Journal of Biomedical Informatics 41 (2008) 1021–1027

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Segmentation of cDNA microarray images by kernel density estimation

Tai-Been Chen [a,d], Henry Horng-Shing Lu [a,*], Yun-Shien Lee [b,c], Hsiu-Jen Lan [a]

[a] Institute of Statistics, National Chiao Tung University, 1101 Ta Hsueh Road, Hsinchu 30010, Taiwan, ROC
[b] Department of Biotechnology, Ming Chuan University, Taiwan, ROC
[c] Genomic Medicine Research Core Laboratory, Chang Gung Memorial Hospital, Taiwan, ROC
[d] Department of Medical Imaging and Radiological Sciences, I-Shou University, Taiwan, ROC

ARTICLE INFO

Article history:
Received 26 October 2007
Available online 7 March 2008

Keywords:
Microarray
Segmentation
Kernel density estimation
Concordance correlation coefficient
Gaussian mixture model

ABSTRACT

The segmentation of cDNA microarray spots is essential in analyzing the intensities of microarray images for biological and medical investigation. In this work, nonparametric methods using kernel density estimation are applied to segment two-channel cDNA microarray images. This approach groups pixels into both a foreground and a background. The segmentation performance of this model is tested and evaluated with reference to 16 microarray data. In particular, spike genes with various contents are spotted in a microarray to examine and evaluate the accuracy of the segmentation results. Duplicated design is implemented to evaluate the accuracy of the model. The results of this study demonstrate that this method can cluster pixels and estimate statistics regarding spots with high accuracy.

1. Introduction

The microarray is a high throughput technique for exploring the expression profiles for thousands of genes in the studies of genomics for biology and medicine. Although high-density oligonucleotide arrays are currently available, custom-made or spotted cDNA microarrays have also been used until now because of their favorable cost, ease of preparation and ease of analysis in the design of co-hybridization experiments [1].

Studies of the functionality of genes in this new era of post-genomics are important [2]. Analyzing the microarray images with a high degree of accuracy is essential to measure the expression profiles of genes based on the microarray. Advanced analysis for selecting significant genes, clustering, classification, and network reconstruction of gene expression profiles can proceed on a solid foundation following complete, accurate measurements [3,4].

cDNA microarray images are typically noisy. Therefore, various approaches have been presented to improve the calibration of scanning efficiencies: alignment and detection of spotting errors, de-noising of background noise from images, marking of dust, gridding, moving, hybridization and other affected factors [3,5,6]. Different methods have been proposed for segmenting cDNA microarray images. Markov random field (MRF) modeling has been proposed to segment spots in microarray images [1]. This MRF-based approach has a high computational cost and relies on the prior assumption of the class labeling of all pixels [7]. The region-growing approach relies on the selection of initial seeds that influence its performance [8]. The Gaussian mixture model (GMM) relies on the assumption of normality for the application to this segmentation problem [9]. Accordingly, this study is motivated to investigate the segmentation of cDNA microarray images using the nonparametric methods that can relax the parametric assumption of normal distribution. In particular, we will consider the nonparametric methods using kernel density estimation (KDE) with data-driven selection of bandwidth [10]. Thus, automatic segmentation can be performed for different types of pixel distributions in microarray images.

In this investigation, KDE is utilized to classify pixels in a spot into background and foreground based on their estimated density function by finding the local minimum point to be the cut-off point. Empirical studies are conducted on microarray data that involve 256 spike genes with known contents. The segmentation results obtained by the KDE and the related method are compared with those obtained using the adaptive irregular segmentation method used in the current version of GenePix Pro software 6.0 (at http://www.moleculardevices.com/pages/software/gn_genepix_pro.html, with an accompanying user manual).

Microarrays with various sources and experimental designs are needed to monitor the variations of gene expressions. Spike spots of the corresponding spike mRNAs with a range of concentrations are used to monitor the variability of fluorescence intensities and determine the consistency of hybridization among arrays. The spike spots also reveal variations of pins in an array. Duplicated

* Corresponding author. Fax: +886 3 5728745.
E-mail address: hslu@stat.nctu.edu.tw (H. Horng-Shing Lu).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.02.007

spots within each array are used to assay the hybridization process of the arrays. Swapped experiments are typically used to assay the labeling efficiency of Cy3 and Cy5 fluorescence dyes.

In this study, microarray images with (1) spike spots with various ratios of Cy5–Cy3 intensities, (2) duplicated spots in an array, and (3) the swapping of microarray experiments, are applied to evaluate the performance and accuracy of the segmentation method. The results are reported in next sections.

