New Normalization Methods using Support Vector Machine Regression Approach in cDNA Microarray Analysis

Insuk Sohn[1], Sujong Kim[2], Changha Hwang[3], Jae Won Lee[1]

[1]Department of Statistics, Korea University, Seoul, Korea
[2]Department of Biochemistry, College of Medicine, Hanyang University, Seoul, Korea
[3] Division of Information and Computer Sciences, Dankook University, Seoul, Korea
Email : sis46@korea.ac.kr, sundance@amorepacific.com, chwang@dankook.ac.kr, jael@korea.ac.kr

ABSTRACT: There are many sources of systematic variations in cDNA microarray experiments which affect the measured gene expression levels like differences in labeling efficiency between the two fluorescent dyes. Print-tip lowess normalization is used in situations where dye biases can depend on spot overall intensity and/or spatial location within the array. However, print-tip lowess normalization performs poorly in situation where error variability for each gene is heterogeneous over intensity ranges. We proposed the new print-tip normalization methods based on support vector machine regression (SVMR) and support vector machine quantile regression (SVMQR). SVMQR was derived by employing the basic principle of support vector machine (SVM) for the estimation of the linear and nonlinear quantile regressions. We applied our proposed methods to previous cDNA microarray data of apolipoprotein-AI–knockout (apoAI-KO) mice, diet-induced obese mice, and genistein-fed obese mice. From our statistical analysis, we found that the proposed methods perform better than the existing print-tip lowess normalization method.

1 INTRODUCION

The technique of cDNA microarray is a new tool in biotechnology, which allows the simultaneous monitoring of thousands of gene expression in cells [1]. This technology has important applications in pharmaceutical and clinical research. By comparing gene expressions in normal and tumor tissues, for example, we can use microarrays to identify tumor-related genes and targets for therapeutic drugs [2].

In a cDNA microarray experiment, two mRNA samples (to be compared) are reverse transcribed into cDNA, labeled using two different fluorescent dyes (usually a red fluorescent dye, Cy5, and a green fluorescent dye, Cy3) and then hybridized simultaneously to the arrayed DNA sequences or probes on the glass slide. Intensity values generated from hybridization to individual DNA spots are indicative of gene expression levels, and the relative abundance of each transcript in the two samples is derived from the resulting intensity ratios [3].

The main idea of normalization for dual labeled arrays is to adjust artifactual differences in intensity of the two labels. Such differences result from differences in affinity of the two labels to DNA, differences in amounts of sample and label used, differences in photomultiplier tube and laser voltage settings, and differences in photon emission response to laser excitation. Although normalization alone cannot control all systematic variations, a choice of the normalization method is important in the earlier stage of microarray data analysis because subsequent analyses, such as differential expression testing, clustering, and gene networks are quite dependent on the pre-processing steps such as image analysis and normalization procedure [4, 5].

Yang et al. [6, 7] summarizes a number of normalization methods for dual labeled microarrays, such as intensity dependent normalization and print-tip lowess normalization. Some other nonlinear normalization methods have been employed, such as B-splines and Gaussian-kernel fitting [8, 9]. Recently, Baird et al. [10] proposed the normalization using spatial mixed models which include splines and Eckel et al. [11] proposed semiparametric normalization procedure utilizes a linear model. However, to date, evaluation of normalization methods applicable to the microarray data where error variability for each gene is heterogeneous over intensity ranges, has not been investigated. We propose some new print-tip normalization methods based on support vector machine regression (SVMR) and support vector machine quantile regression (SVMQR), which perform well in microarray data with heterogeneous error variability depending on signal intensity.

