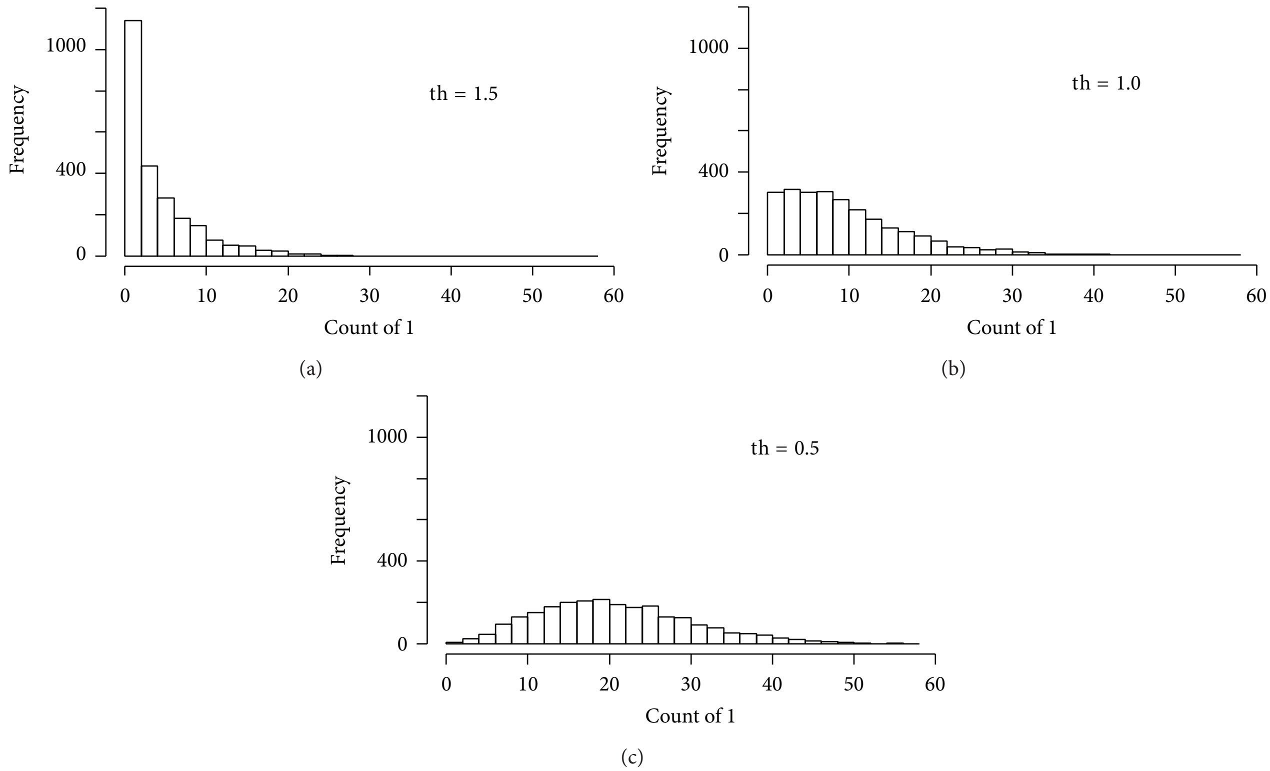


Figure 1: Distribution of the genes with respect to the count of 1 in their profiles in the context of the digitized matrix.

