Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

Multichannel Segmentation of cDNA Microarray Images using the Bayes Classifier

Nikolaos Giannakeas, *Student member, IEEE* and Dimitrios I. Fotiadis, *Senior Member, IEEE*

***Abstract*—Microarray technology provides a powerful tool for the quantification of the expression level for a large number of genes simultaneously. Image analysis is a crucial step for data extraction of microarray gene expression experiments. In this paper we propose a supervised method for the segmentation of microarray images. The Bayes classifier is employed for a pixel by pixel classification. The proposed method classifies the pixels of the image in two classes, foreground and background pixels. For this task, an informative set of features is used from both green and red channels. The method is evaluated using a set of 5184 spots (consisting of ~15000000 pixels), from the Stanford Microarray Database (SMD) and the reported classification accuracy is 82 %.**

I. Introduction

DNA microarray technology yields expression profiles for thousands of genes, in a single hybridization experiment [1]. The whole process begins with the biological experiment. Two mRNA samples are reverse transcribed into cDNA, and labeled with two difference fluorescent dyes. Next, the samples are hybridized with the known genes at the same time, on a glass slide. These known genes are chosen according to the current biological problem, and are printed on the slide by a robotic arrayer.

The microarray slide is scanned in both wave lengths that the two dyes emit, generating two grayscale images. These images contain several blocks which consist of a number of spots, placed in rows and columns as it is shown in Fig.1. The intensity of each spot provides a measure of hybridization, between the sample and the corresponding gene. Image processing techniques are required to extract the intensities of red and green channels. The processing of microarray images [2], usually consists of the following steps: (i) spot finding and gridding in order to find the location of each spot in the image, (ii) segmentation, that groups the pixels with similar features (separation of foreground and background pixels), (iii) intensity extraction to calculate the red and green foreground fluorescence intensity pairs and background intensities.

Manuscript received March 27, 2007.

N. Giannakeas is with Laboratory of Biological Chemistry, Medical School, University of Ioannina, Ioannina, Greece, GR 45110. (e-mail: me01310@cc.uoi.gr).

D.I. Fotiadis is with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Ioannina, Greece, GR 45110 (0030-26510-98803; fax: 0030-26510-97092; e-mail: fotiadis@cs.uoi.gr).

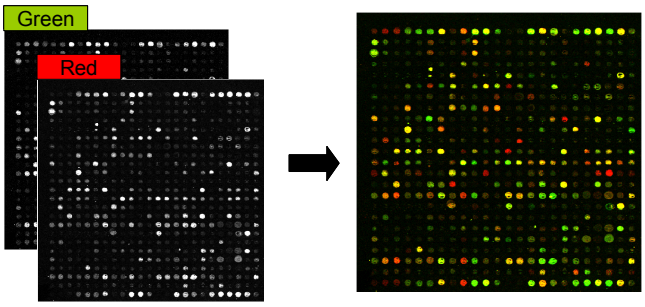

Fig.1: A block of a color microarray image consisting of 24x24 spots.

Segmentation in microarray images results in characterizing the pixels of the image as foreground or background. The intensities of background pixels are used to adjust the foreground intensities for local noise, resulting in corrected red and green intensities for each spot.

Several methods have been proposed for the segmentation of microarray images. These methods can be categorized into four categories: (i) Fixed and adaptive circle, (ii) histogram-based, (iii) adaptive shape, (iv) clustering. Earliest approaches fit a circle (with fixed or adaptive size) around each spot, characterizing the pixels in the circle as signal pixels and the pixels out of the circle as background pixels. Such an approach is used by Scanalyse [3] and Dapple [4]. Histogram-based techniques estimate a threshold [5-6] such that pixels with intensity lower than the calculated threshold are characterized as background pixels, whereas pixels with higher intensity as signal pixels. The adaptive shape segmentation methods are usually based on the watershed transform [7], seed region growing [8-9] and Markov random field [10]. Finally, the most recent techniques employ clustering algorithms like K-means [11], Fuzzy C means (FCM) [12], Expectation-Maximization (EM) [13], Partitioning Around Medoid (PAM) [14] and model-based clustering [15].

In this paper we introduce a novel segmentation method that classifies the pixels of the image into two categories (foreground and background) using the Bayes classifier. A classification-based segmentation method is employed, instead of clustering or other segmentation techniques. Clustering techniques generate groups of pixels, characterized as signal or background using a set of rules, i.e. the group with the maximum mean intensity value is characterized as signal. Instead of this, the proposed approach directly classifies each pixel to the designated category. Another important advantage of this method is the

simultaneous processing of both channels of the image. Moreover, an informative set of features of each pixel is used as input for the classification, in order to deal with the artefacts of the image.

