Improving data reliability on oligonucleotide microarray

Yeoin Yoon[1], Young-Hak Lee[2], Jin Hyun Park[1*]

[1] Bioinfomatics Research Center, PNI Inc., Pohang, Korea
[2] Automation and Systems Research Institute and School of Chemical Engineering, Seoul National University, Seoul, Korea

*To whom correspondence should be addressed. E -mail: pcanda@postech.ac.kr

Abstract

The advent of microarray technologies gives an opportunity to moni tor the expression of ten thousands of genes, simultaneously. Such microarray data can be deteriorated by experiment al errors and image artifacts, which generate non-negligible outliers that are estimated by 15% of typical microarray data. Thus, it is an important issue to detect and correct the se faulty probes prior to high-level data analysis such as classification or clustering. In this paper, we propose a systematic procedure for the detection of faulty probes and its proper correction in Genechip array based on multivariate statistical approaches. Principal component analysis (PCA), one of the most widely used multivariate statistical approaches, has been applied to construct a statistical correlation model with 20 pairs of probes for each gene. And, the faulty probes are identified by inspecting the squared prediction error (SPE) of each probe from the PCA model. Then, the outlying probes are reconstructed by the iterative optimization approach minimizing SPE. We used the public data presented from the gene chip project of human fibroblast cell. Through the application study, the proposed approach showed good performance for probe correction without removing faulty probes, which may be desirable in the viewpoint of the maximum use of data information.

Introduction

The appearance of cDNA and oligonucleotide arrays has made it possible to monitor the thousands of genes in the parallel. The analysis of such technologies brings about a revolution in many different areas such as clinical and pharmaceutical research. And also, these array techniques may be used to analyze the clinical outcome and identify its related genes by comparing gene expression in cured patients with fatal ones. Although it is a promising tool for understanding functional genomics, the complexity of the data produced by these technologies is a major obstacle to analyze these

This work is supported by National R&D Program for Fusion Strategy of Advanced Technologies of MOST.

kinds of data.

In general, data analysis involved in microarray technology can be roughly divided into two categories such as low-level and high-level data analysis. The low-level data analysis refers to a standardization of the microarray data performed before entering the substantial analysis such as a feature extraction, contamination filtering and a data normalization, etc. Since the high-level data analysis can be performed only after the low-level analysis is accomplished well, many researchers on low-level analysis are conducted and the low-level analysis can be treated as a fundamental technology. The high-level data analysis is already established and well-known. It refers to identifying disease related genes and unrecognized tumor subtypes [1]. As the performance of high-level analysis is dependent on the results of low-level data analysis, the developed countries invest human resource and money tremendously in developing the fundamental technologies of the low-level data analysis.

Affymetrix GeneChip® arrays are most popular oligonucleotide array and consist of probe sets indicating unique genes. It is hundred thousands of probes composed of 20~25mer oligonucleotides are directly synthesized by a base on the array in different forms. Each probe set contains 20 probe pairs, so it has high reproducibility. The probe pair consists of perfect match (PM) and mismatch (MM). The sequence of PM and MM are the same except for one sing nucleotide difference at the middle of the sequence [2].

The data obtained form the oligonucleotide microarray is huge, and therefore, computer aided and statistical analysis is required. Most of microarray data analysis are performed at the gene expression level rather than the probe expression level, and as the oligonucleotide chip represents transcription information of one gene by probe expression data of about 20 probe pairs, so that it is complicated to analyze because probe data of about 20 pairs have to be simplified to one value through a specific process. A gene expression value can be calculated by the proper heuristic or model based methods using the corresponding probe data [2-6].

The oligonucleotide chip experiment requires very experienced technique, and exposed to many error or noise through whole process. It is very important to detect the faulty probe as well as to correct proper value based on a statistical background because the number of outlier can be estimated by 15% of typical microarray data [2]. It can be occurred in the following case: i) a sample itself is contaminated, ii) the spot is spoiled, iii) defects in experimental equipment and process, iv) hybridization is not completed correct, v) cross-hybridization occurs and so forth [3].

