PAPER

# An Identification Method of Data-Specific GO Terms from a Microarray Data Set

**Yoichi YAMADA**[†a)], **Ken-ichi HIROTANI**[†], ***Nonmembers***, **Kenji SATOU**[†], ***and*** **Ken-ichiro MURAMOTO**[†], ***Members***

**SUMMARY** Microarray technology has been applied to various biological and medical research fields. A preliminary step to extract any information from a microarray data set is to identify differentially expressed genes between microarray data. The identification of the differentially expressed genes and their commonly associated GO terms allows us to find stimulation-dependent or disease-related genes and biological events, etc. However, the identification of these deregulated GO terms by general approaches including gene set enrichment analysis (GSEA) does not necessarily provide us with overrepresented GO terms in specific data among a microarray data set (i.e., data-specific GO terms). In this paper, we propose a statistical method to correctly identify the data-specific GO terms, and estimate its availability by simulation using an actual microarray data set.

***key words:*** *microarray data set, differentially expressed genes, data-specific GO terms, cell cycle*

## 1. Introduction

Microarray technology is a method in genetic engineering which can monitor the expression of thousands of genes simultaneously [1]. The expression ratio of genes between two samples belonging to distinct conditions is measured in the microarray. For example, when genes display higher expression in a diseased patient than a healthy control, it is conceivable that the function of the genes is more induced in the diseased patient. However, if we know only the name of the genes upregulated in the diseased patient, we can not easily understand what biological events specifically occur in the patient. We therefore need to identify biological terms commonly associated with the upregulated genes.

A number of genes from organisms of over 50 species are annotated to terms, which are defined by Gene Ontology (GO) [2]–[4]. The GO provides us with common terms (i.e., GO terms) for identical biological conception between different organisms or distinct research organizations. The GO terms form a directed acyclic graph (DAG), where each node is a GO term, the root has the most abstract term, and nodes have more concrete terms as going to tips along branch. The GO consortium manages the structured, precisely defined, common and controlled vocabulary for describing the roles of genes in any organism.

Manuscript received May 12, 2008.
Manuscript revised October 31, 2008.

[†]The authors are with the Division of Electrical Engineering and Computer Science, Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa-shi, 920–1192 Japan.
a) E-mail: youichi@t.kanazawa-u.ac.jp

DOI: 10.1587/transinf.E92.D.1093

A preliminary step to computationally analyze microarray data is to identify the induced or suppressed genes (i.e., differentially expressed genes) between two samples in single microarray data or between microarray data. In a next step, each GO term annotation to the differentially expressed genes is statistically evaluated. For instance, when a GO term significantly annotates differentially expressed genes between an object sample and a control in single microarray data, we can understand that the GO term is deregulated between the object sample and the control. On the other hand, the identification of GO terms that significantly annotate differentially expressed genes between microarray data has also been frequently performed. For example, when we compare several microarray data from a control versus a diseased patient with those from a control versus a control, the identification of the differentially expressed genes between two groups of microarray data and their related biological events allows us to determine deregulated biological phenomena in diseased patients compared to controls [5]–[8].

Similar analyses are also frequently conducted between time-course microarray data from several organisms or cultured cells. The main objective of the experiment is to determine what kind of genes and biological events are induced or suppressed in specific time points. Finding deregulated genes and biological events in specific time points has been reported to be useful for the identification of periodically oscillated or stimulation-dependent genes and biological events, etc. [9], [10]. Here we have referred to such deregulated biological terms in specific microarray data as "data-specific GO terms".

Several algorithms and tools have ever been developed to determine the differentially expressed genes and calculate a statistical significance of GO term annotation to those genes [11]–[14]. The simplest method for identifying the differentially expressed genes is based on fold change of gene expression between an object and a control sample or between microarray data. However, it is difficult to choose only one threshold of expression ratio for selecting the differentially expressed genes. If researchers determine only one threshold of the expression ratio to extract the differentially expressed genes, important or unnecessary GO terms may not be or may be detected as statistically overrepresented GO terms, respectively. Recently, the utilization of methods based on fold change decreases due to lacking biological grounds to determine one threshold for selecting the

differentially expressed genes. In this context, we applied a series of multiple thresholds of the ratio to select the differentially expressed genes, and showed it useful to correctly identify their related GO terms.

On the other hand, gene set enrichment analysis (GSEA) has also been developed for identification of deregulated gene sets between microarray data [15]. GSEA can test whether gene sets annotated to GO terms are differentially expressed between microarray data. Therefore it does not require pre-identification of differentially expressed genes for detection of deregulated GO terms between microarray data.

However, above-mentioned approaches are not necessarily enough to find the data-specific GO terms. Because when separate gene groups showing the differential expression in different microarray data share overrepresented annotation of the same GO term, the GO term may be recognized being deregulated in most of a microarray data set: the overrepresentation of this GO term is ubiquitous rather than specific to particular data among a microarray data set. Although several algorithms have been proposed to identify the deregulated GO terms between microarray data, there are few reports referring to this serious problem for identification of the data-specific GO term.

In this paper, we propose a statistical method to correctly identify data-specific GO terms from a microarray data set using multiple thresholds of the expression ratio.

## 2. Materials and Methods

### 2.1 Microarray Dataset

The microarray dataset used in this study was produced by Spellman et al [9]. They synchronized yeast cells by three independent methods: $\alpha$ factor arrest, elutriation, and arrest of a cdc15 temperature-sensitive mutant. After release, yeast cells were periodically recovered and their RNAs were extracted. Control cells were also recovered from asynchronous yeast cells growing in the same culture condition at the same time points and their RNAs were extracted in the same way. Fluorescently labeled cDNA was synthesized from each extracted RNA and the ratio of experimental to control cDNA was measured every recovery time points. The expression ratio of each gene in obtained data was subjected to logarithmic conversion. For our analysis, these logarithmic values were returned to former values by exponential function with base 2.

### 2.2 Proposed Method for Identification of Data-Specific GO Terms from a Microarray Dataset

To determine the data-specific GO terms from a microarray dataset, the following steps 1-3 were carried out.

Step 1: Preparation of a pair-wise comparison matrix between microarray data for each GO term.

| | I | II | III | IV | V | VI | VII | VIII |
|---|---|---|---|---|---|---|---|---|
| I | | | | | | | | |
| II | * | | | | | | | |
| III | | | | | | | | |
| IV | | | | | | | | |
| V | | | | | | | | |
| VI | | | | | | | | |
| VII | | | | | | | | |
| VIII | | | | | | | | |

**Fig. 1** The matrix of pair-wise comparison between microarray data for GO term "A". Each microarray data (I-VIII) has relative fold-inductions of genes in a given sample to a control. The asterisk describes the cell where the calculation in Fig. 2 is conducted.

