Noise Analysis of Duplicated Data on Microarrays Using Mixture Distribution Modeling

Masaru Takeya, Takehiro Matsuda[1], Masao Iwamoto[2],
Norimichi Tsumura[1], Toshiya Nakaguchi[1], and Yoichi Miyake[1]

Division of Genome and Biodiversity Research,
National Institute of Agrobiological Sciences,
2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan
[1]*Graduate School of Science and Technology, Chiba University,*
1-33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan
[2]*Division of Plant Sciences, National Institute of Agrobiological Sciences,*
2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan

ABSTRACT

We propose a technique for estimating gene expression values for duplicated data on cDNA microarrays. In the scatter plots, the distribution is constructed from a mixture of normal two-dimensional distributions, which represent fluctuations in gene expression values due to noise. An EM algorithm is used for estimating the modeling parameters. The probability that duplicated data is shifted by noise is calculated using Bayesian estimation. Six data sets of rice cDNA microarray assays were used to test the proposed technique. Genes in the data sets were subjected to clustering based on probability of true value. Clustering successfully identified candidate genes regulated by circadian rhythms in rice.

Key words: cDNA microarray, mixture distribution model, duplicated data, circadian rhythms

1 Introduction

Microarray techniques have recently allowed biological and medical researchers to simultaneously investigate thousands of hybridizations.[1,2] However, down-scaling an experiment makes it more sensitive to internal and external fluctuations,[3] and microarray experiments involve a large number of error-prone steps that lead to a high level of noise in the resulting data.[4-7] To compensate for fluctuations in the values observed in cDNA microarray, genes are frequently duplicated and arranged at different spots on the glass slide.[8,9] Duplicated data measured at two spots for a given gene is often averaged in order to determine gene expression values. However, average values do not always properly account for the influence of noise, and thus taking noise into consideration while determining gene expression values is important.

Scatter plots of double-spotted signals obtained from cDNA microarrays represent the magnitude of fluctuations and correlations between the duplicated spot data (Fig. 1). In the microarray used in this study, double-spotted genes were arranged on both the left and right sides of the slide glass. The points in Fig. 1 indicate double-spotted pairs on one side. Ideally, the points form a straight line along the diagonal. The scatter plot is actually represented as a two-dimensional, rather than a linear, distribution. Deviations from this ideal behavior essentially reflect random fluctuations in multi-noise sources.[4]

In microarray noise modeling, Dror *et al.* represented the transformation between true transcript concentration and observed value as a model form.[7] The true transcript level was successfully estimated from the noise model and the prior distribution of true transcript levels using Bayesian theorem. However, numerous repetitions of microarray assay are required for determining noise parameters. Cho

and Lee successfully estimated a large number of model parameters and true transcript levels from several microarray assays utilizing the iterated computation of Markov chain Monte Carlo techniques.[10] However, the error effect was assumed to be independent and identical to normal distribution, and the parameter distributions were limited to representations as normal or gamma distribution.

In this article, we propose representing the distribution of gene expression values from only one set of duplicated data on cDNA microarrays using a normal mixture distribution of true values. Gene expression values are estimated from the signal values of duplicated data. For this estimation, distribution of all duplicated data on one slide is modeled as a probability distribution. In the scatter plot, assuming that true values exist on the first principal component of the duplicated data and that the probability of noise in each true value has two-dimensional normal distribution, the plot distribution of duplicated data can be represented as a mixture distribution model of multiple normal two-dimensional distributions. Figure 2 shows a schematic diagram of a mixture distribution model for the duplicated data in Fig. 1.

The mixture model provides a flexible and powerful tool to model various random phenomena.[11] In microarray studies, several uses of this modeling technique have been proposed. Allison *et al.* proposed representing the distribution of *p*-values arising from testing the differences between two conditions using a mixture of multiple beta distributions.[12] Pan *et al.* proposed estimating the distribution of a *t*-type test statistic and its null statistic using normal mixture models in order to identify genes with significantly altered expression.[13]

An EM algorithm [14,15] is used for estimating the parameters of the model. The probability that the duplicated data is shifted by noise from one gene expression

value is calculated from the model using Bayesian estimation. Total RNAs extracted from rice leaves at 4-hour intervals on the same day were used in 6 cDNA microarray experiments. These 6 data sets were used to investigate the present technique. The scale was introduced in order to evaluate the accuracy of the estimation for probability distribution function of noise. The genes in the data sets were then applied to clustering based on probability of true value.