The support vector machine (SVM) was initially developed by Vapnik [12, 13] and his group to solve classification problems and has been successfully applied to a number of real world problems, such as: handwritten character and digit recognition; face detection; text characterization and object detection in machine vision. Recently, its applications have been extended to the domain of regression problems. SVM is based on the structural risk minimization (SRM) principle, which has been shown to be superior to traditional empirical risk minimization (ERM) principle. SRM minimizes an upper bound on the expected risk while ERM minimizes the error on the training data. By minimizing this bound, high generalization performance can be achieved. In particular, for the SVM regression case, SRM results in the regularized ERM with the ε-insensitive loss function. An introduction and overview of recent developments of SVM regression can be found in Cristianini and Shawe-Taylor [14], Gunn [15], Smola and Scholkopf [16] and Vapnik [12, 13]. Recently, Takeuchi and Furuhashi [17] propose non-crossing quantile regression curves via SVM.

In this article, we present an estimation method for linear and nonlinear quantile regressions using the basic principle of SVM. Following this, we propose new print-tip normalization methods based on SVMR and

SVMQR in order to adjust systematic variations in situations where dye biases can depend on spot overall intensity and/or spatial location within the array. We also evaluate the performance of these normalization methods by using the mean of variance in gene expression for each gene separately, within each experimental group as a measure.

A summary of the paper proceeds as follows. In the SYSTEMS AND METHODS section, we present SVMQR and the related normalization methods. In the RESULTS section, we apply our proposed normalization methods to cDNA microaray data of apolipoprotein AI (apo AI) knockout mice, diet-induced obese mice, and genistein-fed obese mice, and present the comparison results. Finally, CONCLUSION and DISCUSSION is given.

2 Support Vector Machine Quantile Regression (SVMQR)

Takeuchi and Furuhashi [17] address the quantile-crossing problem using SVMR approach. With the commonly used kernel trick, they derive a non-crossing conditional quantile estimator in the form of a constrained maximization of a piecewise quadratic function. In a similar setting, we derive linear and nonlinear quantile regression methods by implementing the idea of SVM. In particular, consider a random sample

$$(\mathbf{x}_i, y_i) \in R^d \times R,\ i = 1,...,n,$$

where the output variable y_i is related to the vector $\mathbf{x}_i$ of covariates, possibly including a constant term.

2.1 Linear SVMQR

In the linear quantile regression model introduced by Koenker and Bassett [18], the quantile function of the response y_i for a given $\mathbf{x}_i$ is assumed to be linearly related to the input vector $\mathbf{x}_i$ as follows:

$$Q(\theta \mid \mathbf{x}_i) = \beta(\theta)^t \mathbf{x}_i \quad \text{for } \theta \in (0, 1),$$

where $\beta(\theta)$ is the $\theta - th$ regression quantile and its estimator is defined as any solution to the optimization problem,

$$\min_{\beta} \sum_{i=1}^{n} \rho_\theta (y_i - \beta(\theta)^t \mathbf{x}_i) \text{ for } \theta \in (0, 1),$$

where ρ_θ is the check function defined as

$$\rho_\theta(r) = \theta r I(r \geq 0) + (\theta - 1) r I(r < 0).$$

We now describe how to implement the idea of SVM for the linear quantile regression. Since quantile regression is in principle based on absolute deviation loss, we adopt the procedures of the case $\varepsilon = 0$ in a standard SVM to derive quantile regression using the idea of SVM. Because Vapnik's $\varepsilon -$ insensitive loss function described by

$$|\xi|_\varepsilon = \begin{cases} 0 & if\ |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases},$$

and the case $\varepsilon = 0$ corresponds to considering standard absolute deviation loss function. Furthermore, we make the intercept term invisible in the expression by including it in the regression coefficient vector, in order to follow the above basic idea of quantile regressions and to avoid computation of the intercept $\mathbf{b}$ in the same manner as SVM. We do so for the sake of convenience. Indeed, reexpress $\mathbf{w}$ and $\mathbf{x}_i$ as $\mathbf{w} = (\mathbf{b}, \mathbf{w}^t)^t$ and $\mathbf{x}_i = (1, \mathbf{x}_i^t)^t$ (with an abuse of notation, we use the same notation for the resulting new vector). Then, we can express the linear quantile regression problem by implementing the formulation for SVM.

