Journal of Signal Processing 50, 305–320, 2008

DOI: 10.1007/s11265-007-0123-0

A Comparison of Fuzzy Clustering Approaches for Quantification of Microarray Gene Expression

YU-PING WANG AND MAHESWAR GUNAMPALLY
School of Computing and Engineering, University of Missouri, Kansas City, MO 64110, USA

JIE CHEN
Department of Mathematics and Statistics, Kansas City, MO 64110, USA

DOUGLAS BITTEL AND MERLIN G. BUTLER
Children's Mercy Hospital and Clinics, School of Medicine, University of Missouri, Kansas City, MO 64108, USA

WEI-WEN CAI
Department of Human Molecular Genetics, Baylor College of Medicine, Houston, TX 77005, USA

Received: 1 June 2007; Revised: 19 February 2007; Accepted: 13 June 2007

Abstract. Despite the widespread application of microarray imaging for biomedical imaging research, barriers still exist regarding its reliability for clinical use. A critical major problem lies in accurate spot segmentation and the quantification of gene expression level (mRNA) from the microarray images. A variety of commercial and research freeware packages are available, but most cannot handle array spots with complex shapes such as donuts and scratches. Clustering approaches such as *k*-means and mixture models were introduced to overcome this difficulty, which use the hard labeling of each pixel. In this paper, we apply fuzzy clustering approaches for spot segmentation, which provides soft labeling of the pixel. We compare several fuzzy clustering approaches for microarray analysis and provide a comprehensive study of these approaches for spot segmentation. We show that possiblistic *c*-means clustering (PCM) provides the best performance in terms of stability criterion when testing on both a variety of simulated and real microarray images. In addition, we compared three statistical criteria in measuring gene expression levels and show that a new asymptotically unbiased statistic is able to quantify the gene expression level more accurately.

Keywords: microarray, segmentation, fuzzy clustering, image segmentation, microarray gridding

1. Introduction

As an important and powerful technique recently developed in functional genomics for large-scale gene expression analysis, microarray imaging has been widely used in drug testing and for disease diagnosis [1, 2]. Typical microarray experiments utilize two-color hybridization in which the Cy5 (red) labeled test sample is hybridized against Cy3 (green) labeled control to the same arrays. The ratio of expression levels between the test and control samples can then be derived from the intensity values of the two channel images, which are captured at different wavelengths using a scanner. Figure 1 shows a pseudo-color array spot image (c) that is composed of the two Cy3 (a) and Cy5 images

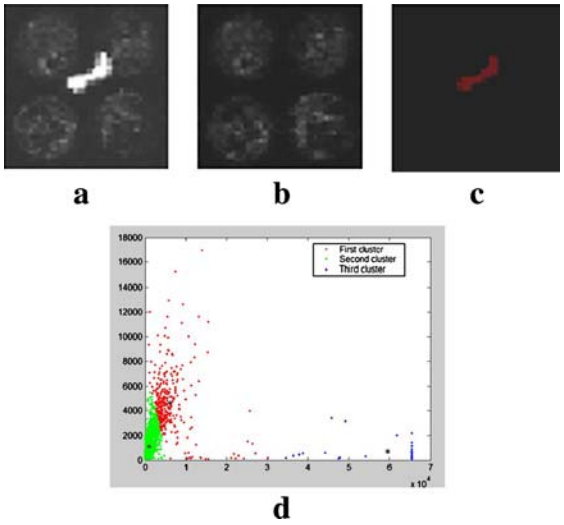

Figure 1. A two-channel microarray spot with a large artifacts. In the scatter plot, three clusters correspond to foreground, background and artifacts. **a** Cy3, **b** Cy5, **c** composite image, and **d** the scatter plot of Cy3 versus Cy5.

(b). The microarray technique measures the mRNA abundance on the test sample with respect to the reference sample by computing the ratio of expression levels between the two channel images. The measurements of RNA levels in cells provide rich information on the overall cell function and the function of individual genes, which can be used for medical diagnosis and treatment and for drug design.

The analysis of microarray experiments in general involves two steps. The first step is to extract quantitative information from array images. Based on the gene expression, the second step is to study the function of the individual genes as well as their relationship. There are many quality issues that impinge the quality of the experiment [3]. A microarray spot is produced by depositing equal amounts of liquid onto the slide [18]. Ideally the spots should be circular and have an equal size. However, deviations do occur due to malfunction of the printers, precipitation on the slide, impurities and debris in printing solution. Deviations can also be caused by an inferior quality of printing needles. Other sources of variations include misalignment of an array of spots, low amount of nucleic acids in the spot, uneven or incomplete hybridization, signal bleaching and low sensitivity of the scanner [4]. Therefore, image processing is crucial for accurately extracting and quantitatively analyzing the gene expression levels. An erroneous measurement can have a drastic effect on the subsequent analysis and interpretation of cellular function.

The rest of the paper is organized as follows. In Section 2, we first introduce the background of microarray image processing with a focus on spot segmentation and then review the existing clustering approaches for microarray analysis. Section 3 presents three fuzzy clustering based approaches for spot segmentation and the validation of these clustering models. In Section 4, we present experiments on both simulated with known "ground-truth" and real microarray images and compare them with existing software packages. The paper concludes with a discussion on the advantage and comparison of three fuzzy clustering approaches and future studies.

2. Background on Microarray Spot Segmentation

2.1. *Spot Segmentation*

After some necessary preprocessing steps such as the registration of two channel arrays and spot finding, a critical problem in microarray image processing is spot segmentation [14–16]. Spots segmentation is an approach to determine which pixels in the target region are due to the actual spot signal, and which pixels are due to background. Ideally, every spot on a microarray should have a circular shape. However, ideal spots seldom happen. Spots have a variable size and contours. They have smeared and incorrectly segmented areas caused by dirt on the slide or from slide treatment. A number of algorithms have been developed, such as the fixed circle segmentation [9, 10], intensity based segmentation [6, 23], Mann–Whitney segmentation [5] and seeded region growing [8] to address these problems. These methods can be classified into spatial and distributional methods [3]. The fixed and adaptive circle methods are implemented in most software packages such as ScanAlyze [10], GenePix [9] and QuantArray (GSI Lumonics, 1999). However, these methods do not account for the occurrence of irregular shapes. Adaptive shape segmentation methods such as watershed and seeded region growing are implemented in the Spot software [11] using the statistical language R. The Mann–Whitney test [5] imple-

mented in QuantArray is a non-parametric method that takes the intensity information into consideration. This method can overcome many difficulties but relies greatly on selecting a good background region. The hybrid approach integrates both the spatial and intensity information [3], as proposed by ImaGene (BioDiscovery Inc., 1997). This method offers a certain advantage but is still non-adaptive from gene spot to gene spot. In general, these available methods work well with good, clean, high contrast images, but have problems in dealing with noise contamination and irregular spot shape.

In this paper we introduce an algorithm that can take advantage of the following information: (1) the integration of shape information with the intensity information; and (2) the two spectral channel information. Clustering based approaches can fully make use of this information. In the scatter plot of R versus G (see example of Fig. 1), it can be seen that the target pixels and background pixels form two separate clusters. In addition, noises or artifacts in the image can form another cluster. Clustering method is shape independent and can be used for processing spot images that have many variations as discussed in the following subsection.