To compare the expression of genes between microarray data, we first made a matrix for each GO term. As an example, matrix linked to GO term "A" is shown in Fig. 1 where eight microarray data (I-VIII) are compared with each other.

Step 2: Identification of differentially expressed genes between two microarray data and their commonly associated GO terms.

To identify differentially expressed genes between microarray data in a matrix, we calculated the expression ratio of genes in the row index to the column index of each cell excluding those marked by the diagonal line. For instance, the calculation result of the asterisked cell in Fig. 1 is shown in Fig. 2 where the expression ratio of genes in microarray data II to I is calculated and sorted.

Then, genes showing expression ratio over a threshold (i.e., differentially expressed genes between microarray data) were selected from the calculation result in each cell and subjected to statistical testing based on the following equation:

$$p - value = \sum_{j}^{min(n,M)} \frac{{}_{M}C_{j} \cdot {}_{N-M}C_{n-j}}{{}_{N}C_{n}} \quad (1)$$

where $N$ is the number of genes examined by the microarray experiment which we refer to as "population gene set", $M$ is the number of genes annotated to the matrix-linked GO term in the population gene set, $n$ is the number of differentially expressed genes between microarray data, and $j$ is the number of genes assigned to the matrix-linked GO term in the differentially expressed genes. Based on the hypergeometric distribution, this testing examines whether the matrix-linked GO term significantly annotates the differentially expressed genes compared to the population gene set.

Since there are no biological grounds to determine one threshold for selecting the differentially expressed genes between microarray data, the threshold was increased by a certain interval from 1.0 to possible maximum value, and the same statistical testing was repeated for genes showing the

| Gene | Ratio of gene expression between microarray data (II/I) |
|---|---|
| CTT1 | 29.1 (25.4/0.9) |
| SCS7 | 12.3 (49.4/4.0) |
| YDR070C | 10.2 (27.8/2.7) |
| HSP26 | 6.8 (3.8/0.6) |
| SPI1 | 4.9 (8.0/1.6) |
| HSP12 | 3.5 (3.2/0.9) |
| YDR533C | 2.9 (2.0/0.7) |
| MSC1 | 1.7 (5.0/3.0) |
| SNG1 | 1.0 (1.1/1.1) |
| YGR052W | 0.2 (0.6/2.7) |
| YBL048W | 0.1 (1.6/15.9) |
| TFS1 | 0.01 (0.1/9.6) |
| ▪ | ▪ |
| ▪ | ▪ |
| ▪ | ▪ |
| ▪ | ▪ |

**Fig. 2** An example of gene expression ratios which were calculated and sorted in the asterisked cell of Fig. 1.

expression ratio over each threshold.

When the annotation of the matrix-linked GO term showed $p$-value below 0.05 to the differentially expressed genes obtained from at least a threshold, the corresponding cell in the matrix was shaded. For instance, when GO term "A" significantly annotates any differentially expressed genes in microarray data I versus II in Fig. 1, the asterisked-cell in Fig. 1 is shaded gray. Similarly, the same statistical testing preceded by selection of the differentially expressed genes is repeated in the other cells of Fig. 1 except those marked by the diagonal line.

Although results of multiple comparisons need to be corrected, no correction was applied to those results because Bonferroni correction or false discovery rate (FDR) correction was so strict that some matrices had few gray cells in important rows (see Fig. 5). Results of FDR or Bonferroni correction for two GO terms-linked matrices are shown in Fig. 5. FDR and Bonferroni corrections were conducted by the following Eq. (2) and (3), respectively:

$$False\ discovery\ rate\ (\%) = \frac{100 \times N \times P}{n} \tag{2}$$

where $N$ is the number of multiple comparisons, $P$ is the $p$-value in Eq. (1), $n$ is the number of the differentially expressed gene set which displayed the $p$-value under 0.05 in Eq. (1).

$$Bonferroni - corrected\ p - value = N \times P \tag{3}$$

where $N$ is the number of multiple comparisons, $P$ is the $p$-value in Eq. (1). In these corrections, FDR (< 20%) and Bonferroni-corrected $p$-value (< 0.05) were considered as

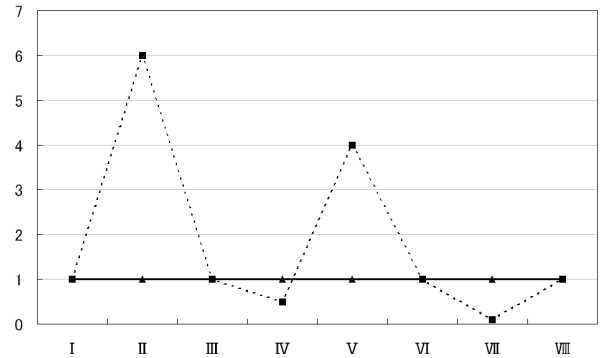



**Fig. 3** The average expression of the population gene set and genes annotated to GO term "A" in each microarray data. Each number (I-VIII) in the horizontal axis represents the microarray data. The vertical axis describes the average expression of genes. The solid line indicates the average expression of the population gene set in each microarray data. The dotted line describes the average expression of genes annotated to GO term "A" in each microarray data.

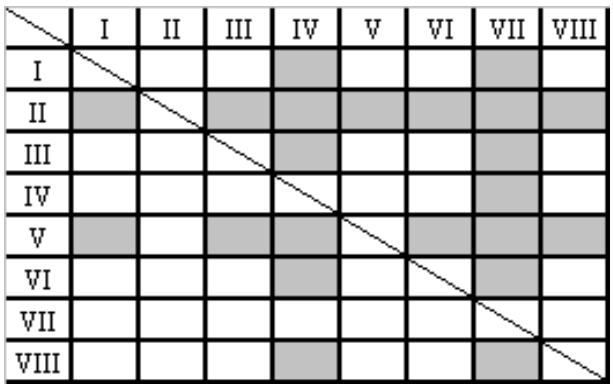



**Fig. 4** The matrix of GO term "A" expected from Fig. 3. The gray cell shows that GO term "A" significantly annotates more upregulated genes in its row index than its column index.

significant annotation of a GO term.

