

Segmentation of microarray cDNA spots using MRF-based method

O. Demirkaya¹, M. H. Asyali¹, M. M. Shoukri¹, K. S. Abu-Khabar²

¹Department of Biostatistics, Epidemiology and Scientific Computing

²Department of Biological and Medical Research

King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia

Abstract—Segmentation or separation of spots from the background in cDNA microarray images is one of the earlier steps in gene expression data analysis. Performance of the segmentation method may profoundly impact the performance of the subsequent stages of data extraction and analysis. Several methods have already been suggested to segment microarray spots. In this study, we propose a new approach based on the Markov random field modeling of the microarray spot regions. Initial parameters were estimated using an entropy-based thresholding algorithm. The proposed method was applied to actual microarray images, and our preliminary results indicate that the proposed segmentation method performed well.

Keywords—cDNA, microarray, image segmentation, MRF

I. INTRODUCTION

cDNA microarray technology provides a powerful and efficient means for measuring the relative abundance of mRNA from a normal and a diseased tissue simultaneously. A comprehensive review of the biological and technological aspect of the microarray technology can be found in [1]. Gene expression information acquired in the form of an image has to be extracted from the image using the techniques of image processing and analysis. In this data acquisition and processing pipeline, segmentation of cDNA spots is the most challenging task.

Segmentation is a process that divides an image into mutually exclusive regions. Each region is homogeneous with respect to a region property such as gray-level intensity. Microarray images are difficult to handle in general, due to factors such as the variability from experiment to experiment, noise and image artifacts. Segmentation of spots in microarray images can further be complicated with nonuniform shape and surface-intensity distribution.

Several methods have been suggested to segment microarray spots and incorporated into commercial microarray image analysis software packages. Gradient-based spot segmentation is used in the software Dapple [2]. We observed that gradient-based segmentation methods did not work well for the microarray images produced in our laboratory. The spots occasionally tend to have a doughnut or crescent-like patches that exhibit high-gradient boundaries.

Region growing based segmentation methods [3] relies on the selection of a seed point. Although the approximate location of spot centers are known a priori, the selection of a

seed point is critical for the performance of a region growing method, especially when spot surfaces have nonuniform intensity patches.

QuantArray (Boston, MA) uses lower and higher range of percentile values for the background and spot. By default, for instance, 80th and 95th percentiles are used in QuantArray to separate the spots from the background.

Fixed and adaptive circle segmentation techniques are also among the methods employed. The former is implemented in ScanAlyze [4] while the latter is implemented in the software GenePix for the Axons scanner [5].

Another approach, which is implemented in the QuantArray software for the GSI Lumonics Scanner and DeArray by Scanalytics, (Fairfax, VA) computes a threshold based on a Mann-Whitney test [6]. This method picks up 8 samples randomly from the known background region and sorts the pixels of the potential spot region, and then iteratively performs a Mann-Whitney test comparing the background samples and the lowest 8 samples of the sorted list. If the two means are not significantly different, one of the 8 samples taken from the potential spot region is discarded and the remaining lowest 8 samples of the sorted list are selected. This test is repeated until the two means are not significantly different from each other. Then the lowest of the last 8 samples is used as the threshold.

In our study, we address the spot segmentation issue only. The spot centers are generally determined manually or automatically in an earlier stage known as gridding or addressing. We used a Markov random field (MRF) based method to segment microarray spots. The proposed method is unique, because, unlike earlier methods, it utilizes the contextual information in addition to the intensity information. The initial estimation of the parameters of the intensity distribution was obtained using an entropy-based thresholding algorithm.