2.2. *Clustering Based Approaches*

Several clustering approaches [19–21] were introduced for microarray spot segmentation. In [19] two clustering approaches derived from *k*-means and Partitioning Around Medoids (PAM) can process spots with variable size and contour. In these two methods, the model selection capabilities were not provided. Specifically, the number of the clusters were not identified and usually assumed to be two. *K*-means based clustering is more favored over PAM because of less computational time [19]. In [20] a robust Guassian mixture model (GMM) was used to model the intensity of the array data, and the number of clusters was determined by the Bayesian information criterion (BIC). In this work, the sum of the intensities of the Cy3 and Cy5 were used, which therefore discarded important two-channel information. The work of Blekas et al. in [21] also employed the bi-variant Gaussian mixture model and took advantage of two-channel data, with the number of clusters or the identification of artifacts determined through a cross-validated likelihood evaluation. When comparing GMM with *k*-means and PAM method for certain cases, *k*-means and PAM methods tend to overestimate the background clusters, while the GMM approach provides more uniform spots [21].

From the analysis of the above approaches, we have identified the following problems. These approaches implicitly assume that the array data follows a Gaussian or Gaussian mixture distribution. More specifically, each cluster is defined by a circle or ellipsoid. Whether an unknown pixel belongs to a foreground, background and/or due to artifacts is determined by which circle or ellipsoid it falls into. The partition of the feature space (e.g., scatter plot) is through a crisp membership function, i.e., 0 or 1. The fuzzy clustering is a more sophisticated approach [7], which groups the pixels according to the degree of similarity between 0 and 1. This way of partition is more realistic in labeling the regions of foreground spots from the background as well as from possible artifacts. The fuzzy c-means (FCM) based approaches have been introduced for several microarray data analysis [26–30]. In [26, 30] FCM was used for grouping biologically relevant genes. The study in [27] compared FCM and Gaussian normal mixture model approach in classifying microarray data into reliable and unreliable populations, showing FCM is computationally more efficient. The work in [28] introduced a new fuzzy approach and compared with FCM and SOM for gene expression profile analysis. None of these fuzzy methods have been used for spot segmentation. In this paper, we present a comprehensive study of fuzzy clustering for spot segmentation. We show that a possiblistic c-means clustering (PCM) is more accurate in measuring gene spots, which has never been used before.

2.3. *Data Quantification*

After segmentation, the key information to be recorded from microarrays is the expression intensity strength of each target or clone. The common measurements computed after segmentation are the total signal intensity, mean signal intensity, median of signal intensity, mode, volume of intensity, intensity ratio, and correlation ratio as discussed in [3]. The choice of which parameter to use is based on how well each of these measurements correlates with the amount of DNA probe present at each spot location. The most commonly used method is the log intensity ratio, which is obtained from the mean, median or mode of the intensity measurement for

each channel. We will introduce two more new statistics [17] when calculating the log intensity ratio in Section 3.5. This new statistic is expected to give a more accurate estimate of gene expression level.

3. Fuzzy Clustering Approaches

3.1. Fuzzy Clustering Models

Our approach takes advantage of the two channel information. The pixel intensities of the Cy3 and Cy5 images constitute distinct features to classify a pixel in the image into foreground, background and artifacts. Specifically, the input to a clustering approach is a vector consisting two channel intensity values

$$X_j = \left(x^1, x^2\right)_j^t, \quad j = 1,\ 2, \cdots, n \tag{1}$$

where n is the number of pixels and x^1, x^2 represent the pixel intensity of Cy3 and Cy5 image respectively. This input is then fed into a classifier and assigned to a specific class. The clustering algorithm partitions the feature space into c clusters, Ω^i, $i = 1,2, ..., c$. The *K*-means or PAM clustering [19, 25] finds the cluster centroids by minimizing a dissimilarity function that measures the overall dissimilarity between the data and centroids given by

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{x_j \in \Omega^i} u_{ji} d_{ji}^2 \tag{2}$$

where $d_{ji} = d(X_j, C_i)$ is the distance between ith centroid (C_i) and jth data point; and Ω^i is the ith cluster set;. If the Manhattan distance or city block distance

$$d\left(X_j, C_i\right) = \left|x_j^1 - c_i^1\right| + \left|x_j^2 - c_i^2\right| \tag{3}$$

is used, we obtain the PAM method [19]. Euclidian distance is another metric that is mostly used and expressed as

$$d\left(X_j, C_i\right) = \sqrt{\left|x_j^1 - c_i^1\right|^2 + \left|x_j^2 - c_i^2\right|^2} \tag{4}$$

In this case, the minimization of objective function (2) leads to the *k*-means clustering. It is also called the hard *c*-means (HCM) clustering [7], where the membership element u_{ji} is 1 if the jth data point x_j belongs to class Ω^{i}, and 0 otherwise, i.e.,

$$u_{ji} = \begin{cases} 1 & \text{if } x_j \in \Omega^i, \text{for each } j \neq i, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The feature space is then partitioned by a membership matrix (U) of dimension $n \times c$ with the above element of u_{ji} and each pixel is labeled by the element u_{ji}.

The *k*-means clustering or HCM is a crisp partition of the feature space according to the binary membership function (5). The fuzzy *c*-means clustering (FCM) is an improvement over the HCM by employing a fuzzy partitioning such that a data point can belong to all classes with different membership grades between 0 and 1 [7]. The dissimilarity function used in FCM is given by

$$J(U, C_1, C_2, \ldots, C_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ji}^m d_{ji}^2 \tag{6}$$

where C_i is the centroid of cluster i, and $d_{ji} = d(X_j, C_i)$ is the Euclidian distance between ith centroid (C_i) and jth data point (X_j). u_{ji} is the membership degree with the value between 0 and 1, with the following constraint

$$0 < u_{ji} < 1, \quad \sum_{j=1}^{n} u_{ji} = 1, \quad i = 1, 2, \cdots c \tag{7}$$

The parameter m is called fuzzifier that controls the degree of fuzziness. The model defines a soft partition of feature space because each pixel is labeled by a fuzzy value or membership function of value from 0 to 1. The larger of the membership function, the more certain the class to which the pixel belongs. The non-crisp labeling offers a more realistic model in partitioning the feature space.

3.2. Possibilistic *c*-Means (PCM) Clustering

The minimization of the FCM objective function (6) might result in a null solution. The FCM can be

further improved. Krishnapuram and Keller [24] relaxed the constraints in Eq. (7) to facilitate a possibilistic interpretation of the memberships. The object function in Eq. (6) is modified as follows

$$J_m(U, C_1, C_2, \ldots, C_c) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ji}^m d_{ji}^2 + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} \left(1 - u_{ji}\right)^m \quad (8)$$

where $0 \leq \eta_i \leq 1$ is a parameter that can be adjusted. When $\eta_i = 0$, Eq. (8) reduces to be the same as Eq. (6). The last term in Eq. (8) forces the membership values to be as close to one as possible during the minimization process. The minimization according to this objective function gives the possiblistic c-means clustering (PCM) [7, 24], which allows a more flexible and accurate partition of the feature space.