2. Materials and methods

2.1. Materials

Sixteen microarray images used herein are obtained by swapping Cy3 and Cy5 dyes. Every array has 32 blocks, 15488 spots with 7744 genes. Two replicated spots are designed in one array, of which the upper 16 blocks are duplicated as the lower 16 blocks. Meanwhile, eight spike genes are designed in each block to evaluate the performance and accuracy of segmentation methods. The arrays from (1,1s) to (4,4s) have eight designed spikes which had known Cy5–Cy3 ratios of {0.1, 0.1, 0.2, 0.2, 0.4, 0.4, 1.0, 1.0} located at the 22nd column and from the 3rd to 10th rows in all blocks, whereas the arrays from (5,5s) to (8,8s) have eight designed constant ratios of 0.2. A typical spot diameter on each microarray in this study is approximately 160 μm. Sixteen microarray experiments were conducted in Genomic Medicine Research Core Laboratory of Chang Gung Memorial Hospital, Taiwan. The details of the microarray experiment procedure and probe information are available on the webpage of the laboratory, http://www.cgmh.org.tw/intr/intr2/c32a0/chinese/corelab_intro/genetics/files/03OctClone_information_F.zip, http://www.cgmh.org.tw/intr/intr2/c32a0/chinese/corelab_intro/genetics/files/MIAME%20(GMRCL%20Human%207K)_ver01.zip, and in [11]. These eight pairs of swapped microarrays were used for cancer research. Some of the results have been published [12]. The image data, algorithm, and computation software are available by contacting the authors.

Fig. 1 displays the segmentation of one Cy3 spot by using GenePix, ScanAlyze (http://rana.lbl.gov/EisenSoftware.htm), and our presented methods. Fig. 1 shows the results of segmentation on one spike gene with a known Cy5–Cy3 ratio of 1.0–1.0 in GenePix 6.0 for spot images of Cy3 and Cy5 dyes using three different adaptive segmentation methods: irregular, circular, and rectangular. The estimated spot feature using the adaptive irregular method is the closest to the target ratio. However, the segmentation region using an irregular method may be inaccurate, leading to an over- or under-estimate of the statistics on spot intensities. Fig. 2 plots the estimated kernel density curves from spot images of Cy3 and Cy5 dyes using the R 2.4.0 software [10, http://finzi.psych.upenn.edu/R/library/stats/html/density.html and http://www.r-project.org/]. These estimated densities typically have two distributions in the foreground and background regions.

2.2. Kernel density estimation (KDE)

The KDE with automatic bandwidth selection [10] is used to estimate the density function of pixel intensities for each spot. Gaussian kernel functions and 128 grid points are used for the KDE for each spot as Eq. (1).

$$\hat{f}(y_j) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}h}\exp\left(-0.5\cdot\left(\frac{y_j - x_i}{h}\right)^2\right), \quad (1)$$

where x_i is the ith sample in a spot, y_j is the jth grid point, h is a bandwidth used in the Gaussian kernel to estimate a spot probability density function (pdf), n is the sample size of pixels in a spot, and $j = 1,2,\ldots,128$. The details are reported in the following algorithm.

Algorithm 1 (*Segmenting one spot by the KDE*).

Step 1: Input data $X = \{x_1, x_1, \ldots, x_n\}$.

Step 2: Find 128 grid points that are equally spaced as Eq. (2).

$$y_j = \mathrm{Min}(X) + j\cdot(\mathrm{Max}(X) - \mathrm{Min}(X))/m, \quad \text{for } j = 1,2,\ldots,m, \text{ and } m = 128. \quad (2)$$

Step 3: Calculate the data-driven bandwidth for KDE as Eq. (3).

$$h = 0.9\cdot\mathrm{Min}\left\{\mathrm{Std}, \frac{\mathrm{IQR}}{1.34}\right\}\cdot n^{-1/5}, \quad (3)$$

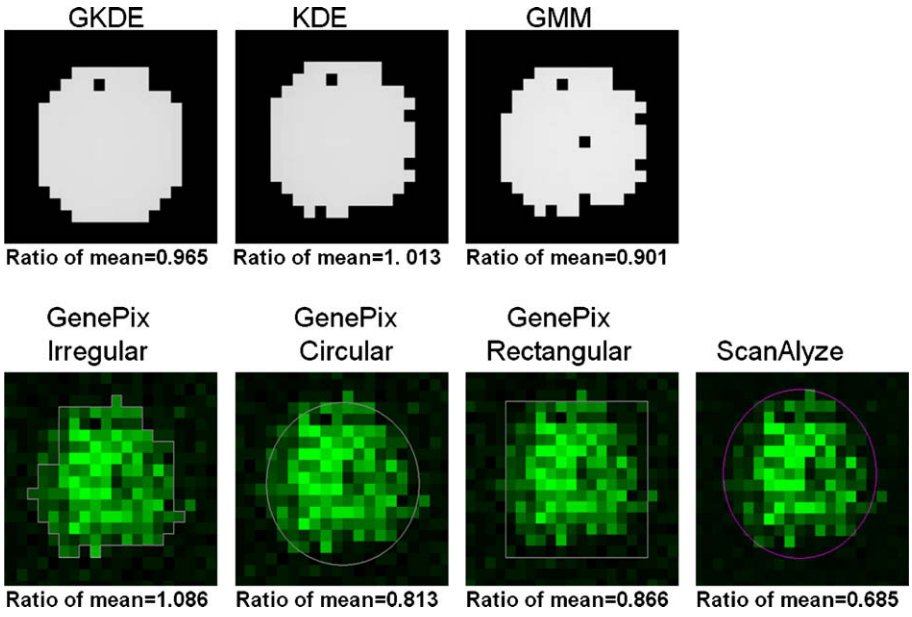

Fig. 1. The results of segmentation by using GenePix, ScanAlyze, and our presented methods on one Cy3 spot which had a known spike ratio of 1.0. Top row shows segmentation results and estimated features by using GKDE, KDE, and GMM. Bottom row shows segmentation results and estimated features by using the methods in GenePix and ScanAlyze. The irregular method in GenePix was close to the known ratio.