II. METHOD

A. Image acquisition and preprocessing

The microarray used in this study was prepared in the Interferons and Cytokines laboratory at the Department of Biological and Medical Research, King Faisal Specialist Hospital and Research Center, Riyadh. The cDNA microarrays are scanned at two different wavelengths (channels), one for Cy3-labeled (green) sample and another for Cy5-labeled (red) sample. The images used in this paper were acquired at 10- μ m resolution by GenePix 4000 scanner, and saved in 16-bit tagged image file format

(TIFF). During image acquisition, microarrays may be positioned slightly rotated. This may result in error in locating the centers of spots during the addressing process. We corrected our images for rotation when necessary to prevent such errors. The test images presented in this paper were rotated about 0.28° in clockwise direction before processing. Then, the two channels were processed independently for locating spots centers and defining rectangular spot regions whose centers coincide with those of spots. Spot regions (i.e., rectangular bounding-boxes concentric with spots) were identified using the method in [7]. It produces horizontal and vertical profiles by averaging rows and columns of the image. The periodicity was determined automatically by autocorrelation. The spot centers and the gaps between were determined from the autocorrelation as well. The method performed reasonably well on our images, but it may perform poorly when there are too many low-contrast spots. We have used this method for the purpose of convenience; other methods could also be used.

B. Segmentation

MRF-based segmentation

This segmentation method has been extensively studied and found widespread use in medical and non-medical imaging applications. This method provides a convenient way to combine both observed intensity and the contextual information in an image using Bayesian theory. If we assume the observed image y is a realization of a random field Y , and let x^* indicate the true unknown label of the observed pixel and \hat{x} indicate the estimate of x^* . Then the aim is to find the estimate x^* given y . This can be found by using maximum a posteriori (MAP) estimation

$$\hat{x} = \max_x p(y|x)p(x) \quad (1)$$

The prior density of X is modeled as a region process by a Markov random field and is given by the following form:

$$p(x) \propto \exp(\beta u) \quad (2)$$

where β is a positive constant that controls the size of clustering, u is the number of neighbors in the local neighborhood belonging to a certain class. In general, conditional density of the image is modeled as a Gaussian:

$$p(y|x) = (1/(\sigma\sqrt{2\pi})) \exp(-(y-\mu)^2/2\sigma^2) \quad (3)$$

where μ and σ are the mean and the standard deviation of the Gaussian function. Taking the logarithm of the posterior and substituting (3), the MAP estimate can be formulated as

$$Q_i = (1/2\sigma_i^2)(y_0 - \mu_i)^2 - \beta u_i + \log(\sigma_i) \quad i = 1, \dots, k, \quad (4)$$

where y_0 is the intensity of the pixel to be classified, μ_i and σ_i are the mean and sigma of the class i , and k is the number of classes. Finding the global minimum of (4) is not easy as the number of possible configurations for pixel labels is too many. We used the iterated conditional modes (ICM) algorithm [9] that finds the local minimum by sequentially updating the labels by minimizing (4). In spot segmentation, the target locations normally contain pixels belonging to spot and the background regions (i.e., $k = 2$). A 3×3 neighborhood is used and the β was set equal to 1 during all experiments. After each iteration step, the labeled image was smoothed by median filter as suggested by Mardia [10].

Initial segmentation

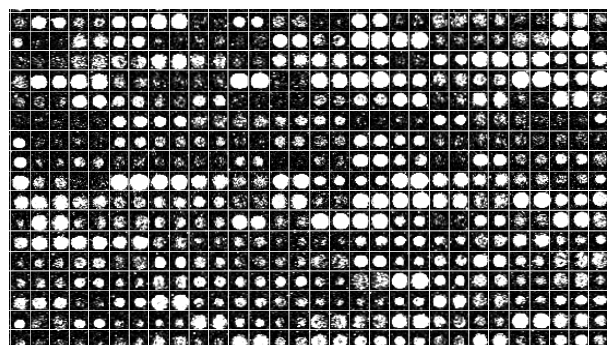
In MRF-based methods, initial segmentation is necessary to initialize the parameters of the intensity distributions. Initial segmentation can be performed by using manual, or automated means. The manual approach involves using a single user-defined threshold. An automated thresholding method can also be used to compute an optimal threshold for each spot region. We have tested several automated thresholding algorithms and Kapur's entropy-based algorithm [8] was found to be performing best based on our visual assessments. Instead of using a separate threshold for each spot region, which will be referred to as the adaptive scheme, we have found that using the average of the thresholds of all spots was also a reasonable choice. In this paper, we will present the results obtained using these two schemes.