3.3. Implementation of FCM and PCM

A necessary condition on the minimization of objective function of in Eq. (6) and Eq. (8) is to make the derivatives of objection functions with respect to u_{ji} to be zeros. This leads to an alternate iterative solution. For FCM clustering, the centroids and membership elements are alternatively updated as follows [7]:

$$C_i = \frac{\sum_{j=1}^{n} u_{ji}^m X_j}{\sum_{j=1}^{n} u_{ji}^m}, \quad u_{ji} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ji}}{d_{jk}}\right)^{2/(m-1)}} \quad (9)$$

For PCM, the centroids and membership elements are derived by:

$$C_i = \frac{\sum_{j=1}^{n} u_{ji}^m X_i}{\sum_{j=1}^{n} u_{ji}^m}, \quad u_{ji} = \frac{1}{1 + \left(\frac{d_{ji}^2}{\eta_i}\right)^{\frac{1}{m-1}}} \quad (10)$$

where the parameter η_i determines the relative degree of importance of the second term when compared with the first term. The value of η_i can be fixed for all iterations. When it is varied in each iteration care must be exercised as it may lead to instabilities. Dynamic variation of η_i is derived as in [24]:

$$\eta_i = K \frac{\sum_{j=1}^{n} u_{ji}^m d_{ji}^2}{\sum_{j=1}^{n} u_{ji}^m} \quad (11)$$

The implementation of the alternative numerical iteration algorithm involves the following steps:

1. Initialize the membership matrix (U) by assigning a random value to each of its element.
2. Calculate centroids (C_i) according to the first equation of formulae (9) or (10).
3. Compute dissimilarity metric (6) or (8) between the centroids and data points. If its improvement over previous iteration is below a given threshold, stop the iteration.
4. Update the membership matrix U according to the second equation defined in Eqs. (9) or (10). Go to step 2.

3.4. Validation of Clustering

Clustering validation is an important but difficult problem associated with the clustering algorithm. For example, what is the optimal number of clusters c? Whether the soft partition is more advantageous over the hard partition? One often used approach is computing the partition coefficient V_{PC} given below [7]:

$$V_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ji}^2 = \frac{1}{n} trace(UU^t) \quad (12)$$

which measures the amount of overlap between clusters. In this definition, V_{PC} is inversely proportional to the overall average overlap between pairs of fuzzy subsets. If $V = 1$, there is no membership sharing between any pairs of fuzzy clusters. A valid number of clusters corresponds to the solution of $\max_c\{\max_{\Omega^c}\{V\}\}$. In our work, we assume two and/or three clusters ($c = 2$ or 3) because the array image contains background, foreground spot region and/or artifacts. In the current

Table 1. Mean and standard deviation of NMSE using different clustering approaches for different spots.

Mean/variance	HCM	FCM	PCM
Circles	0.2043	0.2041	0.2041
	0.1616	0.1623	0.1613
Donuts	0.5527	0.5509	0.5509
	0.5407	0.5377	0.5377
Scratches	0.2326	0.2327	0.2327
	0.1889	0.1883	0.1883

work, we adopt another more reliable validity measure [22], S, defined by

$$S = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ji}^2 \left\| C_i - X_j \right\|^2}{n \min_{i,j} \left\| C_i - C_j \right\|^2} \tag{13}$$

which is independent of the clustering algorithm and easy to calculate. A smaller S indicates a partition in which all the clusters are overall compact and separate to each other. Therefore, $\min_c\{\min_\Omega\{S\}\}$ yields the most valid clustering of array data X. Our experiment in Section 4.3 has confirmed this conclusion.

3.5. *Measurement of Gene Expression Level*

The result of segmentation indicates the region of foreground spot and the background in the Cy3 (green) and Cy5 (red) images. The intensity with background corrections from these two channels can be obtained by subtracting the background intensities as $R_{fg} - R_{bg}$ and $G_{fg} - G_{bg}$ respectively. Let X and Y are the random variable representing the gene expression in Cy3 and Cy5, the relative gene expression level or measurement of mRNA is then calculated as the ratio or the log ratio between two random variables X and Y. More specifically, the ratio is

$$\frac{X}{Y} \quad or \quad R = \log\frac{X}{Y} \tag{14}$$

The calculation of the above ratio can be realized using the following three statistics.

3.5.1. *The First Ratio Estimate r_1.*

$$r_1 = \left(\frac{\overline{x}}{\overline{y}}\right) \tag{15}$$

where $\overline{x}$ and $\overline{y}$ are the estimates of μ_x and μ_y respectively, which represent the mean or average of the intensity values of the spot in the Cy 3 and Cy 5 channels.

3.5.2. *The Second Ratio Estimate r_2.*

$$r_2 = \overline{\left(\frac{x}{y}\right)} \tag{16}$$

This metric first calculate the ratio of each pixel and then take the average.

3.5.3. *The Third Ratio Estimate r_3.*

$$r_3 = r_2 - \frac{1}{n}\left(\frac{\overline{x}}{\overline{y}^3} s_y^2 - \frac{s_{xy}}{\overline{y}^2}\right) \tag{17}$$

where s_y^2 is the sample variance of y, and s_{xy} represents the sample covariance of x and y. The last metric provides an asymptotically unbiased estimate of the ratio as $E\{r_3\} \approx \mu_x/\mu_y$ [17].

The first statistic r_1 and the second statistic r_2 are biased estimates of the ratio between two channel signals. The third statistic r_3 is a more robust ratio estimate and has been used in the two channel FISH imaging [17]. We have evaluated these metrics in quantifying the gene expression and have shown r_3 is more accurate as will be validated in Section 4.

4. Results

4.1. *Data Collection*

We have used both the simulated microarray data and real data in our test of the algorithm. We have

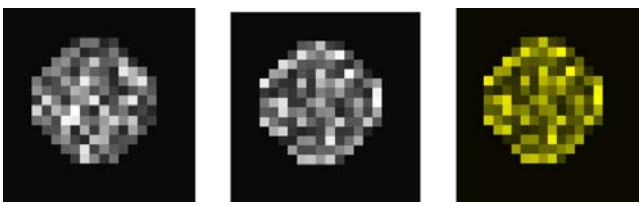

Figure 2. A simulation of two-channel microarray spots with varying intensities. The *rightmost* image is the composite pseudo-color image.

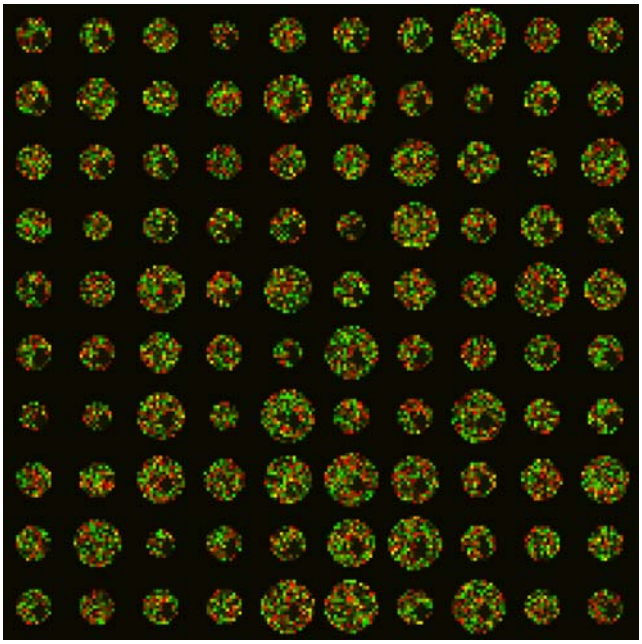

Figure 3. A simulation of a microarray subgrid containing donut spots of varying size and noise.

used a microarray model [13] (described in Section 4.3) to generate simulated data, which can be used to quantitatively evaluate the proposed algorithm because the ground-truth is known.

