Chem-Bio Informatics Journal, Vol. 3, No. 1, pp.30-45(2003)

Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification

Koji Kadota*, Daisuke Tominaga, Yutaka Akiyama, and Katsutoshi Takahashi

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST)

**E-mail: koji-kadota@aist.go.jp*

(Received November 22, 2002; accepted January 30, 2003; published online February 28, 2003)

Abstract

Among samples analyzed for gene expression, samples incorrectly labeled or identified as likely contaminated are those whose expression patterns are markedly different. Such samples should be designated outliers, since they can exert a negative effect on the selection of informative genes for sample classification. We developed a method based on Akaike's Information Criterion (AIC) to detect such outliers. Our method is advantageous because it is free from a significance level and it facilitates objective decision-making. We applied our method to analyze the public microarray data of Alon *et al.* (1999) and found that some of the detected outlying samples coincided with samples considered as likely contaminated. Application of our method produced a higher discrimination level for informative genes in tumor- and normal tissues and, upon exclusion of the outliers, yielded higher classification accuracy. The detection of outlying samples prior to sample classification is essential, and the method described here serves as a valuable check.

Key Words: outlier detection, molecular classification, DNA microarray, AIC, expression analysis

Area of Interest: Genome Wide Experimental Data Analyses

1. Introduction

Sample classification by using microarray data is attracting much interest; its aim is to assign tissue samples to phenotypically characterized categories [1]. Feasible classification methods include the weighted-voting algorithm (WVA) [2], support vector machines (SVM) [3], clustering [4][5], and *k*-nearest neighbors (*k*NN) [6][7]. Hierarchical clustering has been widely used to analyze or group samples based on similarities in their expression patterns. However, this method may not fully extract the information required for sample classification based on gene expression data corrupted by high-dimensional noise [6]. Therefore, perfect classification accuracy remains a

goal to be attained, even when a sophisticated algorithm such as SVM is employed.

While some misclassified samples resulting from incorrect labeling, possible contamination, or heterogeneity [3][6][7][8] may be clearly recognized as outliers [9], recourse to statistical tests remains necessary because of the importance of determining the number of outliers prior to analysis [10].

Statistical means to detect outliers have been proposed [9][10][11][12][13]. The procedure of Kitagawa [13] is based on Akaike's Information Criterion (AIC), means for identifying an optimal model (in this case, a subset of tissues) from among a class of competing models [14]. Its most significant merits are that (i) it allows the simultaneous determination of the number of outliers and testing by comparing the results with AIC values, and (ii) it is independent of a significance level and permits objective decision-making [13]. Ueda's simplification of the method [15] did not alter its performance.

We now describe the application of our simple method for detecting outlying samples from gene expression data. The data used are publicly available colon microarray data obtained from 40 tumor- and 22 normal tissue samples [5]. We focused on the differences in a subset of genes whose expression profiles are significantly different in normal and tumor tissues. We demonstrate the difference in sample classification performance obtained upon comparison of the initial gene expression matrix (in which some samples were identified as likely contaminated and thus termed the 62-heterogeneous-matrix) comprised of 40 tumor- and 22 normal tissue samples, with the "55-homogeneous-matrix," in which no outlying samples were identified.

We document here that our simple method, termed the AIC procedure for detecting Outlying Samples (AICOUS), is able to detect outliers; and we demonstrate the importance of identifying outliers and of excluding them from subsequent sample classification procedures. Our 3 major findings are: (i) some of the 7 outlying samples we detected were identical to those considered contaminated, and/or located in another tissue cluster, or misclassified by other classification methods [3][5][6][7]. (ii) The distinction level of a subset of genes selected from the 55-homogeneous-matrix without outliers was consistently higher than the level in the 62-heterogeneous-matrix and the average levels in 1,000 randomly-selected 55-heterogeneous-matrices. (iii) Cross-validation with WVA and *k*NN showed that the classification accuracy with the genes from the 55-homogeneous-matrix was consistently higher than that for the 62-heterogeneous-matrix and the average of 1,000 55-heterogeneous-matrices.

2. Materials and Methods

2.1 Dataset

Publicly available colon microarray data were used. The data are taken from 40 tumor- and 22 normal samples, each of which contains expression values for 2,000 genes with the highest minimal intensity across the 62 samples (obtained from http://www.molbio.princeton.edu/colondata) [5]. The expression values were subjected to logarithmic transformation (base 10). Some groups have reported sample classification results by using the dataset [3][6][7][16]; some of the samples in the dataset may have been contaminated [6]. Our intent was to document the importance of detecting outlying samples (not misclassified samples) before proceeding to the next step. Hence, only intensively investigated datasets were considered suitable for this study.

2.2 Outlier detection method

A simple method described by Ueda [15] was used to detect outliers in samples from both normal and tumor tissues. The method is a simplified version of the method developed by Kitagawa [13]. Since both methods are based on the AIC, they have several important characteristics: (i) they allow simultaneous determination of the number of outliers and testing, (ii) they do not require selection of a significance level, and (iii) they facilitate objective decision-making [13].

Akaike [14] proposed an information criterion, the AIC, for the identification of an optimal model from a class of competing models. It is approximated as AIC = -2 (maximum log likelihood for n regular observations) + 2(number of independently adjusted parameters). Then, the major part in the probabilistic model of observations can be represented by

$$n!\prod_{j=1}^{n} f(x_j),$$

where n denotes the number of regular (not outlying) samples in each tissue state (tumor and normal) and $f(x_j)$ the probability density function of observations x_j from the normal distribution. Accordingly, the AIC is reflected in the following equations [15]:

$$AIC = 2(n\log\hat{\sigma} - \log n! + s), \tag{1}$$

where (n+s) denotes the total number of samples in each state (tumor and normal), s the number of outlier candidates, and $\hat{\sigma}$ the standard deviation of scores assigned to each of n samples, excluding outlier candidates.

During the development of a simple method for the detection of outliers, Ueda [15] recognized that $\log n!$ could roughly be approximated by $x \times n$. (*e.g.*, x=1 for n=5~9, x=2 for n=10~28). Then, $s - \log n!$ in equation (1) could also be approximated by $2s - const.$ for n=5~9, $3s - const.$ for n=10~28, and so on. Variation of the approximated terms (2s, 3s, …), which are discrete and different from n, is less useful in the actual application, because the total number of samples (n+s) is constant. Therefore, Ueda [15] developed a substitute, $\sqrt{2} \times s \times \frac{\log n!}{n}$, for the terms. The value of the substitute is continuous as it depends on n. Hence, a statistic, U, to identify the outliers is defined as

$$U = n\log\hat{\sigma} + \sqrt{2} \times s \times \frac{\log n!}{n}, \tag{2}$$

The statistic U has a clear interpretation in outlier detection. A low value for the first term in equation (2) indicates that we can predict true outlier(s) from within a combination of outlier candidate(s); this cannot be done in the presence of a high value. The second term in eq. (2) indicates increased unreliability owing to an increased number of parameters. For example, if all observations are derived from the normal distribution, the lowest value for U is the case of s=0 because $\hat{\sigma} \approx 1$ for any number of s. Also, if there are s outliers (s>0) in a set of observations, the value of $\log\hat{\sigma}$ in the first term becomes the lower value, steering U into the minimum. The best approximating combination is one that achieves the lowest value for U and is termed the Minimum AIC Estimate (MAICE). The procedure aimed at obtaining the MAICE for the models is called the minimum AIC procedure. We identified outlying samples corresponding to s by using eq. (2), because Ueda [15] demonstrated the utility of the statistic for the detection of outliers.