$$\text{Minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} (\theta\, \xi_i + (1-\theta)\, \xi_i^*) \quad \text{for } \theta \in (0, 1),$$

$$\text{subject to} \begin{cases} y_i - \mathbf{w}^t \mathbf{x}_i \leq \xi_i \\ \mathbf{w}^t \mathbf{x}_i - y_i \leq \xi_i^* \\ \xi, \xi_i^* \geq 0 \end{cases},$$

where the $\theta - th$ regression quantile $\beta(\theta)$ is expressed in terms of $\mathbf{w}$, ξ is upper training error, and ξ^* is lower training error. The constant $C > 0$ determines the trade off between the flatness of f and the amount up to which deviations larger than 0 are tolerated. We construct a Lagrange function as follows:

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} (\theta\, \xi_i + (1-\theta)\, \xi_i^*) - \sum_{i=1}^{n} \alpha_i (\xi_i - y_i + \mathbf{w}^t \mathbf{x}_i)$$
$$- \sum_{i=1}^{n} \alpha_i^* (\xi_i^* + y_i - \mathbf{w}^t \mathbf{x}_i) - \sum_{i=1}^{n} (\eta_i \xi_i + \eta_i^* \xi_i^*) \qquad (1)$$

We notice that the positivity constraints $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$ should be satisfied. After taking partial derivatives of equation (1) with regard to the primal variables $(\mathbf{w}, \xi_i, \xi_i^*)$ and substituting them into equation (1), we have the optimization problem

$$\max_{\alpha, \alpha^*} -\frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_i - \alpha_i^*)\, \mathbf{x}_i^t \mathbf{x}_j + \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) y_i$$

with constraints $\alpha_i \in [0, \theta C]$ and $\alpha_i^* \in [0, (1-\theta)C]$.

Solving this optimization problem with the constraints determines the optimal Lagrange multipliers, $\hat{\alpha}_i, \hat{\alpha}_i^*$, the $\theta - th$ regression quantile estimators, and the $\theta - th$ quantile function predictors of the input vector $\mathbf{x}$ where the latter two are defined respectively as follows:

$$\hat{\mathbf{w}} = \sum_{i=1}^{n} (\hat{\alpha}_i - \hat{\alpha}_i^*)\, \mathbf{x}_i \quad \text{and} \quad \hat{Q}(\theta \mid \mathbf{x}) = \sum_{i=1}^{n} (\hat{\alpha}_i - \hat{\alpha}_i^*)\, \mathbf{x}_i^t\, \mathbf{x}.$$

2.2 Nonlinear SVMQR

In nonlinear quantile regression, the quantile function of the response y_i for a given $\mathbf{x}_i$ is assumed to be nonlinearly related to the input vector $\mathbf{x}_i \in R^d$. For nonlinear quantile regression, the input vector $\mathbf{x}_i$ is nonlinearly transformed into a potentially higher dimensional feature space R^f by some function $\phi(\cdot)$. Here, similar to SVM for nonlinear regression, the nonlinear regression quantile estimator cannot be given in explicit form since we use the kernel function of the input vector instead of the dot product of their feature mapping function. The quantile function of the response y_i for a given $\mathbf{x}_i$ can be given as

$$Q(\theta \mid \mathbf{x}_i) = \beta(\theta)^t \phi(\mathbf{x}_i) \text{ for } \theta \in (0,\ 1),$$

where $\beta(\theta)$ is the $\theta - th$ regression quantile. Then, by constructing the Lagrangian with kernel $K(\cdot, \cdot)$, we obtain the optimal problem similar to the linear quantile regression case as follows

$$\max_{\alpha,\ \alpha^*} -\frac{1}{2}\sum_{i,\ j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\mathbf{x}_i,\ \mathbf{x}_j) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*)$$

with constraints $\alpha_i \in [0, \theta C]$ and $\alpha_i^* \in [0, (1-\theta)C]$. Solving the above optimization problem with the constraints we obtain the optimal Lagrange multipliers, $\hat{\alpha}_i$, $\hat{\alpha}_i^*$, so that the $\theta - th$ quantile function predictor given the input vector $\mathbf{x}$ can be obtained as

$$\hat{Q}(\theta \mid \mathbf{x}) = \sum_{i=1}^{n}(\hat{\alpha}_i - \hat{\alpha}_i^*)K(\mathbf{x}_i,\ \mathbf{x}).$$

2.3 Normalization methods.

In this section, we propose new print-tip normalization methods based on both Vapnik's support vector machine regression (SVMR) and our proposed support vector machine quantile regression (SVMQR). The new print-tip SVM quantile median regression (SVMQMR) and SVM interquantile mean regression (SVMIQMR) normalizations are derived.