The real images were collected at the Dr. Cai's Laboratory at the Department of Human Molecular Genetics of Baylor College of Medicine. The data have been processed using commercial software by

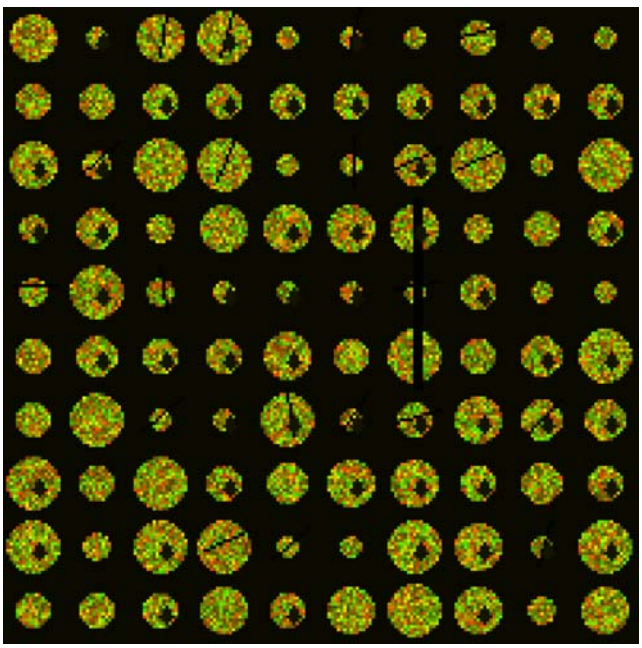

Figure 4. A simulated microarray subgrid, which contains donut and scratched spots with varying radii.

Table 2. Mean and standard deviation of SSD values using different clustering approaches for different spots.

Mean/variance	HCM	FCM	PCM
Circles	0.7462	0.7603	0.6497
	0.8339	0.8811	0.5265
Donuts	0.9997	1.0275	0.8662
	1.8386	1.9695	1.1970
Scratches	0.9005	0.8956	0.7781
	1.5109	1.4536	0.9530

Biodiscovery Inc. The data are from a normal female versus male comparative genomic hybridization on mouse whole genome bacterial artificial chromosome (BAC) arrays, in which the two genomic DNA samples were differentially labeled with fluorescent dyes Cy3 and Cy5. The images were collected sequentially using a microarray scanner.

4.2. Evaluation Criterion

For simulated image with known ground-truth, we quantitatively measured the performance of the proposed approach. One metric is the normalized mean square error (NMSE), which was defined as follows:

$$NMSE = \frac{\sqrt{\frac{1}{MN}\sum_{j}^{M}\sum_{i}^{N}\left(X_{ji} - \overline{X}_i\right)^2}}{\frac{1}{MN}\sum_{j}^{M}\sum_{i}^{N}X_{ji}} \qquad (18)$$

where M and N are the dimensions of the image. X_i refers to each pixel in the input image and $\overline{X}$ is the image after clustering. NMSE was calculated for varying noise levels in the input image for all the different spots.

Table 1 shows the variation of NMSE with the change in noise factor for K-means, FCM and PCM. It was observed that NMSE changes with respect to the noise added to the image. Noise level is directly proportional to NMSE. After certain noise level it was observed that NMSE decreases, as it caused due to complete distortion of the spot (input image).

We also used the stability criterion [20]. The stability of estimated expression levels across replicates is measured by the sum of squared difference (SSD), defined as

$$SSD = \sum_{i=1}^{N}\sum_{r=1}^{R}\left(l_{i,r} - \bar{l}_i\right)^2 \qquad (19)$$

where N is the total number of spots in the microarray, R is the total number of replicates, $l_{i,r}$ is the log ratio of ith spot on the rth replicate, and $\bar{l}_i$ is the mean of the log ratios across all the replicates for the ith spots. The formula given in Eq. (19) calculates the variation in the log ratio estimate. A smaller value of SSD or the less variation of the estimation, the more stable the method is.

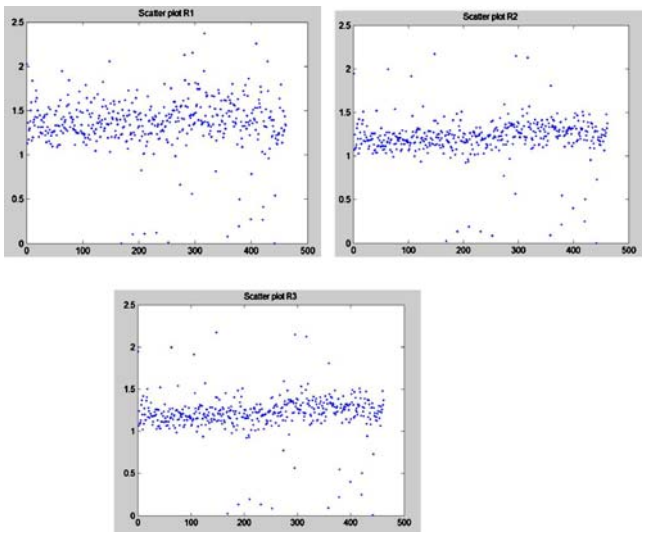

Figure 5. The logR/G ratios using three different statistics defined in Eqs. (15–17).

Table 3. SSD values for simulated images using three different statistics by FCM with nine different parameters.

$[\eta_i]$	R_1	r_2	r_3
[0.1, 0.9]	5.4633	4.2959	4.3084
[0.2, 0.8]	6.0449	4.5021	4.5118
[0.3, 0.7]	6.7890	4.7921	4.7981
[0.4, 0.6]	9.5896	5.6349	5.6327
[0.5, 0.5]	11.1448	5.9851	5.9782
[0.6, 0.4]	10.0514	5.4691	5.4691
[0.7, 0.3]	9.7354	5.3422	5.3346
[0.8, 0.2]	6.0449	4.5021	4.5118
[0.9, 0.1]	5.4633	4.2959	4.3084
Average	7.8141	4.9799	4.8937

4.2.1. Simulated Data Set. We generated synthetic two-color microarray data according to the model proposed by Rocke and Durbin [13].

$$\begin{aligned} y_t &= \alpha_t + \mu e^{\eta_t} + \varepsilon_t \\ y_r &= \alpha_r + \mu e^{\eta_r} + \varepsilon_r \end{aligned} \tag{20}$$

where the notations r and t represent reference and test channels, respectively. In the above equations, y represents the intensity measurement; μ is the expression level contributed by the quantity of interest; α is the mean background intensity and

$$\eta \sim N\left(0, \sigma_\eta\right) \text{ and } \varepsilon \sim N(0, \sigma_\varepsilon) \tag{21}$$

and ε represents the additive error that always exists. The mean background intensities α will be estimated by the proposed algorithms and then the intensities of test and reference signals will be estimated by

$$x_t = y_t - \alpha_t \quad x_r = y_r - \alpha_r \tag{22}$$

From this value, the log ratio is calculated using one of the statistics of Eqs. (15–17). This model was also used in [12].