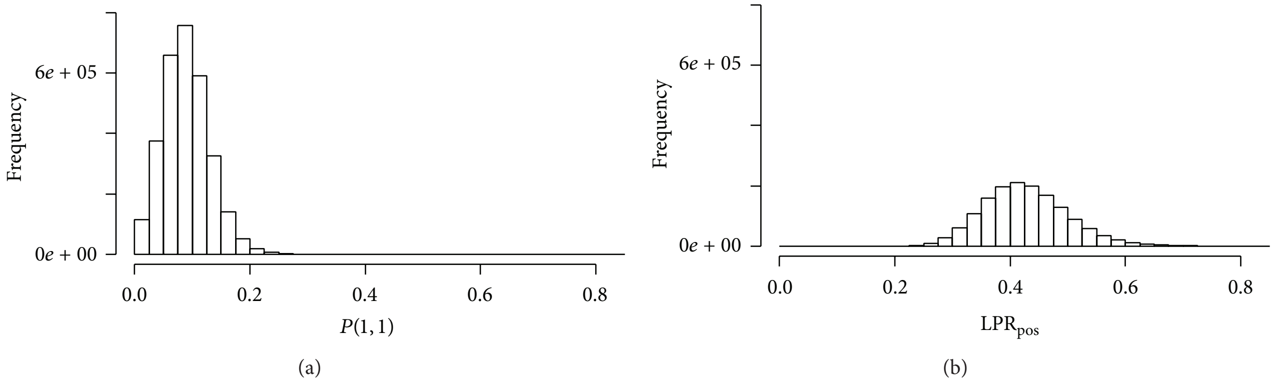


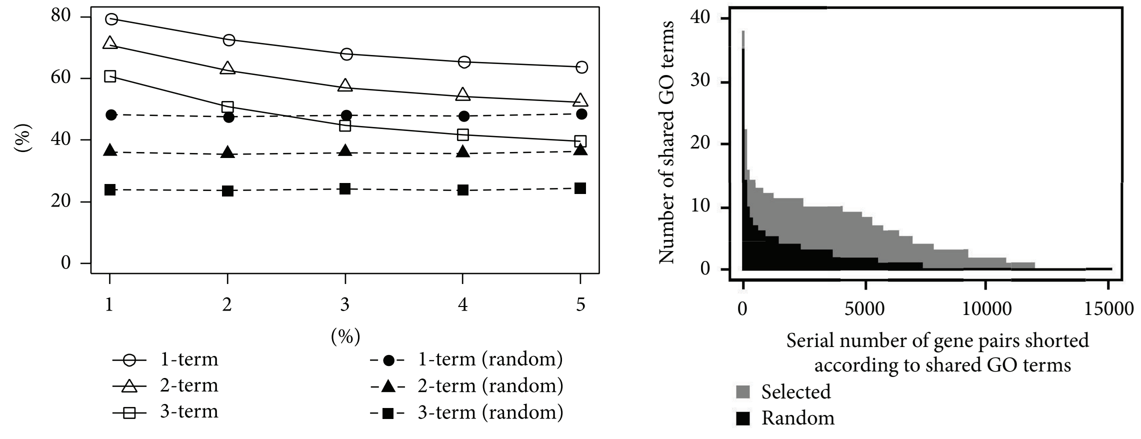
Figure 2: Distribution of gene pairs in the context of (a) $P(1,1)$ and (b) $\mathrm{LPR}_{\mathrm{pos}}$.

probabilities $P(b = 1, a = 1) + P(b = 0, a = 0) + P(b = -1, a = -1)$ actually indicates an integrated measure of both linear and probabilistic relations between the profiles of two genes and this term will be referred to as positive linear and probabilistic relation ($\mathrm{LPR}_{\mathrm{pos}}$) in the following. To our knowledge this is the first approach to measure similarity between two multivariate entities based on joint probability density masses in selected points giving emphasis on both linear and probabilistic relations.

3. Results

3.1. Effectiveness of LPR_{pos}. The distribution of all gene pairs in the context of $P(1,1)$ is shown in Figure 2(a). The average value of $P(1,1)$ is 0.0819. We calculated $\mathrm{LPR}_{\mathrm{pos}}$ for the gene pairs for which $P(1,1)$ is larger than the average value. The distribution of those gene pairs with respect to $\mathrm{LPR}_{\mathrm{pos}}$ is shown in Figure 2(b). The average value of $\mathrm{LPR}_{\mathrm{pos}}$ is 0.429. Initially we selected the highest 1%, 2%, 3%, 4%, and 5% gene pairs from the distribution of Figure 2(b), that is, gene pairs with higher $\mathrm{LPR}_{\mathrm{pos}}$ values, and determined the number of GO terms [23] shared by both the genes of each pair.