Since the statistic was originally developed for one-dimensional observations (scalar data), we expanded it for a two-dimensional array of numbers placed in rows and columns (vector data). This expanded method is termed AICOUS. Consider the following expression vector $E^i = (E^i_1, E^i_2, \ldots, E^i_j)$ for i = 1, 2, …, n+s, where j indicates the number of genes. A distance score D^i (=1-$r^i_{average}$)

for a sample *i* is assigned, where $r^{i}_{average}$ is an average of the Pearson correlation coefficients between the expression vectors of all but *i* samples in the same state versus the expression vector of sample *i*,

$$D^{i} = 1 - \frac{1}{n+s-1} \sum_{k=1,k\neq i}^{n+s} \frac{\sum_{l=1}^{j} (E_l^k - \overline{E^k})(E_l^i - \overline{E^i})}{\sqrt{\sum_{l=1}^{j} (E_l^k - \overline{E^k})^2 \sum_{l=1}^{j} (E_l^i - \overline{E^i})^2}}, \quad (3)$$

where $\overline{E^k}$ is an average expression value for a sample *k* (*i.e.*, $\overline{E^k} = \sum_{l=1}^{j} E_l \Big/ j$).

Next, according to the original methods [13][15], the scores are normalized by subtracting the mean and dividing the result by the standard deviation. The samples are then sorted in order of increasing magnitude of their *Z* scores, such as $Z^1 \leq Z^2 \leq \ \ldots \ \leq Z^{n+s}$. As Ueda did in eq. (2), we assume a normal distribution for sample scores, including those with average scores. It would be ideal if MAICE were decided by considering various combinations of outlier candidates starting from both sides of the *Z* scores. In practice, however, we regard the samples with high scores (*i.e.,* Z^{n+s}, Z^{n+s-1},…) as outlier candidates, because such samples have different expression profiles from the others. Accordingly, we search for the number of outlying samples by starting only from the high side of the *Z* scores in descending order (for example, case 1: Z^{n+s} as outlier, case 2: Z^{n+s} and Z^{n+s-1}, etc) and set the maximum number of the outlier candidates to be half of the (*n*+*s*) samples.

Although we assume that D^i is the realization of a random variable, some high *D* values may have an unknown and/or independent distribution. Moreover, the expression vectors of outlying samples can result in an unfavorable assignment of all *D* values, suggesting that *D* cannot be used directly. Therefore, we adopt the order of samples, but not their values, in searching for the best approximating combination (MAICE). Specifically, in combinations of *s* outlier candidates whose respective values of *D*s are D^{n+s}, D^{n+s-1},…, and D^{n+1}, the other *D* values are calculated by not including the corresponding expression vectors, indicating that the *Z* scores for a sample vary among the combinations. The procedure of recalculating each of the possible combinations can reduce the disadvantageous effect of the expression vectors of outlier candidates without producing artifacts.

2.3 Feature selection and calculation of the distinction level

While many measures have been reported for scoring genes, we used the neighborhood analysis method proposed by Golub *et al.* [2]. We focused on differences in a selected subset of genes and differences in classification accuracy using the subset rather than relevance measures. The measure, *P*(*j*) for a gene *j* is calculated by

$$P(j) = \frac{\mu_{normal}(j) - \mu_{tumor}(j)}{\sigma_{normal}(j) + \sigma_{tumor}(j)}, \quad (4)$$

where [$\mu_{normal}(j)$, $\sigma_{normal}(j)$] and [$\mu_{tumor}(j)$, $\sigma_{tumor}(j)$] denote the mean and standard deviation of log-transformed expression values of gene *j* for samples of normal and tumor tissue, respectively. To evaluate the effects of outliers detected in 40 tumor and 22 normal samples, we compared the following evaluation score *S(m)* between a matrix (row: 2,000 genes, column: 62 samples, called 62-heterogeneous-matrix) and a homogeneous-matrix without outliers:

$$S(m) = \frac{1}{m} \left\{ \sum_{j=1}^{m/2} \left| P_{normal}(j) \right| + \sum_{j=1}^{m/2} \left| P_{tumor}(j) \right| \right\}, \quad (5)$$

where m indicates the number of genes that are well distinguishable between samples of different states (called informative genes); and $P_{normal}(j)$ and $P_{tumor}(j)$ denote the highest measures of the absolute values of $P(j)$ in normal and tumor samples, respectively. The informative genes consist of $m/2$ genes with the highest P_{normal} and of $m/2$ genes with the highest P_{tumor}. The higher $S(m)$ value indicates that the set of normal and tumor samples has a higher normal-tumor distinction level for the m informative genes, with m set at 50, 100, 150, …, 2000. We also compared the values of $S(m)$ calculated for a homogeneous-matrix with those calculated for 1,000 randomly selected sub-heterogeneous-matrices, to determine whether the observed values correlated more highly with the normal tumor distinction than would be expected by chance.

2.4 Classification methods

We used two traditional classification methods to determine the feasibility of our strategy: WVA proposed by Golub *et al.* [2] and *k*NN proposed by Massart *et al.* [17]. WVA employs a voting procedure for classification of a new sample X with m-gene predictors (m = 50, 100, …, 2000) selected by eq. (4). The vote for gene j is performed by

$$v(j) = P(j)\left\{ x(j) - \frac{\mu_{normal}(j) + \mu_{tumor}(j)}{2} \right\}, \tag{6}$$

where $x(j)$ denotes an expression value for gene j in sample X. The total vote V_{normal} for the normal state is obtained by summing the positive votes over the m informative genes. The total vote V_{tumor} for the tumor state is obtained by summing the absolute values of the negative votes. Since we performed leave-one-out cross-validation (LOOCV) tests to distinguish between tumor and normal samples in the dataset, we regarded a positive value of $(V_{tumor} - V_{normal})/(V_{tumor} + V_{normal})$ as a correct prediction, if the unknown sample to be predicted was indeed a tumor sample, and considered the negative value with the unknown normal sample as correct. The other cases were regarded as incorrect.

Since it only considers the neighborhoods of an unknown sample to be predicted, *k*NN is a local method. While there are many variants of the *k*NN algorithm, we used the following conditions: the Pearson correlation coefficient to identify the nearest neighbors of k (arbitrarily set to 3), weight = 1, and the majority vote of the 3-nearest neighbors. Accordingly, for example, an unknown sample X was predicted as being a tumor in a case where two samples of the tumor state and one of the normal state existed among the 3 nearest neighbors of X (and vice versa). The feasibility of our strategy is demonstrated by applying these two algorithms to 3 cases: (i) the 62-heterogeneous-matrix, (ii) a homogeneous-matrix without outliers, and (iii) 1,000 sub-heterogeneous-matrices, whose numbers correspond to those of the homogeneous matrix.