The bimodal gray-level histogram is normalized and regarded as a discrete probability distribution function $p(i)$. Each pixel in the image assumes a gray-level value from the set $\{0, \dots, N-1\}$ where N is the number of gray levels. If the image histogram is divided into two classes by the gray-level intensity t then the entropy function to compute an optimal threshold is given by

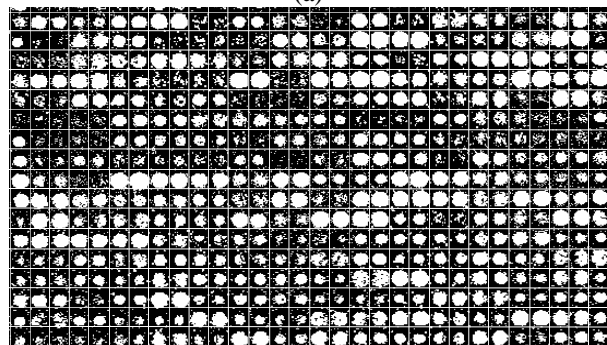
$$\varphi(t) = \ln P_t(1 - P_t) + H_t / P_t + (H_n - H_t) / (1 - P_t) \quad (5)$$

where $P_t = \sum_{i=0}^t p(i)$ and $H_t = \sum_{i=0}^t p(i) \ln p(i)$, are the total probability and the entropy of the class 1 consisting of intensities less than or equal to the threshold t , and H_n is the entropy of the entire image. The intensity t that maximizes (1) is regarded as the optimal threshold that separates the two classes. Here the entropy algorithm was applied to the log-transformed intensities.

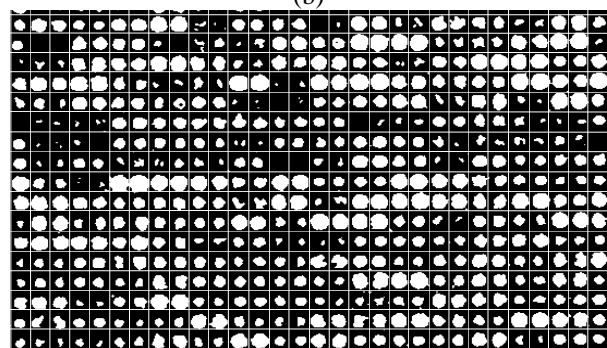
The entire method was implemented and tested in Matlab (The Mathworks, Inc., Natick, MA, USA) environment.



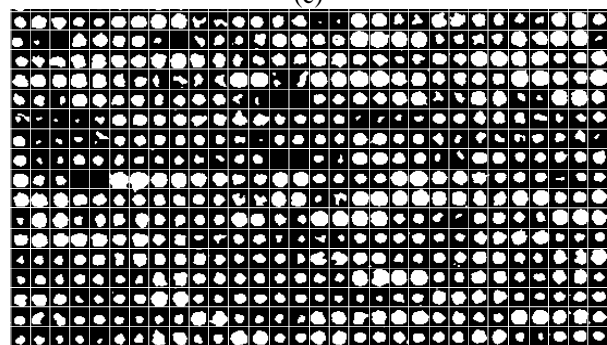
(a)



(b)

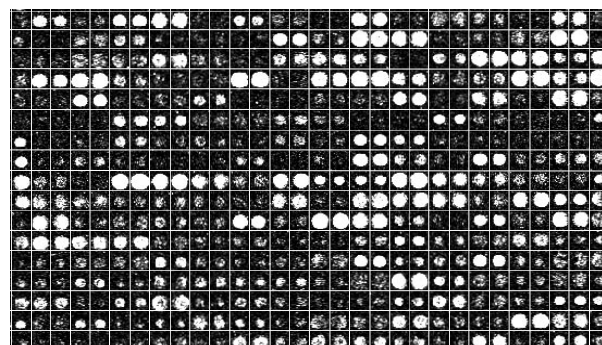


(c)