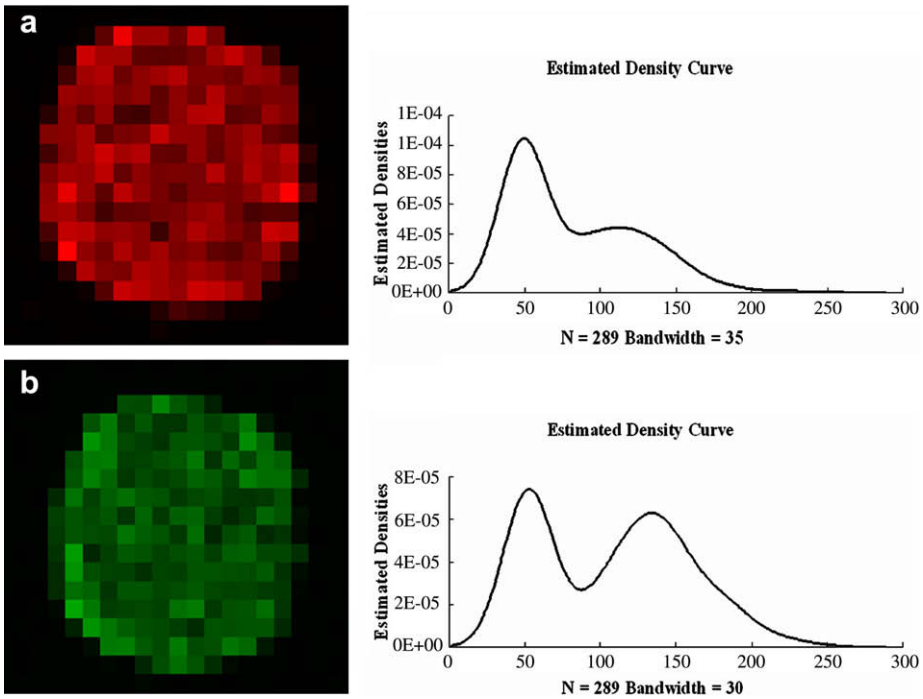

Fig. 2. Two estimated density curves for spot of Cy5 (a) and Cy3 (b) dyes. Both Cy3 and Cy5 images have two intensity distributions for background and foreground pixels. The local minimum is used to be the cut-off point for segmentation the spot.

where Std is the standard deviation of X and IQR is the interquartile range of X [13].

Step 4: Calculate the KDE using Eq. (1).

Step 5: Find a cut-off point that is the first local minimum of the KDE at y_j^* and let $CP = y_j^*$.

Step 6: Segment the pixel x_i into foreground if $x_i > \text{CP}$, else into background.

2.3. Gaussian mixture model (GMM)

The GMM assumes that the distribution of foreground intensities is a Gaussian distribution $f_1(\mu_1, \sigma_1^2)$ with mean μ_1 and variance σ_1^2; while the distribution of background intensities is another Gaussian distribution $f_2(\mu_2, \sigma_2^2)$ with mean μ_2 and variance σ_2^2. Hence, the distribution of the intensity at every pixel x_j in a spot can be modeled as a mixture of two Gaussian distributions as Eq. (4).

$$f(x_i;\phi) = \pi_1 f_1(x_i;\mu_1,\sigma_1^2) + \pi_2 f_2(x_i;\mu_2,\sigma_2^2), \quad i = 1,\ldots,n, \tag{4}$$

where $f_m(x_i;\mu_m,\sigma_m^2) = \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left(\frac{-(x_i-\mu_m)^2}{2\sigma_m^2}\right)$, $m = 1,2$, $\phi = \{\pi_m, \mu_m, \sigma_m^2, m = 1,2\}$ and π_m is the mixing (or prior) probability for the foreground and the background constrained by $0 \leqslant \pi_m \leqslant 1$ and $\pi_1 + \pi_2 = 1$. The foreground intensities typically include those of the signals and noise. Therefore, the mean foreground intensity usually exceeds the mean background intensity. Accordingly, the condition $\mu_1 \geqslant \mu_2$ is considered in this study, as it is also commonly used in the literature [14]. The log-likelihood of the observed data in the model of two mixtures is Eq. (5).

$$\log(L(\phi|x)) = \sum_{i=1}^{n} \log\left(\sum_{m=1}^{2} \pi_m f_m(x_i;\mu_m,\sigma_m^2)\right). \tag{5}$$

To estimate above parameters, the EM algorithm can be applied [14]. The segmentation algorithm of one spot using the GMM is listed below.

Algorithm 2 (*Segmenting one spot by the GMM*).

Step 1: Input initial parameters: $k = 0$ and $\phi^{(k)} = \{\pi_m^{(k)}, \mu_m^{(k)}, \sigma_m^{2(k)}, m = 1,2,\ldots\}$. In this study, the initial parameters are set as follows. Initial μ_1 and μ_2 are set to the first and third quartiles of pixel intensities in one spot. Initial σ_1 and σ_2 are the standard deviations of the pixel intensities below the first quartile and above the third quartile, respectively. Initial π_1 and π_2 values are set to 0.5.

