*Mathematical and Computational Applications*, Vol. 15, No. 2, pp. 240-247, 2010.


# MICROARRAY IMAGE SEGMENTATION USING CLUSTERING METHODS

Volkan Uslan and İhsan Ömür Bucak
Department of Computer Engineering,
Fatih University, 34500,
B.Çekmece, İstanbul, Turkey
vuslan@fatih.edu.tr and ibucak@fatih.edu.tr

**Abstract-** Microarray image processing is a technology for viewing and computationally measuring thousands of genes at the same time. Gene expressions provide information about the cell activity in an organism. Observing a substantial change in gene expressions between the cDNA (complementary DNA) microarray experiments of an organism can be a sign of a disease. The goal of this study is to make a fine distinction against the gene expressions in the microarray image processing. For this reason, two clustering methods have been experimented and compared. In this study we have specifically investigated the segmentation step of the microarray image. Other than the segmentation methods used in commercial packages we have used the clustering techniques. We have applied fuzzy *c*–means and *k*-means methods and observed the results.

**Keywords-** Microarray Image, Image Segmentation, Clustering

## 1. INTRODUCTION

Microarray is a rapidly growing technology used in biological processes. There are many uses of existing microarrays in the area of cancer, diabetic and genetic diagnoses, gene and drug discovery in molecular biology, etc. Microarrays aid computation of hundreds of thousands of genes simultaneously. Microarray image processing is the process of extraction and interpreting gene information. Gene expressions provide information about the cell activity in an organism. cDNA spots are derived from experimental or clinical samples [1]. Firstly, RNAs are isolated from samples, and then reverse transcription process is used to convert the RNAs into cDNAs. cDNAs are labelled with fluorescent probes Cy3 (green) for the control and Cy5 (red) for the experimental channel. A probe represents a DNA sequence. The probes are attached on a glass slide by a robotic arm in the form of a grid. As a result a microarray slide containing hundreds of thousands of spots is generated. Each spot in a microarray expresses a different DNA sequence.

There exist three main steps in microarray image processing. These steps are discussed in detail [2]. The first step is gridding. Gridding is used for locating the centers and bounding boxes of each spot. The second step is segmentation. Segmentation is the classification of pixels either as signal or background. The third step is information extraction. Information extraction calculates signal intensity for each spot of the array. In this paper, we have studied the segmentation step of microarray image processing and experimented fuzzy *c*-means and *k*-means clustering approaches in the segmentation process and compared the observed results [3, 4].

## 2. MATERIALS AND METHODS

### 2.1. Gridding

In order to find out where the spots are located, gridding is crucial in microarray image processing. In this study, gridding procedure is used as described in [5].

Spots have different sizes and intensities in a microarray image. However these spots are located in the image in an order. To estimate the spacing between spots, autocorrelation have been used. Autocorrelation is a mathematical tool for finding repeating patterns. The mean intensity has been calculated for both, horizontally and vertically. Then autocorrelation has been applied to enhance the self similarity of the horizontal and vertical means. Peak values have been obtained by differentiating left and right slopes of the means. Once the peak values are found, the centroids of the peaks have been extracted. These centroids correspond to the centres of the spots. The midpoint between two centres gives the grid locations. Thus, grid lines pass through these grid locations.

### 2.2. Segmentation

The segmentation step is important, because it considerably affects the precision of microarray data [6]. There are many segmentation methods that are available, and some are already used in commercial packages.

There are two main techniques in microarray image segmentation: (a) Image processing techniques (b) Machine learning techniques [7]. Image processing techniques involve three methods. Fixed or adaptive circle segmentation considers the spots that are circle shaped. Fixed circle segmentation assumes that the diameter of circles is fixed [8]. On the other hand, adaptive circle segmentation adjusts the diameter of the circle dynamically and seeded growing region is one of the well-known uses of this method [9, 10]. Histogram based segmentation method computes a threshold value. According to the computed threshold value, pixels are assigned to foreground and background classes [11]. Machine learning techniques employ clustering and classification methods.

Microarray image segmentation is a pixel-based segmentation by clustering the cDNA image pixels into either spots or image background. Clustering is the grouping of the objects that are more similar to each other. Foreground pixels represent the signal and background ones represent the surrounding area. The pixels of the microarray image have been clustered to determine whether they are part of the foreground or background classes.