Table 4. SSD values for simulated images using three different statistics and clustering approaches.

Method	r_1	r_2	r_3
K-means	9.6122	5.6284	5.6252
FCM	10.1599	5.7850	5.7789
PCM	8.8789	5.3746	5.3743

Table 5. The calculation of metrics *V*2/*V*3 and *S*2/*S*3 using Eqs. (12) and (13) in validating the clustering algorithms, where *V*2/*V*3 and *S*2/*S*3 correspond to $c = 2$ and $c = 3$ respectively.

No. of iteration	*V*2	*V*3	*S*2	*S*3
1	1.0046	1.9265	0.0417	0.5627
2	1.0096	1.3500	0.0417	0.3580
3	1.0096	1.9864	0.0417	0.0370
4	1.0096	1.1350	0.0417	0.3580
5	1.0096	1.9265	0.0417	0.5627

4.2.2. Creating Various Image Replicates. We have generated the expression levels to each spot of every image based on the Eq. (21). Since parameters in Eq. (21) follow a normal distribution, the distribution is varied by changing the parameters μ and σ according to the following equation.

$$x = \sigma z + \mu \tag{23}$$

By varying the values of z, the spots in each grid vary in each pixel while follow a distribution defined in Eq. (21).

4.2.3. Spot Simulation. The simulation of spots should be able to model a variety of real microarray spots. The array spots are generally circular in shape, but their radius might not be constant over the whole microarray, due to possible difference in the spotting robot operating at two different locations of the slide [18]. To simulate this effect, we allowed the radius of a spot to vary by dilating or eroding the circle using mathematical morphological operation. In the robot printing process, lesser amount of DNA may be deposed at the center of the spot [18]. This is

Table 6. A comparison of R/G ratios using HCM and FCM. The two methods give slightly different values.

Sample	HCM 2 clusters	FCM 2 clusters	HCM clusters	FCM 3 clusters
1	0.6640	0.6406	0.6332	0.6414
2	6.4008	6.4008	6.7127	6.7200
3	0.6261	0.6253	0.5428	0.5555
4	0.8181	0.8181	0.7963	0.7950
5	0.6195	0.6195	0.592	0.595
6	2.3410	2.3419	2.4373	2.4588
7	0.1356	0.1355	0.1121	0.1121

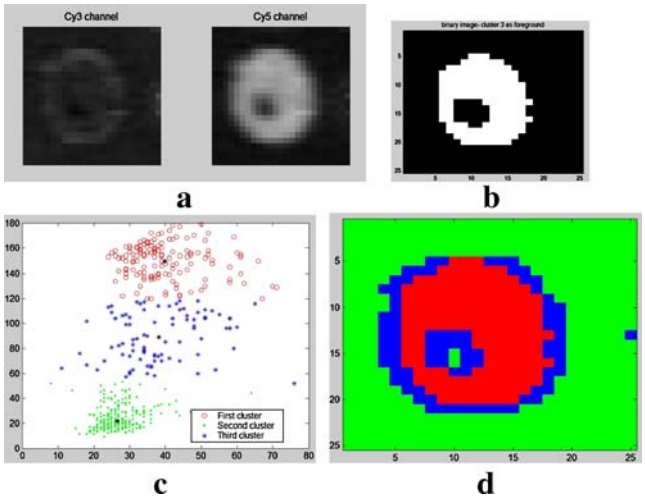

Figure 6. Segmentation of donut spots from a real image using clustering. **a** Cy3 and Cy5 images; **b** segmented regions using two clusters; **c** scatter plot of Cy3 versus Cy5, three clusters were shown; **d** the region corresponding to clustering **c**. Pixels in blue indicate the region of donuts and boundary of the donuts, which can be excluded in the calculation.

known as the doughnuts effect; a hole is generated at the center of certain spots. In addition, some parts of the spots can disappear during the microarray fabrication as a result of the washing process or surface tension on the glass during the drying process of the biological material [18]. The spot generated will take on an almost random shape.

Our simulation algorithm produces spots at grid locations that resemble the actual microarray. Using the model of Eq. (20) we can have the ground-truth of the array spots. Figure 2 shows a spot image with different intensities. The two gray scale images are the two channels Cy3 and Cy5 images. The pseudo-color images are generated from these two grayscale images. Our algorithm can generate spots with various radii in a grid.

4.2.4. Doughnut Hole. In certain situations, because of the impact of the print tip on the glass surface, a smaller amount of cDNA can be attached to the center of the targets. As a result, the center of these targets emits less fluorescent photons, giving a target with the doughnut shape [18]. Our simulation allows the spot to have a hole in the center with varying size and shape. In addition, noise is added. The pixels around the hole of doughnut are set to have smaller intensity values. A threshold radius for these doughnuts is set to 3 pixels. Figure 3 shows such an example of microarray subgrids, containing donut spots with varying size and noise. The percentage of the donut spots in the subgrid is set to be 50% in the figure.

4.2.5. Scratched Spots/Chord Removals. Part of the spots can be washed off due to various physical effects during hybridization and the drying process, resulting in low intensity levels in the array spots [19]. This irregularity is modeled by cutting the spots at irregular positions. Some of these scratches remove part of the spots for more than one spot.

In our simulated microarray images, all the variations and defects are put together to simulate the real array image. Figure 4 shows such an example of simulating a variety of distorted and noisy spots. Doughnut shaped spots with varying radius and scratches at various positions.

4.2.6. Evaluation of Segmentation Region Using NMSE. For the simulated image with known ground truth, we can compare the difference between the proposed algorithm and ground truth. In order to test the reliability of the algorithm under noises, we added different levels of variations to the simulated spots. We then computed the NMSE according to Eq. (18) between the calculated region and theoretical spot area. It was observed that NMSE changes with the level of noise. Higher the noise level, then larger is the NMSE which is consistent with our intuition. Table 1 compares the statistics (mean and standard deviation) of the NMSE when using three