Step 2: Calculate $\tau_{im}^{(k)} = \frac{\pi_m^{(k)} f_m(x_i;\mu_m^{(k)},\sigma_m^{2(k)})}{\sum_{l=1}^{2} \pi_l^{(k)} f_l(x_i;\mu_l^{(k)},\sigma_l^{2(k)})}$.

Step 3: Calculate new estimates of

$$\phi^{(k+1)} = \{\pi_m^{(k+1)}, \mu_m^{(k+1)}, \sigma_m^{2(k+1)}, m = 1,2,\ldots\} = \left\{\frac{1}{n}\sum_{i=1}^{n}\tau_{im}^{(k)}, \frac{\sum_{i=1}^{n}\tau_{im}^{(k)} x_i}{\sum_{i=1}^{n}\tau_{im}^{(k)}}, \frac{\sum_{i=1}^{n}\tau_{im}^{(k)}(x_i-\mu_m^{(k+1)})^2}{\sum_{i=1}^{n}\tau_{im}^{(k)}}, m = 1,2\right\}.$$

Step 4: If $\log(L(\phi^{(k+1)}|\,x)) - \log(L(\phi^{(k)}|x)) < \text{tol}$ and the tolerance parameter of tol is set to 10^{-2}, then the iteration is terminated. Otherwise, $k \leftarrow k+1$, $\phi^{(k)} \leftarrow \phi^{(k+1)} = \{\pi_m^{(k+1)}, \mu_m^{(k+1)}, \sigma_m^{2(k+1)}, m = 1,2\}$, and the iteration proceeds to Step 2.

Step 5: Segment the pixel x_i into foreground or background according to the maximum of posterior probabilities with the final values of the parameters, $\tau_{im}^{(k+1)} = \frac{\pi_m^{(k+1)} f_m(x_i;\mu_m^{(k+1)},\sigma_m^{2(k+1)})}{\sum_{l=1}^{2} \pi_l^{(k+1)} f_l(x_i;\mu_l^{(k+1)},\sigma_l^{2(k+1)})}$.

2.4. GMM incorporated with KDE (GKDE)

We can combine the methods of GMM and KDE, which will be abbreviated as GKDE. The GMM method can provide the initial segmentation and the KDE method can further improve the segmentation by relaxing the assumption of normality in the GMM method.

Once the foreground and background are found using GMM, the KDE can be applied to find their estimated densities. Then, a cut-off point for segmenting a spot into two clusters is determinate by the equality of two estimated densities. The details are reported below.

Algorithm 3 (*Segmenting one spot by the GKDE*).

Step 1: Segment a spot initially using the GMM in Algorithm 2.
Step 2: Estimate the kernel densities for foreground ($\hat{f}_f$) and background ($\hat{f}_g$) similar to Eqs. (1)–(3).
Step 3: Find a cut-off point CP that is close to the equality of $\hat{f}_f$ and $\hat{f}_g$.
Step 4: Segment a spot as follows.

$$x_i \in \begin{cases} \text{foreground}, & \text{if } x_i \geqslant \text{CP}; \\ \text{background}, & \text{elsewhere}. \end{cases}$$

2.5. Microarray studies

Spike genes (or spots) with known contents on microarrays are used in the empirical studies. The target ratios of spike genes thus represent the gold standard for evaluating the accuracy of segmentation methods investigated in this study. The sum of squared relative error (SSRE) and the sum of squared error (SSE) are used to evaluate accuracy according to Eqs. (6) and (7).

$$\text{SSRE} = \sum_{j=1}^{M}\sum_{b=1}^{B}\left\{\frac{\hat{T}_{j,b} - T_j}{T_j}\right\}^2, \tag{6}$$

$$\text{SSE} = \sum_{j=1}^{M}\sum_{b=1}^{B}(\hat{T}_{j,b} - T_j)^2, \tag{7}$$

where $\hat{T}_{j,b}$ is the feature estimated from the ratio of means between Cy3 and Cy5 arrays for the jth spike gene in the bth block, and T_j is target ratio of the jth spike gene. The number of blocks is $B = 32$ and the number of spike genes is $M = 8$. The smallness of SSRE and SSE indicate closeness to the target ratio. For those two types of spike genes, four sets of microarrays are produced and each set of microarrays consists of one pair of two dye swapped microarrays. Therefore, each type of spike gene is associated with eight microarrays, of which a total of 16 microarrays are tested herein.

The concordance correlation coefficient [15] of two random variables Y_1 and Y_2 is shown as Eq. (8).

$$\rho_c = \frac{2\text{Cov}(Y_1, Y_2)}{\text{Var}(Y_1) + \text{Var}(Y_2) + (E(Y_1) - E(Y_2))^2}. \tag{8}$$

It is also used in this study to measure the accuracy and precision between the expression pattern of every gene and that of its duplicated spot using the log ratios of means in Cy5–Cy3 dyes from one microarray image. The concordance correlation coefficient can be used to determine the degree of similarity, agreement and reproductively in expression between duplicated spots of all genes in one microarray, which is expected to be close to 1.

The concordance correlation coefficients of the swapped microarrays are also considered to evaluate the performance with reference to selected features with high log ratios of means in Cy5–Cy3 dyes. The dyes of Cy3 and Cy5 in the swapped arrays are exchanged. Accordingly, the negative concordance correlation coefficient is obtained from the features of the swapped arrays and is expected to be close to −1.