#### 2.2.1 Fuzzy C-Means Method

The method in reference [3] has been implemented to evaluate the fuzzy *c*-means (FCM) clustering method. This implemented FCM algorithm works as follows:

*i. Make random initialization for the membership matrix,*

*ii. Loop through the following steps until a stopping condition is satisfied,*

- *a. Compute the centroid values for each cluster,*
- *b. Compute the membership values belonging to clusters for each pixel.*

By implementing FCM, we have clustered the pixels such that each pixel has a degree of membership belonging to foreground or background clusters. In the nature of fuzzy logic, each point has a degree of membership to clusters rather than belonging to only one cluster. The membership degree of a pixel is a value such that $u_{ij} \in [0,1]$. The sum of membership values of a pixel belonging to clusters equals to 1:

$$\sum_{j=1}^{C} u_{ij} = 1, \ \forall i = 1,2,...n. \tag{1}$$

In this study, an objective function for fuzzy c-means method can be defined as follows:

$$J_m^f = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2, \ m \in [1,\infty), \tag{2}$$

where $m$ is a real number greater than 1 and is chosen 2, and $u_{ij}$ is the degree of the membership of pixel $x_i$ belonging to designated cluster. The $\left\| x_i - c_j \right\|$ above expresses the distance measured between data and the center. An absolute value of the difference between two consecutive objective functions, $J_{m+1}^f$ and $J_m^f$, is sought to be minimized iteratively until a stopping condition that is less than a user-specified parameter $\varepsilon_f$ is reached , i.e.,

$$\left| J_{m+1}^f - J_m^f \right| \le \varepsilon_f . \tag{3}$$

At each iterative step, the membership $u_{ij}$ is updated as follows:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left( \frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}}, \tag{4}$$

and the cluster centers $c_j$ are updated according to the following:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m}. \tag{5}$$

### 2.2.2 K-Means Method

For the evaluation of the *k*-means clustering, we have implemented the method detailed in reference [4]. *K*-means is one of the basic methods in clustering. We have

also reviewed the *k*-means approach used for the microarray image segmentation [12]. *K*-means algorithm in this study has been implemented as following:

i. *Initialize the cluster means, so that one mean is the minimum value and the other is the maximum value among the pixels,*
ii. *Loop through the following steps until a stopping condition is satisfied:*
    a. *Compute the nearest cluster for each pixel and classify it to that cluster,*
    b. *Compute new means after all the pixels classified.*

Two clusters have been considered; one is for foreground and the other one is for background. *K* has simply been chosen 2. Minimum value for the background cluster and maximum value for the foreground cluster have been initialized. After the initialization, the Euclidean distance for each pixel to each of the means has been calculated. Each pixel to the cluster to which it is closest has been classified. New means for each cluster have been calculated upon the completion of the classification process. The classification step until a stopping condition is reached has been repeated. An objective function for *k*-means method can be defined as follows:

$$J_m^k = \sum_{i=1}^{N} \sum_{j=1}^{C} \left\| x_i - u_j \right\|^2 , \tag{6}$$

where $u_j$ is the mean of the designated cluster. An absolute value of the difference between two consecutive objective functions in the *k*-means method, $J_{m+1}^k$ and $J_m^k$, is sought to be minimized iteratively until a stopping condition that is less than a user-specified parameter $\varepsilon_k$ is reached, i.e.,

$$\left| J_{m+1}^k - J_m^k \right| \leq \varepsilon_k . \tag{7}$$

The pixel values in this study are used as intensity values which are 1-D points. So the Euclidean distance is calculated as follows:

$$\sqrt{(x_i - u_j)^2} = \left| x_i - u_j \right| . \tag{8}$$

## 3. IMPLEMENTATION AND EXPERIMENTAL RESULTS

A sample microarray slide shown in Fig. 1 has been used for the experimental purposes [13]. The sample microarray slide contains a 4*6 number of microarray blocks, each of which has 22*20 spots. The slide has been read and the first block of the slide shown in Fig. 2 has been cropped for use in the experiment.