2 Systems and Methods

2.1 Mixture distribution model of noise using duplicated data

A schematic diagram of the duplicated data on the microarray used in this study is shown in Fig. 3. This microarray has $2n$ spots duplicated from n genes. The vector $\boldsymbol{x}_i$ of duplicated data i derived from i^{th} gene is expressed as

$$\boldsymbol{x}_i = (x_{i1}, x_{i2})^T, \tag{1}$$

where x_{i1} and x_{i2}, $i = 1,2, \ldots, n$ denote an observed value from the left and right sides of the slide, respectively, and the superscript T represents the transpose of a vector. The observed value is a logarithm of the sum of pixel values measured within the corresponding spot. The $2n$-dimensional vector $\boldsymbol{x}$ of observed data in the microarray is expressed as

$$\boldsymbol{x} = \left(\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \Lambda\ , \boldsymbol{x}_n^T\right)^T. \tag{2}$$

Principal component analysis of $\boldsymbol{x}_i$, $i = 1,2, \ldots, n$, was applied to understand the two-dimensional distribution of duplicated data in Fig. 1. The first principal component expresses the level of gene expression of duplicated data, as it has the biggest variance among the data. In this paper, original gene expression, which is not influenced by noise, is assumed to exist on the first principal component axis.

We call this original gene expression the true value. The second principal component, which is orthogonal to the first, expresses the deviation between duplicated data pairs. Multiple noise levels due to various factors lead to deviations in duplicated data. It is very difficult to estimate the influence of noise from only one pair of duplicated data; however, it is efficient to analyze the features of the distribution using all duplicated data in a microarray. To estimate the influence of noise on each gene, we propose a noise modeling technique using the distribution of duplicated data. The duplicated data distribution is modeled using a mixture of noise distributions centering on the true values. In this paper, the true value was considered to be a discrete value to reduce computation time.

From the proposed model, the probability density function (p.d.f.) of a random vector with an observed random sample $\boldsymbol{x}_i$ is expressed as

$$f(\boldsymbol{x}_i;\boldsymbol{\psi}) = \sum_{j=1}^{g} \pi_j f_j(\boldsymbol{x}_i;\boldsymbol{\theta}_j), \quad , \tag{3}$$

$$\sum_{j=1}^{g} \pi_j = 1 \quad ,$$

where g, π_j, and $f_j(\boldsymbol{x}_i;\boldsymbol{\theta}_j)$ denote the number of the discrete true value, j^{th} mixing proportion and probability distribution function of noise distribution center on the j^{th} true value, respectively. The parameter vectors are represented by $\boldsymbol{\psi}$ and $\boldsymbol{\theta}_j$.

The probability distribution function $f_j(\boldsymbol{x}_i;\boldsymbol{\theta}_j)$ is considered to be normal two-dimensional distribution due to the general definition of random noise as follows,

$$f_j(\boldsymbol{x}_i;\boldsymbol{\theta}_j)=(2\pi)^{-1}\left|\Sigma_j\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{t}_j)^T\Sigma_j^{-1}(\boldsymbol{x}_i-\boldsymbol{t}_j)\right\}, \tag{4}$$

where

$$\boldsymbol{t}_j=(t_{j1},t_{j2})^T,$$

$$\Sigma_j=\begin{pmatrix}\sigma_{j11} & \sigma_{j12}\\ \sigma_{j12} & \sigma_{j22}\end{pmatrix},$$

t_{j1_1}, t_{j2}, σ_{j11}, σ_{j12}, and σ_{j22} denote the corresponding values of the j^{th} true value on the left and right sides, the variance on the left, the covariance of both, and the variance on the right, respectively. The parameters t_{j1_1} and t_{j2} are fixed in order to indicate the true value, which exists on the first principal component axis at regular intervals, while the noise parameters σ_{j11}, σ_{j12}, and σ_{j22} are unknown. The j^{th} parameter vector $\boldsymbol{\theta}_j$ is then expressed as

$$\boldsymbol{\theta}_j=(\sigma_{j11},\sigma_{j12},\sigma_{j22})^T. \tag{5}$$

The parameter vector $\boldsymbol{\psi}$ containing g-1 mixing proportions π_1,Λ,π_{g-1} is expressed as

$$\boldsymbol{\psi}=(\pi_1,\Lambda,\pi_{g-1},\boldsymbol{\theta}_1^T,\Lambda\,\boldsymbol{\theta}_g^T)^T. \tag{6}$$