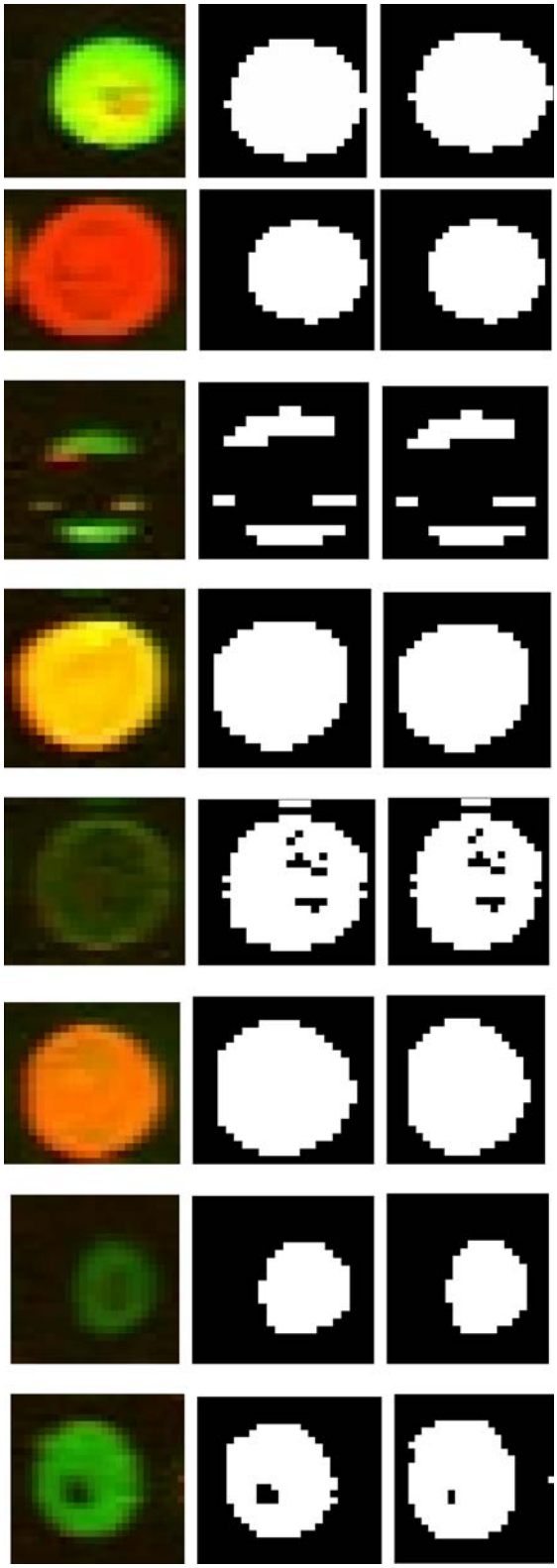

Figure 7. Processing of a variety of spots using HCM (*second column*) and FCM based clustering (*third column*).

different clustering approaches for three types of spots when adding 13 different levels of variation. It can be concluded that HCM, FCM and PCM generally produce the comparable NMSE and PCM gives a little lower mean error and deviation.

4.2.7. Evaluation of Stability Using SSD Value. The stability value or the SSD quantifies the robustness of the proposed algorithm. We have performed four replications of three different types of spots using Eq. (23) and calculated the statistical analysis of the stability quantization according to Eq. (19). Table 2 compares the mean and standard deviation of the stability values. It can be concluded that PCM based approaches produces the lowest SSD values, indicating the highest stability among three clustering approaches, irrespective of the shapes of the array spots.

4.2.8. Evaluation of the Ratio Statistics. We have compared the three statistics r_1, r_2, r_3 in Eqs. (15–17) when quantifying the gene expression levels or amount of mRNA. Figure 5 displays the log ratios calculated for a simulated array spot. The distribution of using r_2 is quite similar to that of r_3 because theoretically the variance of r_2 estimate is the same as that of r_3 [17].

We compared the stability values of using the three different statistics. From Table 3 it can be seen that r_3 produces the smallest SSD values or the highest stability. In the experiment four replicates, i.e., eight images were used. The simulated images had 6 spots in a $[8 \times 8]$ sub-grid. The FCM clustering was performed and the number of clusters is taken to be two.

We also compared the stability using three different statistics and three different clustering approaches, as listed in Table 4. We can draw the conclusion that, whatever statistics used the PCM based approach always gives the lowest SSD value, or highest stability. This is consistent with the conclusion drawn from Table 2, where the statistic r_1 was used.

4.2.9. Validation of Clustering. According to the analysis in Section 3.4, the result of the clustering can be validated through the calculation of the partition coefficient V_{PC} [Eq. (12)] or the metric S [Eq. (13)].

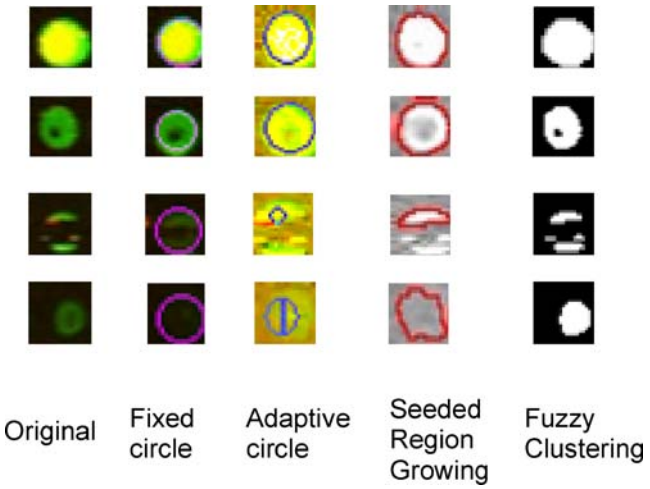

Figure 8. A comparison of different segmentation approaches on clean spots, donut spots, spots with scratches and low intensity, indicating that fuzzy clustering is the best approach.

We have tested the example in Fig. 1. The scatter plot of Cy3 versus Cy5 indicates three clusters, with the third one corresponding to artifacts. We implemented the FCM clustering algorithms by giving different number of clusters and initial values. The solution of the $\max_c\left\{\max_{\Omega^c}\{V\}\right\}$ or $\max_c\left\{\max_{\Omega^c}\{S\}\right\}$ is $c=3$ (see Table 5), which is consistent with our observation from the scatter plot of Fig. 1d. Therefore, the metric S is a more valid criterion in identifying the number of clusters.

4.3. *Real World Data: Case Example*

We present the case study on the processing of a variety of spots which we might encounter in the realistic microarray slides. The images were collected from *Baylor College of Medicine* described in Section 4.1. Table 6 shows the log ratios calculated using HCM clustering and FCM clustering with different number of clusters, showing slightly different values. Because of the low intensities of these array images, they were enhanced before applying our algorithms.

We have conducted a series of experiments on different types of data sets using the proposed clustering algorithms. Figure 6 shows an example of processing donut shape spots using the clustering approach. Using third cluster, we can easily identify the inner circle of donuts and the boundary (shown in blue in Fig. 6d). This region can be excluded in the calculation of gene expression.

The proposed clustering approach can find regions of various complex shapes. Figure 7 shows such an example of segmenting a variety of spots using both the HCM and FCM clustering. Both approaches are adaptive to the complex spot shape and are insensitive to variations in intensities. These spots can not be well segmented using fixed circle or adaptive circle approaches.

We have compared our fuzzy clustering based approach with some existing methods such as fixed circle, adaptive circle (implemented in ScanAlyze

Table 7. A comparison of SSD values using different approaches for a real array spot image.

Method	SSD value
FCM clustering	244.65
K-means clustering	254.62
Seeded region growing	751.04
Fixed circle	1,029.4

Table 8. The calculation of SSD to analyze stability for real array images.