The measurement errors affect a specific probe to get an abnormal expression value, and therefore, the probe shows a different pattern from the expression pattern of other normal probes. The sample having an abnormal probe expression value is called "outlier" in the oligonucleotide chip.

Therefore, it is a critical issue to detect and correct the faulty probes before high level data analysis was performed. Despite the importance

of low-level data analysis, few research efforts have been devoted to detect the outlier of oligonucleotide microarray data [1]. Li and Wong [4] proposed a statistical model for the probe-level data and they can detect the outlier based on this model.

The aim of this paper is to identify the faulty probes and correct them based on multivariate statistical approach. In order to identify the samples including the genes deteriorated by faulty probes, we used a confirmatory clustering using the supervised class information because we knew the exact class in this data. And then, we used squared error prediction (SPE) method for detecting the gene including faulty probes. After identifying probes, we reconstruct the faulty probes based on measurement correlation between 20 pair probe intensities for each gene in oligonucleotide microarray data.

Methods

The methodology for correcting chip data is largely consists of the five steps. First, a statistical significance of the chip data should be identified. The status of data corrosion is monitored and the possibility of the correction is checked in this step. Second, if the raw data is considered as one to be corrected in the first step, the correcting models are constructed. An individual model is generated to each gene expressed with the assembled intensities of probes according to the number of arrays. Then, a corrupted gene and the faulty probes in the gene are detected. After the detected intensities are corrected using the PCA models, validating the results terminates the correction processing.

Theoretical background

Principal Component Analysis

Principal Component Analysis (PCA) is an effective tool in multivariable data analysis. It transforms the high-dimensional problems into lower dimensional problems with the minimum loss of information. This method is particularly useful in analyzing the large set of correlated data. It transforms correlated variables into new uncorrelated orthogonal variables, called principal components (PCs). Each PC is a linear combination of original variables. The coefficients of each linear combination are obtained from the corresponding eigenvector of the covariance matrix of original variables [10].

Let $x \in R^m$ denote a array vector of m variable. Assuming there are n samples for each variable, a data matrix $X \in R^{n \times m}$ is composed with each row representing a array. X can be represented by the product of a loading matrix P which shows the influence of variables and a score matrix T that summarizes the X variables. E is a residual matrix representing the deviations between the original values and the projections:

$$X = \hat{X} + E \tag{1}$$

$$\hat{X} = TP^T = \sum_{i=1}^{l} t_i p_i^T \tag{2}$$

where t and p are the score and loading vector, respectively, l is the number of principal components.

Then, a particular array vector x can be projected into the principal component subspace (S_p), which is spanned by the first l loading vectors, and the residual space (S_r), respectively. The projection of x on the principal component subspace

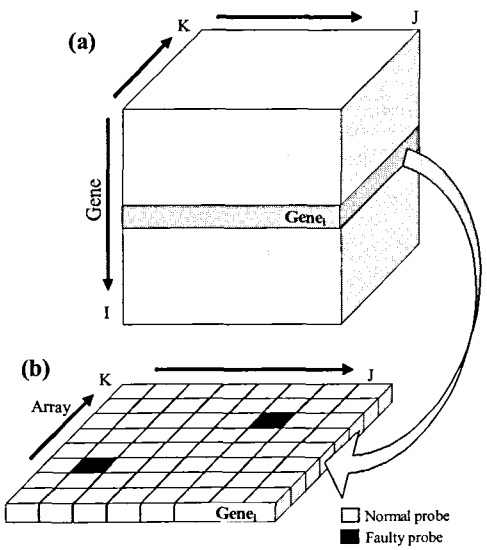


Figure 1. Model data structure for correction in oligonucleotide microarray. (a) is three dimensional structure of oligonucleotide data. (b) is two dimensional structure data for building the model in each gene.

is represented by

$$\hat{x} = PP^T x \equiv Cx \quad (3)$$

The projection of x on the residual space is

$$\tilde{x} = (I - C)x \quad (4)$$

Since S_p and S_r are orthogonal, two projected amounts satisfy

$$\tilde{x}^T \hat{x} = 0 \quad (5)$$

In order to determine the number of PCs, we used the variance of reconstruction error (VRE) [6]. The VRE can be decomposed into a portion in the principal component subspace and a portion in the residual subspace. As a result, the VRE always has a minimum which points to the optimal number of PCs [11].