2.3.1 Print-tip SVMR normalization.

Each M-value ($\log_2(R/G)$) is normalized by subtracting from it the corresponding value of the tip group SVMR curve. The normalized log-ratios N are the residuals from the tip group SVMR, i.e.,

$$N = M - SVMR_i(A),$$

where $SVMR_i(A)$ is SVMR curve as a function of A-value ($\log_2(R/G)$) for the i th tip group.

We used the R implementation svm () that is based on LIBSVM [19] for the implementation of SVMR. The regularization parameter C, the kernel parameter σ, and ε are chosen using a 10-fold cross validation.

2.3.2 Print-tip SVMQMR normalization.

Each M-value is normalized by subtracting from it the corresponding value of the tip group SVM $0.5 - th$ quantile regression curve. The normalized log-ratios N are the residuals from the tip group SVM $0.5 - th$ quantile regressions, i.e.,

$$N = M - SVMQMR_i(A),$$

where $SVMQMR_i(A)$ is SVM $0.5 - th$ quantile regression curve as a function of A –value for the i th tip group.

The regularization parameter C and the kernel parameter σ are chosen using 10-fold cross validation for the implementation of SVM quantile median regression.

2.3.2 Print-tip SVMIQMR normalization

Each M-value is normalized by subtracting from it the corresponding average value of the tip group SVM $0.25 - th$ quantile regression curve and the tip group SVM $0.75 - th$ quantile regression curve. The normalized log-ratios N are the residuals from the average of the tip group SVM $0.25 - th$ quantile regressions and the tip group SVM $0.75 - th$ quantile regressions, i.e.,

$$N = M - (SVMQ1_i(A) + SVMQ3_i(A))/2,$$

where $SVMQ1_i(A)$ is SVM $0.25 - th$ quantile regression curve as a function of A –value for the i th tip group and $SVMQ3_i(A)$ is SVM $0.75 - th$ quantile regression curve as a function of A –value for the i th tip group.

The regularization parameter C and the kernel parameter σ are chosen using 10-fold cross validation for the implementation of SVM quantile regression.

3 RESULTS

We apply our print-tip SVMR, SVMIQMR, and SVMQMR normalization methods to previous cDNA microarray data of apolipoprotein-AI –knockout (apoAI-KO) mice, diet-induced obese mice, and genistein-fed obese mice.

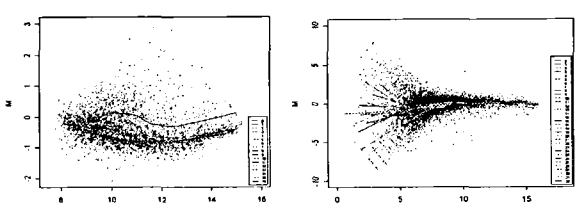
The cDNA microarray data of apoAI-KO mice, which compares the gene expression profiles between a test group of six mice with the apoAI gene knocked out and a control group of six normal mice, was previously reported in Callow et al. [20]. Target cDNA is obtained from mRNA of six apoAI-KO mice and six normal C57BL/6J mice by reverse transcription and labeled with a red fluorescent dye, Cy5. The control sample used in all hybridizations is prepared by pooling cDNA from the eight normal C57BL/6J mice and labeled with green fluorescent dye, Cy3. Here, we analyze the data from 6 different hybridizations performed with target cDNA from three apoAI-KO mice and three normal C57BL/6J mice. Probes are spotted onto glass slides using 4×4 print head.