Method	r_1	r_2	r_3
HCM	1,144.8	123.6	131.2
FCM	1,253	123.4	130.8
PCM	2,024.7	101.8	102.5

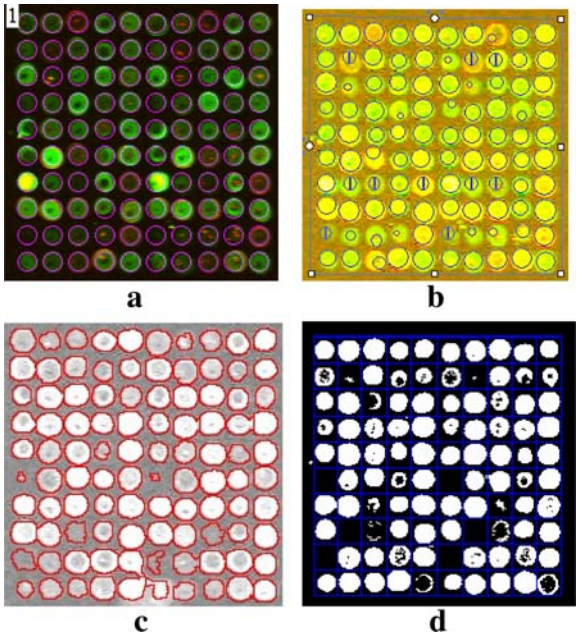

Figure 9. A comparison of segmentation using different approaches. **a** ScanAnalyze, *fixed circle*; **b** GenePix, *adaptive circle*; **c** spot with seeded region growing; and **d** FCM clustering.

[10] and GenePix [9]) and seeded region growing [8] (Spot software [11]). Figure 8 compares these approaches in processing a variety of real array spots. The fuzzy clustering based approach can treat the spots with various complex shapes. The SSD values shown in Table 7 indicate that the fuzzy clustering approach gives the smallest SSD value or the highest stability.

We also compared the three statistics in quantifying the gene expression levels using three different approaches as shown in Table 8. Again, the clustering approach using PCM along with the use of statistic r_3 gives the best stability.

We performed the segmentation results on the subgrid using different approaches. Figure 9 compares the segmentation of subgrid array with several different approaches. Unlike the ScanAnalyze and GenePix software, the fuzzy clustering based approach can process the spots with various shapes. When we compare using the quantitative criteria defined in Eq. (19) (listed in Table 8), it shows that the PCM based approach generally gives the best stability values with metrics r_2 and r_3, indicating that PCM performs better than HCM and FCM clustering.

Our observations have shown that the clustering using improper number of clusters can miss some regions. Figure 6 illustrates this effect. Using two clusters, the spots of the lower intensities are usually considered as background pixels. When the number of clusters is three, one can detect the regions of lower intensities (shown in blue). With the increase of the number of clusters to three and then merge the two clusters, these spots can be attributed to foreground. This region can be excluded in the calculation of gene expression levels, resulting more accurate quantification of gene expression.

The clustering approaches can be performed on the whole subgrid or the single spot following gridding procedure. We have conducted extensive experiments on various microarray images, showing better performance of the proposed approach. Figure 10 shows an example of segmenting an array grid. Because of the low intensity of the image, an image enhancement or background correction was applied before performing clustering. The result of segmentation is compared with existing approaches such as an adaptive circle based approaches implemented in GenePix [9] and ScanAlyze [10] software (Fig. 11). As can be seen from Fig. 11, fixed circle based segmentation has many spots incorrectly segmented as non-foreground.

5. Discussion and Conclusions

In this work we have compared several fuzzy clustering based approaches to segment microarray spots. In addition, we have evaluated three statistics in the quantization of gene expression levels. The technique has been evaluated with the stability metric using both the simulated array spots and real world microarray images collected from biological experiments. In comparison with several existing array segmentation approaches, our proposed fuzzy clustering approaches demonstrate better performance in terms of stability. Our major conclusions are:

1. Overall, the PCM based fuzzy clustering segmentation provides better segmentation than existing clustering based approaches such as k-means or HCM. This is because fuzzy clustering approaches use soft labeling, providing more accurate discrimination of foreground, background and artifacts.
2. The clustering based approaches fully take advantage of two-channel image information. The existing intensity based approaches use the average of the two-channel image data, so they

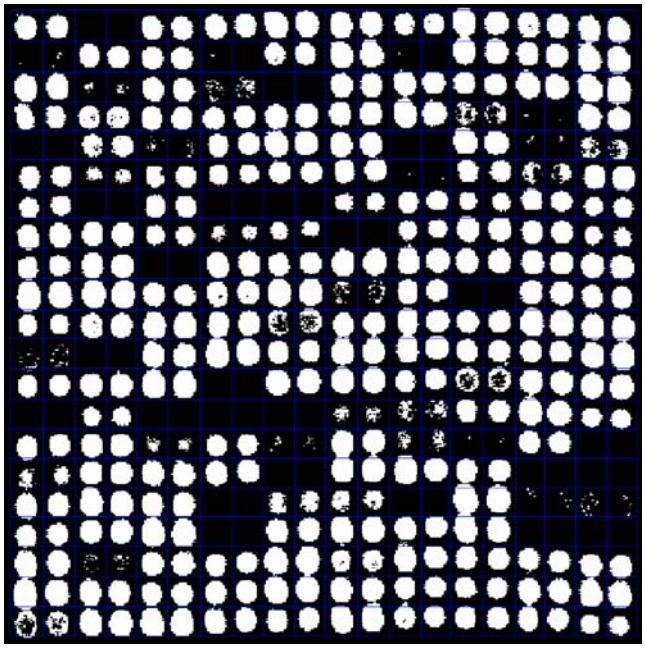

Figure 10. Segmentation of subgrids using FCM with three clusters.

cannot process the spots containing large artifacts like Fig. 1. Using two-channel information, this artifact can be easily identified to correspond to the outlier, as shown in Fig. 1d.

3. The clustering approaches use the two-channel imaging data as features to find the homogenous region. Therefore, they can process array spots of complex shapes including donuts and scratches.
4. A difficulty of the Gaussian mixture model based approaches is the selection of the number of Gaussian components. For the fuzzy clustering approaches, the model can be easily validated

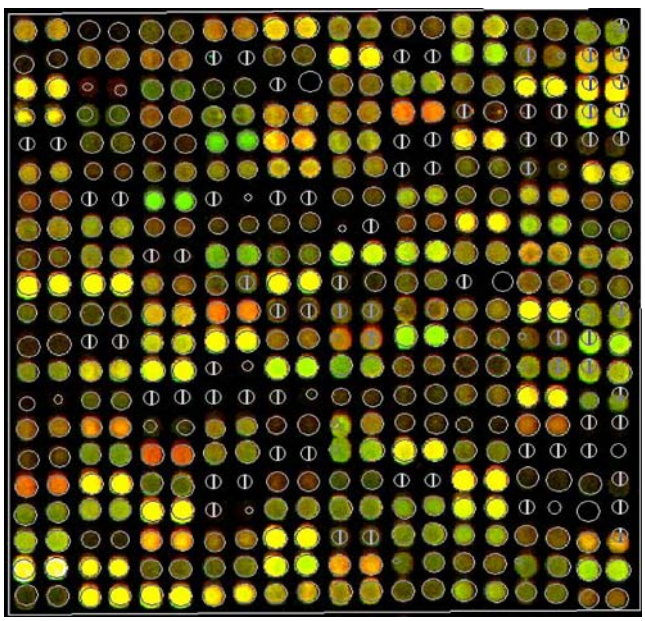

Figure 11. Segmentation of spots in subgrids with the GenPix software, which uses adaptive circle algorithm. The non-spots are specified as *white circles*. Some spots are overlooked.

using metrics defined in Eq. (12) or Eq. (13). The computational complexity is relatively low.