SPE

SPE represents the squared perpendicular distance between a new multivariate observation and the corresponding reconstructed observation obtained by removing from the principal subspace. It offers way to test whether the correlations between measurements are valid or not. For *i*th array,

$$\mathrm{SPE}_i = \sum_{j=1}^{m} (x_j - \hat{x}_j)^2 \quad (6)$$

Jackson [7] developed a statistical significance test for the residuals obtained from Eq. (6). Such a test suggests that the correlation among variables is not valid if

$$SPE \geq cl(\alpha) \quad (7)$$

where $cl(\alpha)$ is the confidence limit for the 1-α percentile in a normal distribution [11].

Clustering

Clustering is a technique used for combining observation into groups according to grouping objectives. It can provide informal information for identifying outliers or tumor subtypes. Many researchers have been implemented clustering techniques to extract useful information from microarray data [12, 13]. There are various methods for partitioning data into meaningful sub-groups. In this study, we mainly used the hierarchical clustering methods because the graphical visualization of these methods, called dendrogram, enables to recognize outlying samples easily.

Formation of data matrix for correcting the chip data

Oligonucleotide chip data have three-dimensional I x J x K structure: array, gene and probe intensities in gene as Shown Figure 1(a). Data matrix for each gene is generally composed of the intensities for 40 probes in all arrays, which

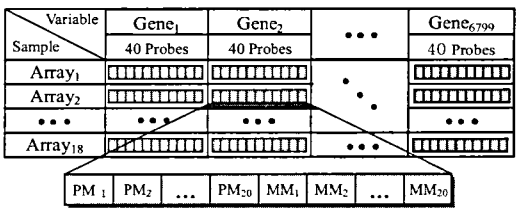
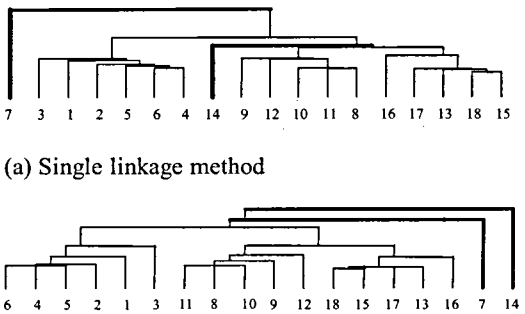


Figure 2. Data Structure of human fibroblast



(a) Single linkage method

(b) Single linkage method

Figure 3. Two hierarchical clustering for outliers detection

are strongly correlated, as shown Fig. 1(b). The correction model is based on the correlation structure of probes in each gene. Data matrix for the correction modeling for an individual gene constructed in the J□K matrix for each gene. Faulty probe intensities can be identified as those violating correlation between probes in each gene.

Finding out the array to be corrected

In general, microarray data include the considerable biased information considerable due to ambiguity in image processing, instrumental trouble, and so on. The target arrays to be corrected are detected as ones violating ones to our objective. PCA has been applied to visualizing the arrays composed of a large number of genes and clustering the arrays to represent similarity of those quantitatively. The score plot from PCA and the dendrogram from clustering are efficient for representing shows array similarities, groups, trends, outliers and so on.

