Normalization of DNA Microarray Data with BIC Model Comparison

Takeo Okazaki

Faculty of Engineering, University of the Ryukyus, Okinawa, 903-0213 Japan

Summary
DNA microarray data, which are efficient for estimation and identification of genetic network, have a large variety due to those experimental environments and measurement. Standardization by an appropriate bias correction is needed for the comparison and the data integration between two or more experiments. On the grounds of normality of distribution in the ideal expression data, some adjustment methods that consider a specific bias for specific expression data have been proposed. In this research, after all combinations of assumed multiple bias and the adjustment methods were modeled, and the appropriate model by BIC was selected for normalization of expression data. The proposal method was applied to a Yeast, Escherichia coli, and Homo sapiens microarray data from Stanford Microarray Database, and the comparative experiment results with previous methods were shown so far.

Key words:
DNA Microarray, Gene Expression Data, Normailization, BIC

1. Introduction

The effective experimental methods for estimation and identification of a gene network model involve the microarray method. The microarray method made thousands to tens of thousands of genes fix as a different spot on slide glass. Hybridization of mRNA or cDNA which is compounded with the model of mRNA lets us know all the gene expression patterns simultaneously on a genome scale.

On the other hand, since the observations in this experiment have a large variety of experimental circumstances and measuring method, reproducibility is low and a repeat experiment is required. However, it is rare to conduct the repeat experiment of actually sufficient number of times because of a cost and the priority in the plan of a large amount of the experiment patterns. Therefore, accurate genome analysis and comparison of multiple experiments and data integration are difficult.

It is common that log transformed value of the observation fluorescence ratio at each spot should have normality as a guideline for standardization towards a comparison and an integration. According to this guideline, there are some traditional correction method for specific data and bias as follows.

Yang *et al.* [1][2][3] took up the mouse case and proposed the correction method for the bias by the variation of dye difference, the bias depending on signal strength, the bias by the variation of print-tip and the scale variation between print-tip group. Uchida *et al.* [4] took up the yeast case and proposed the correction method for the bias by the variation of dye difference, the bias depending on signal strength and the bias by the variation of print-tip. Smyth *et al.* [5] took up the mouse case and proposed the correction method for the variation in scale between microarrays. These correction methods are aimed at specific expression data and specific bias, we do not have a guarantee to get a similar correction result for other data.

In this study, in order to build the correction or normalizing method independent of a specific experiment, the bias generating factor over the conventional microarray data and its correction method were systematized. Then the global normalization with the judgment of Bayesian Information Criteria (BIC) concerning the existence of each generating factor, selection of correction method and applying order were realized.

2. Bias of gene expression data and its correction

Even if we conduct the repeat experiment of a microarray to the same experiment system, mRNA or the amount of gene expression is not necessarily in agreement. The lowness of reproducibility depends on bias mainly. Bias here corresponds to the one of Kohane *et al.* [6] definitions: "the physical effect of the outside independent of the target living system, if the levels exceed a measurement sensitivity, observations by the side of a living body system or measuring instrument are influenced." The parameters which influence the experiment about a living body system are wide in scope and it is very difficult to control them all completely. Therefore, in a microarray experiment all bias that arises from both of the living body systems and measuring instrument side are intricately interwoven with each other.

Manuscript received March 5, 2014
Manuscript revised March 20, 2014

We can divide bias roughly into two categories, one is what is seen by each spot in the same array and the other depends on the difference in the experimental environment between different microarrays. The result is as follows when organizing each bias taken up with the conventional correction method from the viewpoint of the generating part.

Table 1: Bias generating part in microarray experiments

	Microarray Internal	Between Microarray
Microarray Fabrication	✓	✓
Reagent Adjustment	✓	
Fluorescent Marker Characteristics	✓	
Hybridize	✓	✓
Scan	✓	

Conventional correction methods for each generating factor are as follows.