An example of the process described above is shown in Fig. 3 and Fig. 4. Figure 3 depicts the average expression of genes assigned to GO term "A" and the population gene set in each microarray data. The matrix of GO term "A" expected from Fig. 3 is shown in Fig. 4. As is apparent from Fig. 3, the average expression of genes annotated to GO term "A" is higher in microarray data II and V than that of the population gene set. Accordingly, microarray data II and V in the first column of Fig. 4 represent significant annotation of GO term "A" to microarray data I, III, IV, VI, VII and VIII in the first row. Furthermore, since the average expression of GO term "A"-annotated genes is higher in microarray data II than V (see Fig. 3), II in the first column of Fig. 4 shows significant annotation of GO term "A" to V in the first row but not *vice versa*.

In contrast, the average expression of genes annotated to GO term "A" is lower in microarray data IV and VII than that of the population gene set (see Fig. 3). Accordingly, microarray data I, II, III, V, VI and VIII in the first column of Fig. 4 exhibit significant GO annotation to IV and VII in

the first row. The significant GO annotation in IV of the first column to VII of the first row but not in VII to IV results from higher average expression of IV compared to VII. This expression difference between IV and VII also leads to no significant GO annotation in VII of the first column to IV of the first row.

Step 3: Identification of data-specific GO term.

To determine GO terms deregulated in specific microarray data (i.e., data-specific GO term), we examined whether gray cells are enriched in any rows compared to whole cells in each GO term-linked matrix. To examine whether gray cells significantly concentrated in any rows compared to whole cells, a statistical testing was performed in each row by the following equation:

$$p - value = \sum_{j(m)}^{min(n(m),M(m))} \frac{{}_{M(m)}C_{j(m)} \cdot {}_{N(m)-M(m)}C_{n(m)-j(m)}}{{}_{N(m)}C_{n(m)}} \quad (4)$$

where *N(m)* is the number of all cells except those with self-comparison in the matrix, *M(m)* is the number of gray cells in the *N(m)*, *n(m)* is the number of cells in a row, *j(m)* is the number of gray cells in the row. FDR correction was also applied to the results of these multiple comparisons by the following equation:

$$False\ discovery\ rate\ (\%) = \frac{100 \times N(r) \times P(r)}{n(r)} \quad (5)$$

where *N(r)* is the number of multiple comparisons, *P(r)* is the *p*-value in Eq. (4), *n(r)* is the number of rows which displayed the *p*-value under 0.05 in Eq. (4). Consequently, rows which showed false discovery rate under 5% were identified as significantly concentrated rows of gray cells.

Thus, a matrix-linked GO term showing significantly concentrated gray cells in specific rows is the data-specific GO term which is deregulated in specific microarray data.

### 2.3 Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) software was downloaded from the web site: http://www.broad.mit.edu/gsea/downloads.jsp [15]. The yeast cell cycle microarray dataset synchronized by $\alpha$ factor was used as GCT-formatted gene expression data. Each recovery time point name (i.e., 0 min, 7 min, 14 min, ·····., 105 min, 112 min and 119 min) was attached to the microarray data from the same time point name as phenotype label in CLS-format, respectively. GO terms (i.e., "*mitotic cell cycle*", "*DNA replication*", "*sulfur metabolic process*", "*chromatin assembly or disassembly*", "*cytokinesis, completion of separation*", "*telomere maintenance via recombination*", "*DNA unwinding during replication*" and "*response to pheromone*") and gene sets annotated to them were prepared in GMT-format.

For GSEA, "1000", "gene_set", "weighted" and "*log*2 Ratio of Classes" were selected as "Number of permutations", "Permutation type", "enrichment statistic" and "Metric for ranking genes", respectively. In the parameter of "phenotype labels", each microarray data vs the rest was selected. For instance, "0 min vs REST" calculates fold change of each gene in 0 min-labeled microarray data vs the rest of microarray data (i.e., 7 min, 14 min, ·····., 105 min, 112 min and 119 min). Gene sets (i.e., Genes annotated to GO terms), which were upregulated in each microarray data and showed FDR under 20%, were identified as deregulated gene sets (i.e., deregulated GO terms) compared to the other microarray data.

## 3. Experimental Results

To test whether our proposed method can identify the data-specific GO terms from an actual microarray dataset, we used the yeast cell cycle microarray data set produced by Spellman et al [9]. In the dataset, yeast culture was synchronized by $\alpha$ factor and collected every 7 minutes after release. It is generally known that many genes periodically oscillate in the expression during the cell division cycle. Spellman et al. identified about 800 genes regulated in a cell cycle-dependent manner by the combination of a Fourier algorithm and a correlation algorithm. Those genes were furthermore classified by their expression pattern and divided into eight groups termed G1 (CLN2), G1 (Y'), S (Histone), G2 (MET), M (CLB2), M/G1 (MCM), M/G1 (SIC1) and M/G1 (MAT). Since they show periodically oscillated expression and the expression peak in the same time point, we thought that candidates for the data-specific GO terms could be identified from genes within these groups. On the basis of this idea, we explored GO terms that significantly annotate genes within each group by free software, GO term finder (http://www.yeastgenome.org/). The most overrepresented GO term (i.e., GO term with the lowest *p*-value) in each gene cluster is summarized in Table 1. Note that one may not determine these GO terms as the most overrepresented GO term in each gene cluster because annotations for genes change day by day. Then matrices like Fig. 1 were made for each overrepresented GO term. Here, each microarray data in Fig. 1 corresponds to that obtained from each recovery

**Table 1** The most overrepresented GO term in each gene cluster.

| Cluster name | Overrepresented GO term | Rate in cluster | Rate in population | *p*-value |
|---|---|---|---|---|
| G1 (CLN2) | DNA replication | 21 of 58 | 133 of 7291 | 3.75E-20 |
| G1 (Y') | Telomere maintenance via recombination | 5 of 31 | 19 of 7291 | 1.87E-07 |
| S (Histone) | Chromatin assembly or disassembly | 8 of 9 | 118 of 7291 | 1.71E-12 |
| G2 (MET) | Sulfur metabolic process | 12 of 20 | 68 of 7291 | 1.65E-18 |
| M (CLB2) | Mitotic cell cycle | 11 of 35 | 268 of 7291 | 2.05E-06 |
| M/G1 (MCM) | DNA unwinding during replication | 4 of 38 | 10 of 7291 | 3.25E-05 |
| M/G1 (SIC1) | Cytokinesis, completion of separation | 6 of 27 | 11 of 7291 | 1.16E-10 |
| M/G1 (MAT) | Response to pheromone | 6 of 11 | 94 of 7291 | 2.16E-07 |

"Overrepresented GO term" indicates the most overrepresented GO term annotated to genes within each cluster. "Rate in cluster" shows the rate of genes annotated to each overrepresented GO term in each cluster to genes in each cluster. "Rate in population" describes the rate of genes annotated to each overrepresented GO term in the population gene set to the population gene set. "*p*-value" is calculated from the rate in cluster and the rate in population.