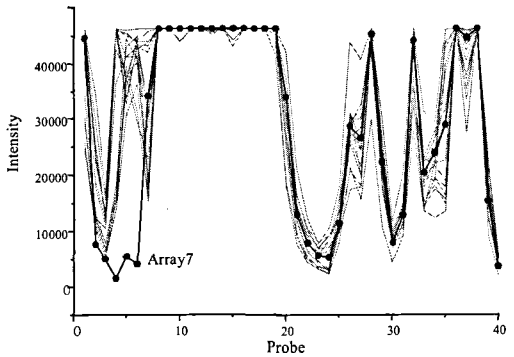
Two types of the outliers are divided according to the characteristics of outliers in a group or among groups. Type I is assigned to the outlying samples far from the normal ones that are identified by the statistical confidence region in PC space. Type II is the case that an array is assigned as the other cluster to be not done. Both cases could make a bad effect on analysis of gene expression data. Two criteria are well-known to find out the array to be corrected. One is to investigate whether the arrays are deviated from a statistical boundary of array-to-array variations. A statistical boundary is constructed based on the Hotelling T^2 statistics [14]. The Hotelling T^2 for array i, based on A PCs is

$$T_i^2 = \sum_{\alpha=1}^{A} \frac{t_{i\alpha}^2}{\lambda_\alpha} \qquad (8)$$

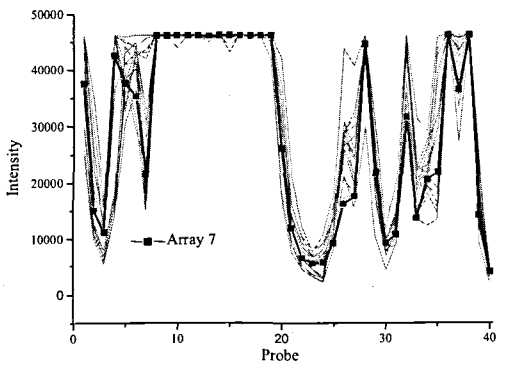
where λ_α is the variance of α th score. T2 value indicates the Mahalanobis distance of a point from origin in the PC space. $T_i^2 \times N(N\text{-}1) / A(N^2\text{-}1)$ is F distributed with A and N-A degrees of freedom where N is the number of arrays in the PCA model data and A is the number of PC in the model. Hence, if

$$T_i^2 > A(N^2-1)/N(N-A) \times F_c(p=0.05) \qquad (9)$$

then the array i is outside the 95% confidence region of the model. The F_c (p=0.05) is the critical value of F distribution with 0.05 as p value. The other is to detect the misclassified arrays in the predefined clusters in the case of utilizing the supervised information such as tumor type, the extent of disease, and so on.



(a) Before correction array 7



(b) After correction array 7

Figure 4. The probe expression level of the raw data of ribosomal protein L8 The upper plot shows raw data before correction of array7. The lower plot shows raw data after correction of array7.

Construction of PCA models using the intensities of probes in genes

The covariance structure of probes in a gene can be identified using data matrix in Figure 1(b). Faulty probes can be detected readily by comparing the covariance structure of probes . In order to correct the faulty probes, the PCA models are constructed using the data without the outlying arrays detected in the previous step. From PCA and clustering analysis, if the arrays are not classified into various groups, it prefers to construct a model with total data except outliers. If the arrays are clearly classified into various groups, it prefers to construct a model with selected group data.

Detecting gene expressions including the faulty probes using SPE

SPE is applied for detection of abnormal gene expressions due to faulty probe intensities. SPE is the index that detects a correlation breakdown. In other words, SPE index goes out of the confidence level as described in Eq. (7) when a array contains the probe with abnormal intensity. In general, the 15 % of expressed genes are known as contaminant information [7]. This index can detect the contaminant expressions based on strong correlation between probe intensities. A contribution plot approach based on SPE has been used to identify the faulty probes in a gene [14, 15]. Comparing the relative contribution of the probe intensities in SPE when a faulty array is detected identifies the faulty probes.