3. Results

3.1. Microarrays with spike genes

There are 256 spike genes on any array with different target ratios between Cy5 and Cy3. Those spike genes are used to detect the performance of GKDE, KDE, GMM, ScanAlyze, and GenePix 6. Tables 1 and 2 show that all of the SSEs and the SSREs obtained from KDE are smaller than those obtained by the irregular segmentation method in GenePix 6.0, according to the test based on 16 cDNA microarray images. The relative improvements of these two segmentation methods are defined as the percentages of the evaluation values in (Min{GenePix, ScanAlyze} − Methods)/Min{GenePix, ScanAlyze}. Since the first eight arrays are produced according to varying target ratios, the relative improvements measured by SSRE and SSE are different according to Eqs. (6) and (7). The last eight arrays are produced according to a constant ratio, and the relative improvements measured by SSRE and SSE are the same according to Eqs. (6) and (7). Tables 1 and 2 show that the average relative improvements of GKDE, KDE, and GMM associated with the compared segmentation methods in GenePix 6.0 and ScanAlyze for SSRE and SSE are at levels of (50.55%, 45.36%), (50.16%, 48.59%), and (49.98%, 45.23%). These results reveal that the features estimated by GKDE, KDE, and GMM are closer to the designed target ratios for the spike genes (spots) than those obtained by the

Table 1
The comparisons of SSEs obtained by GMM, GKDE, KDE, ScanAlyze, and GenePix 6.0 for spike genes are listed

Array	Sum of square of errors							Relative performance		
	GKDE	KDE	GMM	ScanAlyze	GenePix irregular	GenePix circular	GenePix rectangular	GKDE	KDE	GMM
1	2.868	2.781	2.869	25.696	22.523	27.509	29.040	87.266	87.654	87.263
2	3.024	3.019	3.027	21.155	12.082	17.091	18.741	74.971	75.013	74.950
3	5.432	5.408	5.439	42.612	33.806	39.260	41.237	83.932	84.004	83.910
4	9.391	9.290	9.700	9.944	10.446	9.915	10.198	5.286	6.301	2.165
5	0.412	0.316	0.416	0.610	0.804	0.643	0.789	32.544	48.161	31.765
6	0.305	0.309	0.306	2.136	2.304	2.203	2.340	85.710	85.525	85.678
7	2.436	2.375	2.437	3.605	4.621	176.431	4.581	32.418	34.106	32.405
8	4.439	4.076	4.440	6.549	8.792	7.877	8.293	32.220	37.761	32.213
1s	4.414	3.464	4.398	17.577	13.413	16.293	16.882	67.094	74.172	67.211
2s	2.062	2.675	2.265	3.261	3.201	2.966	3.401	30.489	9.811	23.614
3s	12.308	14.816	12.309	44.033	30.953	40.024	39.269	60.236	52.135	60.233
4s	88.786	99.959	86.532	151.779	106.721	132.938	131.203	16.805	6.335	18.917
5s	0.488	0.484	0.489	0.521	0.929	0.582	0.690	6.295	7.049	6.162
6s	0.270	0.262	0.271	3.794	4.295	4.078	4.192	92.879	93.093	92.845
7s	0.509	0.497	0.510	1.195	2.142	1.765	1.803	57.400	58.419	57.371
8s	0.399	0.401	0.400	0.703	1.020	0.859	0.989	43.201	42.942	43.105
Average relative performance								50.547	50.155	49.988

Array 1s is obtained by swapping the dyes of Array 1. Relative improvement is specified by (Min{GenePix, ScanAlyze} − Methods)/Min{GenePix, ScanAlyze} as a percentage.

Table 2
The comparisons of SSREs obtained by GMM, GKDE, KDE, ScanAlyze, and GenePix 6.0 for spike genes are listed

Array	Sum of square of relative errors							Relative performance		
	GKDE	KDE	GMM	ScanAlyze	GenePix irregular	GenePix circular	GenePix rectangular	GKDE	KDE	GMM
1	85.482	82.495	85.482	243.383	301.408	320.264	258.886	64.878	66.105	64.878
2	55.009	45.899	55.025	117.817	152.598	123.584	128.127	53.310	61.042	53.296
3	80.421	77.148	80.421	286.480	317.267	303.845	317.147	71.928	73.070	71.928
4	29.861	28.021	30.170	36.042	31.277	34.409	35.664	4.528	10.409	3.539
5	10.401	7.908	10.410	15.256	20.094	16.070	19.737	31.823	48.161	31.765
6	7.605	7.729	7.647	53.392	57.603	55.068	58.491	85.757	85.525	85.678
7	60.911	59.383	60.916	90.118	115.513	115.779	114.534	32.411	34.106	32.405
8	110.991	101.908	110.992	163.735	219.798	196.928	207.335	32.213	37.761	32.213
1s	33.005	31.740	32.980	130.681	147.429	132.150	132.289	74.744	75.712	74.763
2s	26.900	27.211	26.905	32.285	31.157	33.630	37.984	13.662	12.664	13.648
3s	149.074	130.196	149.739	244.790	286.494	261.726	272.590	39.101	46.813	38.830
4s	675.010	648.212	674.388	769.750	826.411	761.916	767.239	11.406	14.923	11.488
5s	12.202	12.106	12.222	16.525	23.215	14.541	17.244	16.086	16.745	15.950
6s	6.781	6.550	6.786	94.839	107.370	101.951	104.794	92.850	93.093	92.845
7s	12.705	12.425	12.739	29.883	53.558	44.113	45.079	57.484	58.419	57.371
8s	9.910	10.025	9.997	17.570	25.496	21.485	24.715	43.599	42.942	43.105
Average relative performance								45.283	48.593	45.231