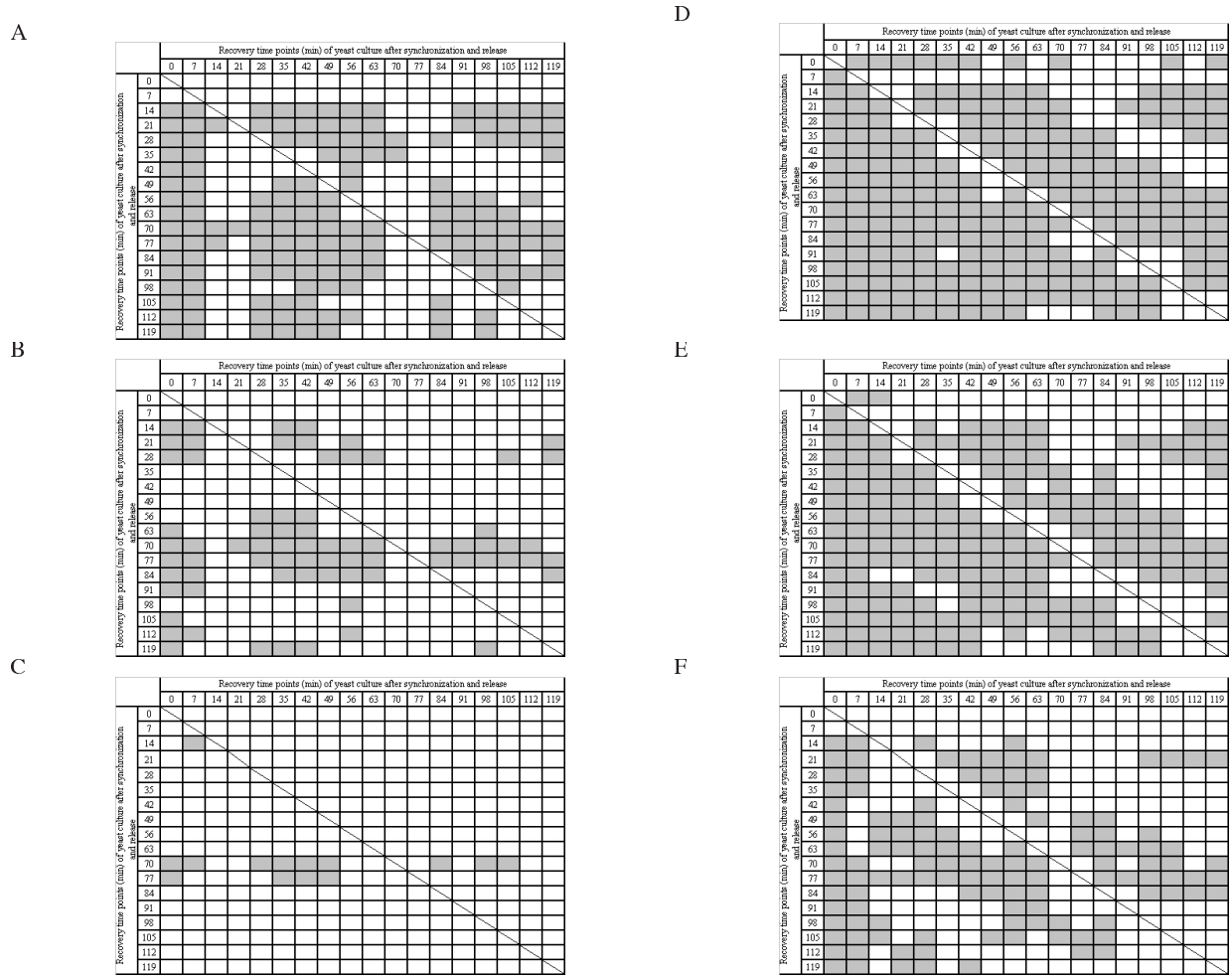



**Fig. 5** The effect of false discovery rate (FDR) correction or Bonferroni correction for matrices of GO term *DNA unwinding during replication* and *mitotic cell cycle*. The A, B and C indicate the application results of no correction, FDR correction and Bonferroni correction in the matrix of GO term *DNA unwinding during replication*, respectively. The D, E and F describe the application results of no correction, FDR correction and Bonferroni correction in the matrix of GO term *mitotic cell cycle*, respectively. The threshold increase interval 0.01 of the expression ratio was used for identification of differentially expressed genes in GO term *DNA unwinding during replication* and *mitotic cell cycle*. In B and E, cells showing FDR under 20% for GO term annotation were shaded. Cells showing Bonferroni-corrected $p$-value under 0.05 for GO term annotation were shaded in C and F.

time point of yeast culture after synchronization and release. As described in "Materials and Methods", gene expression ratios of each microarray data in the first column to each one in the first row were calculated. Every threshold that increases by 0.01 from 1.0 to maximum value, genes showing expression ratio over each threshold were isolated for the succeeding statistical testing. When an overrepresented GO term significantly annotates any of the isolated genes, the corresponding cell in the overrepresented GO term-linked matrix was marked in gray.

Multiple comparisons require any corrections for the calculation results. Figure 5 describes the results of no correction, FDR correction and Bonferroni correction in two GO terms ("DNA unwinding during replication" and "*mitotic cell cycle*")-linked matrices. Numbers of the gray cell in no correction, FDR correction and Bonferroni correction for GO term *DNA unwinding during replication*-linked matrix were 160, 73 and 14, respectively. On the other hand, numbers of gray cell in no correction, FDR correction and Bonferroni correction for GO term *mitotic cell cycle*-linked matrix were 239, 197 and 125, respectively. Thus effects of multiple testing correction for GO term annotaion seemed to vary according to GO terms. In "DNA unwinding during replication"-linked matrix of no correction, rows of 21 min, 70 min and 77 min showed the statistical significance (FDR $< 5\%$ in Eq. (5)) in concentration testing of gray cells in rows (see A in Fig. 5). When FDR or Bonferroni correction was applied to "DNA unwinding during replication"-linked matrix, the row of 21 min did not show the statistical significance (FDR $< 5\%$ in Eq. (5)) in concentration testing of gray cells in the row (see B and C in Fig. 5). Here actual FDRs of 21 min, 28 min, 70 min and 77 min rows in concen-