II. Materials and Methods

The proposed method begins with the spot addressing and gridding procedure, in order to isolate each spot from the image. Next, a segmentation approach based on the Bayes classifier is applied on each spot separately. For this task, a set of features is selected using both the red and green grayscale images. This set of features is used for the classification of the pixels of the image into foreground and background. The flowchart of the proposed method is shown in Fig.2

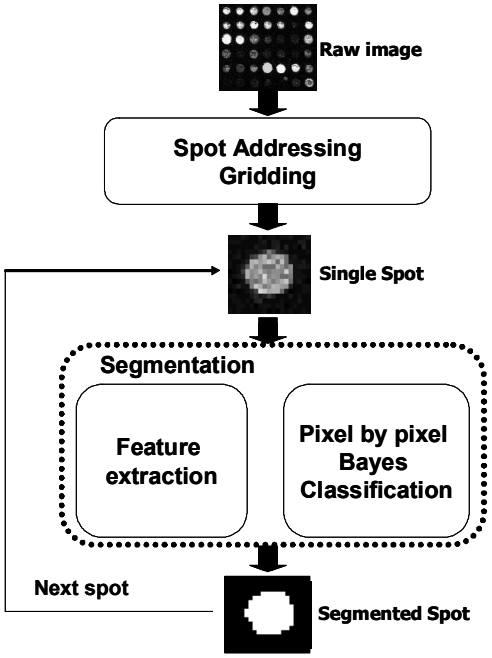

Fig.2: The flowchart of the proposed method.

A. SPOT ADDRESSING AND GRIDDING

The applied spot addressing and gridding procedure consists of several steps, dealing with the noise and the artefacts of the image. Briefly (details can be found in [16]), in the first step, the raw microarray image is preprocessed with a template matching technique. In a second step, the blocks of the image are located. The third step is the spot finding in each block. In this step outlier detection is applied on each row and column of spots in order to remove the artefacts. The next step is the detection of the non-expressed spots. Finally, a grid is fit on the image using a Voronoi diagram. Fig.2 shows the grid that the above procedure produces.

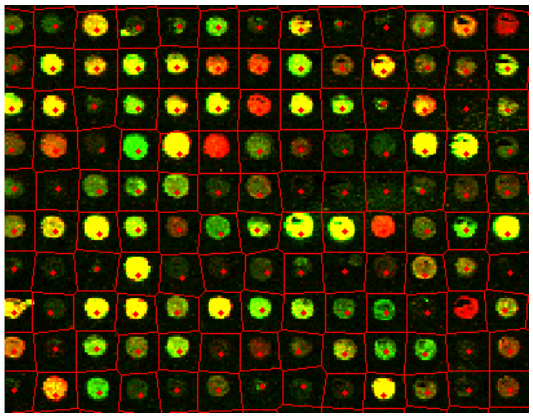

Fig.3: Voronoi diagram in a microarray image.

B. SEGMENTATION

After the gridding procedure, each Voronoi cell determines an area around each spot. In this area, the segmentation technique is applied. Each pixel in the area, is represented by a feature vector and then, it is classified as background or foreground.

1) Feature extraction

The features which are used can be categorized in three main categories: (i) features related to the intensity value of each pixel. The intensity of the pixel, the mean value of the neighborhood of the pixel and the standard deviation of the neighborhood of the pixel, are used as intensity features. (ii) spatial features, such as the Euclidean distance between the position of the pixel and the center of the corresponding spot and finally, (iii) features related to the shape of the theoretical spot. This category contains the correlation of the pixel's 11x11 neighbourhood with an 11x11, 2D-gaussian template. This is a measure of similarity between the neighborhood and the shape of a typical spot. All the above features, except the spatial ones, are used in both channels of the image, in order to segment the two channels at once. Spatial features are the same for both channels. A detailed description of the above features is given in Table I.