2.1 Bias correction for the variation of dye difference

Two methods were proposed in order to rectify to bias resulting from the marker and detection efficiency between the fluorescence pigment so that the average ratio of the signal strength of dye Cy3 and Cy5 may become 1.

In total intensity normalization [7], G_i and R_j are the signal strength of Cy3 and Cy5 for j-th gene respectively, and N is the number of genes spotted on the microarray. Then the total intensity ratio is derived as follows.

$$T_{total} = \frac{\sum_{j=1}^{N} R_j}{\sum_{j=1}^{N} G_j} \tag{1}$$

We can correct the signal strength with this ratio as follows.

$$G_j' = T_{total} G_j, \quad R_j' = R_j \tag{2}$$

In global normalization [1], assuming the constant relationship, such as $R=kG$ between the signal strength of Cy3 and Cy5, we can correct the log ratio of signal strength as follows.

$$M' = M - c, \quad M = \log_2 \frac{R}{G} \tag{3}$$

In this correction, the median or average of M is used for constant c.

2.2 Bias correction depends on the absolute amount of signal strength

As the spot which expression level is low may tend to be affected by bias which occurs during an experiment, a regression formula for expression level is estimated and applied for correction.

$$M' = M - c(A) \tag{4}$$

For constructing the regression formula $c(A)$, two methods were proposed.

In Lowess method [1], data smoothing is done by local weighted linear regression. After data smoothing locally, outlier observations are removed and regression process is repeated. In Loess method [5], data smoothing is done by local weighted linear regression same as Lowess method, but Loess method uses 2nd order polynomial for the weighted linear least squares regression.

2.3 Bias correction for the variation of print-tip

Print-tip group is the gene groups spotted by the same print-tip. As the difference of the length, width of the hole and aged state of print-tip causes systematic difference to expression level, three correction method were proposed as follows.

In within-print-tip-group normalization [1], systematic bias may exist from being spotted by the same print-tip at all blocks of microarray. For each print-tip groups, we can correct by following formula.

$$M_i' = M_i - c_i(A), \quad i = 1,2,\cdots,I \tag{5}$$

$c_i(A)$ is the regression formula with Lowess method for i-th print-tip group, I is the number of print-tip. In linear regression line equations [4], for each print-tip group, linear regression formula $c_i(A)$ is applied same as within-print-tip-group normalization. In robust fitting of linear models [8], the least squares method is sensitive to a gap of the standard residual of linear regression. Even one outlier causes big influence, robust linear regression by M-estimator which is extended maximum likelihood method is applied for the correction.

2.4 Bias correction for the spotting order in print-tip

There is another bias result from print-tip, that is spotting order in a print-tip, print-order normalization [4] is proposed. Bias which arises in the stage where each probe was spotted on the microarray is rectified as follows from the assumption of the relationship to the used spot pin.

$$M_{kn}' = M_{kn} - c_k(n) \tag{6}$$

M_{kn} is M-value for the n-th spot in the k-th print-tip. $c_k(n)$ is the linear regression equation of the n-th spot in the k-th printer-order group.

2.5 Variations of the spread of the distribution among print-tip group

There are cases where the spread of the distribution varies among the different print-tip group. To correct for this, within-slide scale normalization was proposed. We assume that a_i^2 is the scale factor for the i-th print-tip group and M-value of the i-th print-tip group follows the normal distribution $N(0, a_i^2\sigma^2)$. At this time, the maximum likelihood estimator of a_i is given as follows.

$$\hat{a}_i = \frac{\sum_{j=1}^{n_i} M_{ij}^2}{\sqrt[I]{\prod_{i=1}^{I}\sum_{j=1}^{n_i} M_{kj}^2}}, \quad i = 1,2,\cdots,I \tag{7}$$

M_{ij} is log ratio of the j-th signal intensity in the i-th print-tip group, and I is the total number of print-tip on the microarray.

2.6 Correction between the microarray

There is two point of view for the data comparison between the different microarray.