The second data is from a cDNA microarray experiment of diet-induced obese mice(E/F) reported previously in Kim et al. [21]. The experimental group consists of 6 mice supplemented with a high-fat diet (HFD) for 12 weeks and a control group consists of age/weight-matched 6 mice

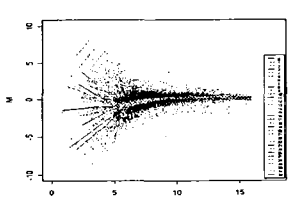
supplemented with low-fat diet (LFD) for 12 weeks. Equal amounts of RNA from six mice of each group are pooled and each sample is equally divided; one half is used to generate Cy3-labeled cDNA, and the other half is used to generate Cy5-labeled cDNA for dye swapping. Six technical replicates of hybridization are performed, and three of these are repeated with the fluorophores reversed to prevent dye-bias. The Cy5 and Cy3 probes are mixed and hybridized to microarray containing 10,336 cDNA probes. Probes are spotted onto glass slides using 4×8 print head. The two fluorescent images (Cy3 and Cy5) are scanned separately by a GMS 418 Array Scanner (Affymetrix), and the signal intensity values are obtained from ImaGene 4.2 (Biodiscovery) and MAAS (Gaiagene, Seoul, Korea) software.

The third data is from a cDNA microarray experiment of genistein-fed obese mice (E/B) reported previously in Kim et al. [21]. Thirty mice were randomly assigned to one of three dietary groups (n=10) for 12 wk; a low-fat diet (LFD), a high-fat diet (HFD) and the HFD supplemented with genistein (2 g/kg diet) (HFD+GEN). Equal amounts of the RNA from six mice of each group were pooled, and each pooled RNA sample was equally divided; one half was used to generate Cy3-labeled cDNA, the other half was used to generate Cy5-labeled cDNA for dye swapping. We analyze the data from 6 different hybridizations performed with target cDNA from HFD+GEN mice and control cDNA from HFD mice.

Figure 1 shows the M versus A plot of each dataset (Figure 1) (in the same fashion as Figure 1 in Yang et al (2001)), where the log-ratio is given by $M = \log_2(R/G)$ and average log-intensity is given by $A = \log_2\sqrt{RG}$. Plots (b) and (c) show a tendency of increasing dispersion of the log-ratio M as the spot intensity A decreases. Lowess regression curves appear to have narrower spacing in regions of high average log-intensity A and wide spacing in regions of low average log-intensity A. The conditional distribution of the log-ratio M may be asymmetric and heteroscedastic.



(a) ApoAI-KO mice data (b) Diet-induced obese mice data



(c) Genistein-fed obese mice data

Figure 1: M versus A plot of each dataset displaying the lowess curve for each of the print-tips. Here "g" denotes the lowess curve for the entire dataset.

Generally, diagnostic plots such as M versus A plots, density plots, box plots and spatial plots can be used for visually comparing different normalization methods or for checking whether the artifacts have been removed by normalization in microarray analysis. In addition to M versus A plots, we adopt the mean of variance in gene expression for each gene separately, within each experimental group in order to compare different normalization methods. The variance in gene expression for each gene separately, within each experimental group is estimated as

$$\hat{\sigma}_g^{\,2} = \frac{1}{l(m-1)} \sum_{i=1}^{m} \sum_{j=1}^{l} (M_{ij} - \overline{M}_{.j})^2,$$

where $\overline{M}_{.j} = \frac{1}{m} \sum_{i=1}^{m} M_{ij}$, g is an index over n genes, i is an index over m replicates, and j is an index over l experimental groups. The smaller variance estimates provide better normalization methods. Table 1 shows the mean of $\hat{\sigma}_g^{\,2}$ values for apoAI-KO mice data, diet-induced obese mice data, and genistein-fed obese mice data, respectively.