2.2 Parameter estimation using EM algorithm

Noise parameters of the model are estimated according to eqs. (3) and (4) using observed data. The likelihood function $L(\boldsymbol{\psi})$ for $\boldsymbol{\psi}$ of eq. (3) is expressed as

$$L(\boldsymbol{\psi})=\prod_{i=1}^{n}f(\boldsymbol{x}_i;\boldsymbol{\psi}). \tag{7}$$

The log likelihood function for $\boldsymbol{\psi}$ in eq. (3) is given by

$$\begin{aligned}\log L(\boldsymbol{\psi}) &= \sum_{i=1}^{n} \log f(\boldsymbol{x}_i;\boldsymbol{\psi}) \\ &= \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{g} \pi_j f_j(\boldsymbol{x}_i;\boldsymbol{\theta}_j)\right\},\end{aligned} \tag{8}$$

Differentiating eq. (8) with respect to $\boldsymbol{\psi}$ does not yield an explicit solution, and the following unobservable data vector $\boldsymbol{z}$ is introduced in order to solve the problem,

$$\boldsymbol{z} = \left(\boldsymbol{z}_1^T, \Lambda\ , \boldsymbol{z}_n^T\right)^T, \tag{9}$$

where

$$\boldsymbol{z}_i = \left(z_{i1}, \Lambda\ , z_{ij}, \Lambda\ z_{ig}\right)^T,$$

$$z_{ij} = \begin{cases} 1 & \cdots\ i^{\text{th}} \text{ data comes from } j^{\text{th}} \text{ true value} \\ 0 & \cdots\ \text{otherwise} \end{cases}$$

For example, when the i^{th} data comes from the second true value among three true values,

$$\boldsymbol{z}_i = \left(z_{i1}, z_{i2}, z_{i3}\right)^T = (0,1,0)^T. \tag{10}$$

If z_{ij} were observable, the maximum likelihood estimate $\hat{\pi}_j$ for π_j is given by

$$\hat{\pi}_j = \sum_{i=1}^{n} \frac{z_{ij}}{n}. \tag{11}$$

However, eq. (11) cannot be immediately calculated because z_{ij} is unobservable. In this paper, the EM algorithm is used for estimating the parameters of the model. This algorithm is a computation technique for obtaining maximum likelihood estimates by iterating the E-step and M-step. In the E-step, the conditional expectation of complete-data log likelihood is calculated using the observed data $\boldsymbol{x}$ on the current estimated parameter. The complete-data $\boldsymbol{y}$ is expressed as

$$\boldsymbol{y} = \left(\boldsymbol{x}^T, \boldsymbol{z}^T\right)^T. \tag{12}$$

The complete-data likelihood function $L_c(\boldsymbol{\psi})$ for $\boldsymbol{\psi}$ can be expressed as

$$L_c(\boldsymbol{\psi}) = \prod_{i=1}^{n}\left[\prod_{j=1}^{g}\left\{\pi_j f_j\left(\boldsymbol{x}_i;\boldsymbol{\theta}_j\right)\right\}^{z_{ij}}\right], \tag{13}$$

The complete-data log likelihood function for $\boldsymbol{\psi}$ is expressed as

$$\begin{aligned}\log L_c(\boldsymbol{\psi}) &= \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\log\pi_j f_j\left(\boldsymbol{x}_i;\boldsymbol{\theta}_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\log\pi_j + \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\log f_j\left(\boldsymbol{x}_i;\boldsymbol{\theta}_j\right),\end{aligned} \tag{14}$$