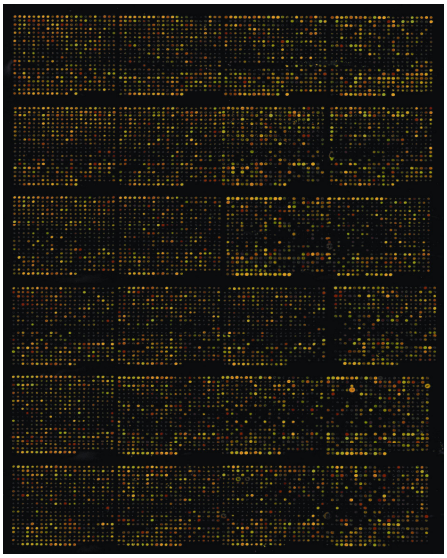

**Fig. 1:** Microarray slide [13].

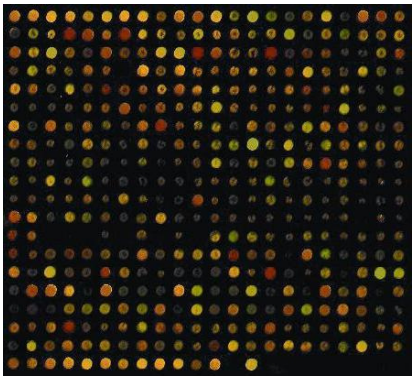

**Fig. 2:** Sample microarray block consisting of 22*20 spots cropped from the microarray slide.

The first 5*5 spots of the first block as shown in Fig. 3 have been cropped for computational simplicity. Then coloured image have been converted to greyscale image to be used in the microarray image segmentation process. Gridded image as shown in Fig. 3c has been obtained according to the details given in Section 2.1.

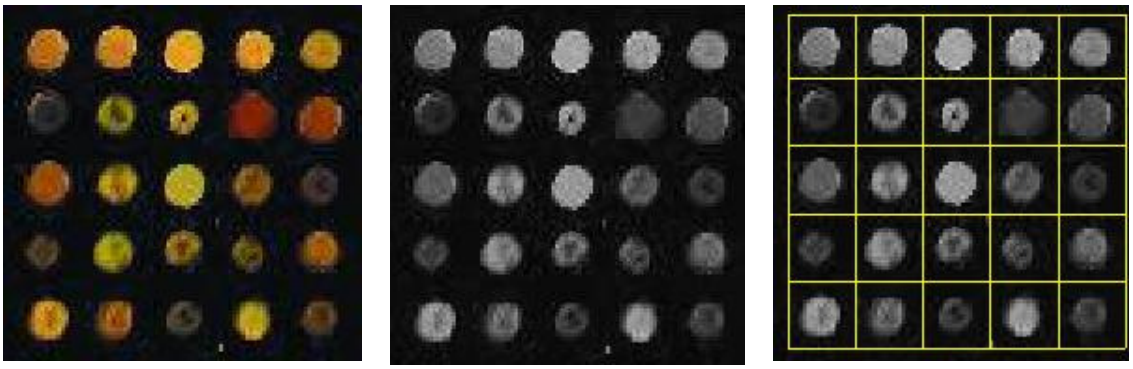

**Fig. 3:** 5*5 spotted microarray image **a**. Image, **b**. Greyscale image, **c.** Gridded image.

The goal of this study is to experiment and compare two clustering methods for the microarray image segmentation process. The resulted images are shown in Fig. 4. The 5*5 spotted sample image is a 106*106 pixel image which contains a total of 11236 pixels. The mean values and the number of pixels that belong to foreground and background classes have been obtained for experimenting purposes as shown in Table 1.

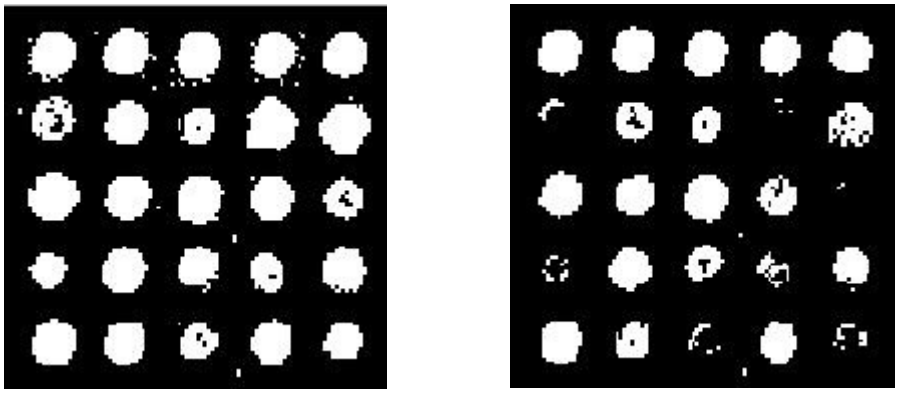

**Fig. 4:** The segmented microarray image after the clustering methods applied: **a.** fuzzy *c*-means clustering, **b.** *k*-means clustering.