3. Results

Outlying samples have a negative impact on sample classification and present one of the main issues confronting microarray analysis. To resolve this issue, we introduced our AICOUS method for detecting outliers from a gene expression matrix (62-heterogeneous-matrix) that consists of the measurement of 2,000 genes in 22 normal- and 40 tumor samples. We evaluated the irrelevant effects of outlying samples in light of a distinction level for subsets of predictor genes and of classification accuracy.

3.1 Detection of outlying samples using AICOUS

Our method for detecting outliers is based on AIC [14] and was developed for one-dimensional data [13][15]. The most outstanding characteristic of our method is that it allows the output of an objective decision since the procedure is free from the significance level [13]. To deal with two-dimensional microarray data, we performed a modification. To detect outlying samples, we applied our method to two gene expression matrices consisting of 22 normal- and 40 tumor samples.

Table 1 shows the distance scores (D) calculated by using eq. (3) for normal and tumor samples. The lower D^i in sample i is in higher harmony with the other samples of the same state, indicating that it is not an outlier. Samples with a high D value have a high potential to be outliers. Hence, we considered samples with distance scores higher than the mean to be outlier candidates (i.e., 11 combinations for the normal samples and 20 combinations for the tumor samples, Table 1).

Table 2 presents the results of our search. Outliers are detected by searching for a combination of the lowest U [15]. Accordingly, we detected 7 outlying samples: N36, N8, N34, N2, T2, T37, and T6; 5 of the 7 samples coincided with samples reported as misclassified outlying samples, and are likely to have been contaminated [3][5][6][7]. For example, Alon *et al.* [5], who used a clustering method, detected 8 samples as outliers: 3 were normal samples in the created cluster comprised mostly of tumor samples and 5 were tumor samples in the cluster comprised primarily of the normal samples. Of their 8 outlying samples, 4 (N8, N34, T2,

Table 1. Distance scores for each sample.

Normal samples			Tumor samples					
Serial	Name	D	Serial	Name	D	Serial	Name	D
1	N36	0.311	1	T2	0.396	21	T38	0.233
2	N8	0.292	2	T37	0.361	22	T10	0.233
3	N34	0.290	3	T6	0.329	23	T1	0.232
4	N2	0.274	4	T5	0.305	24	T4	0.232
5	N9	0.262	5	T25	0.296	25	T34	0.228
6	N10	0.261	6	T33	0.287	26	T8	0.224
7	N4	0.256	7	T12	0.281	27	T27	0.222
8	N39	0.245	8	T36	0.276	28	T21	0.221
9	N32	0.242	9	T20	0.271	29	T39	0.220
10	N12	0.240	10	T29	0.267	30	T35	0.220
11	N3	0.237	11	T32	0.250	31	T24	0.219
12	N35	0.236	12	T9	0.249	32	T3	0.217
13	N29	0.235	13	T17	0.246	33	T18	0.213
14	N27	0.226	14	T26	0.241	34	T14	0.209
15	N33	0.224	15	T28	0.239	35	T7	0.207
16	N28	0.216	16	T31	0.237	36	T13	0.203
17	N11	0.215	17	T30	0.236	37	T16	0.200
18	N1	0.214	18	T19	0.234	38	T15	0.195
19	N40	0.214	19	T11	0.234	39	T23	0.187
20	N5	0.213	20	T40	0.234	40	T22	0.185
21	N6	0.208						
22	N7	0.201						

Scores in normal and tumor samples were calculated by using equation 3. Samples are sorted in order of the score magnitude. High-scoring samples, such as N36 and T2, can be considered outlier candidates. Shaded samples are regarded as outlier candidates.

Table 2. Search results of outliers in normal and tumor samples.

Normal samples		Tumor samples			
Combination	U	Combination	U	Combination	U
-	-0.51	-	-0.51	1-12	9.29
1	-1.47	1	-3.98	1-13	12.40
1-2	-0.59	1-2	-6.86	1-14	15.45
1-3	-1.99	1-3*	-7.03	1-15	20.67
1-4*	-2.14	1-4	-6.29	1-16	22.88
1-5	0.43	1-5	-5.55	1-17	24.84
1-6	3.64	1-6	-4.12	1-18	26.59
1-7	0.07	1-7	-1.49	1-19	31.08
1-8	3.56	1-8	0.51	1-20	32.10
1-9	8.70	1-9	2.46		
1-10	10.17	1-10	4.23		
1-11	10.02	1-11	7.75		

The statistic U for detecting outliers was calculated by using equation 2. Numbers in the "Combination" column correspond to those in the "Serial" column in Table 1. A combination "-" indicates that nothing is regarded as an outlier. We searched outlying samples from 11 combinations for normal samples and 20 combinations for tumor samples as practically considerable. A combination with the lowest statistic is the solution. Hence, we detected 7 samples in the combination with an asterisk (N36, N8, N34, and N2 in normal samples and T2, T37, and T6 in tumor samples) as outliers.

and T37) coincided with ours; 6 samples (N8, N34, N36, T30, T33, and T36) were misclassified by the SVM method [3]. Li *et al.* [6] reported that 5 identified samples (N34, N36, T30, T33, and T36) were likely contaminated. Li *et al.* [7] also reported that when these 5 questionable samples were excluded, all but one (N8) of the remaining 57 samples were correctly classified. Three of our samples (N8, N34, and N36) were identical to samples reported by Li *et al.* [6][7]. The overall relationships are shown in Table 3. Our results document that our method can detect some true outliers.

3.2 Differences between informative genes with and without outliers

To investigate the unfavorable/negative effects of the 7 outliers we detected, we scored each of the genes selected from the 62-heterogeneous-matrix and the 55-homogeneous-matrix without the outliers by neighborhood analysis [2]. We then evaluated the distinction level between the different states on *m*-gene subsets (m = 50, 100, …, 2000) selected from the following three matrices: (i) the 62-heterogeneous-matrix, (ii) the 55-homogeneous-matrix, and (iii) 1,000 55-heterogeneous-matrices consisting of 18 samples randomly selected from among 22 normal samples and 37 samples randomly selected from among 40 tumor samples. We used an evaluation score $S(m)$ to determine the distinctness of an *m*-gene subset (see Methods). High evaluation scores indicate a strong correlation between the gene expression profile and the distinction level.

Table 3. Comparison of outliers detected by other reports.

	Outliers	
	tumor samples	normal samples
Alon et al. 1999	T2, T30, T33, T36, T37	N8, N12, N34
Furey et al. 2000	T30, T33, T36	N8, N34, N36
Li et al. 2001a	T30, T33, T36	N34, N36
Li et al. 2001b		N8
Our result	T2, T6, T37	N2, N8, N34, N36

Five of seven detected outliers coincided with samples reported as being unfavorable by at least one of four other reports.