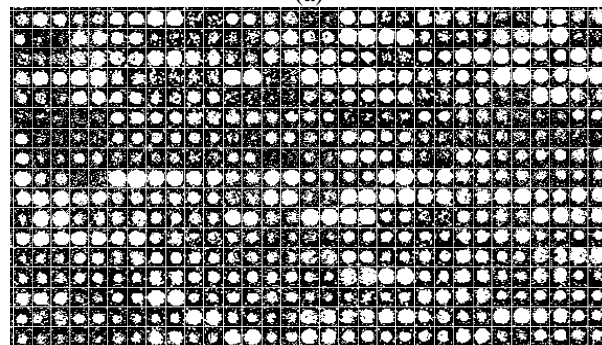


(d)

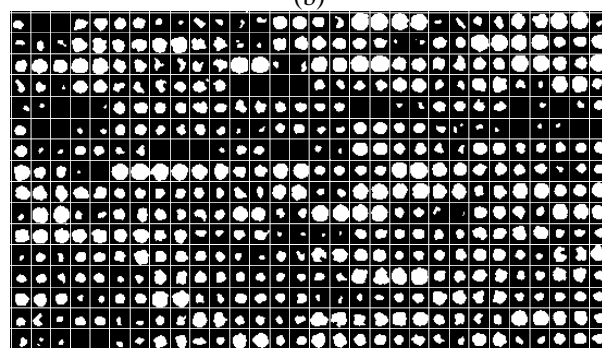
Fig. 1. (a) Original spot image (Cy5) cropped from a larger microarray image. (b) Initial segmented image (thresholded with the average of the all thresholds, 203). (c) Final segmented image using the initial segmented image above. (d) Final segmented image using entropy thresholds calculated for each spot (i.e., adaptive).



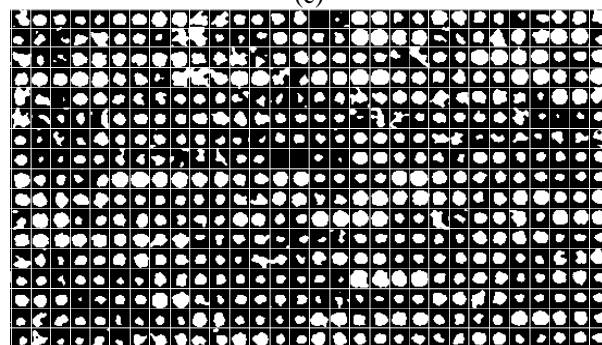
(a)



(b)



(c)



(d)

Fig. 2 (a) Original spot image (Cy3) cropped from a larger microarray image. (b) Initial segmented image (thresholded with the average of the all thresholds, 138). (c) Final segmented image using the initial segmented image above. (d) Final segmented image using entropy thresholds calculated for each spot (i.e., adaptive).

III. RESULTS

cDNA images acquired at two different wavelengths and referred to as Cy5 and Cy3 are shown in Fig. 1a and Fig. 2a, respectively. The superimposed grid lines were computed using the gridding algorithm mentioned in the image acquisition and preprocessing section. They were used to define the rectangular spot regions. Fig. 1b and 2b show the image thresholded using a single threshold that is the average of all the thresholds calculated for the spots using the entropy method. Figure 1c and 2c show images that were segmented using the MRF method whose initial parameters were computed using the images on the second rows. Fig. 1d and 2d show images that were segmented using the MRF method but the initial parameters were computed using the adaptive scheme where each spot was undergone an initial segmentation using an optimal threshold. The graphs in Fig. 3 show the effect of the two different initial segmentation schemes applied to Cy5 (top plot) and Cy3 (bottom plot) images. In both plots, the x-axis shows the median intensity within the segmented spot when the adaptive method was used to compute the initial parameters, while y-axis shows the median intensity when a single threshold (average of the all the thresholds) was used. The solid lines are the fitted lines, whose equations are also shown, to the data.