5. We have compared three different statistics defined in Eqs. (15–17) when quantifying the gene expression levels. In the current microarray data analysis, the statistics using r_1 or r_2 is popularly used. We have shown the metrics using r_2 and r_3 produce more robust estimation than r_1. The results using r_2 and r_3 are similar but r_3 appears better in several cases. We recommend the use of the new statistic r_3 for two channel imaging analysis, because it is an asymptotically unbiased estimation.

Like other existing approaches, the fuzzy clustering based approaches also have limitations. When the two neighing spots are connected due to scratch or other noise, the approach has the difficulty in discriminating the right spot. Also, when the array image has low intensities, an enhancement procedure is necessary. We are studying new ways of employing spot geometric features such as wavelet representations to improve the segmentation of array spots.

Acknowledgment

The work was partially supported by the University of Missouri Research Board, Faculty Research Grant and Kansas City Area Life Sciences Institute (KCALSI) Research Development Grant Award.

References

1. C.R. Cantor and C.L. Smith, "Genomics: The Science and Technology Behind the Human Genome Project," Wiley, 1999.
2. W.W. Cai, J.-H. Mao, W.-W., Chow, S. Damani, A. Balmain and A. Bradley, "Genome-Wide Detection of Chromosomal Imbalances in Tumors Using BAC Microarrays," *Nat. Biotechnol.*, vol. 20, 2002, pp. 393–396.
3. Petrov and S. Shams, "Microarray Image Processing and Quality Control," *J. VLSI Signal Process.*, vol. 38, 2004, pp. 211–225.
4. Y.H. Yang, M. Buckley, S. Dudoit and T. Speed, "Comparison of Methods for Image Analysis on cDNA Microarray Data," *J. Comput. Graph. Stat.*, vol. 11, 2002, pp. 108–136.
5. Chen, E.R. Dougherty and M. Bittner, "Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images," *Biomed. Opt.*, vol. 2, 1997, pp. 364–374.
6. A.J. Jain et al., "Fully Automatic Quantification of Microarray Image Data," *Genome Res.*, vol. 12, 2002, pp. 325–332.
7. Sutton, J.C. Bezdek and T.C. Cahoon, "Image Segmentation by Fuzzy Clustering: Methods and Issues," in *Handbook of Medical Imaging*, I.N. Bankamn (Eds.), Academic, 2000.
8. R. Adams and L. Bischof, "Seeded Region Growing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, 1994, pp. 641–647.
9. Axon Instruments Inc., "GenePix 400A User's Guide," Axon Instruments Inc., 1999.
10. M.B. Eisen, "ScanAlyze," http://rana.stanford.edu/software, 1999.
11. M.J. Buckley, "The Spot User's Guide." CSIRO Mathematical and Information Sciences, http://www.cmis.csiro.au/IAP/Spot/spotmanua.html, 2000
12. C.L. Myers, M.J. Dunham, S.Y. Kung and O.G. Troyanskaya, "Accurate Detection of Aneuploidies in Array CGH and Gene Expression Microarray Data," *Bioinformatics*, vol. 20, no. 18, 2004, pp. 3533–3543.
13. D.M. Rocke and B. Durbin, "A Model for Measurement Error for Gene Expression Arrays," *J. Comput. Biol.*, vol. 8, no. 6, 2001, pp. 557–569.
14. H.Y. Jung and H.G. Cho, "An Automatic Block and Spot Indexing with *k*-Nearest Neighbors Graph for Microarray Image Analysis," *Bioinformatics*, vol. 18, Suppl. 2, 2002, pp. 141–151.
15. M. Katzer, F. Kummert and G. Sagerer, "Methods for Automatic Microarray Image Segmentation," *IEEE Trans. Nanobiosci.*, vol. 2, 2003, pp. 202–214.
16. A.W.C. Liew, H. Yan and M. Yang, "Robust Adaptive Spot Segmentation of DNA Microarray Images," *Pattern Recogn.*, vol. 36, 2003, pp. 1251–1254.
17. G.M. van Kempen and L.J. van Vliet, "Mean and Variance of Ratio Estimators Used in Fluorescence Ratio Imaging," *Cytometry*, vol. 39, 2000, pp. 300–305.
18. Y. Balagurunathan, E.R. Dougherty, Y. Chen, M.L. Bittner and J.M. Trent, "Simulation of cDNA Microarrays Via a Parameterized Random Signal Model," *J. Biomed. Opt.*, vol. 7, no. 3, 2002, pp. 507–523 (July).
19. D. Bozinov and J. Rahnenführer, "Unsupervised Technique for Robust Target Separation and Analysis of DNA Microarray Spots Through Adaptive Pixel Clustering," *Bioinformatics*, vol. 18, no. 5, 2005, pp. 747–756 (June 15).
20. Q. Li, C. Fraley, R.E. Bumgarner, K.Y. Yeung and A.E. Raftery, "Donuts, Scratches and Blanks: Robust Model-Based Segmentation of Microarray Images," *Bioinformatics*, vol. 21, no. 12, 2005, pp. 2875–2882 (Jun 15).
21. K. Blekas, N.P. Galatsanos, A. Likas and I.E. Lagaris, "Mixture Model Analysis of DNA Microarray Images," *IEEE Trans. Med. Imag.*, vol. 24, no. 7, 2005, pp. 901–909.
22. X. Xie and G. Beni, "A Validity Measure for Fuzzy Clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, 1991, pp. 841–847.
23. R. Nagarajan, "Intensity Based Segmentation of Microarray Images," *IEEE Trans. Med. Imag.*, vol. 22, 2003, pp. 882–889.
24. R. Krishnapuram and J. Keller, "A Possibilistic Approach to Clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, 1993, pp. 98–110 (Feb).
25. J. Chen, X. He and L. Li, "Identifying the Patterns of Hematopoietic Stem Cells Gene Expressions Using Clustering

Methods: Comparison and Summary," *J. Data Sci.*, vol. 2, 2004, pp. 297–299.
26. D. Dembele and P. Kastner, "Fuzzy *c*-Means Method for Clustering Microarray Data," *Bioinformatics*, vol. 19, no. 8, 2003, pp. 973–980.
27. M. Asyali and M. Alci, "Reliability Analysis of Microarray Data Using Fuzzy *c*-Means and Normal Mixture Modeling Based Classification Methods," *Bioinformatics*, vol. 21, no. 5, 2005, pp. 644–649.
28. L. Fu and E. Medico, "FLAME, A Novel Fuzzy Clustering Method for the Analysis of DNA Microarray Data," *BMC Bioinformatics*, vol. 8, p. 3, 2007.
29. A.P. Gasch and M.B. Eisen, "Exploring the Conditional Coregulation of Yeast Gene Expression Through Fuzzy *k*-Means Clustering," *Genome Biol.*, vol. 3, no. 11, pp. 1–22, 2002.
30. N. Belacel, M. Cuperlovic-Culf, M. Laflamme and R. Ouellette, "Fuzzy J-Means and VNS Methods for Clustering Genes from Microarray Data," *Bioinformatics*, 20, 2004, pp. 1690–1701.