Correction of the faulty probe intensities and its validation

From the highly contributed probe to the large residual, one may estimate the ith faulty probe intensity from the orthogonal intensity using Eq. (3) [16] . In this approach, the probe is used to reconstruct itself. To eliminate the effect the faulty probe, we used the alternative approach suggested by Dunia et al. However we find the ith predicted probe intensity by iterating

$$z_i^{new} = c_{ii} z_i^{old} + [x_{-i}^T 0 \, x_{+i}^T] \, c_i = [x_{-i}^T \; z_i^{old} \; x_{+i}^T] \, c_i \quad (10)$$

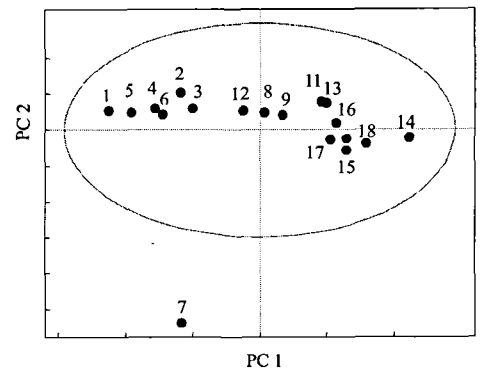
until it converges to a value z_i. Here

$$C = PP^T = [c_1 \; c_2 \; c_3 \cdots c_m] \quad (11)$$

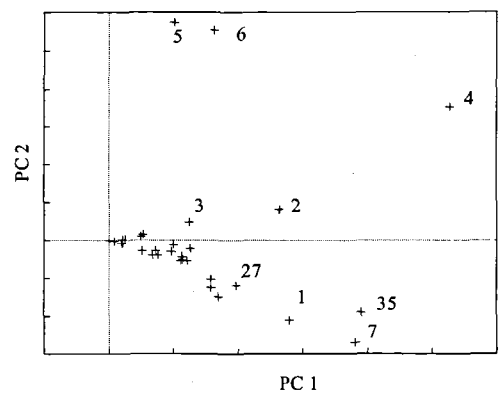
and

$$c_i^T = [c_{1i} \; c_{2i} \; \cdots \; c_{mi}] \quad (12)$$

and x^T represents a row vector composed of probe intensities of the model data matrix and the subscripts $-i$, $+i$ denote a vector formed by the first i-1 and the last m-i elements of the original vector, respectively.



(a) Score plot



(b) Loading plot

Figure 5. Score and loading plot for probe validation in Ribosomal protein L8

Optimization approach was implemented by Wise and Ricker (1991) for a set of corrected intensities. For the reconstruction of the *i*th variable, the optimization procedure reduces to

$$\min_{z_i} \|\breve{x}_i\|^2_{I-C} \qquad (13)$$

where $\breve{x}$ is identical to x except for the ith entry, z_i. Since ***I-C*** is positive semi-definite, the derivative of the objective function with respect to z_i leads to a necessary condition for a minimum:

$$\xi_i^T (I-C)\breve{x}_i = [x_{-i}^T \quad z_i^{old} \quad x_{+i}^T](\xi_i - c_i) = 0$$

From equation (10), if $c_i = \xi_i$, $z_i^{new} = z_i^{old}$.

This expression leads to the following solutions.

$$z_i = [x_{-i}^T \quad 0 \quad x_{+i}^T] c_i \; / (1-c_{ii}) \qquad (14)$$

The corrected probe data should be eval uated from posterior analysis using supervised information.