**Table 1.** The results obtained for the 5*5 spotted microarray image after *k*-means and fuzzy c-means clustering methods have been applied.

| 106*106 pixel Image | # of Pixels | Foreground Pixels | Background Pixels | Mean Fore | Mean Back |
|---|---|---|---|---|---|
| *k-means* | 11236 | 1836 | 9400 | 116.57 | 18.26 |
| *fuzzy c-means* | 11236 | 2826 | 8410 | 93.40 | 14.47 |

Table 1 shows the comparison of two methods according to which fuzzy c-means seems to be more efficient than *k*-means. Execution of *k*-means algorithm has

given rise to a hard classification in which each pixel has been assigned to either foreground or background clusters. This classification has assigned 1836 pixels to the foreground cluster. On the other hand, fuzzy c-means execution has led to more sensitive classification in which each pixel has belonged to both foreground and background clusters at the same time but with a different degree. Then, the fuzziness of a pixel's membership to a cluster has been defuzzified by selecting the cluster with the highest membership. This classification has assigned 2826 pixels to the foreground cluster. The number of foreground pixels has increased greatly when compared with the *k*-means. Fuzzy c-means has ensured a relatively higher clustering quality. This sensitive classification has resulted in to the more precise classification of weak spots.

Both methods do not ensure the optimal solution. The performances of both methods have also been observed in terms of time. Although the performance depends on some other factors such as determining the initial centroids, it is observed that fuzzy c-means has converged sharply and each iteration has run almost four times faster than *k*-means.

## 4. CONCLUSION

In this paper two clustering methods have been used to make a fine distinction against the gene expressions in the microarray image processing. The clustering methods used are fuzzy c-means and *k*-means. The segmented images and measured values have been obtained and compared each other. One can conclude that fuzzy c-means is more efficient than the *k*-means in terms of clustering the signal pixels. This is because fuzzy c-means has ensured a sensitive classification when compared with the *k*-means. This has resulted in to the more precise classification of the weak spots. However, there is too much noise found in the segmented microarray image obtained through the fuzzy c-means method. As for the future work, the noise removal has to be addressed to get much smoother image. The segmentation step is important, because it considerably affects the precision of the microarray data. Intensity extraction step is the next one which follows the segmentation step. In the future, the efficiency of the clustering methods can also be scaled by observing the signal values in the intensity extraction step of the microarray image processing.

## 5. REFERENCES

1. M. Schena, D. Shalon, Ronald W. Davis, and Patrick O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, **270**, 467-470, 1995.
2. L. Qin, L. Rueda, A. Ali, and A. Ngom, Spot Detection and Image Segmentation in DNA Microarray Data, *Applied Bioinformatics*, **4**, 1-11, 2005.
3. J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.
4. J.A. Hartigan and M.A. Wong, A K-means clustering algorithm, *Applied Statistics*, **28**, 100-108, 1979.
5. B. Alhadidi, H. N. Fakhouri, and O. S. Al Mousa, cDNA Microarray Genome Image Processing Using Fixed Spot Position, *American Journal of Applied Sciences*, **3**, 1730-1734, 2006.

6. A. A. Ahmed, M. Vias, N. Gopalakrishno Iyer, C. Caldas, and J. D. Brenton, Microarray segmentation methods significantly influence data precision, *Nucleic Acids Research*, **32**, no.5 e50, 2004.
7. N. Giannakeas and D.I. Fotiadis, *Image Processing and Machine Learning Techniques for the Segmentation of cDNA Microarray Images*, Handbook of research on advanced techniques in diagnostic images and biomedical application, 2008.
8. M.B. Eisen and P.O. Brown, DNA Arrays for Analysis of Gene Expressions, *Methods Enzymol*, **303**, 179-205, 1999.
9. R. Adams and L. Bischof, Seeded Region Growing, *IEEE Trans. On Pattern Analysis and Machine Intelligence*, **16**, 641-647, 1994.
10. M.J. Buckley, *Spot's User Guide*, CSIRO Mathematical and Information Sciences, Australia, 2000.
11. GSI Lumonics, *QuantArray Analysis Software*, Operator's Manual, 1999.
12. S. Wu and H. Yan, Microarray Image Processing Based on Clustering and Morphological Analysis, *Proc. Of First Asia-Pasific Bioinformatics Conference*, Adelaide, Australia, 111-118, 2003.
13. National Human Genome Research Institute, http://www.genome.gov, 2009.