Figure 3(a) shows the percentage of selected gene pairs that share at least 1, 2, and, 3 GO terms and also that of equal number of randomly selected gene pairs. In the context of minimum number of shared GO terms the percentage of selected gene pairs is always much higher compared to that



(a)

(b)

FIGURE 3: (a) x-axis is percentage of gene pairs of the distribution of Figure 2(b) selected based on higher LPR_{pos} values and y-axis is percentage of selected gene pairs that share at least 1, 2, or 3 GO terms. Empty markers correspond to gene pairs selected by the proposed method and filled markers corresponding to equal number of randomly selected gene pairs. (b) Actual number of GO terms shared by selected and random gene pairs corresponding to the 1% point of (a).

of randomly selected pairs. Figure 3(a) further shows that the higher the lower cutoff value of LPR_{pos} for a group of gene pairs is, the higher proportion of the gene pairs share common GO terms. To further illustrate the result we show in Figure 3(b) the actual number of shared GO terms for the highest 1% selected gene pairs and the equal number of random gene pairs which implies that the gene pairs selected based on LPR_{pos} share much more GO terms. Thus LPR_{pos} is a good measure to determine functional relation between genes.

3.2. FDR Analysis. We conducted FDR (false discovery rate) [24, 25] analysis to statistically assess the false positive rates among the selected gene pairs based on LPR_{pos}. For each pair of genes for which $P(1,1)$ is above average we did the following.

(i) The numbers of 1 s, 0 s, and −1 s in the digital profile of both genes are counted.

(ii) Random profiles of both the genes are constructed by randomly imputing the same numbers of 1 s, 0 s, and −1 s. This process is repeated 100 times.

(iii) Then, $C(1,1)$, $C(0,0)$, and $C(-1,-1)$ are calculated for both real and random profile pairs. $C(k,k)\{k \in 1,0,-1\}$ is the total number of profile points for which the expression level of both genes is k. In case of random profiles the average values corresponding 100 random profile pairs were considered.

(iv) A chi-square value is calculated as follows where N is the width of the expression matrix:

$$\chi^2 = \left[\sum_{k=1,0,-1} \frac{\{C(k,k)_{\text{real}} - C(k,k)_{\text{random}}\}^2}{C(k,k)_{\text{random}}}\right] + \frac{\{\sum_{k=1,0,-1} C(k,k)_{\text{real}} - \sum_{k=1,0,-1} C(k,k)_{\text{random}}\}^2}{N - \sum_{k=1,0,-1} C(k,k)_{\text{random}}}. \tag{5}$$

(v) Based on the chi-square value, a P-value for the gene pair is determined using R statistical software. Note that LPR_{pos} is directly proportional to $\sum_{k=1,0,-1} C(k,k)_{\text{real}}$.

Figure 4(a) shows the distribution of the gene pairs with respect to the P-values with a P-value interval of 0.05. For any given cutoff P-value the FDR is calculated as follows:

$$\text{FDR} = \frac{(\text{Total \# of gene pairs}) \times (P\text{-value})_{\text{cut-off}}}{\text{\# of gene pairs with } P\text{-value less than } (P\text{-value})_{\text{cut-off}}}. \tag{6}$$

Figure 4(b) shows the plot of FDR with respect to cutoff P-values. As the cutoff P-value decreases, FDR decreases rapidly and becomes roughly constant at P-value of 0.001. There are 25559 gene pairs for which the P-value is less than 0.001.

3.3. Network and Modules of the Selected Gene Pairs. Based on the FDR analysis of the above section, we selected 25559 gene pairs having highest LPR_{pos} values. Such selected gene pairs make a network consisting of 2131 nodes. We determined high density modules in that network using the network clustering algorithm DPClusO [26] and found 1154 modules of size 3 or more (see Supplementary File 1 in supplementary material available online at http://dx.doi.org/10.1155/2014/154594).

3.3.1. Richness of Similar Function Genes. To evaluate the richness of similar function genes in the modules we calculated their hypergeometric P-values by using the R package GOstats [27] in the context of all three types of GO terms: biological process (BP), cellular compartment (CC), and molecular function (MF). Figures 5(a), 5(b), and 5(c) show the distribution of the modules with respect to