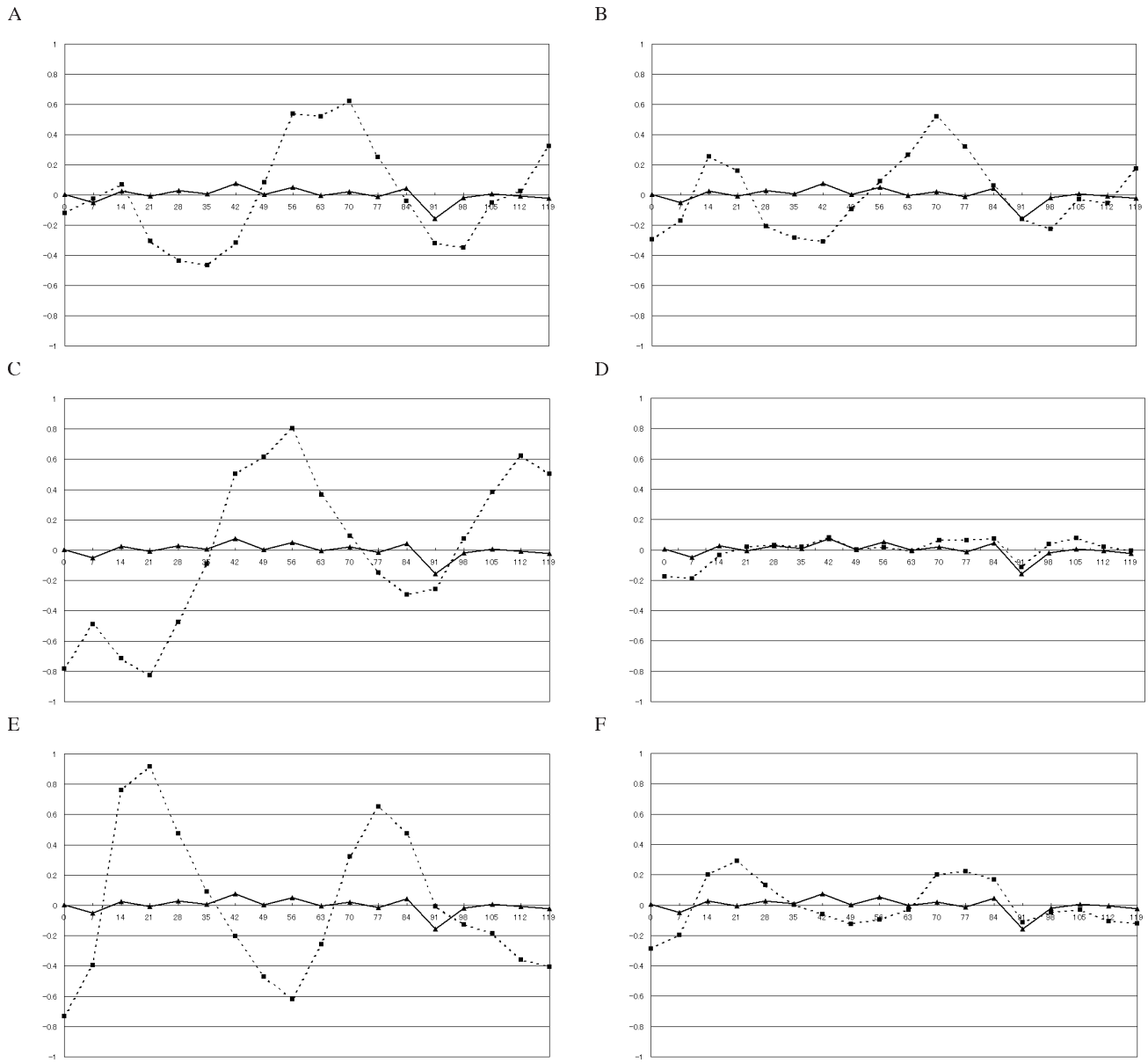



**Fig. 6** Comparison of the average expression between genes in cell cycle clusters and those assigned to overrepresented GO terms in Table 1. The horizontal axis represents the recovery time points (min) of yeast culture after synchronization and release. The vertical axis describes the average expression of genes. The solid line indicates the average expression of the population gene set at each time point. The dotted lines in graphs on the left describe the average expression of genes in each cluster: A, M/G1 (MCM); C, M (CLB2); E, G1 (CLN2). The dotted lines in graphs on the right denote the average expression of genes annotated to the overrepresented GO term in the population gene set: B, genes associated with "*DNA unwinding during replication*"; D, genes associated with "*mitotic cell cycle*"; F, genes annotated to "*DNA replication*".

tration testing of gray cells were 117%, 49%, 0.00015% and 0.00015%, respectively (see B in Fig. 5).

However the average expression of genes assigned to "*DNA unwinding during replication*" was much higher in 14 min, 21 min, 63 min, 70 min and 77 min than that of population gene set (see B in Fig. 6). Considering these results, FDR and Bonferroni corrections seemed to be so strict that the significant GO annotation was not detected in 14 min and 21 min. The same problem to "DNA unwinding during replication" was also observed in "sulfur metabolic process" (data not shown). However, even if Bonferroni correction was applied to "*DNA replication*"-linked matrix, the same problem did not occur at the matrix (data not shown).

Thus application of FDR or Bonferroni correction to matrices may bring about serious problems in a part of matrices. Because of this, we did not apply any corrections to results of the multiple comparisons of GO annotation.