ApoAI-KO mice data

Raw	print-tip lowess	print-tip SVMR	print-tip SVMIQMR	print-tip SVMQMR
0.135	0.095	0.094	0.094	0.098

Diet-induced obese mice data

Raw	print-tip lowess	print-tip SVMR	print-tip SVMIQMR	print-tip SVMQMR
0.877	0.759	0.737	0.729	0.743

Genistein-fed obese mice data

Raw	print-tip lowess	print-tip SVMR	print-tip SVMIQMR	print-tip SVMQMR
1.125	0.758	0.729	0.702	0.738

Table 1: The mean of variance in gene expression for each gene separately, within each experimental group in each dataset.

For apoAI-KO mice data, print-tip SVMR and SVMIQMR normalization methods provide the lowest the mean of $\hat{\sigma}_g^{\,2}$ value and print-tip lowess normalization method also provide low the mean of $\hat{\sigma}_g^{\,2}$ value. For diet-induced obese mice and genistein-fed obese mice, print-tip SVMIQMR normalization method provide the lowest the mean of $\hat{\sigma}_g^{\,2}$ value and print-tip lowess normalization method provide the largest the mean of $\hat{\sigma}_g^{\,2}$ value. This result implies that print-tip SVMR and SVMIQMR normalization methods provide superior performance as compared to print-tip lowess regression normalization method. We observe that print-tip SVMR and SVMIQMR normalization methods provide consistently good performance with apoAI-KO mice data, diet-induced

obese mice and genistein-fed obese mice, while print-tip lowess normalization method perform poorly with diet-induced obese mice and genistein-fed obese mice. In particular, print-tip SVMIQMR normalization method seems to be the best.

Broberg [22] compared the average ranks of the four testing methods such as B-statistic, SAM, samroc and T-statistic. We also use the average ranks of SAM method (significance analysis of micoarrays) [23] of the reference genes. The selection of reference gene is not an easy task, since it is usually not known which genes are true positives for a specific biological sample. Even the "verification" of microarray results by an conventional technique such as quantitative RT-PCR is just replacing one error-prone method by another. For the diet-induced obese mice, we used 106 reference genes which selected differentially expressed genes to be detected by previous our work [21]. For the Apo AI-KO mice data, we use 8 reference genes previously verified as differentially expressed genes.

ApoAI-KO mice data (8 genes)

	print-tip lowess	print-tip SVMR	print-tip SVMIQMR	print-tip SVMQMR
Average	5.12	5.12	4.5	5.25

Diet-induced obese mice data (106 genes)

	print-tip lowess	print-tip SVMR	print-tip SVMIQMR	print-tip SVMQMR
Average	59.56	58.77	58.08	58.75

Table 2: Average ranks of SAM of the reference genes in each dataset.

For the Apo AI-KO mice data, print-tip SVMIQMR normalization method provides the lowest average rank and print-tip lowess and print-tip SVMR normalization methods also provide low average rank. For diet-induced obese mice data, print-tip SVMIQMR normalization method provides the lowest average rank and print-tip lowess normalization method provides the largest average rank.

4 CONCLUSION AND DISCUSSION

In this paper, we propose new print-tip normalization methods based on SVMR and SVMQR in order to adjust systematic variations in situations where dye biases can depend on spot overall intensity and/or spatial location within the array. One of the problems of lowess regression is that it performs poorly in situations where error variability for each gene is heterogeneous over intensity ranges. Diet-induced obese mice and genistein-fed obese mice data presented in the RESULT section are shown to be heteroscedastic. In essence, for such data with heteroscedasticity, SVM and SVMQ regressions perform robustly since both methods use a version of absolute. It turns out that our proposed print-tip SVMR and print-tip SVMIQMR normalization methods perform superior to print-tip lowess normalization method for microarray data with heteroscedasticity such as diet-induced obese mice and genistein-fed obese mice data. It is established that print-tip SVMR and SVMIQMR normalization methods give consistently superior performance in apoAI-KO mice data, diet-induced obese mice and genistein-fed obese mice, while print-tip lowess normalization method perform poorly with diet-induced obese mice and genistein-fed obese mice. Print-tip SVMIQMR normalization method appears to be the best and print-tip SVMR normalization method also performs well.