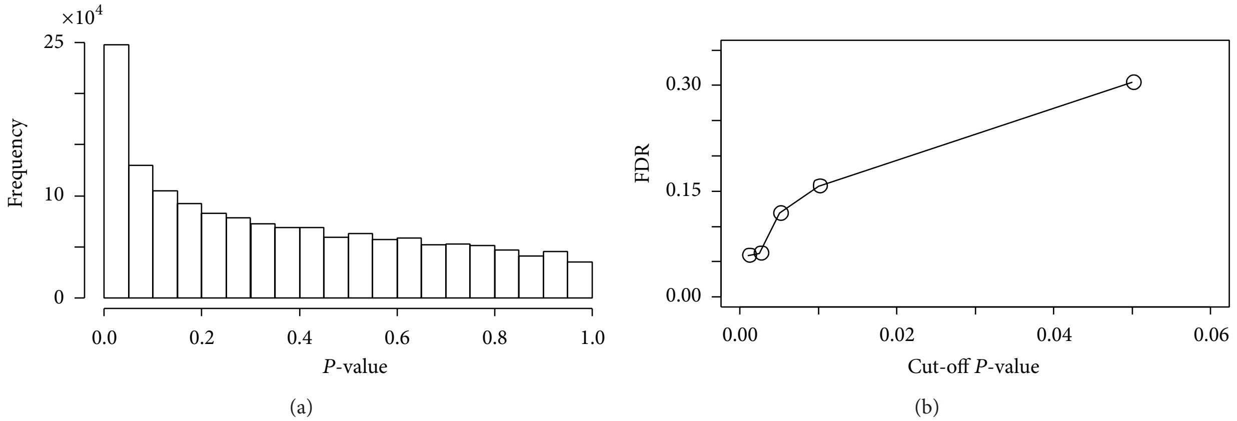


FIGURE 4: (a) Distribution of the gene pairs with respect to the χ-square P-values. (b) Plot of FDR with respect to cutoff P-values.

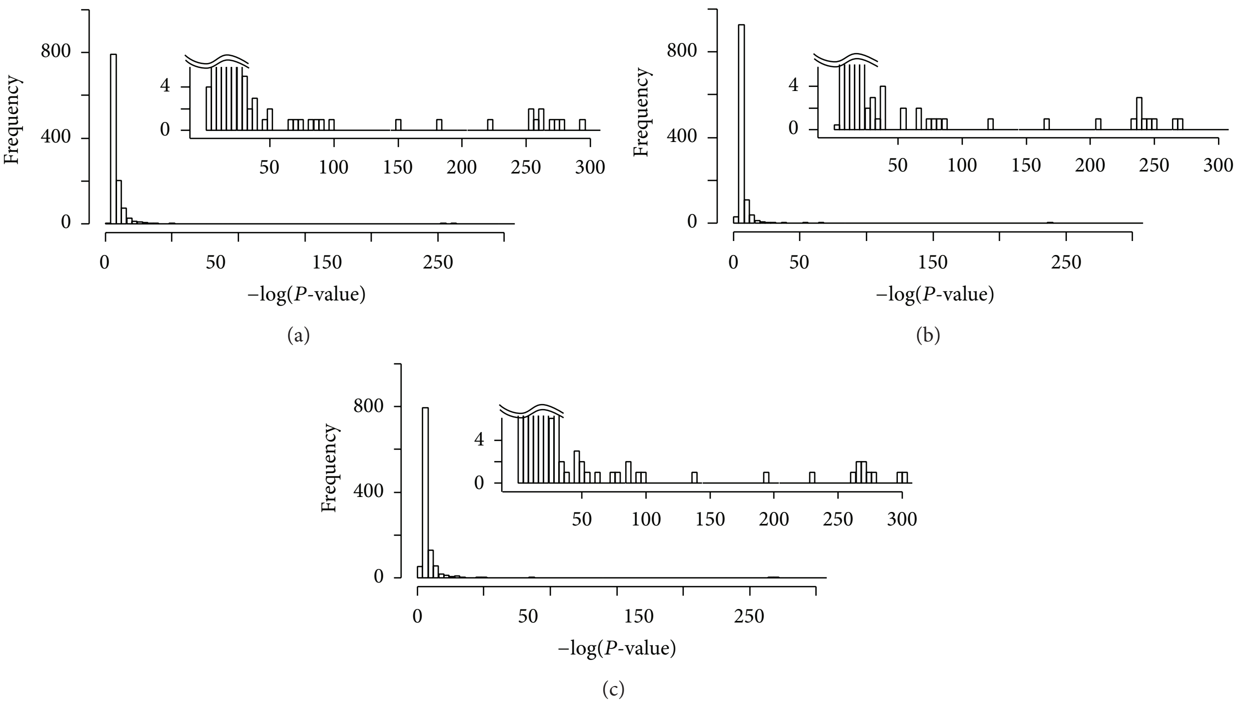


FIGURE 5: Distribution of the modules with respect to −log(P-value). P-values determined in the context of all three types of GO terms (a) biological process (BP), (b) molecular function (MF), and (c) cellular compartment (CC). The lower part of each graph is enlarged in the insets.