Array 1s is obtained by swapping the dyes of Array 1. Relative improvement is specified by (Min{GenePix,ScanAlyze} − Methods)/Min{GenePix,ScanAlyze} as a percentage.

segmentation methods in GenePix 6 and ScanAlyze. Among these methods, the segmentation results by GKDE have the greatest improvement.

3.2. Duplicated spots and dye swapped arrays

The numbers of spots (excluding spike genes and bad spots) in each array are used to evaluate the accuracy and the performance of presented methods.. The bad spots are defined by having negative values for foreground mean minus background mean, as provided from GKDE, KDE, GMM, ScanAlyze, and GenePix 6. Those genes are used to investigate performance of GKDE, KDE, GMM, ScanAlyze, and GenePix 6. Fig. 3 shows agreement scatter-plots of two replicate gene expression and swapped arrays produced by GKDE, KDE, GMM, ScanAlyze, and GenePix 6, respectively. The scatter-plot of KDE has less variation than those by GKDE, GMM, ScanAlyze, and GenePix 6. Meanwhile, the scatter-plots of GKDE and GMM are similar, which have less variation than those by GenePix 6. Fig. 4 shows the concordance correlation coefficients, Pearson's correlations and standard deviations between replicates

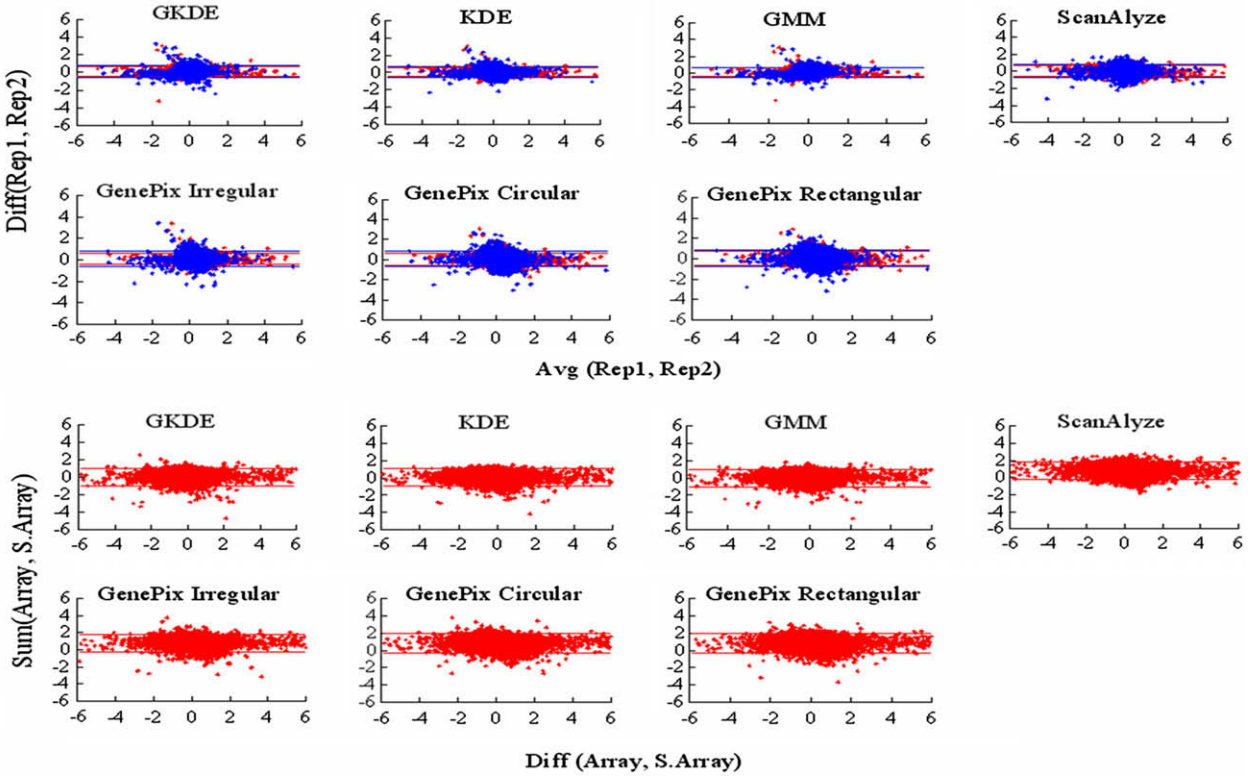

Fig. 3. Top row shows seven methods to evaluate duplicated spots for 3rd (red) and swapped 3rd (blue) arrays. The *x*-axis and *y*-axis represent the average and the difference between duplicated spots. Bottom row shows seven methods to evaluate swapped arrays (3rd,3rds). The *x*-axis and *y*-axis represent the summation and the difference between swapped arrays. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

gene expression of the 16 arrays and eight swapped arrays. The KDE method typically produced higher correlation and lower standard deviation than those by other methods tested on 16 arrays with duplicated genes. For swapped arrays, the KDE method also provided lower standard deviation and higher correlation between the eight tested and swapped arrays. Meanwhile, GKDE and GMM both have higher correlations and lower standard deviations.