Although our print-tip normalization has a little complex than print-tip lowess normalization, for data with in the presence of heteroscedasticity, our print-tip normalization can be obtained more informative than print-tip lowess normalization. Lowess normalization is actually concerned with the estimation of conditional mean of response variable given input variables. By the way, quantile regression deals with the estimation of conditional median or the $\theta - th$ quantile of response variable given input variables. Therefore, if we apply informations on several quantiles of response variables to normalization, for data with in the presence of heteroscedasticity due to systematic variations, our print-tip normalization methods based on support vector machine quantile regression gives much better normalization than the existing print-tip lowess method only using overall central tendency.

ACKNOWLEDGEMENTS

This work was supported by Korea Science and Engineering Foundation Grant (R14-2003-002-01000).

REFERENCES

[1] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. The chipping forecast, 21: 33--37, 1999.

[2] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown and L. M. Staudt. Different type of diffuse large b-cell lymphoma identified by gene expression profiling. Nature, 403: 503--511, 2000.

[3] M. Taniguchi, K. Miura, H. Iwao and S. Yamanaka. Quantitative assessment of DNA microarrays comparison with northern blot analyses. Genomics: 71, 34--39, 2001.

[4] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci., 95: 14863--14868, 1998.

[5] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh,M, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286: 531--537. 1999.

[6] Y. H. Yang, S. D. Dudoit, P. Luu and T. P. Speed. Normalization for cDNA microarray data. In SPIE BioE . 2001.

[7] 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.

[8] C. Workman, L. J. Jensen, H. Jarmer, R. Berka, L. Gautier, H. B. Nielsen, H. H. Saxild, C. Nielsen, S. Brunak and S. Knudsen. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biology, 3:1--16, 2002.

[9] Y. Zhou, F. G. Gwadry, W. C. Reinhold, L.D. Miller, L. H. Smith, U. Scherf, E. T. Liu, K. W. Kohn,Y. Pommier and J. N. Weinstein. Transcriptional regulation of mitotic genes by Camptothecin-induced DNA damage: Microarray analysis of dose and time-dependent effects. Cancer Research, 62: 1688--1695, 2002.

[10] D. Baird, P. Johnston, T. Wilson. Normalization of microarray data using a spatial mixed model analysis which includes splines. Bioinformatics, 20: 3196--3205, 2004,

[11] J. E. Eckel, C. Gennigs, T. M. Therneau, L. D. Burgoon, D. R. Boverhof, and T. R. Zacharewski. Normalization of two-channel microarray experiments: a semiparametric approach. Bioinformatics, 21: 1078--1083, 2005.

[12] V. N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.

[13] V. N. Vapnik. Statistical Learning Theory. Springer, New York, 1998.

[14] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press, 2000.

[15] S. Gunn. Support Vector Machines for *Classification and Regression. ISIS Technical* Report, University of Southampton, 1998.

[16] A. Smola and B. Scholkopf. On a kernel-based method for Pattern Recognition, Regression, approximation and operator inversion, Algorithmica. 22: 211--231, 1998.

[17] I. Takeuchi and T. Furuhashi. Non-Crossing Quantile Regression Curves by Support Vector Machine and Its Efficient Implementation. the proceeding of International Joint Conference on Neural Networks 2004.

[18] R. Koenker and G. Bassett Regression Quantiles. Econometrica, 46, 33-50, 1978.

[19] C. C. Chang and C. J. Lin. LIBSVM: a library for support vector machines, available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[20] M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed and E. M. Rubin. Microarray expression profiling identifies genes with altered expression in HDL deficient mice. Genome Research, 10: 2022--2029, 2000

[21] S. Kim, I. Sohn, J. –I. Ahn, K. –H. Lee, Y. –S. Lee and Y. –S. Lee. Hepatic gene expression profile in long-term high-fat diet-induced obesity mouse model. Gene (in press), 2004

[22] P. Broberg. Ranking genes with respect to differential expression. Genome Biology 3: preprint0007.1-0007.23, 2002

[23] V. G. Tusher, R. Tibshirani, G. Chu.Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U.S., 98: 51116--5121, 2001.