As described above, we next examined whether gray

| GO term | Recovery time points (min) of yeast culture after synchronization and release | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 7 | 14 | 21 | 28 | 35 | 42 | 49 | 56 | 63 | 70 | 77 | 84 | 91 | 98 | 105 | 112 | 119 |
| A | | | X | X | X | | | | | | X | X | X | X | | | | |
| A (>=12) | | | X | X | X | | | | | | X | X | X | X | | | | |
| A (GSEA) | | | X | X | X | | | | | | X | X | X | X | | | | |
| B | | | X | X | X | | | | | | X | X | X | X | | | | |
| B (>=12) | | | X | X | X | | | | | | X | X | X | X | | | | |
| B (GSEA) | | | X | X | X | | | | | | X | X | X | X | | | | |
| C | | | | X | X | X | X | | | | | X | X | X | X | X | X | X |
| C (>=12) | | | | X | X | X | X | | | | | X | X | X | X | X | X | X |
| C (GSEA) | | | | X | X | X | X | | | | | X | X | X | X | X | X | X |
| D | | X | X | | | X | X | X | X | | | | | | | | X | |
| D (>=12) | | X | X | | | X | X | X | X | | | | | | | | X | |
| D (GSEA) | | X | X | | | X | X | X | X | | | | | | | | X | |
| E | | | | X | X | X | X | | | | X | X | X | X | X | X | X | X |
| E (>=12) | | | | X | X | X | X | | | | X | X | X | X | X | X | X | X |
| E (GSEA) | | | | X | X | X | X | | | | X | X | X | X | X | X | X | X |
| F | | | X | X | | | | | X | X | X | X | X | | | | | X |
| F (>=12) | | | X | X | | | | | X | X | X | X | X | | | | | X |
| F (GSEA) | | | X | X | | | | | X | X | X | X | X | | | | | X |
| G | X | X | | | | | | | | | X | X | X | X | X | X | X | X |
| G (>=12) | X | X | | | | | | | | | X | X | X | X | X | X | X | X |
| G (GSEA) | X | X | | | | | | | | | X | X | X | X | X | X | X | X |
| H | X | X | X | | | | | | | | | | | | | | | |
| H (>=12) | X | X | X | | | | | | | | | | | | | | | |
| H (GSEA) | X | X | X | | | | | | | | | | | | | | | |

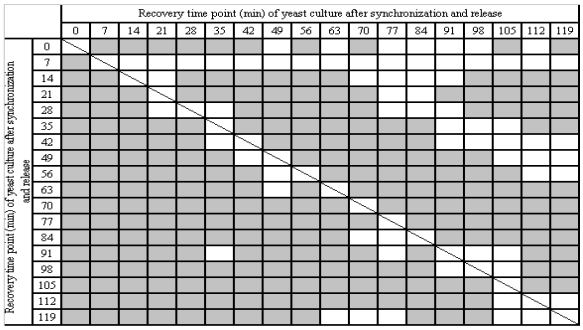
**Fig. 7** The verification of deregulated time points of overrepresented GO terms identified by three methods. The overrepresented GO terms in Table 1 are shown as follows: A, DNA replication; B, telomere maintenance via recombination; C, chromatin assembly or disassembly; D, sulfur metabolic process; E, mitotic cell cycle; F, DNA unwinding during replication; G, cytokinesis, completion of separation; H, response to pheromone. "X" shows the time point where the average expression of genes annotated to overrepresented GO terms is higher than that in the population gene set. For rows of A-H, a shaded square means the time point in which each representative GO term showed a statistical significance in Eq. (5). In rows of A-H (>=12), a shaded square describes that the time point (i.e., the row) in the GO term-linked matrix has 12 gray cells and over. For rows of A-H (GSEA), a shaded square means that the time point showed FDR < 20% in GSEA analysis.

cells were significantly enriched in any rows of the matrix for each overrepresented GO term in Table 1. According to Eq. (4) and (5), statistical test and corrections of multiple comparisons were performed in each row of the overrepresented GO term-linked matrix. When gray cells significantly concentrated (FDR < 5%) in any rows of the overrepresented GO term-linked matrix, the rows were identified as time points at which genes annotated to the matrix-linked GO term have higher expression than others.

The overrepresented GO terms shown in Table 1 are candidates for the data-specific GO term, because they annotate significant number of genes with periodically oscillated expression. If these overrepresented GO terms are actually data-specific GO terms, they will show significant concentration (FDR < 5%) of gray cells around higher expression time points of genes within each cluster in Table 1. We therefore searched the time points in which the average expression of genes within each cluster is higher than that of the population gene set, and marked them by X, as shown in Fig. 7. When increase interval 0.01 of the threshold for isolation of the differentially expressed genes was used, rows (i.e., time points) showing significant concentration of gray cells in the overrepresented GO term-linked matrices were shaded at rows of GO term A-H in Fig. 7. Consequently, shaded cells coincided with ones marked by X at high rates for all overrepresented GO terms except GO term E *mitotic cell cycle*. Moreover although most of gene sets annotated to each overrepresented GO term have two time points showing the expression peak, our method could identify both of the two time points. For instance, since genes annotated to GO term A (i.e., *DNA replication*) display the expression peak in time points of 21 min and 77 min (see F in Fig. 6), two expression peak time points (i.e., 21 min and 77 min) were identified as shaded cells in the row of the GO term A (see Fig. 7).

These results suggested that our proposed algorithm can correctly identify the data-specific GO terms from an actual microarray data set. Furthermore we compared these results with results (i.e., GO term A-H (>=12) in Fig. 7) from our proposed method without application of Eq. (4) and (5) or those (i.e., GO term A-H (GSEA) in Fig. 7) from GSEA method. In our method without using Eq. (4) and (5), rows having 12 gray cells and over were identified as significant concentration of gray cells. GSEA is a method which can test whether gene sets annotated to GO terms are differentially expressed between microarray data [15]. As a result, in both methods, high coincidence between shaded and X-marked cells was observed for all overrepresented GO terms including GO term E *mitotic cell cycle*. Here FDRs of 21 min, 70 min, 77 min, 98 min and 105 min in GSEA of

A




B

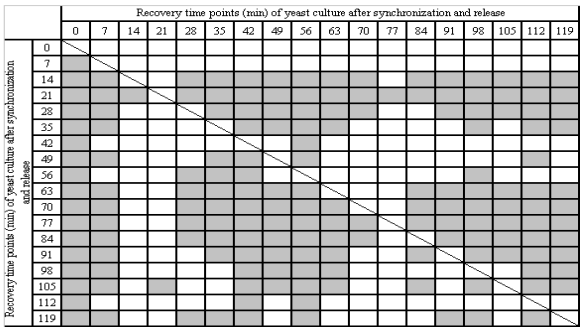



**Fig. 8** Matrices for GO term *mitotic cell cycle* and *DNA replication*. The A and B indicate the matrices for the GO term *mitotic cell cycle* and *DNA replication*, respectively. The shaded square represents that each matrix-linked GO term significantly annotates more induced genes at the time point of its row index than that of its column index.

GO term E *mitotic cell cycle* were 12%, 10%, 1.6%, 17% and 19%, respectively. However our method without Eq. (4) and (5) tended to detect non X-marked cell. These results may result from no correction for multiple comparisons of GO term annotation testing. In addition, GSEA could not identify 0 min of GO term G as a statistical significance different from other two methods. This result suggested that the threshold of 20% for FDR was not so mild in this GSEA.