(1) Comparison between microarray by repeated experiments with same control

(2) Comparison between microarray by experiments with different control

In repeated experiments, there is a possibility of bias due to the variation of the experimental environment occurs. In different control experiments, there is a possibility of bias due to the variation in the experimental environment and the difference in hybridization protocols. After you make a correction in the microarray, each median log ratio of the signal intensity of each slide is substantially the same. However, since there are variations in the spread of the distribution, scale normalization between microarray method has been proposed. In this method, $median_j(M_{ij})$ of $MAD_i = median_i\{|M_{ij} - median_j(M_{ij})|\}$ is set to 0, we can correct the scale by equalizing the median of absolute value of M-value and other microarray's.

3. Normalization of gene expression data

Bias is generated from a variety of factors in the process microarray experiment. Table 2 shows summaries of the correspondence between conventional normalization methods and the target factor and correction methods.

Table 2: Correspondence between bias and conventional normalization methods

	Yang (mouse)	Uchida (yeast)	Smyth (mouse)
Variation of dye difference	Global normalization	Total intensity normalization	
Signal strength	Lewess method	Lowess method	
Variation of print-tip	Lowess method	Linear regression	Loess method
Spotting order in print-tip		Linear regression	
Scale between print-tip groups	With-in slide scale normalization		
Scale between microarrays			Scale normalization between arrays

As this table shows, the normalization methods have been proposed so far, it can be seen that it aims for a particular biological and does not consider all bias factors. Further, since the determination of the presence or absence of the bias factor is not clear, the correction depends on the data. It can be said, therefore, when it is used to correct other data, there is no guarantee either be similarly corrected.

In this study, three conditions for normalization were picked up as follows.

- Considering all bias factors
- Integrated model for each correction method
- Determination of the presence or absence of each bias

It was considered that normalization method which does not depend on the observation data can be constructed by satisfying these conditions.

In order to perform the correct global, it is necessary to express the integrated bias all. Then the observed data was re-defined as follows.

i : index of print-tip group (Total number is I)

k : index of print-order group (Total number is K)

j : index of microarray

n : spotting order in print-order group

a_i : scaling factor for i-th print-order group

a^*_j : scaling factor for j-th microarray

$G_{(ikn)}$: green signal strength

$R_{(ikn)}$: red signal strength

$$M_{(ikn)} = \log_2\left(\frac{R_{(ikn)}}{G_{(ikn)}}\right)$$

$$A_{(ikn)} = \log_2\sqrt{G_{(ikn)}R_{(ikn)}}$$

$$T_{total} = \frac{\sum_{j=1}^{N} R_j}{\sum_{j=1}^{N} G_j}$$

With these definitions, each correction method can be re-formulated as a function as follows.

(1) Correction of the bias caused by the difference in the change of the dye

Global normalization:

$$f_1^{(G)}(M_{(ikn)}) = M_{(ikn)} - c, \quad c: median\ or\ average\ of\ M_{(ikn)} \tag{8}$$

Total intensity normalization:

$$f_1^{(T)}(G_{(ikn)}, R_{(ikn)}) = \log_2\left(\frac{R_{(ikn)}}{T_{total} G_{(ikn)}}\right) \tag{9}$$

(2) Correction of the bias that depends on the absolute amount of signal strength

Linear regression:

$$f_2^{(L)}(M_{(ikn)}, A_{(ikn)}) = M_{(ikn)} - c^{(L)}(A_{(ikn)}) \tag{10}$$

Lowess method:

$$f_2^{(LW)}(M_{(ikn)}, A_{(ikn)}) = M_{(ikn)} - c^{(LW)}(A_{(ikn)}) \tag{11}$$

Loess method:

$$f_2^{(LE)}(M_{(ikn)}, A_{(ikn)}) = M_{(ikn)} - c^{(LE)}(A_{(ikn)}) \tag{12}$$

Robust linear regression:

$$f_2^{(R)}(M_{(ikn)}, A_{(ikn)}) = M_{(ikn)} - c^{(R)}(A_{(ikn)}) \tag{13}$$

(3) Correction of the bias caused by the variation in the print-tip

Linear regression:

$$f_3^{(L)}(M_i, A_i) = M_i - c_i^{(L)}(A_i) \tag{14}$$

Lowess method:

$$f_3^{(LW)}(M_i, A_i) = M_i - c_i^{(LW)}(A_i) \tag{15}$$

Loess method:

$$f_3^{(LE)}(M_i, A_i) = M_i - c_i^{(LE)}(A_i) \tag{16}$$

Robust linear regression:

$$f_3^{(R)}(M_i, A_i) = M_i - c_i^{(R)}(A_i) \tag{17}$$

(4) Correction of the bias caused by the spotting order in the print-tip

Linear regression:

$$f_4^{(L)}(M_{(kn)}, n_k) = M_{(kn)} - c_k^{(L)}(n) \tag{18}$$

Lowess method:

$$f_4^{(LW)}(M_{(kn)}, n_k) = M_{(kn)} - c_k^{(LW)}(n) \tag{19}$$

Loess method:

$$f_4^{(LE)}(M_{(kn)}, n_k) = M_{(kn)} - c_k^{(LE)}(n) \tag{20}$$

Robust linear regression:

$$f_4^{(R)}(M_{(kn)}, n_k) = M_{(kn)} - c_k^{(R)}(n) \tag{21}$$

(5) Correction of the scale between print-tip groups

$$f_5(M_i) = \frac{M_i}{a_i} \tag{22}$$

(6) Correction of the scale between microarrays

$$f_6(M_j) = \frac{M_j}{a_j^*} \tag{23}$$

Because each correction was expressed as a function, a combination of correction can be expressed as a composite function. At this time, the synthetic sequence of the function corresponds to the order of applying bias correction. For example, the composite function in the case of applying equation (8), (10) and (14) is expressed as follows.

$$f_3^{(L)} \circ f_2^{(L)} \circ f_1^{(G)} = M_{(ikn)} - c^{(L)}(A_{(ikn)}) - c_i^{(L)}(A) - c$$

Order to apply the correction corresponds to the structure in which each bias occurs in the microarray experiment. However, in the case of the following, we have no control over the order of occurrence.

- Since the dyes continue to degrade until the microarray is scanned, the order of occurrence of the bias due to differences in the fluctuations of the dye and bias that depends on the signal strength cannot be determined.
- The bias due to the spotting order in the print-tip and bias due to the variation in the print-tip, order of occurrence cannot be determined because both occur in microarray during production.

Therefore, the order of correction when all bias is present can be considered the following 4 cases.

Adjustment Order #1

[1] Correction of bias due to dye variations
[2] Correction of bias in signal intensity dependent
[3] Correction of bias caused by fluctuations in print-tip
[4] Correction of bias by the spotting order in print-tip
[5] Correction of scale between print-tips

Adjustment Order #2

[1] Correction of bias due to dye variations
[2] Correction of bias in signal intensity dependent
[3] Correction of bias by the spotting order in print-tip
[4] Correction of bias caused by fluctuations in print-tip
[5] Correction of scale between print-tips

Adjustment Order #3

[1] Correction of bias in signal intensity dependent
[2] Correction of bias due to dye variations

[3] Correction of bias caused by fluctuations in print-tip

[4] Correction of bias by the spotting order in print-tip

[5] Correction of scale between print-tips

Adjustment Order #4

[1] Correction of bias in signal intensity dependent

[2] Correction of bias due to dye variations

[3] Correction of bias by the spotting order in print-tip

[4] Correction of bias caused by fluctuations in print-tip

[5] Correction of scale between print-tips

Bias shown so far is not intended to be necessarily generated. And, in the cases of the Adjustment Order #3 and #4, we cannot apply Total intensity normalization after correcting the bias of the signal intensity dependent. Considering these constraints, the total number of the correction candidate model is 5000. Further, comparing the expansion formulas of the synthetic, there are some equivalent models that cannot be distinguished. Then, we have 2853 candidate models.