4. Conclusion and discussion

The effect of expression profiling on prognostic and predictive testing for cancer has been recently discussed [16]. However, the low reproducibility of microarray experiments [17,18] impedes the scheduler from using a microarray to prognose and predict the outcome of cancer. The proposed GKDE, KDE, and GMM methods can improve the reproducibility in duplicated spots, in swapped arrays and in the spike gene spots. This will be useful for the advanced utilization of microarrays in biology and medicine.

In this study, the GKDE, KDE, and GMM were applied to segment cDNA microarray images, and performance evaluations were conducted. First, spike genes with known contents were designed on microarrays, and the criteria of SSRE and SSE measured accuracy and performance. The GKDE, KDE, and GMM methods more accurately estimated the features of spots than the segmentation methods in GenePix 6 and ScanAlyze. Secondly, duplicated spots are utilized to examine expression variation on a microarray image. The GKDE, KDE, and GMM methods also have better average relative performances, as measured by the concordance correlation coefficients, Pearson's correlation coefficients and standard deviations of expression values of duplicated spots. Finally, swapped microarray experiments are conducted to study the variation among dyes. The correlation coefficients measure the linear relationship for the selected spots with significantly differentially expressed levels. Again, the GKDE, KDE, and GMM methods are more accurate when tested on eight pairs of swapped cDNA microarray images.

Sixteen microarray images were used to determine the accuracy and performance, in comparison with the segmentation method in GenePix 6 and ScanAlyze. The ratio of means is used to estimate features in segmented spots. Other statistics could be studied. Improved methods for segmenting images can be studied further [19–21].

The GKDE, KDE, and GMM programs were run in less than 1000 seconds to test one cDNA microarray image on a personal computer with Intel CPU 2.6 GHz and 2 GB RAM. The parametric meth-