To examine the reason for these different outcomes between methods, we compared the matrix of "*mitotic cell cycle*" with that of "*DNA replication*", as shown in Fig. 8. Interestingly, most time points in the matrix of "*mitotic cell cycle*" had the statistical significance to most of the others (Fig. 8A). For example, although the row index of 70 min exhibited the statistical significance to the other time points (i.e., gray cells in the row of 70 min), all the row indexes other than 7 min, 14 min, 84 min and 119 min also displayed the statistical significance to the column index of 70 min (i.e., gray cells in the column of 70 min). In contrast, although there were some contradictions between the same time point-labeled row-column pairs, typical contradictions such as *mitotic cell cycle* were not found in the matrix of "*DNA replication*" (Fig. 8B).

The contradictions in the matrix of "*mitotic cell cycle*" suggested that GO term *mitotic cell cycle* annotates separate gene clusters showing high expression in different time points and different expression cycles. In other words, distinct lists of genes with the same annotation may contribute to the statistical significance of cells in the row and the column marked by the same time point, respectively.

In order to investigate whether our interpretation is correct, we compared the average expression of genes in each cluster with that of genes assigned to each overrepresented GO term in the population gene set (Fig. 6). Genes in cluster M (CLB2) showed the periodically oscillated expression through time course (dotted line in Fig. 6C). In contrast, genes associated with GO term *mitotic cell cycle* in the population gene set showed flat expression pattern (dotted line in Fig. 6D) similar to that of the population gene set (solid line in Fig. 6D). On the other hand, genes in cluster G1 (CLN2) exhibited periodic expression pattern in a manner different from those of cluster M (CLB2) (dotted line in Fig. 6E). Moreover, the average expression of genes with annotation of *DNA replication* in the population gene set was also similar, but mild oscillated expression pattern in comparison to that in cluster G1 (CLN2) (dotted line in Fig. 6F). These results strongly supported our idea mentioned above: since GO term *mitotic cell cycle* annotates separate gene clusters showing expression peaks in different time points and different expression cycles, the average expression of genes associated with *mitotic cell cycle* in the population gene set is averaged at most time points. In contrast, the other overrepresented GO terms (e.g., *DNA replication*) are composed of uniform genes showing similar expression patterns, so that the average expression of genes associated with those terms in population gene set is less averaged than *mitotic cell cycle* and maintain periodically oscillated expression through the time course.

Thus, since GO term *mitotic cell cycle* seemed not to be the data-specific GO term, it was suggested that calculations of step 3 in our proposed method are essential for correct identification of the data-specific GO term (see E in Fig. 7). Here note that when we applied FDR or Bonferroni correction to the matrix of "*mitotic cell cycle*", rows of 70 min and 77 min displayed significant concentration of gray cells in FDR-corrected matrix, and rows of 21 min, 70 min, 77 min and 84 min displayed significant concentration of gray cells in Bonferroni-corrected matrix (see E and F in Fig. 5). These results also suggested that the correction for multiple comparisons of GO term annotation may also lead to mis-identification of a part of data-specific GO terms.

Moreover, we also changed the threshold increase interval from 0.01 to 2 for isolation of the differentially expressed genes and examined their effects, as shown in Fig. 9. When we used the threshold increase intervals of 1 and 2, no shaded cell was observed in 77-119 min time points of GO term C which are around the second expression peak point. In addition, threshold increase intervals of 1 and 2 identified GO term E *mitotic cell cycle* as the statistical significance in

| GO term | Recovery time points (min) of yeast culture after synchronization and release | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 7 | 14 | 21 | 28 | 35 | 42 | 49 | 56 | 63 | 70 | 77 | 84 | 91 | 98 | 105 | 112 | 119 |
| A (0.01) | | | X | X | X | | | | | | X | X | X | X | | | | |
| A (0.1) | | | X | X | X | | | | | | X | X | X | X | | | | |
| A (1) | | | X | X | X | | | | | | X | X | X | X | | | | |
| A (2) | | | X | X | X | | | | | | X | X | X | X | | | | |
| B (0.01) | | | X | X | X | | | | | | X | X | X | X | | | | |
| B (0.1) | | | X | X | X | | | | | | X | X | X | X | | | | |
| B (1) | | | X | X | X | | | | | | X | X | X | X | | | | |
| B (2) | | | X | X | X | | | | | | X | X | X | X | | | | |
| C (0.01) | | | | X | X | X | X | | | | | X | X | X | X | X | X | X |
| C (0.1) | | | | X | X | X | X | | | | | X | X | X | X | X | X | X |
| C (1) | | | | X | X | X | X | | | | | X | X | X | X | X | X | X |
| C (2) | | | | X | X | X | X | | | | | X | X | X | X | X | X | X |
| D (0.01) | | X | X | | | X | X | X | X | | | | | | | | X | |
| D (0.1) | | X | X | | | X | X | X | X | | | | | | | | X | |
| D (1) | | X | X | | | X | X | X | X | | | | | | | | X | |
| D (2) | | X | X | | | X | X | X | X | | | | | | | | X | |
| E (0.01) | | | | X | X | X | X | | | | X | X | X | X | X | X | X | X |
| E (0.1) | | | | X | X | X | X | | | | X | X | X | X | X | X | X | X |
| E (1) | | | | X | X | X | X | | | | X | X | X | X | X | X | X | X |
| E (2) | | | | X | X | X | X | | | | X | X | X | X | X | X | X | X |
| F (0.01) | | | X | X | | | | | | X | X | X | X | | | | | X |
| F (0.1) | | | X | X | | | | | | X | X | X | X | | | | | X |
| F (1) | | | X | X | | | | | | X | X | X | X | | | | | X |
| F (2) | | | X | X | | | | | | X | X | X | X | | | | | X |
| G (0.01) | X | X | | | | | | | | | X | X | X | X | X | X | X | X |
| G (0.1) | X | X | | | | | | | | | X | X | X | X | X | X | X | X |
| G (1) | X | X | | | | | | | | | X | X | X | X | X | X | X | X |
| G (2) | X | X | | | | | | | | | X | X | X | X | X | X | X | X |
| H (0.01) | X | X | X | | | | | | | | | | | | | | | |
| H (0.1) | X | X | X | | | | | | | | | | | | | | | |
| H (1) | X | X | X | | | | | | | | | | | | | | | |
| H (2) | X | X | X | | | | | | | | | | | | | | | |

**Fig. 9** Exploration of more suitable threshold increase interval for isolation of differentially expressed genes. The overrepresented GO term in Table 1 is described as follows: A, DNA replication; B, telomere maintenance via recombination; C, chromatin assembly or disassembly; D, sulfur metabolic process; E, mitotic cell cycle; F, DNA unwinding during replication; G, cytokinesis, completion of separation; H, response to pheromone. Numbers in parentheses describe threshold increase intervals for isolation of the differentially expressed genes. The shaded cell means that the time point in each overrepresented GO term displayed a statistical significance in Eq. (5). "X" shows the time point where the average expression of genes annotated to overrepresented GO terms is higher than that in the population gene set.

time points of 70 min and 77 min. As described above, since GO term *mitotic cell cycle* seemed not to be the data-specific GO term, it was suggested that the threshold increase intervals of 1 and 2 are not suitable for extraction of the differentially expressed genes. Similarly, when the threshold increase interval of 0.1 was used, no shaded cell was observed in 14 min and 21 min time points of GO term F which are around the first expression peak point. These results suggested that 0.01 or less increase interval of threshold is more suitable for extracting the differentially expressed genes.