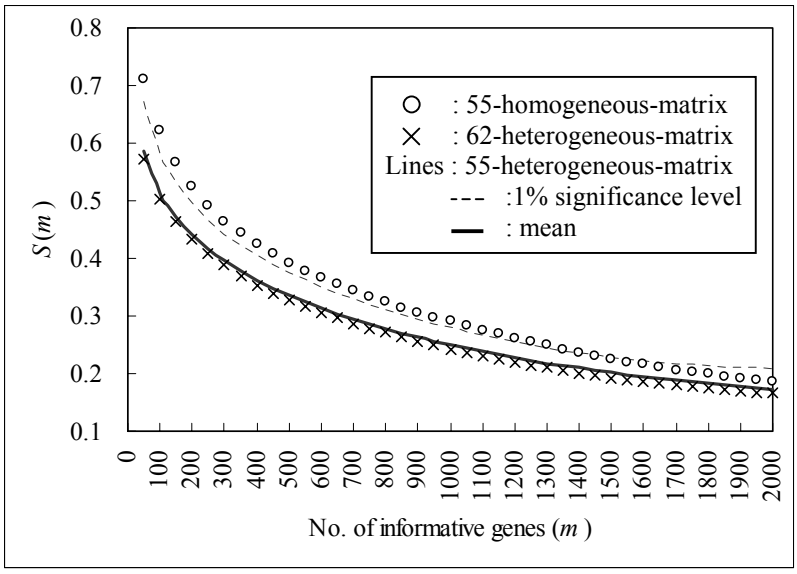

Figure 1. Plot of evaluation scores $S(m)$ for 62-heterogeneous-matrix, 55-homogeneous-matrix, and 55-heterogeneous-matrix.

Circles and crosses designate scores for *m* genes selected in the 55-homogeneous-matrix and the 62-heterogeneous-matrix, respectively. Thick and broken lines represent mean score and the 1% significance level, respectively, for the *m* genes in the 1,000 55-heterogeneous-matrices. Note the higher scores of the 55-homogeneous-matrix than of the 62-heterogeneous-matrix.

As shown in Figure 1, the scores for *m*-gene subsets (m = 50, 100, …, 1400) from the 55-homogeneous-matrix were above the 1% significance level (broken line) from 1,000 randomly selected 55-heterogeneous-matrices. The scores were also higher than those from the 62-heterogeneous-matrix, suggesting that we could correctly detect outlying samples.

Table 4 shows the 50 top-ranking genes from the 62-heterogeneous- and the

55-homogeneous-matrix. $P(j)$ values were clearly higher in the genes from the 55-homogeneous-matrix and 13 genes differed between the two matrices. For example, cyclin-dependent kinase is up-regulated in tumor samples in the 62-heterogeneous-matrix but not in the 55-homogeneous-matrix. This is reasonable because the gene is known to be a negative prognostic marker in colorectal tumors [18].

Table 4. Comparison of 50 high-ranking genes in 55-homogeneous-matrix and 62-heterogeneous-matrix.

Genes highly expressed in tumor samples						Genes highy expressed in normal samples					
with outliers			without outliers			with outliers			without outliers		
rank	gene j	$P(j)$	gene j	$P(j)$		rank	gene j	$P(j)$	gene j	$P(j)$	
1	1042	-0.740	513	-0.907		1	493	0.834	493	0.996	
2	1772	-0.727	1042	-0.818		2	249	0.711	249	0.971	
3	1671	-0.722	780	-0.802		3	1423	0.688	765	0.931	
4	625	-0.696	1671	-0.798		4	377	0.681	245	0.890	
5	513	-0.688	625	-0.768		5	897	0.656	267	0.863	
6	1771	-0.668	1772	-0.760		6	765	0.639	1423	0.852	
7	1582	-0.666	365	-0.732		7	1635	0.635	66	0.836	
8	780	-0.620	1060	-0.731		8	245	0.613	377	0.832	
9	1060	-0.620	241	-0.725	27	9	66	0.589	1635	0.810	
10	964	-0.607	1730	-0.715		10	267	0.575	897	0.807	
11	365	-0.606	1771	-0.714		11	1843	0.567	1494	0.760	
12	138	-0.600	1153	-0.707		12	1494	0.544	1411	0.654	
13	399	-0.599	26	-0.704		13	822	0.539	1843	0.635	
14	1730	-0.580	964	-0.702		14	1668	0.480	33	822	0.631
15	1153	-0.578	1002	-0.695	31	15	1967	0.453	27	1387	0.611
16	75	-0.557	1582	-0.690		16	1411	0.449		1892	0.580
17	515	-0.556	44	75	-0.669		17	415	0.441	31	1943	0.576	49
18	1325	-0.554	27	1414	-0.666	52	18	1884	0.437		1884	0.567	
19	26	-0.548		1770	-0.665	29	19	1674	0.433	32	1897	0.554	26
20	1900	-0.547		138	-0.664		20	739	0.420		824	0.549	30
21	1406	-0.547	39	495	-0.662	42	21	1387	0.417		1258	0.548	27
22	1648	-0.545	36	1900	-0.661		22	286	0.409		739	0.547	
23	43	-0.543	26	992	-0.660	28	23	67	0.401	28	286	0.544	
24	1346	-0.542	62	31	-0.659	38	24	143	0.401	29	1058	0.523	38
25	391	-0.541	30	399	-0.650		25	1892	0.395		1111	0.523	36

Numbers in the "gene" column indicate the serial numbers of the genes. Genes that are shaded indicate they had disappeared or were newly emerged among the top-ranking genes in the 55-homogeneous-matrix. Numbers to the right of the shading for a case indicate the rank of the gene in another case.

3.3 Prediction accuracy

The validity of the 7 samples detected as outliers by our method and the potentiality of m-gene predictors from the 55-homogeneous-matrix can also be explained by the prediction accuracy. We employed two supervised learning methods, WVA and kNN, and evaluated prediction accuracy by using the LOOCV test. In LOOCV, one constructs m-gene predictors only with training (n-1) samples using eq. (4), and then applies the m-gene predictors to assign the remaining sample to one of the states (normal or tumor). We preset m at values ranging from 50 to 2,000 for the most differentially expressed genes and another parameter, n, as 62 or 55 (after eliminating the 7 detected outliers). We compared the classification accuracy for 3 cases: (i) 62-heterogeneous-matrix, (ii) 55-homogeneous-matrix, and (iii) 1,000 55-heterogeneous-matrices.

The results of the LOOCV test calculated by WVA and kNN are shown in Figures 2a and 2b. The prediction accuracies of the 55-homogeneous-matrix were clearly higher than those of the

62-heterogeneous-matrix for all *m* (50, 100, …, 2000). The average accuracy difference between the two matrices with WVA and *k*NN was 9.43% and 6.88%, respectively. Additionally, the accuracy of the 62-heterogeneous-matrix was close to the average of the accuracy of the 1,000 55-heterogeneous-matrices. This is reasonable because each of the 1,000 55-heterogeneous-matrices was a sub-matrix randomly constructed from the 62-heterogeneous-matrix. Remarkably, most of the prediction accuracies for the 50 – 1200-gene predictors of the 55-homogeneous-matrix were higher than the 5% significance level of the prediction accuracy of 1,000 55-heterogeneous-matrices and slightly below the 1% significance level. Furthermore, most of the accuracies for about 400-gene predictors of the 55-homogeneous-matrix were higher than the 1% significance level of the 1,000 55-heterogeneous-matrices. These results indicate that the relatively low accuracies from the 62-heterogeneous-matrix and from the 1,000 55-heterogeneous-matrices compared with the 55-homogeneous-matrix are actually due to the 7 outliers.