It is reasonable that the determination of the presence or absence and the order of correction of each bias may be selected those that best fits the observations in the correction candidate model. So, Bayesian Information Criterion [9] is used as a criterion for goodness of fit judgment.

$$BIC = -2\log(Maximum\,Likelihood) + p\log(n) \qquad (24)$$

It is assumed that the correction candidate model with minimum BIC value fits most to the observations. This is the same meaning to judge the presence or absence of each bias, to determine the correction order and to determine the regression equation for the correction.

Therefore, the proposal procedure of gene expression data normalization is shown as follows.

(Step1) Correct microarray data in each correction candidate model

(Step2) Calculate the BIC of each correction candidate model

(Step3) Select the model with minimum BIC value, and output the corrected value as the normalized data

Because the conventional correction methods are also included in the correction candidate model, proposal procedure should be the more versatile approach.

4. Verification Experiments

In order to verify the normalization of actual gene expression data with the proposal, experiments were conducted using the expression data of Yeast, Escherichia coli and Homo sapiens, which are published by Stanford MicroArray Database [10].

Normality checking of the corrected, be measured by the fit of a normal distribution is appropriate. BIC value as a comprehensive index is at a minimum because it is self-evident, the kurtosis and skewness, which is a feature of the distribution shape were added for evaluation. Under the exact normal distribution, skewness is 0 and kurtosis is 3. MA-plot is a visual representation of gene expression data which has been transformed onto the M (log ratio) and A (mean average) scale, distributes around the 0 if the ideal state.

4.1 Yeast case

From genetic data of 10752 samples obtained, 9851 pieces were applied, excluding the invalid value. The model selected by BIC minimum value was due to the following correction.

$$f_5 \circ f_4^{(LE)} \circ f_3^{(L)} \circ f_2^{(LE)} \circ f_1^{(T)} \qquad (25)$$

Table.3: Normalization index (Yeast)

	Original	Proposal	Yang	Uchida
BIC	25633.44	-61290.67	30195.46	16492.63
skewness	1.8018	-0.0693	0.8420	1.9520
kurtosis	9.4975	3.4720	8.4624	12.6900

We could confirm the normality because both the skewness and kurtosis were sufficiently close to the ideal value. Comparing with conventional methods, better result was obtained.

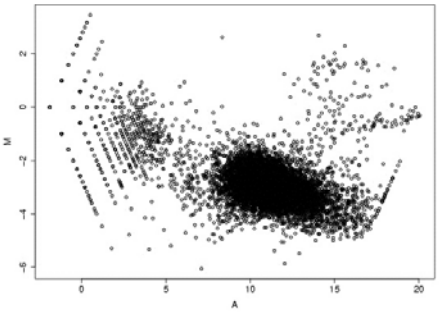


Fig.1: MA-plot before correction (Yeast)

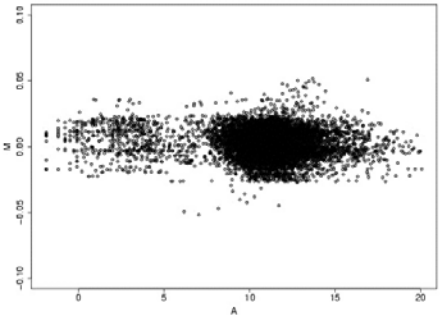


Fig.2: MA-plot after proposed correction (Yeast)

4.2 Escherichia coli case

From genetic data of 6384 samples obtained, 5603 pieces were applied, excluding the invalid value. The model selected by BIC minimum value was due to the following correction.

$$f_5 \circ f_4^{(LE)} \circ f_3^{(L)} \circ f_2^{(LE)} \circ f_1^{(T)} \quad (26)$$

Table.4: Normalization index (Escherichia coli)

	Original	Proposal	Yang	Uchida
BIC	14201.22	-20742.05	13684.78	9532.93
skewness	0.09945	-0.05894	-0.48203	-0.43780
kurtosis	6.72031	4.88145	7.84636	7.43943

As well as Yeast case, we could confirm the normality because both the skewness and kurtosis were sufficiently close to the ideal value.