## 4. Conclusion

In this paper, we have proposed novel method to identify the data-specific GO terms from a microarray data set. In spite of considerable attention for identification of differentially expressed genes or deregulated GO terms, the identification method of data-specific GO terms has not been deliberated by researchers. However, for instance, the availability of the cancer classification based on expression pattern of genes annotated to a GO term makes us realize the importance of data-specific GO terms [7]. In this context, we estimated our proposed method using the yeast cell cycle microarray data set in which many genes are differentially expressed between microarray data. Consequently, it was shown that our proposed algorithm can correctly identify the data-specific GO terms from an actual microarray data set, and application of a series of multiple thresholds to the identification of differentially expressed genes between microarray data allows us to correctly identify the data-specific GO terms. Moreover we compared our method with the method without step 3 or GSEA. As a result, it was concluded that any processes (e.g., step 3 in our method) after the identification of deregulated GO terms are essential to correctly determine data-specific GO terms.

Thus, even if we identify deregulated GO terms, data-specific GO terms will not be correctly identified. Since GO terms in higher levels of the DAG generally annotate a number of genes, it is possible that a GO term annotates separate gene clusters showing the differential expression in different microarray data. In such case, a part of GO terms would be misunderstood to be deregulated in most of a microarray data set. At this point, our proposed method may exert important effect for deleting such false positive data-specific GO terms.

## Acknowledgments

This work was in part supported by research grants from Ministry of Education, Science, Sports, Culture and Technology, Japan.

### References

[1] D. Shalon, S.J. Smith, and P.O. Brown, "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization," Genome Res., vol.6, pp.639–645, 1996.
[2] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, "Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium," Nat. Genet., vol.25, pp.25–29, 2000.
[3] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N, Harte, R. Lopez, and R. Apweiler, "The gene ontology annotation (GOA) database: Sharing knowledge in uniprot with gene ontology," Nucleic Acids Res., vol.32 (Database issue), D262–266, 2004b.
[4] Gene Ontology Consortium, "Creating the gene ontology resource: Design and implementation," Genome Res., vol.11, pp.1425–1433, Aug. 2001.
[5] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science, vol.286, pp.531–537, Oct. 1999.
[6] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," Proc. Natl. Acad. Sci. USA, vol.96, pp.6745–6750, June 1999.
[7] R. Maglietta, A. Piepoli, D. Catalano, F. Licciulli, M. Carella, S. Liuni, G. Pesole, F. Perri, and N. Ancona, "Statistical assessment of functional categories of genes deregulated in pathological conditions by using microarray data," Bioinformatics, vol.23, pp.2063–2072, May 2007.
[8] A. Barrier, P.Y. Boelle, F. Roser, J. Gregg, C. Tse, D. Brault, F. Lacaine, S. Houry, M. Huguier, B. Franc, A. Flahault, A. Lemoine, and S. Dudoit, "Stage II colon cancer prognosis prediction by tumor gene expression profiling," J. Clin. Oncol., vol.24, pp.4685–4691, Oct. 2006.

[9] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," Mol. Biol. Cell, vol.9, pp.3273–3297, 1998.
[10] H. Yoshimoto, K. Saltsman, A.P. Gasch, H.X. Li, N. Ogawa, D. Botstein, P.O. Brown, and M.S. Cyert, "Genome-wide analysis of gene expression regulated by the calcineurin/Crz1p signaling pathway in Saccharomyces cerevisiae," J. Biol. Chem., vol.277, pp.31079–31088, Aug. 2002.
[11] J.M. Claverie, "Computational methods for the identification of differential and coordinated gene expression," Hum. Mol. Genet., vol.8, pp.1821–1832, 1999.
[12] P. Khatri and S. Draghici, "Ontological analysis of gene expression data: Current tools, limitations, and open problems," Bioinformatics, vol.21, pp.3587–3595, 2005.
[13] T. Beissbarth and T.P. Speed, "GOstat: Find statistically overrepresented gene ontologies within a group of genes," Bioinformatics, vol.20, pp.1464–1465, Sept. 2004.
[14] E.I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J.M. Cherry, and G. Sherlock, "GO: TermFinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes," Bioinformatics, vol.20, pp.3710–3715, Dec. 2004.
[15] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," Proc. Natl. Acad. Sci. USA, vol.102, pp.15545–15550, Sept. 2005.

**Yoichi Yamada** received B.S. degree from Tsukuba University in 1996. He received M.S. and D.M. degrees in biological science and medical science from Tokyo University in 1998 and 2002, respectively. He is a research associate in the Division of Electrical Engineering and Computer Science at Kanazawa University, Japan. His research interests include molecular biology and bioinformatics.

**Ken-ichi Hirotani** received B.E. and M.E. degrees from Kanazawa University in 2005 and 2007, respectively. His research interests include computer science and bioinformatics.

**Kenji Satou** received the B.E., M.E., and D.E. degrees in computer science and communication engineering from Kyushu University, in 1987, 1989, and 1996, respectively. He was a research associate of Kyushu University, Fukuoka, Japan (1989–1994) and the University of Tokyo, Tokyo, Japan (1995–1997). From 1997 to 2007, he was an associate professor of Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan. He is currently an associate professor of Kanazawa University, Ishikawa, Japan. His research interests include wide variety of topics in bioinformatics.

**Ken-ichiro Muramoto** received B.E. and M.E. degrees from Toyama University in 1971 and 1973, respectively. He received his doctor of medical science degree from Toyama Medical and Pharmaceutical University in the field of neurophysiology. He has also received Ph.D. degree in Engineering from Kyoto University in the field of image information science. He is a professor in the Division of Electrical Engineering and Computer Science at Kanazawa University, Japan. His research interests include image processing, pattern recognition, human vision and remote sensing. He is a member of IEEE.