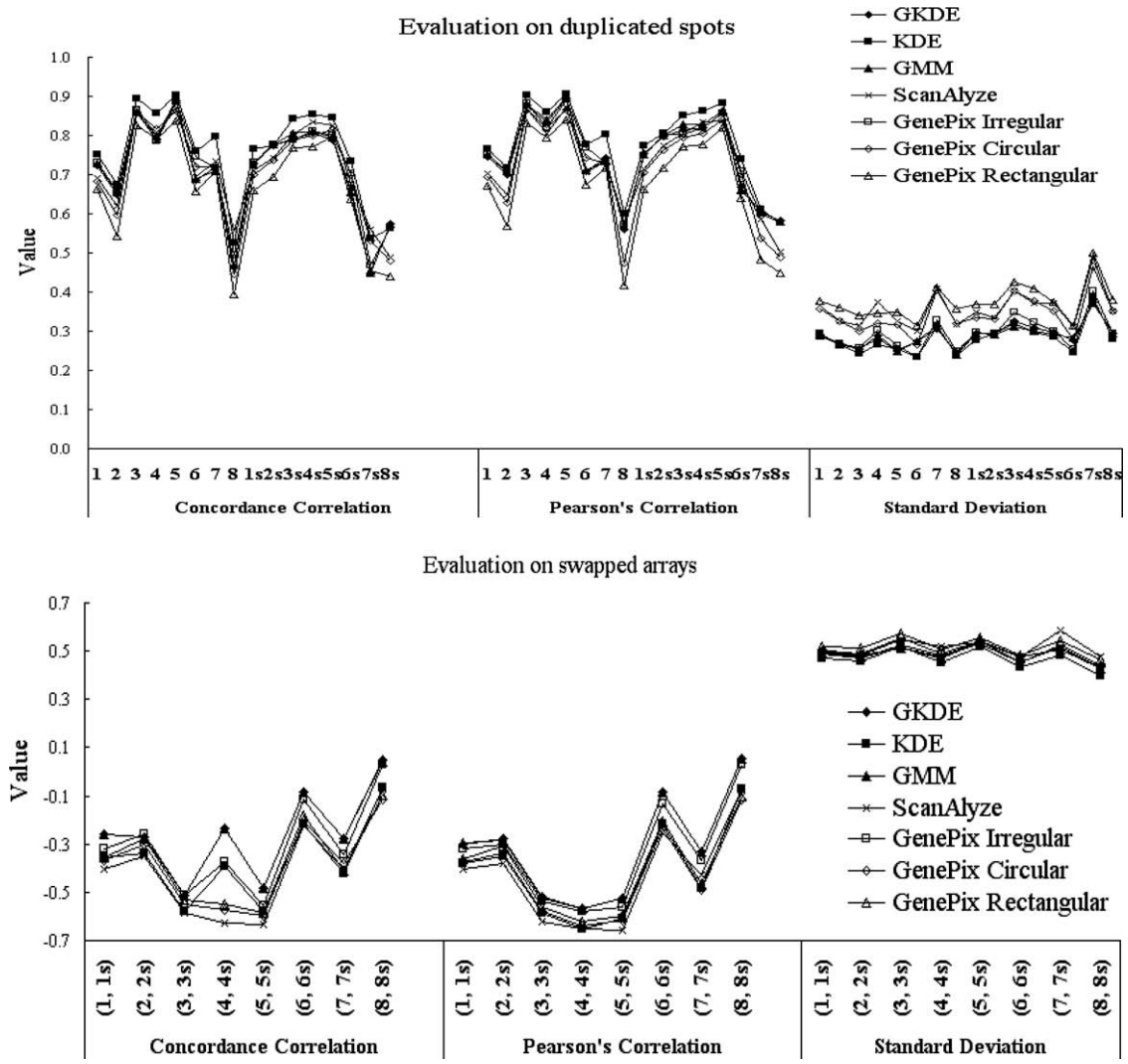

Fig. 4. Top and down figures are concordance correlations, Pearson's correlations and standard deviations between duplicated spots of 16 arrays and between swapped arrays of eight arrays using GKDE, KDE, GMM, ScanAlyze, and GenePix 6.

od of GMM has computational efficiency and effective segmentation performance when the normality assumption holds. Preserving the properties of computational efficiency and effective segmentation performance, the nonparametric methods of GKDE and KDE can further relax the assumption of normality for microarray images that can have pixel distributions that are not normal. The main advantages of both GMM and KDE are incorporated in GKDE. The GMM approach is highly dependent on the initial values of parameters and on the stopping rules of convergence. The GKDE approach can resolve the selection problem of initial values from the KDE approach, but it is still dependent on stopping criteria of convergence. The study of convergence and combination with GMM could be future work.

Acknowledgment

This study was supported in part by the Grant NSC-96-3112-B-001-017 from The National Research Program for Genomic Medicine, National Science Council, Taiwan, ROC.

References

[1] Demirkaya O, Asyali MH, Shoukri MM. Segmentation of cDNA microarray spots using Markov random field modeling. Bioinformatics 2005;21(13):2994–3000.
[2] Yang YH, Speed TP. Design issues for cDNA microarray experiments. Nat Rev Genet 2002;3:579–88.
[3] Yang YH, Buckley MJ, Dudoit S, Speed TP. Comparison of methods for image analysis on cDNA microarray data. J Comput Graph Stat 2002;11:108–36.
[4] Li LM, Lu HHS. Explore biological pathways from noisy array data by directed acyclic Boolean networks. J Comput Biol 2005;12(2):170–85.
[5] Chen Y, Dougherty ER, Bittner ML. Ratio based decisions and the quantitative analysis of cDNA microarray images. J Biomed Opt 1997;2:364–74.
[6] Ho J, Hwang WL, Lu HHS, Lee DT. Gridding spot centers of smoothly distorted microarray images. IEEE Trans Image Process 2006;15(2):342–54.
[7] Yang F, Jiang T. Pixon-based image segmentation with Markov random fields. IEEE Trans Image Process 2003;12(12):1552–9.
[8] Wang X, Wang H. Markov random field modeled range image segmentation. Pattern Recognit Lett 2005;25:367–75.
[9] Blekas K, Galatsanos NP, Likas A, Lagaris IE. Mixture model analysis of DNA microarray images. IEEE Trans Med Imaging 2005;24(7):901–9.
[10] Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. J Roy Stat Soc B 1991:683–90.
[11] Wang TH, Lee YS, Chen ES, Kong WH, Chen LK, Hsueh DW, et al. Establishment of cDNA microarray analysis at the Genomic Medicine Research Core Laboratory (GMRCL) of Chang Gung Memorial Hospital. Chang Gung Med J 2004;27(4):243–60.
[12] Chao A, Wang TH, Lee YS, Hsueh S, Chao AS, Chang TC, et al. Molecular characterization of adenocarcinoma and squamous carcinoma of the uterine cervix using microarray analysis of gene expression. Int J Cancer 2006;119(1):91–8.
[13] Engel J, Herrmann E, Gasser T. An iterative bandwidth selector for kernel estimation of densities and their derivatives. J Nonparametr Stat 1994;4:21–34.
[14] McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 2000;18(3):413–22.
[15] Lin LIK. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45(1):255–8.
[16] Reis-Filho JS, Westbury C, Pierga JY. The impact of expression profiling on prognostic and predictive testing in breast cancer. J Clin Pathol 2006;59:225–31.
[17] Sherlock G. Of fish and chips. Nat Methods 2005;2(5):329–30.
[18] Irizarry RA, Warren D, Spencer F, et al. Multiple-laboratory comparison of microarray platforms. Nat Methods 2005;2(5):345–50.
[19] Chen CM, Lu HHS, Lin YC. An early vision based snake model for ultrasound image segmentation. Ultrasound Med Biol 2000;26(2):273–85.
[20] Chen CM, Lu HHS, Huang YS. Cell-based dual snake model: a new approach to extracting highly winding boundaries in the ultrasound images. Ultrasound Med Biol 2002;28(8):1061–73.
[21] Wu HM, Lu HHS. Supervised motion segmentation by spatial-frequential analysis and dynamic sliced inverse regression. Stat Sin 2004;14:413–30.