The asymptotic value for z_i is

$$z_i = \frac{[x_{-i}^T \quad 0 \quad x_{+i}^T] c_i}{1 - c_{ii}} \qquad (15)$$

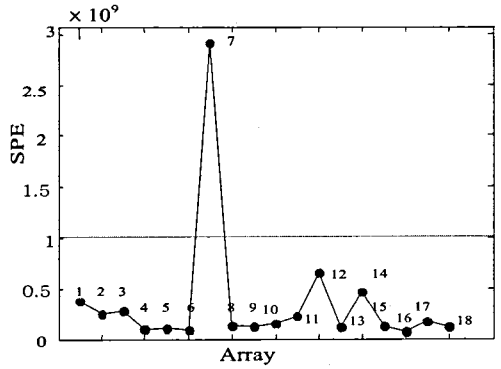
Results

Data description

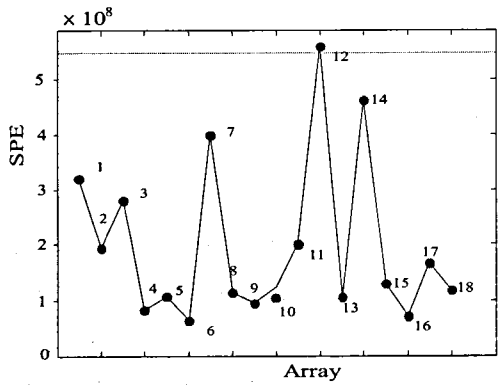
We used the public data presented from the gene chip project of human fibroblast cell using Affymetrix oligonucleotide Genechip® expression array in the Ohio State University [17]. The experimental data involves in the three groups of human fibroblast cells, with six replicate HuGeneFl arrays in each group. The three groups are serum-starved samples (A), serum stimulated samples(C), and a 50:50 mixture of starved/stimulated samples (B). Consequently, the true expression for each gene in array B is the average of the same one in array A and C. Each line of experiment data contains PM and MM data for a probe set. Fig.2 shows the data structure to be identified.

Detection of outlying arrays

The hierarchical clustering of samples is performed using 6799 each genes including 40 probes. 6799 genes were categorized in terms of their temporal response to growth factor se rum in fibroblast cells by using a standard correlation coefficient to define the distance between each pair of genes. Addition of serum induced proliferation such as absence of serum induces a



(a) Before correction array 7



(b) After correction array 7

Figure 6. SPE index for array in of ribosomal protein L8. The upper plot shows SPE before correction of array7. The lower plot shows SPE after correction of array7

non-dividing state termed G_0 and low metabolic activity.

The target arrays to be corrected are detected a using the popular hierarchical clustering. Single linkage and average linkage hierarchical clustering methods were employed. Fig . 3 shows that the array 7 and 14 are far away from the other normal samples.

Detection of the genes including faulty probes in a array

In the previous section we defined the outlying arrays that have probe faults. The faulty probes in a specific gene can be identified known as systematic errors in the arrays u sing PCA model of probe data for each gene. Using the SPE index based on the all array except to the array 7 from SPE index, 1350 genes among total 6799 genes were identified as ones including faulty probe data by the statistical significant test.

As shown in Fig. 4, the probe intensity pattern of Ribosomal protein L8 the array 7 is different from those in the other arrays. The score and loading plots, from PCA the model are shown in Fig. 5. In the PC model space, the score plot describe s the pattern of the uncorrelated new variables in the first two dimensions spanned by the first two PCs. Because the array 5 of score value is out of control limit, the array 7 is considered outliers. The loading plot describes which probes mainly affect the outlying array. The array 7 is largely affected by probe 4, probe 5, and probe 6, which can be identified as the faulty probes. Th e PC2 axis of loading plot helps ones to identify the faulty probes. The expression levels of probe 4, 5 and 6 are lower than those at the other arrays. This unusual phenomenon could result from the possibility of cross hybridization, failure of normal hybridization, and so on. In Fig. 6(a), the SPE plot from the PCA model for gene L8 shows that the array 7 is located on the outsi de the statistical limit which is a criterion whether the correlation structure of probes.

Correction of faulty probe intensities in a specified gene and validation

Correcting faulty probes using correlation structure refreshes the corrupted genes in the previous step. The SPE value before and after correction of array 7 are shown in Fig. 8 . Once the faulty probes are reconstructed, a significant decrease in the residual is expected. After the faulty probe intensities are corrected, the

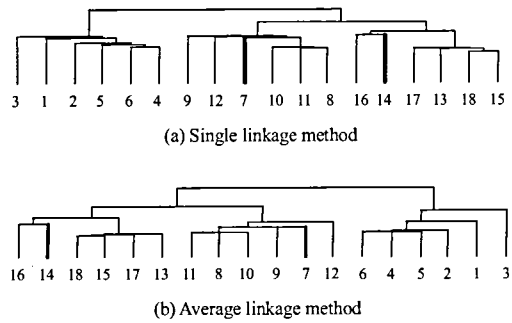


Figure 7. Two hierarchical clustering after correction of fault probe

correction performance can be checked using the hierarchical clustering method previously employed to detect the outliers. Fig. 7 represents the clustering results using two different clustering methods with corrected data. By comparing Fig. 5 with Fig. 7, the outlying array 7 and 14 are rearranged into the three groups, which show that the proposed correction algorism was performed on the oligonucleotide microarray data.