−log(P-value) which implies that almost all the modules are statistically significant. We selected 10 lowest P-value clusters corresponding to different GO terms from each of the three distributions of Figure 5 and their set union resulted in 22 clusters. Some biological information from the SGD database [28] about those 22 clusters is summarized in Table 2. Column 3 in Table 2 shows the P-values and corresponding GO terms determined by GOstats. Column 4 in Table 2 shows other GO terms retrieved from SGD database associated to the clusters covering many genes which implies that almost all the genes of each of the clusters could be associated to important GO terms which confirms the fact that the proposed method is a promising way to establish functional relation between genes based on expression data.

3.3.2. Richness of Similar Binding Sites. Furthermore to verify the presence of similar binding sites in the promoters of the genes included in individual modules we used the tool PRIMA (PRomoter Integration in Microarray Analysis) [29] from the software package EXPANDER [30]. Total 180 modules were found to have P-values less than 10^{-3} in the context of binding site enrichment of 57 various transcription

Table 2: Richness of similar function genes in selected clusters. For each cluster, hypergeometric P-values, corresponding GO terms, and also the actual number of genes of a particular function are indicated.

CID	Total number of genes	P-value/GO ID (From GOstats result)	Some relevant GO terms (corresponding number of genes) (From SGD database)
4	97	$1.20E - 131$/GO:0022626 (CC) $2.62E - 117$/GO:0002181 (BP) $4.80E - 129$/GO:0003735 (MF)	Cytosolic ribosome (94), structural constituent of ribosome (94), cytoplasmic translation (93), ribosome (96)
16	76	$6.42E - 24$/GO:0044391 (CC) $7.23E - 17$/GO:0006412 (BP)	Ribosomal subunit (37), structural molecule activity (38)
19	73	$3.29E - 23$/GO:0030529 (CC)	Ribonucleoprotein complex (47), intracellular part (73)
226	8	$1.50E - 20$/GO:0000788 (CC) $1.93E - 14$/GO:0006333 (BP)	Nuclear nucleosome (8), DNA bending complex (8)
1	113	$1.42E - 17$/GO:0042254 (BP)	Cellular metabolic process (104), intracellular part (109)
44	34	$2.89E - 16$/GO:0005840 (CC)	Cytosolic part (21), cytoplasm (34)
35	44	$3.35E - 16$/GO:0010467 (BP)	Gene expression (41), primary metabolic process (43)
85	17	$4.76E - 14$/GO:0044429 (CC)	Mitochondrial part (14), mitochondrion (16)
155	11	$6.28E - 14$/GO:0051082 (MF) $4.97E - 13$/GO:0006457 (BP)	Protein folding (9), protein binding (11), cellular protein metabolic process (10)
278	7	$3.00E - 13$/GO:0000502 (CC)	Proteasome complex (7), proteasome storage granule (5)
87	16	$5.26E - 13$/GO:0005730 (CC)	Nucleolus (12), non-membrane-bounded organelle (14)
107	14	$1.97E - 12$/GO:0007005 (BP)	Mitochondrion organization (12), cellular component organization (13)
121	13	$5.32E - 12$/GO:0006094 (BP)	Glycolysis (7), generation of precursor metabolites and energy (9)
442	5	$1.55E - 11$/GO:0022904 (BP)	Mitochondrial respiratory chain (5), oxidoreductase complex (5)
173	10	$1.56E - 11$/GO:0006457 (BP) $2.58E - 08$/GO:0051082 (MF)	Protein folding (7), unfolded protein binding (5), protein binding (8)
282	7	$5.58E - 11$/GO:0004298 (MF)	Modification-dependent protein catabolic process (7), roteasomal ubiquitin-independent protein catabolic process (5)
71	15	$5.90E - 11$/GO:0005840 (CC)	Ribosome (13), ribonucleoprotein complex (14)
725	3	$1.61E - 09$/GO:0003993 (MF)	Acid phosphatase activity (2)
214	9	$2.88E - 09$/GO:0008121 (MF)	Hydrogen ion transmembrane transporter activity (5), single-organism metabolic process (7)
736	3	$4.03E - 09$/GO:0004067 (MF)	Asparaginase activity (3)
1092	3	$2.26E - 08$/GO:0015002 (MF)	Heme-copper terminal oxidase activity (3)
270	7	$2.32E - 08$/GO:0015078 (MF)	Ion transmembrane transporter activity (6)

Table 3: Richness of binding sites in the promoters of the module genes corresponding to 10 different transcription factors.