On the $(k+1)^{\text{th}}$ iteration, the conditional expectation of complete-data log likelihood given observed data $\boldsymbol{x}$ is given by

$$\begin{aligned}E_{\boldsymbol{\psi}^{(k)}}\left\{\log L_c(\boldsymbol{\psi})\mid \boldsymbol{x}\right\} &= E_{\boldsymbol{\psi}^{(k)}}\left\{\sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\log\pi_j + \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\log f_j\left(\boldsymbol{x}_i;\boldsymbol{\theta}_j\right)\mid \boldsymbol{x}\right\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{g} E_{\boldsymbol{\psi}^{(k)}}\left(Z_{ij}\mid \boldsymbol{x}\right)\log\pi_j + \sum_{i=1}^{n}\sum_{j=1}^{g} E_{\boldsymbol{\psi}^{(k)}}\left(Z_{ij}\mid \boldsymbol{x}\right)\log f_j\left(\boldsymbol{x}_i;\boldsymbol{\theta}_j\right),\end{aligned} \tag{15}$$

where $E_{\boldsymbol{\psi}^{(k)}}\left(\cdot\mid \boldsymbol{x}\right)$ and Z_{ij} denote the conditional expectation given $\boldsymbol{x}$ on the k^{th} parameter $\boldsymbol{\psi}^{(k)}$ and the random variable corresponding to z_{ij}.

The value of Z_{ij} is 0 or 1 from eq. (9), and $E_{\boldsymbol{\psi}^{(k)}}\left(Z_{ij}\mid \boldsymbol{x}\right)$ is expressed as

$$\begin{aligned}E_{\boldsymbol{\psi}^{(k)}}\left(Z_{ij}\mid \boldsymbol{x}\right) &= 1\cdot P_{\boldsymbol{\psi}^{(k)}}\left(Z_{ij}=1\mid \boldsymbol{x}\right) + 0\cdot P_{\boldsymbol{\psi}^{(k)}}\left(Z_{ij}=0\mid \boldsymbol{x}\right) \\ &= P_{\boldsymbol{\psi}^{(k)}}\left(Z_{ij}=1\mid \boldsymbol{x}\right) \\ &= z_{ij}^{(k)},\end{aligned} \tag{16}$$

where $P_{\boldsymbol{\psi}^{(k)}}\left(\cdot\mid \boldsymbol{x}\right)$ and $z_{ij}^{(k)}$ denote the conditional probability distribution of Z_{ij} given $\boldsymbol{x}$ on the k^{th} parameter $\boldsymbol{\psi}^{(k)}$ and the conditional probability when Z_{ij} is 1, respectively. Substituting eq. (16) into eq. (15) yields

$$E_{\psi^{(k)}}\{\log L_c(\psi)\,|\,x\} = \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}^{(k)} \log \pi_j + \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}^{(k)} \log f_j(x_i;\theta_j). \tag{17}$$

On the $(k+1)^{\text{th}}$ iteration in the EM algorithm, $z_{ij}^{(k)}$ is calculated in the parameter $\psi^{(k)}$ in the E-step, and $\psi^{(k+1)}$ is obtained from the maximum likelihood estimate using $z_{ij}^{(k)}$ in the M-step. These steps are iterated until the difference between $\psi^{(k)}$ and $\psi^{(k+1)}$ becomes less than the stopping criterion. Converged values represent an estimate of parameter ψ. McLachlan *et al.* showed equations the E- and M-steps on the $(k+1)^{\text{th}}$ iteration as follows,[15]

E-step.

$z_{ij}^{(k)}$ can be calculated using Bayes' Theorem according to the following equation,

$$\begin{aligned} z_{ij}^{(k)} &= \frac{\pi_j^{(k)} f_j\left(x_i;\theta_j^{(k)}\right)}{f\left(x_i;\psi^{(k)}\right)} \\ &= \frac{\pi_j^{(k)} f_j\left(x_i;\theta_j^{(k)}\right)}{\sum_{j=1}^{g} \pi_j^{(k)} f_j\left(x_i;\theta_j^{(k)}\right)} \end{aligned} \tag{18}$$

where $z_{ij}^{(k)}$ is a posterior probability, and $\pi_j^{(k)}$ is a prior probability.

M-step.

$\pi_j^{(k+1)}$ is calculated using $z_{ij}^{(k)}$ from eq. (11).

$$\pi_j^{(k+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(k)}}{n} \tag{19}$$

The j^{th} covariance matrix is given by,

$$\Sigma_j^{(k+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(k)} \left(\boldsymbol{x}_i - \boldsymbol{t}_j\right)\left(\boldsymbol{x}_i - \boldsymbol{t}_j\right)^T}{\sum_{i=1}^{n} z_{ij}^{(k)}}, \tag{20}$$