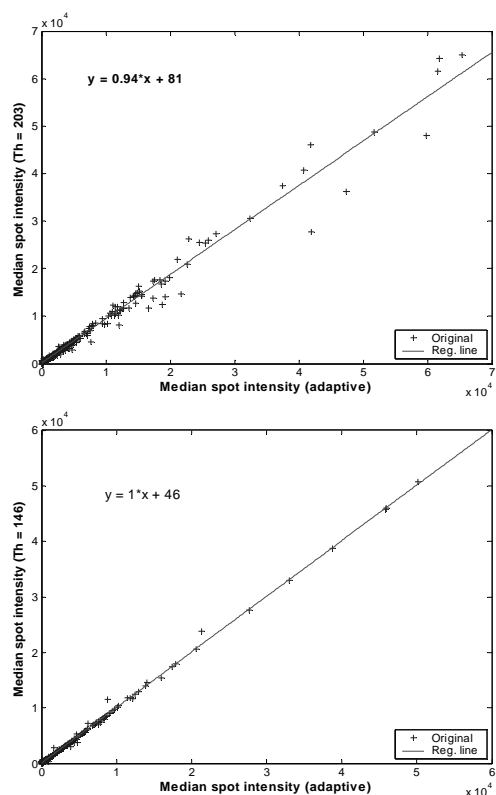


Fig. 3. Effect of the initial segmentation intensity schemes on median spot intensity for the Cy5 (top plot) and Cy3 (bottom plot) image. The solid lines are the fitted lines to the original data. The equations of the lines are also shown in the plot.

IV. DISCUSSION AND CONCLUSION

The preliminary results indicate that the MRF based segmentation method may be a better alternative to segment microarray spots, because they also consider the contextual information. The initial segmentation stage, which is needed to compute the initial estimations of the distribution parameters, seems to have some impact on the performance of the overall method (see Fig. 1c-d and Fig. 2c-d). The entropy-based thresholding performs well when used adaptively and to compute a single threshold, the average of all the thresholds. The plots in Fig. 3 show strong correlation between the two schemes. When assessed visually, the single-threshold approach gives better results, but the adaptive method seems to perform better, when the contrast between the spot and its surrounding region is significantly low. This effect was more profound in the Cy3 image than in the Cy5 image. We have also observed that sometimes no spot was detected in Cy3 or Cy5 while there was a spot in the corresponding Cy5 or Cy3 image. In this case, the Cy5 or Cy3 spot region can be used as a mask to compute the expression level in the corresponding image.

In this ongoing study, we intend to compare the proposed method with the existing methods. We have already developed a method (unpublished) to simulate realistic microarray images with reference spot regions. The future study will also include the validation of the proposed method on simulated images.

REFERENCES

- [1] D. V. Nguyen, A. B. Arpat, N. Wang, and R. J. Carroll, "DNA microarray experiments: biological and technological aspects," *Biometrics*, vol. 58, pp. 701-717, 2002.
- [2] J. Buhler, T. Ideker, and D. Haynor, "Improved techniques for finding spots on cDNA microarrays," *University of Washington*, 2000.
- [3] Y. H. Yang, M. J. Buckley, and T. P. Speed, "Analysis of cDNA microarray images," *Briefings in Bioinformatics*, vol. 2, pp. 341-349, 2001.
- [4] M. B. Eisen, "URL:<http://rana.lbl.gov/manuals/ScanAlyzeDoc.pdf>," 1999.
- [5] GenePix4000, "A User's Guide," *Axon Instruments, Inc.* URL:http://www.axon.com/GN_Genomics.html#software, 1999.
- [6] Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *Journal of Biomedical Optics*, vol. 4, pp. 364-374, 1997.
- [7] R. Bemis, "The Mathworks, Inc., Natic, MA. (<http://www.mathworks.com/matlabcentral/fileexchange/index.jsp>),"
- [8] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, pp. 273-285, 1985.
- [9] J. Besag, "On the statistical analysis of dirty images," *J. R. Statist. Soc. B*, vol. 48, pp. 259-302, 1986.
- [10] K. V. Mardia and T. J. Hainsworth, "A spatial thresholding method for image segmentation," *IEEE PAMI*, vol. 10, pp. 919-927, 1988.