CID	Size	TF	Number of Promo. (PRIMA)	P-value	Known regulatory relations (YEASTRACT)
3	98	YP00066 [SFP1]	58	$2.82E - 42$	98
5	95	M00213 [RAP1]	55	$3.82E - 28$	93
72	18	YP00036 [MBP1]	10	$4.40E - 12$	12
155	11	M00169 [HSF]	7	$2.38E - 09$	11
230	8	YP00068 [SIP4]	5	$7.89E - 09$	4
227	8	YP00064 [RPN4]	8	$1.01E - 08$	8
725	3	M00064 [PHO4]	3	$1.08E - 08$	3
259	7	YP00076 [STB1]	5	$8.97E - 08$	2
736	3	YP00013 [DAL82]	3	$3.65E - 07$	0
233	8	YP00043 [MSN4]	8	$1.03E - 06$	7

factors. The enrichment table generated by EXPANDER is in supplementary material (Supplementary Table 1). Table 3 shows information about 10 modules corresponding to lowest *P*-values involving 10 different transcription factors. We downloaded a list of known regulatory relations from the YEASTRACT database [31] and verified whether the genes in a module have regulatory relation with the associated transcription factor. Column 6 of Table 3 shows that a large number of genes in individual modules are already reported to be regulated by the corresponding transcription factor. Only in case of CID736, though all 3 genes contain in their promoters the binding site of the transcription factor DAL82, no regulatory relation between those genes is reported in the YEASTRACT database presently. However based on our analysis regulatory relations between DAL82 and those three genes may be predicted. Thus the proposed measure can also be integrated to other types of information for developing a method to predict regulatory relations between genes which is one of our future works.

4. Conclusions

In this work we propose a novel measure to determine functional relation between genes based on gene expression data. The present approach first digitizes the log-ratio type gene expression data to a matrix consisting of 1, 0, and −1 indicating highly expressed, no major change and highly suppressed conditions for genes, respectively. Then a probability density mass function table is constructed indicating nine joint probabilities for each pair of genes. Those pairs of genes were considered as functionally related for which the sum of probability density masses in selected points are statistically significant. We applied the method to a sample gene expression data of *S. cerevisiae*. It was found that substantial majority of the selected gene pairs share many GO terms. Also the network consisting of the selected gene pairs contains high density modules. It was shown that those modules were rich with similar function genes. Furthermore, it was verified that for many modules many of the genes contain similar binding sites in their promoters corresponding to known transcription factors of yeast and those transcription factors are known regulators of many of the genes in the corresponding module. Above all this work introduces a new approach for simultaneously measuring both linear and probabilistic relations between multivariate entities which is useful for handling multivariate data and big data biology.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research is supported by the National Bioscience Database Center in Japan, the Ministry of Education, Culture, Sports, science, and Technology of Japan (Grant-in-Aid for Scientific Research on Innovation Areas "Biosynthetic Machinery. Deciphering and Regulating the System for Creating Structural Diversity of Bioactivity Metabolites (2007)").

References

[1] T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal, "Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues," *Bioinformatics*, vol. 18, no. 1, pp. S71–S77, 2002.

[2] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 8, pp. 4285–4288, 1999.

[3] A.-C. Gavin, M. Bösche, R. Krause et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[4] Y. Ho, A. Gruhler, A. Heilbut et al., "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[5] P. Uetz, L. Glot, G. Cagney et al., "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*," *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.

[6] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[7] A. H. Y. Tong, M. Evangelista, A. B. Parsons et al., "Systematic genetic analysis with ordered arrays of yeast deletion mutants," *Science*, vol. 294, no. 5550, pp. 2364–2368, 2001.

[8] I. Miko, "Epistasis: gene interaction and phenotype effects," *Nature Education*, vol. 1, article 197, 2008.

[9] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte, "A probabilistic functional network of yeast genes," *Science*, vol. 306, no. 5701, pp. 1555–1558, 2004.