The conditional probability $p(\boldsymbol{t}_j \mid \boldsymbol{x}_i)$, i=1,…,n, j=1,…,g, of true value $\boldsymbol{t}_j$, given observed value $\boldsymbol{x}_i$, is expressed as

$$\begin{aligned} p(\boldsymbol{t}_j \mid \boldsymbol{x}_i) &= \frac{p(\boldsymbol{t}_j)p(\boldsymbol{x}_i \mid \boldsymbol{t}_j)}{\sum_{j=1}^{g} p(\boldsymbol{t}_j)p(\boldsymbol{x}_i \mid \boldsymbol{t}_j)} \\ &= \frac{\pi_j f_j(\boldsymbol{x}_i;\boldsymbol{\theta}_j)}{\sum_{j=1}^{g} \pi_j f_j(\boldsymbol{x}_i;\boldsymbol{\theta}_j)}, \end{aligned} \tag{21}$$

where, by Bayes' Theorem, $p(\boldsymbol{t}_j)$ and $p(\boldsymbol{x}_i \mid \boldsymbol{t}_j)$ correspond to π_j and $f_j(\boldsymbol{x}_i;\boldsymbol{\theta}_j)$, respectively. Equation (21) has the same form as eq. (18), i.e., the convergent z_{ij}, i=1,…,n, j=1,…,g, also represents the probability that observed value $\boldsymbol{x}_i$ is derived from true value $\boldsymbol{t}_j$. An example of probability distribution of estimated true value is shown in Fig. 4. The j^{th} true value in the figure is represented by the mean of true value t_{j1} on the left side and t_{j2} on the right. The expected probability distribution of a true value is defined as the center expression value of the corresponding gene. The possible range excludes 5% on both sides of the probability distribution of $p(\boldsymbol{t}_j \mid \boldsymbol{x}_i)$.

2.3 Clustering using noise modeling of duplicated data

The estimated probability of true value is used for clustering of the gene expression profile. The profile includes center expression values of gene expression

data obtained from numerous microarray experiments; each gene expression profile can be represented as a vector in the data space having dimensions corresponding to the number of experiments. The clustering is based on center expression value. In this study, adaptive quality-based clustering [16] was applied as the method of clustering. This method does not require that the number of clusters be predefined or that each gene in the data set be forced into a cluster.

This clustering procedure first consists of a step to find the center of the cluster in the data space, followed by a step to determine the radius of the cluster. The subspace within the radius around the cluster center is defined as the cluster region. The gene expression profile inside the cluster region is identified with an element of the corresponding cluster. Gene expression profiles assigned to a cluster are excluded in the next cluster search. This procedure is iterated until the stop criterion is satisfied.

In the first step of adaptive quality-based clustering, the mean profile of all expression profiles is initiated as the cluster center. Iteratively, the mean profile of these expression profiles is calculated within a sphere of decreasing radius and subsequently, the cluster center moves toward this mean profile.

If the initialized cluster center exists on a low-density area of the data set, the clustering procedure may stop mid-course. In this paper, we improved the algorithm in order to efficiently identify the cluster center. The data space was previously divided into small areas and data density was calculated for every area. The highest density data center is then initiated as the cluster center, followed by steps to find the cluster center and to determine the cluster radius. During the clustering procedure, the initial cluster center is iteratively selected from a high-density area in the dividend parts.

In the gene profile clusters identified using the conditional probability $p(\boldsymbol{t}_j \mid \boldsymbol{x}_i)$ obtained from the proposed noise modeling technique, the possible range of probability is utilized. An example of the possible range among gene profiles having six gene expression data points is shown in Fig. 5. The gene profile may fluctuate within the possible ranges around the center expression values. By considering the possible range, the gene profile can be handled not as a point but as a territory in the data space. The center expression value is used to cluster representative gene expressions, and genes that are not included in clusters are clustered again based on the possible range. If at least one possible range overlaps a cluster region, the gene can thereby be identified as a candidate element of the corresponding cluster. A schematic diagram to represent relationship between cluster and possible range in three-dimensional space is shown in Fig. 6.