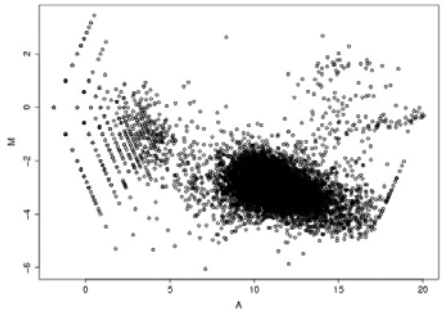


Fig.3: MA-plot before correction (Escherichia coli)

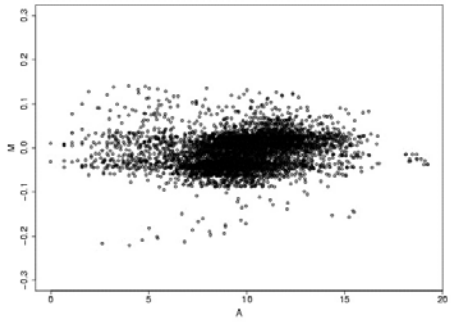


Fig.4: MA-plot after proposed correction (Escherichia coli)

4.3 Homo sapiens case

From genetic data of 44536 samples obtained, 43715 pieces were applied, excluding the invalid value. The model selected by BIC minimum value was due to the following correction.

$$f_5 \circ f_4^{(LE)} \circ f_3^{(R)} \circ f_2^{(LE)} \circ f_1^{(T)} \quad (27)$$

Table.5: Normalization index (Homo sapiens)

	Original	Proposal	Yang	Uchida
BIC	76867.86	-326368.60	77814.04	42957.59
skewness	0.14703	-0.18647	0.00465	-0.23228
kurtosis	6.71133	4.63284	6.34008	6.65077

It has been slightly degraded with respect to skewness, the correction method of Yang was effective. On the other hand, it is close to the ideal value for kurtosis, improvement is greater than the other correction methods. As seen from the histogram shown in Figure 7 and 8, it is considered that symmetry close to the normal distribution is already obtained in the original data, and that improve the kurtosis affected the skewness.

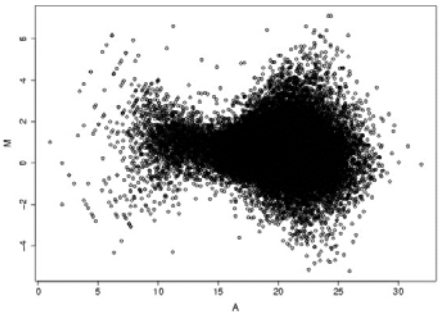


Fig.5: MA-plot before correction (Homo sapiens)

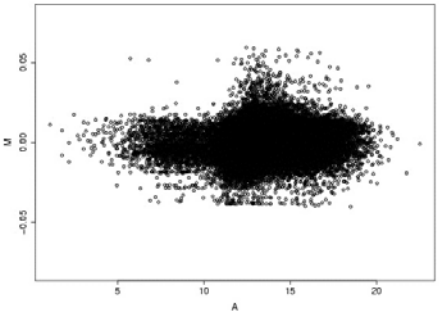


Fig.6: MA-plot after proposed correction (Homo sapiens)

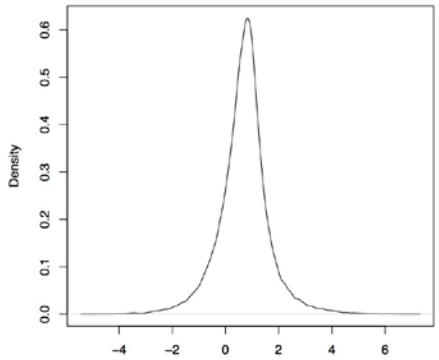


Fig.7: Histgram of original gene expression data (Homo sapiens)