Conclusions

We have proposed a multivariate statistical approach for the faulty probe detection and its correction algorithm on the oligonucleotide microarray. The target arrays to be corrected can be detected using the popular hierarchical clustering with the cluster information. The basic idea for the faulty probe detection and correction is statistical correlation modeling and model based reconstruction skill. The proposed method shows satisfactory results for detecting faulty probes and reconstructing them. Also, we need not discard any probe intensity value, which gives a chance to use maximum information of microarray data. Furthermore, this approach may be extended to estimate missing values in microarray data.

Acknowledgements

This work is supported by National R&D Program for Fusion Strategy of Advanced Technologies of MOST.

References

[1] E.E. Schadt, C. Li, C. Su and W.H. Wong, Analyzing high-density oligonucleotide gene expression array data, *Journal of Cellular Biochemistry*, Vol. 80, 2000, pp. 192-202

[2] Statistical Algorithms Reference Guide, Affymetrix Technical Note 2002

[3] R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs and T.P. Speed, Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Res.*, Vol. 31, 2003, pp. e15

[4] J.C. Mills and J.I. Gordon, A new approach for filtering noise from high-density oligonucleotide microarray datasets, *Nucleic Acids Res.*, Vol. 29, 2001, pp. e72

[5] C. Li and H.W. Wong, Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *Proc. Natl. Acad. Sci.*, Vol. 98, 2001, pp. 31-36

[6] R. Sasik, E. Calvo and J. Corbeil, Statistical analysis of high-density oligonucleotide array: a multiplicative noise model, *Bioinformatics*, Vol. 18, No. 12, 2002, pp. 1633-1640

[7] M.L.T. Lee, F.C. Kuo, G.A. Whitemore and J. Sklar, Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations, *Proc. Natl. Acad. Sci.*, Vol. 97, 2000, pp. 9834-9839

[8] A.A. Hill, E.L. Brown, M.Z. Whitley, G. Tucker-Kellogg, C.P. Hunter and D.K. Slonim, Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls, *Genome Biology,* Vol. 2, 2001, pp. 0055.1-0055.13

[9] R. Nadon and J. Shoemaker, Statistical issues with microarrays: processing and analysis, *Trends in Genetics*, Vol. 18, 2002, pp. 265-271

[10] F.K. Wang and T.C.T. Du, Using principal component analysis in process performance for multivariate data, *Omega*, Vol. 28, 2000, pp. 185-194

[11] S.J. Qin and R. Dunia, Determining the number of principal component for best reconstruction, *Journal of Process Control,* Vol. 10, 2000, pp. 245-250

[12] T.R. Golub and D.K. Slonim, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, Vol. 286, 1999, pp. 531-537

[13] A.A. Alizadeh, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, Vol. 403, 2000, pp. 503-511

[14] T. Kourti and J.F. MacGregor, Process analysis, monitoring and diagnosis, using multivariate projection methods, *Chem. Intell. Lab. Syst.*, Vol. 28, 1995, pp. 3-21

[15] J.E. Jackson, A user's guide to principal components, New York: Wiley, 1991

[16] R. Dunia, S.J. Qin, T.F. Edger and T.J. McAvoy, Identification of faulty sensors using principal component analysis, *AIChE.* Vol. 42, 1996, pp. 2797-2812

[17] W.J. Lemon, J.J.T. Palatini, R. Krahe and F.A. Wright, Theoretical and experimental comparison of gene expression estimators for oligonucleotide arrays, *Bioinformatics,* Vol. 18. No. 11, 2002, pp. 1470-1476