3 Results

3.1 Noise modeling using rice cDNA microarray

Total RNAs extracted from rice leaves at 4-hour intervals on the same day were used for the 6 cDNA microarray assays. The experiments provided 6 sets of duplicated data. In this study, the plotted data included 4515 duplicated data points, including 40 control spots, obtained from a hybridization experiment using a cDNA microarray for rice. The proposed technique was applied to estimate gene expression values based on duplicated values. The probability density distributions of duplicated data and estimation using mixture distribution modeling in array 1 (the 6 microarrays are numbered from 1 to 6 as a matter of convenience) are shown in Figs. 7(a) and (b), respectively. Figure 7(b) shows results of estimation when the number of discrete true values is set to 10, 30, and 60. The scale α is introduced

in order to evaluate the accuracy of estimation for the probability distribution function

$$\alpha = \sum_{p=1}^{m}\sum_{q=1}^{m}\left|\frac{count(p,q)}{n} - \frac{f(\boldsymbol{\beta};\boldsymbol{\psi})}{\sum_{p'=1}^{m}\sum_{q'=1}^{m} f(\boldsymbol{\beta}';\boldsymbol{\psi})}\right|, \tag{22}$$

where

$$\boldsymbol{\beta} = \left(x_{\min} + \left(p-\frac{1}{2}\right)\Delta x, x_{\min} + \left(q-\frac{1}{2}\right)\Delta x\right)^T,$$

$$\boldsymbol{\beta}' = \left(x_{\min} + \left(p'-\frac{1}{2}\right)\Delta x, x_{\min} + \left(q'-\frac{1}{2}\right)\Delta x\right)^T.$$

$x_{\min}$, $x_{\max}$, m, and Δx denote the minimum and maximum of the data set on the left side of a microarray, and the parameters to divide length $|x_{\min}-x_{\max}|$, and length $|x_{\min}-x_{\max}|/m$, respectively. The function $count(p,q)$ outputs the number of duplicated data included within a square region from $x_{\min}+(p-1)\cdot\Delta x$ to $x_{\min}+p\cdot\Delta x$ on the left side and from $x_{\min}+(q-1)\cdot\Delta x$ to $x_{\min}+q\cdot\Delta x$ on the right side. The coefficient α of true values from 1 to 20 is shown in Fig. 8.

3.2 Clustering

Six data sets of the rice cDNA microarray experiments were subjected to clustering using the adaptive quality-based clustering technique. Before the clustering process, the duplicated data was normalized using positive control spots.[17] The results of clustering against 4475 genes except the corresponding 40 control spots are shown in Table 1. Listed are the number of genes having estimated center values of six gene expressions included within the cluster region

and genes having at least one possible range that overlaps the cluster region. In clustering using the adaptive quality-based clustering technique, there are genes that aren't classified into any clusters. The latter column shows more genes can be detected from genes that are classified into other cluster or aren't classified into any clusters in the previous clustering by taking possible ranges into consideration. The proposed method can give biological researchers more genes as candidate of gene function research. To confirm visually the success in the detection of genes that are similar to the target cluster, gene expression behavior in the representative clusters is shown in Fig. 9.

4 Discussion and Conclusion

The estimated mixture probability distributions using the proposed model were very similar to the density distributions of the corresponding experimental data. This indicates the viability of mixture distribution modeling of duplicated data on cDNA microarrays. Figure 7 shows the efficiency of mixture distribution modeling; the use of more than 15 true values provides a more consistent and lower coefficient than non-mixture modeling with one true value.

The objective of the cDNA microarray experiments performed in this study was the discovery of genes regulated by circadian rhythms in rice. Gene expression regulated by circadian rhythms would be represented as a daily fluctuation of RNA levels under continuous dark conditions. In the clustering analysis, cluster 3 exhibited this feature, and thus cluster 3 is believed to include genes regulated by circadian rhythms. The list of genes believed to be regulated by circadian rhythms in cluster 3 is shown in Table 2. Listed are the number of genes assigned to cluster 3 and the genes having at least one possible range included within the cluster

region. The 19 genes in the latter column are a substantial proportion of the 36 in cluster 3. By clustering while taking possible ranges into consideration, more genes can be classified as candidates in biological processes.

In this study, cDNA microarray was used to test our proposed method. This method can also be applied to oligonucleotide arrays because questions of statistical significance and quality control are similar for both types of array.[4)]

Acknowledgement

This work was supported by Grants-in-Aid (Code-No.: 15500202) for Scientific Research from the Japan Society for the Promotion of Science.

References

1) F. C. Holstege, E. G. Jennings, J. J. Wyrich, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander and R. A. Young: Cell **95** (1998) 717.

2) P.O Brown and D. Botstein: Nature Genet. **21** (1999) 33.