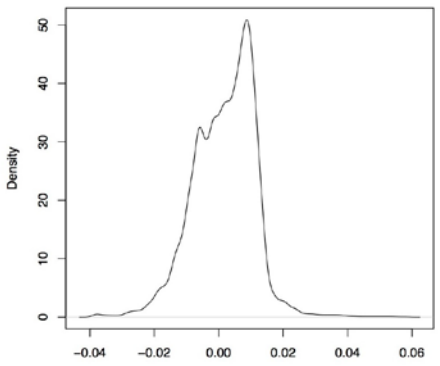


Fig.8: Histgram of corrected gene expression data (Homo sapiens)

5. Conclusion

In this study, I proposed a normalization method of microarray data in order to establish reliability. For that occasion, I organized the bias occurrence factors and correction methods in the measurement of observation. While the conventional method corrected for specific bias and data, the proposed method of normalization is not dependent on a specific data was realized, by determining the presence or absence of the bias globally based on BIC. The results of applying the proposed method for gene expression data Yeast, Escherichia coli, Homo sapiens confirmed the advantages of the proposed method in BIC and the statistical measure of the normality.

The proposed method is positioned prior processing for genomic analysis. Thus, it is possible to integrate multiple experimental results by the normalization of the proposed method and to construct a large-scale experiment pseudo.

Acknowledgment

The authors would like to express their cordial thanks to Ms. Madoka Kamiya for her experimental support.

References

[1] Y. H. Yang, S. Dudoit, P. Luu,and T. P. Speed, “Normalization for cDNA Microarray Data”, Microarrays : Optical Technologies and Informatics, Proceedings of SPIE, Vol.4266, 2001.

[2] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, “Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation”, Nucleic Acids Research, 30(4), e15, 2002.

[3] Y. H. Yang and N. Thorne, “Normalization for Two-color cDNA Microarray Data. Science and Statistics: A Festschrift for Terry Speed”, D. Goldstein (eds.), IMS Lecture Notes, Monograph Series, Vol.40, pp.403-418, 2003.

[4] S. Uchida, Y. Nishida, K. Satou, S. Muta, K. Tashiro, and S. Kuhara, “Detection and Normalization of Biases Present in Spotted cDNA Microarray Data: A Composite Method Addressing Dye, Intensity-Dependent, Spatially-Dependent, and Print-Order Biases”, DNA RESEARCH, Vol.12, No.1, pp.1-7, 2005.

[5] G. K. Smyth and T. Speed, “Normalization of cDNA microarray data, Methods”, 31(4), pp.265-273, 2003.

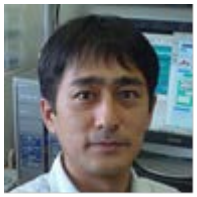
[6] I.S.Kohane, A.T.Kho, A.J.Butte, "Microarrays for an Integrative Genomics”, MIT Press, 2005.

[7] J. Quackenbush, “Maicroarray data normalization and transformation”, Nature Genetics Supplement, Vol.32, pp.496-501, 2002.

[8] Y. Xiao, M. R. Segal and Y. H. Yang, “Stepwise normalization of two-channel spotted microarrays”, Statistical Applications in Genetics and Molecular Biology, 4(1), No.4., 2004.

[9] G. Schwarz, “Estimating the dimension of a model”, Ann. Statist., Vol.6, No.2, pp.461-464, 1978.

[10] Stanford MicroArray Database, http://smd.princeton.edu/



Takeo Okazaki received the B.Sc. and M.Sc. degrees, from Kyushu University in 1987 and 1989, respectively. He had been a research assistant at Kyushu University from 1989 to 1995. He has been an assistant professor at University of the Ryukyus since 1995. His research interests are statistical data normalization for analysis, statistical causal relationship analysis. He is a member of JSCS, IEICE, JSS, GISA, and BSJ Japan.