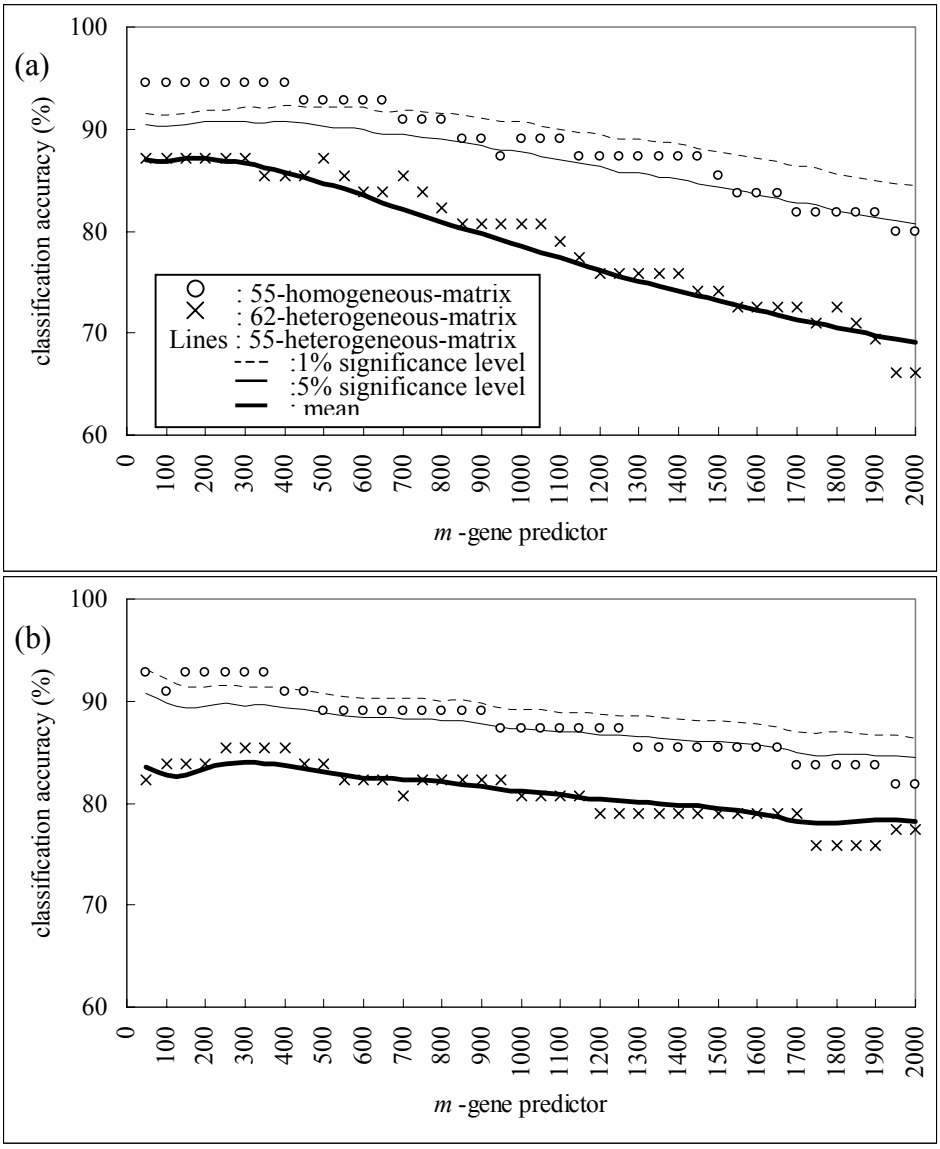

Figure 2. Classification accuracies of two methods: (a) WVA, (b) *k*NN.
(a) Weighted-voting algorithm (WVA). (b) *k*-nearest neighbor method (*k*NN, *k* = 3). Abbreviations are the same as those in Figure 1. In addition, thin lines represent the 5% significance level in the accuracies with *m*-gene predictors in the 1,000 55-heterogeneous-matrices.