3) G. Pietu, O. Alibert, V. Guichard, B. Lamy, F. Bois, E. Leroy, R. Mariage-Sampson, R. Houlgatte, P. Soularue and C. Auffray: Genome Res. **6** (1996) 492.

4) J. Schuchhardt, D. Beule, A. Malik, E. Wolski and H. Eickhoff: Nucleic Acids Research **28** (2000) e47.

5) Y. Balagurunathan, E.R. Dougherty, Y. Chen, M.L. Bittner and J.M. Trent: J. Biomed. Opt. **7** (2002) 507.

6) Y. Tu, G. Stolovitzky and U. Klein: Proc. Natl. Acad. Sci. USA. **99** (2002) 14031.

7) R. O. Dror, J. G. Murnick, N. J. Rinaldi, V. D. Marinescu, R. M. Rifkin and R. A. Young: J. Comput. Biol. **10** (2003) 433.

8) M-.L.T. Lee, F. C. Kuo, G. A. Whitmore and J. Sklar: Proc. Natl. Acad. Sci. U.S.A. **97** (2000) 9834.

9) J. Yazaki, N. Kishimoto, K. Nakamura, F. Fujii, K. Shimbo, Y. Otsuka, J. Wu, K. Yamamoto, K. Sakata, T. Sasaki and S. Kikuchi: DNA Research **7** (2000) 367.

10) H. Cho and J. K. Lee: Bioinformatics **20** (2004) 2016.

11) D. M. Titterington: Statistics **21** (1990) 619.

12) D. B. Allison, G. L. Gadbury, M. Heo, J. Fernández, K-C. Lee, T. A Prolla and R. Weindruch: Comput. Statist. Data Anal. **39** (2002) 1.

13) W. Pan, J. Lin and C. T. Le: Functional & Integrative Genomics **3** (2003) 117.

14) A. P. Dempster, N. M. Laird and D. B. Rubin: J. R. Stat. Soc. B **39** (1977) 1.

15) G. J McLachlan and T. Krishnan: The EM Algorithm and Extensions (John Wiley & Sons, New York, 1996) Chap. 2.

16) F. D. Smet, J. Mathys, K. Marchal, G. Thijs, B. D. Moor and Y. Moreau: Bioinformatics **18** (2002) 735.

17) A. J. Hartemink, D. K. Gifford, T. S. Jaakkola and R. A. Young: Proc. SPIE **4266** (2001) 132.

Figure captions

Fig. 1. Scatter plot of double-spotted signals obtained from cDNA microarray.

Fig. 2. Schematic diagram of mixture distribution modeling for duplicated data.

Fig. 3. Schematic diagram of duplicated data on the microarray used in this study.

Fig. 4. Example of probability distribution of estimated true value.

Fig. 5. Examples of possible range in gene profile.

Fig. 6. Schematic diagram to represent relationship between cluster and possible range

Fig. 7(a) Two-dimensional distribution of probability density of duplicated data on array 1, (b) Estimation using mixture distribution modeling.

Fig. 8. Coefficient α of true values from 1 to 20

Fig. 9. Behavior of gene expressions in clusters 1, 2, and 3. The 20 genes were randomly extracted as samples from each cluster.

Table 1. Clustering of six rice cDNA microarray data sets.

	No. of assigned genes	No. of genes having possible ranges overlapping cluster
Cluster 1	2667	384
Cluster 2	535	255
Cluster 3	183	126
Cluster 4	164	40
Cluster 5	42	25
Cluster 6	49	16
Cluster 7	23	12
Cluster 8	14	10
Cluster 9	15	6
Cluster 10	14	8
Cluster 11	5	3
Cluster 12	7	2
Cluster 13	3	1
Cluster 14	4	1
Cluster 15	4	0
Cluster 16	2	0

Table 2. List of genes potentially regulated by circadian rhythms in cluster 3.

Function	No. of assigned genes	No. of genes having possible ranges overlapping cluster
Amino acid metabolism	1	0
ATP synthesis	1	1
Carbohydrate metabolism	3	0
Cell duplication	0	1
Cell organization	1	3
Lipid metabolism	3	0
Photosynthesis, electron transport	1	1
Protein synthesis & degradation	2	1
Signal transduction	2	3
Transcription	2	1
Transport	1	1
Others	2	0
Unknown	17	7
Total	36	19