[10] Z. Bar-Joseph, G. K. Gerber, T. I. Lee et al., "Computational discovery of gene modules and regulatory networks," *Nature Biotechnology*, vol. 21, no. 11, pp. 1337–1342, 2003.

[11] R. Nagarajan, S. Datta, M. Scutari, M. L. Beggs, G. T. Nolen, and C. a Peterson, "Functional relationships between genes associated with differentiation potential of aged myogenic progenitors," *Frontiers in Physiology*, vol. 1, article 21, 2010.

[12] Q. Wang, J. Sun, M. Zhou et al., "A novel network-based method for measuring the functional relationship between gene sets," *Bioinformatics*, vol. 27, no. 11, pp. 1521–1528, 2011.

[13] A. A. Margolin, I. Nemenman, K. Basso et al., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, supplement 1, article S7, 2006.

[14] A. Kundaje, M. Middendorf, M. Shah, C. H. Wiggins, Y. Freund, and C. Leslie, "A classification-based framework for predicting and analyzing gene regulatory response," *BMC Bioinformatics*, vol. 7, supplement 1, article S5, 2006.

[15] E. Segal, M. Shapira, A. Regev et al., "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nature Genetics*, vol. 34, no. 2, pp. 166–176, 2003.

[16] M. P. S. Brown, W. N. Grundy, D. Lin et al., “Knowledge-based analysis of microarray gene expression data by using support vector machines,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 1, pp. 262–267, 2000.

[17] A. Mateos, J. Dopazo, R. Jansen, Y. Tu, M. Gerstein, and G. Stolovitzky, “Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons,” *Genome Research*, vol. 12, no. 11, pp. 1703–1715, 2002.

[18] A. Lægreid, T. R. Hvidsten, H. Midelfart, J. Komorowski, and A. K. Sandvik, “Predicting gene ontology biological process from temporal gene expression patterns,” *Genome Research*, vol. 13, no. 5, pp. 965–979, 2003.

[19] G. P. S. Raghava and J. H. Han, “Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein,” *BMC Bioinformatics*, vol. 6, article 59, 2005.

[20] J. Tuikkala, L. Elo, O. S. Nevalainen, and T. Aittokallio, “Improving missing value estimation in microarray data with gene ontology,” *Bioinformatics*, vol. 22, no. 5, pp. 566–572, 2006.

[21] M. Ouyang, W. J. Welsh, and P. Georgopoulos, “Gaussian mixture clustering and imputation of microarray data,” *Bioinformatics*, vol. 20, no. 6, pp. 917–923, 2004.

[22] W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig, “pcaMethods—a bioconductor package providing PCA methods for incomplete data,” *Bioinformatics*, vol. 23, no. 9, pp. 1164–1167, 2007.

[23] M. Ashburner, C. A. Ball, J. A. Blake et al., “Gene ontology: tool for the unification of biology,” *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[24] J. D. Storey and R. Tibshirani, “Statistical significance for genomewide studies,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[25] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society B*, vol. 57, no. 1, pp. 289–300, 1995.

[26] M. Altaf-Ul-Amin, M. Wada, and S. Kanaya, “Partitioning a PPI network into overlapping modules constrained by high-density and periphery tracking,” *ISRN Biomathematics*, vol. 2012, Article ID 726429, 11 pages, 2012.

[27] S. Falcon and R. Gentleman, “Using GO stats to test gene lists for GO term association,” *Bioinformatics*, vol. 23, no. 2, pp. 257–258, 2007.

[28] J. M. Cherry, C. Adler, C. Ball et al., “SGD: saccharomyces genome database,” *Nucleic Acids Research*, vol. 26, no. 1, pp. 73–79, 1998.

[29] R. Elkon, C. Linhart, R. Sharan, R. Shamir, and Y. Shiloh, “Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells,” *Genome Research*, vol. 13, no. 5, pp. 773–780, 2003.

[30] R. Shamir, A. Maron-Katz, A. Tanay et al., “EXPANDER—an integrative program suite for microarray data analysis,” *BMC Bioinformatics*, vol. 6, article 232, 2005.

[31] D. Abdulrehman, P. T. Monteiro, M. C. Teixeira et al., “YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in *Saccharomyces cerevisiae* through a web services interface,” *Nucleic Acids Research*, vol. 39, no. 1, pp. D136–D140, 2011.