3.4 Validation of detected outliers

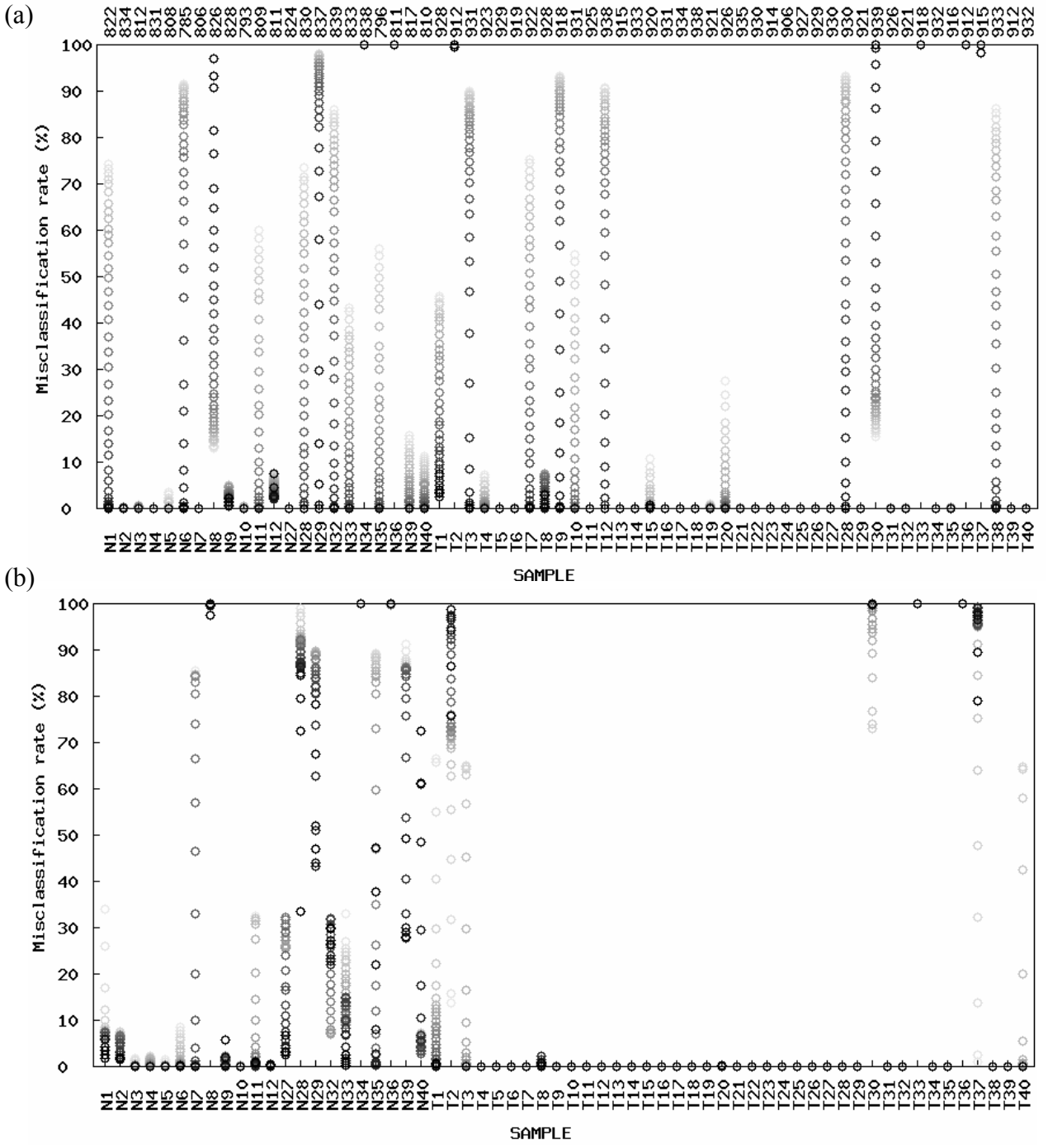

Figure 3. Misclassification rates for 62 samples.

Misclassification rates were calculated by two methods: (a) WVA and (b) *k*NN (*k* = 3). The numbers above the graphs indicate the counts of a sample that emerged in the 1,000 55-heterogeneous-matrices. A black circle denotes the misclassification rate for a sample using a 50-gene predictor, while an almost white circle denotes that using a 2000-gene predictor. The intermediate colored circles (such as gray) represent those using an intermediate number of predictor genes.

To validate each of the outlying samples detected, we investigated the misclassification rate of each of the 62 samples by using *m*-gene predictors (*m* = 50, 100, …, 2000). We performed the investigation based on the results of the LOOCV test on 1,000 55-heterogeneous-matrices.

Figure 3 shows the misclassification rates for each of the samples on *m*-gene predictors ranging from *m* = 50 (black circles) to *m* = 2,000 (almost white circles). We employed WVA (Figure 3a) and *k*NN (Figure 3b) in the validation. The numbers above the graphs represent the number of sets of samples. For example, a normal sample "N1" was included in 822 matrices of 1,000 possible matrices and the misclassification rate was 0% (none of the 822 tests assigned N1 as tumor) when using a 50-gene predictor, whereas the rate was 74.1% (609 of 822 tests assigned N1 as tumor) when a 2000-gene predictor was used.

Again, we detected 7 samples as outliers (T2, T6, T37, N2, N8, N34, and N36). Of these, T6 and N2 showed low misclassification rates. Furthermore, the distance scores for samples T6 and N2 were the lowest scores in both tumors and normal tissue samples (see Table 1). These results suggest potential false-positive errors for T6 and N2; that is, they may in fact not be outliers.

4. Discussion

We used our AICOUS method to detect outlying samples whose aberrant expression profiles could exert deleterious effects on gene expression analyses (especially sample classification). We focused on the importance of detecting/checking outlying samples in advance of sample classification and demonstrated a higher classification accuracy when we used the remaining homogeneous samples. Our method and the importance of method validation and of increasing the degree of homogeneity in samples to be analyzed are discussed.

4.1 Performance of AICOUS

AICOUS is based on the AIC whose information criterion has been used for modeling in various fields of statistics, engineering and numerical analysis [19][20][21]. AICOUS is an improved version of methods previously proposed [13][15] for the detection of outlying samples in two-dimensional microarray data. The main advantage of our method is its ability to output an objective decision about outlying samples (i.e., significance level independence; Table 2).

We demonstrated the feasibility of our method on colon microarray data consisting of 62 samples from 2 states: normal and tumor [5]. A total of 7 samples, 4 of the 22 normal- and 3 of the 40 tumor samples, were detected as outliers (Tables 2 and 3) and 5 of the 7 outlying samples detected by AICOUS coincided with samples misclassified and/or regarded as unfavorable in previous reports (Table 3) [3][5][6][7]. We suspect that the remaining 2 samples (T6 and N2) may not be outliers. AICOUS can be improved, however, by the inclusion of additional term(s) in equation (1). With the improvements, we should be able to detect outliers confidently in other microarray data.

4.2 Comparison of subsets of genes for distinguishing the 2 states of the samples

The clustering technique is an accepted method for microarray analysis. However, it is not sufficiently sensitive for the type of study we performed, because it focuses on group similarities, not differences, within individual genes [22]. To identify such a subset of genes, Golub *et al* [2] proposed selecting *m* genes that individually are highly correlated with the known classification (called neighborhood analysis) and then using a voting procedure for the classification of new samples based on the *m*-gene predictors (called WVA).

We created a measurement, $S(m)$, that shows the distinction level of *m* genes between two states of samples (see Methods). Large values of $S(m)$ indicate that the *m*-gene signature has a high level. Accordingly, $S(m)$ values should be higher in the 55-homogeneous-matrix than in the

62-heterogeneous-matrix, provided that we can detect a majority of the outliers. In Figure 1, measurements from the 55-homogeneous-matrix were actually higher than those in the 62-heterogeneous-matrix and the average of those in the 1,000 randomly selected 55-heterogeneous-matrices.

4.3 Comparison of classification accuracies

While various discrimination methods have been applied for the classification of clinical samples [2][3][23][24][25], Dudoit *et al.* [26] found that the traditional linear classifiers and nearest neighbors perform remarkably well compared with more sophisticated methods. Also, the focus of our study was the detection of outlying samples whose aberrant expression profiles may have negative effects on sample classification rather than the detection of misclassified samples. Hence, the feasibility of AICOUS was demonstrated by two conventional classification methods (WVA and *k*NN). Our results showed that accuracies in the 55-homogeneous-matrix were clearly higher than in the 62-heterogeneous-matrix (see Figure 2) for all selections of *m*. The values in the 55-homogeneous-matrix (circles) were also close to the 5% significance level of accuracy in the 1,000 randomly-selected 55-heterogeneous-matrices. These results indicate that the outlying samples detected here have significant effects on sample classification.

4.4 Validation of outlying samples

A thorough investigation of accuracies in the 1,000 55-heterogeneous-matrices suggested that 2 of the 7 outlying samples (N2 and T6) may not be outliers. They were selected 834 and 919 times, respectively, in the 1,000 55-heterogeneous-matrices (Figure 3). The misclassification rate for the two samples was almost 0%, whereas the rate for the other 5 samples was almost 100%.

Since samples that are always misclassified are likely to be contaminated or mislabeled [3][6], 5 of the 7 outlying samples, N8, N34, N36, T2, and T37 must be outliers. Indeed, N34 and N36 were verified as outliers in light of the sample composition [5][6]. According to Li *et al* [7], only N8 was misclassified when a 57-homogeneous-matrix was used after elimination of the verified outliers N34, N36, T30, T33, and T36 (Table 3) [6]. Sample N8 is also in the cluster containing mostly tumor tissues [5]. Furthermore, T2 and T37 detected by AICOUS are among the 5 tumor tissues described by Alon *et al*. as the cluster containing mostly normal tissues, although supervised discriminant methods (SVM and *k*NN) correctly classified them as tumors [5][6][7].

We have no evidence that verifies N2 and T6 as outlying samples. These samples had the lowest distance scores in each sample state among the 7 outlying samples detected (Table 1), suggesting that they may be false-positives rather than outliers. We were also unable to verify 3 samples (T30, T33, and T36) as outliers; these samples represent false-negatives. The paucity of combinations considered may be the main reason for the false positive/negative errors: to save computation time, we only searched for half the number of samples for each state since the numbers of $\sum_{k=0}^{s} {}_{22}\mathrm{C}_k$ combinations and $\sum_{k=0}^{s} {}_{40}\mathrm{C}_k$ combinations (s denotes the number of outliers for the 22 normal- and 40 tumor samples) are considerable.

4.5 Fluctuation of the different distance metric

In general, the use of a different distance metric yields different results [25]. We used a Pearson correlation coefficient as a distance metric, because this was the metric applied in previous analyses of the colon dataset. To investigate the fluctuation of the outlying samples, we also used average

Euclidean distance as the other distance metric. As a result, 2 samples (N9 and N12) were detected (see Supplementary material 1). Sample N9 was detected anew among the samples shown in Table 3, while N12 was among those samples reported by Alon *et al.* [5]. Overall, the Euclidean distance seems not to be the appropriate metric for this type of analysis, because those 2 samples have not been confirmed as outlying samples. Further validations are underway in our laboratory.

4.6 Selected 50 top-ranking genes

Compared with 2 sets of the 50 top-ranking genes from the 62-heterogeneous-matrix and the 55-homogeneous-matrix, the difference in the genes selected has the advantage of eliminating outlying samples, although there are a few exceptions (Table 5). Ribosomal proteins (accession: T57619, T62947) up-regulated in tumor samples, which disappeared in the 50 top-ranking genes selected in the 55-homogeneous-matrix, are an example of the exception [5][27]. While many reasonable genes were extracted, cyclin-dependent kinase, which is up-regulated in tumor samples and a negative prognostic marker in colorectal cancer, is an example of those genes that disappeared from among the 50 top-ranking genes selected in the 55-homogeneous-matrix [18][28]. On the other hand, other tumor-related proteins such as laminin and NDP kinase emerged [29][30]. Bo and Jonassen [16] stated that a larger difference in the top-ranking genes selected by various methods did not necessarily coincide with larger differences in prediction accuracies. We believe

Table 5. List of genes that appeared or disappeared from the list of the 50 top-ranking genes in the 55-homogeneous-matrix.

state	expression	serial	accession	sequence	description
disappeared	up	43	T57619	3' UTR	40S ribosomal protein S6 (Nicotiana tabacum)
disappeared	up	391	D31885	gene	mRNA (KIAA0069) for ORF (novel proetin), partial cds
disappeared	up	515	T56604	3' UTR	tubulin beta chain (Haliotis discus)
disappeared	up	1325	T47377	3' UTR	S-100P protein
disappeared	up	1346	T62947	3' UTR	60S ribosomal protein L24 (Arabidopsis thaliana)
disappeared	up	1406	U26312	gene	heterochromatin protein HP2Hs-gamma mRNA, partial cds
disappeared	up	1648	T86749	3' UTR	cyclin-dependent protein kinase mRNA, complete cds
disappeared	down	67	T51534	3' UTR	cystatin C precursor
disappeared	down	143	R28373	3' UTR	hemoglobin beta chain
disappeared	down	415	T60155	3' UTR	actin, aortic smooth muscle
disappeared	down	1668	M82919	gene	gamma amino butyric acid (GABAA) receptor beta-3 subunit
disappeared	down	1674	T67077	3' UTR	sodium/potassium-transporting ATPase gamma chain
disappeared	down	1967	T60778	3' UTR	matrix gla-protein precursor (Rattus norvegicus)
appeared	up	31	T61609	3' UTR	laminin receptor
appeared	up	241	M36981	gene	putative NDP kinase (nm23-H3S) mRNA, complete cds
appeared	up	495	H20426	3' UTR	nucleoside diphosphate kinase (Ginglymostoma cirratum)
appeared	up	992	X12466	gene	snRNP E protein
appeared	up	1002	R08183	3' UTR	Q04984 10 kD heat shock protein, mitochondrial
appeared	up	1414	R64115	3' UTR	adenosylhomocysteinase
appeared	up	1770	U17899	gene	chloride channel regulatory protein mRNA, complete cds
appeared	down	824	Z49269	gene	chemokine HCC-1
appeared	down	1058	M80815	gene	a-L-fucosidase gene, exon 7 and 8, and complete cds
appeared	down	1111	D31716	gene	GC box bindig protein, complete cds
appeared	down	1258	R67358	3' UTR	MAP kinase phosphatase-1 (Homo sapiens)
appeared	down	1897	U19969	gene	two-handed zinc finger protein ZEB mRNA, partial cds
appeared	down	1943	D29808	gene	T-cell acute lymphoblastic leukemia associated antigen 1

The terms up and down in the "expression" column indicate the respective up- and down-regulated states in tumor samples. Numbers in the "serial" column are the same as those in the "gene *j*" column in Table 4. UTR = untranslated region.

that there are two reasons for these findings. First, as pointed out by Li *et al* [6], hundreds of genes that discriminate between different sample classes may exist, since typical array data consist of a large number of genes and a small number of samples. A second reason is the existence of outlying samples. In fact, we have observed significant improvements in analysis attributable to improved classification accuracy.

4.7 Application to another (AML/ALL) dataset

The purpose of this study was the detection of outlying samples (not of "consistently misclassified samples" with some classifiers). We evaluated only a colon dataset, because some of the samples in the dataset were reported as being contaminated [6][7]. There is another dataset available for evaluation, the acute myeloid leukemia/acute lymphoblastic leukemia (AML/ALL) dataset [2]. It has no outlying samples. However, unlike solid tumors such as the colon dataset, we can expect greater homogeneity in individual leukemia samples. Not surprisingly, we observed fewer outlying samples in the AML/ALL set than in the colon set: 1 of 25 AMLs and 1 of 47 ALLs (see Supplementary material 2), suggesting that AICOUS is indeed of potential general applicability.

4.8 Scalability of AICOUS

AICOUS was designed to detect outlying samples but in principle can also be applied to detect genes whose expression patterns are markedly high, or markedly low, in some particular tissues compared with the expression level in other tissues or sources. Greller and Tobin [31] proposed a method for identifying genes that are markedly down- or up-regulated only in a specific tissue compared with other tissues. Nevertheless, genes exist whose expression profiles are clearly different depending on the tissue. In fact, clustering results of gene expression data on adult and fetal mouse tissues (a total of 49 tissues) revealed the existence of such genes, i.e., genes specific for digestive organs (colon, cecum, and stomach), smooth muscle-related genes (tongue, heart, and skeletal muscle), etc. [32][33][34]. We believe the high scalability of AICOUS makes our method applicable to a wide variety of areas.

To isolate a particular cell type, micro-dissection techniques are commonly used to prepare samples for microarray experiments [35]. Kitahara *et al.* [36] identified a set of genes involved in colorectal carcinogenesis from normal and tumor samples rendered homogeneous by laser-capture microdissection. Venet *et al.* [8] presented an approach that permits the mathematical separation of samples consisting of many different cell types into their constituents. AICOUS is not meant to be a rival of these methods, but rather an addition to facilitate the evaluation of homogeneity. We strongly recommend the use of an AICOUS strategy for detecting outlying entities in expression data.

We thank T. Ueda, K. Shimizu, S. Nakamura, K. Tsuda, and L. Li for their helpful comments. We also thank M. Sekijima, M. Kadota, and M. Terauchi for their valuable technical assistance. This work was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] A. von Heydebreck, W. Huber, A. Poustka and M. Vingron, *Bioinformatics*, **17**, S107-S114, (2001).

[2] 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, *Science*, **286**, 531-537, (1999).

[3] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer and D. Haussler, *Bioinformatics*, **16**, 906-914, (2000).

[4] M. Eisen, P. Spellman, P. Brown and D. Botstein, *Proc. Natl. Acad. Sci. USA*, **95**, 14863-14868, (1998).

[5] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack and A. J. Levine, *Proc. Natl. Acad. Sci. USA*, **96**, 6745-6750, (1999).

[6] L. Li, T. A. Darden, C. R. Weinberg and L. G. Pedersen, *Comb. Chem. High Throughput Screen*., **4**, 727-739, (2001).

[7] L. Li, C. R. Weinberg, T. A. Darden and L. G. Pedersen, *Bioinformatics*, **17**, 1131-1142, (2001).

[8] D. Venet, F. Pecasse, C. Maenhaut and H. Bersini, *Bioinformatics*, **17**, S279-S287, (2001).

[9] F. E. Grubbs, *Technometrics*, **11**, 1-21, (1969).

[10] G. L. Tietjen and R. H. Moore, *Technometrics*, **14**, 583-597, (1972).

[11] W. J. Dixon, *Biometrics*, **22**, 74-89, (1953).

[12] S. S. Shapiro and M. B. Wilk, *Biometrika*, **52**, 591-611, (1965).

[13] G. Kitagawa, *Technometrics,* **21**, 193-199, (1979).

[14] H. Akaike, 2nd International Symposium on Information Theory, 267-281, (1973).

[15] T. Ueda, *Japanese J. Appl. Stat*., **25**, 17-26, (1996).

[16] T. H. Bo and I. Jonassen, *Genome Biol*., **3**, research0017.1-11, (2002).

[17] D. L. Massart, B. G. Vandeginste, S. N. Deming, Y. Michotte and L. A. Kaufman, Textbook (Data Handling in Science and Technology, Vol2). Elsevier Science, NY, (1988).

[18] T. Tsunoda, T. Nakamura, K. Ishimoto, H. Yamaue, H. Tanimura, N. Saijo, K. Nishio, *Anticancer Res*., **21**, 137-143, (2001).

[19] H. Akaike, *Ann. Inst. Statist. Math.*, **22**, 203-217, (1970).

[20] H. Akaike, *Ann. Inst. Statist. Math.*, **30**, 9-14, (1978).

[21] Y. Sakamoto and H. Akaike, *Ann. Inst. Statist. Math*., **30**, 185-197, (1978).

[22] J. G. Thomas, J. M. Olson, S. J. Tapscott, L. P. Zhao, *Genome Res*., **11**, 1227-1236, (2001).

[23] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein and P. O. Brown, *Genome Biol*., **1**, 0002.1, (2000).

[24] M. Takahashi, D. R. Rhodes, K. A. Furge, H. Kanayaam, S. Kagawa, B. B. Haab and B. T. The, *Proc. Natl. Acad. Sci. USA*, **98**, 9754-9759, (2001).

[25] A. Szabo, K. Boucher, W. L. Carroll, L. B. Klebanov, A. D. Tsodikov, A. Y. Yakovlev, *Math. Biosci*., **176**, 71-98, (2002).

[26] S. Dudoit, J. Friedlyand, T. P. Speed, Tech. Rep. 576, University of California, Berkeley, (2000).

[27] K. Pogue-Geile, J. R. Geiser, M. Shu, C. Miller, I. G. Wool, A. I. Meisler and J. M. Pipas, *Mol. Cell. Biol*., **11**, 3842-3849, (1991).

[28] H. Kawana, J. Tamaru, T. Tanaka, A. Hirai, Y. Saito, M. Kitagawa, A. Mikata, K. Harigaya and T. Kuriyama, *Am. J. Pathol*., **153**, 505-513, (1998).

[29] M. L. Lacombe, X. Sastre-Garau, I. Lascu, A. Vonica, V. Wallet, J. P. Thiery and M. Veron, *Eur. J. Cancer*, **27**, 1302-1307, (1991).

[30] C. Lenander, J. K. Habermann, A. Ost, B. Nilsson, H. Schimmelpenning, K. Tryggvason and G. Auer, *Anal. Cell. Pathol*., **22**, 201-209, (2001).
[31] L. D. Greller and F. L. Tobin, *Genome Res*., **9**, 282-296, (1999).
[32] K. Kadota, R. Miki, H. Bono, K. Shimizu, Y. Okazaki, Y. Hayashizaki, *Physiol. Genomics*, **4**, 183-188, (2001).
[33] R. Miki, K. Kadota, H. Bono, Y. Mizuno, Y. Tomaru, P. Carninci, M. Itoh, K. Shibata, J. Kawai, H. Konno, S. Watanabe, K. Sato, Y. Tokusumi, N. Kikuchi, Y. Ishii, Y. Hamaguchi, I. Nishizuka, H. Goto, H. Nitanda, S. Satomi, A. Yoshiki, M. Kusakabe, J. L. DeRisi, M. B. Eisen, V. R. Iyer, P. O. Brown, M. Muramatsu, H. Shimada, Y. Okazaki and Y. Hayashizaki, *Proc. Natl. Acad. Sci. USA*, **98**, 2199-2204, (2001).
[34] K. Kadota, S.-I. Nishimura, H. Bono, S. Nakamura, Y. Hayashizaki, Y. Okazaki and K. Takahashi, *Physiol. Genomics*, **12**, 251-259, (2003).
[35] M. R. Emmert-Buck, R. F. Bonner, P. D. Smith, R. F. Chuaqui, Z. Zhuang, S. R. Goldstein, R. A. Weiss and L. A. Liotta, *Science*, **274**, 998-1001, (1996).
[36] O. Kitahara, Y. Furukawa, T. Tanaka, C. Kihara, K. Ono, R. Yanagawa, M. E. Nita, T. Takagi, Y. Nakamura and T. Tsunoda, *Cancer Res*., **61**, 3544-3549, (2001).