2) Bayes Classifier

Using the Bayes classifier [17], we classify a set of pixels into two different classes. The classification is based on the above described features of the pixels. In this case, 11 features are used to characterize each pixel. The concept of Bayes classifier is to estimate the *a posteriori* probability of a sample (pixel) to belong in a class. The *a posteriori* probability is given by the Bayes theorem:

$$P(w_i \mid x) = \frac{p(x \mid w_i)P(w_i)}{\sum_{i=1}^{2} p(x \mid w_i)P(w_i)} = \frac{p(x \mid w_i)P(w_i)}{p(x)}, \quad (1)$$

Table I: Extracted features

FEATURE TYPE	CHANNEL		DESCRIPTION
Intensity features	GREEN	1	Intensity of the pixel
		2	Mean intensity value of the 3x3 neighbor of the pixel
		3	Intensity standard deviation of the 3x3 neighbor of the pixel
	RED	4	Intensity of the pixel
		5	Mean intensity value of the 3x3 neighbor of the pixel
		6	Intensity standard deviation of the 3x3 neighbor of the pixel
Spatial Features		7	x-coordinate of the pixel in the image
		8	y-coordinate of the pixel in the image
		9	Euclidean distance between the pixel and the center of the spot
Shape features	GREEN	10	Correlation of the neighbor of the pixel and the Gaussian template
	RED	11	Correlation of the neighbor of the pixel and the Gaussian template

where, $x \in R^{11}$ is the feature vector, $w_i : i = 1,2$ are the two classes, $P(w_i)$ is the *a priori* probability that an arbitrary sample belongs to class w_i, $P(w_i \mid x)$ is the *a posteriori* conditional probability that a specific sample belongs to a class, $p(x)$ is the density distribution of all samples, and $p(x \mid w_i)$ is the conditional density distribution of all samples belonging to w_i.

The theorem is applicable for all probability density functions, however, it depends on the nature of the data. The Gaussian density function is often used to model the distribution of feature values of a particular class. The general multivariate Gaussian density function is given as:

$$p(x) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}, \quad (2)$$

where D is the dimension of the feature vector (D=11 in our case). μ_i and Σ_i are the mean vector and the covariance matrix of the features of the corresponding class respectively:

$$\mu_i = \frac{1}{N_i}\sum_{x \in w_i} x, \qquad \mu_i \in \Re^{11}, \quad (3)$$

$$\Sigma_i = \frac{1}{N_i}\sum_{x \in w_i} xx^T - \mu_i \mu_i^T, \qquad \Sigma_i \in \Re^{11\times 11}, \quad (4)$$

where N_i is the number of pixels belonging to class w_i

In the training stage, the proposed approach estimates the mean vector and the covariance matrix for each class. Given the mean vector, the covariance matrix and the gaussian density distribution, the *a posteriori* probability is estimated for each sample and each class in the testing stage. The class where the object belongs is given by the Bayes Decision Rule which is:

$$P(w_i \mid x) > P(w_j \mid x), \forall j \neq i. \quad (5)$$

III. DATASET AND RESULTS

For the evaluation of the proposed approach, we have randomly chosen 11 blocks from the SMD [18]. Every block consists of 576 spots, forming 24x24 rows and columns. Two of the blocks are used for the training (1152 spots and ~350000 pixels) and 9 blocks for the testing (5184 spots and ~1500000 pixels). Scanalyse [5] has been used for the initial annotation of the images. In order to extract pixel by pixel information from the annotation, we simulate the fixed circle segmentation that is used by Scanalyse. For this task, the known radius of the fixed circle and the coordinates of the centres of each spot are used. A binary map is generated for the whole block, characterizing the pixels inside the circle as signal pixels and the pixel outside of the circle as background.

Fig. 4 shows several high or low expressed spots. It contains the two raw greyscale images, which are generated from the two channels, the extracted annotation image and the result of the classification-based segmentation.

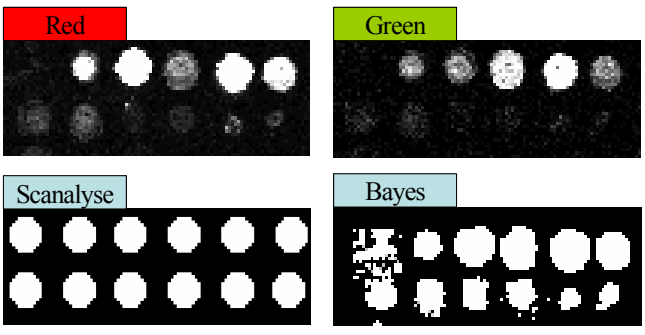

Fig.4. Segmentation results in several spots using the Bayes classifier approach.

In order to validate our method, sensitivity, specificity and accuracy are employed:

$$Se = \frac{\#of\ correctly\ identified\ signal\ pixels}{total\ \#\ of\ signal\ pixels}, \quad (6)$$

$$Sp = \frac{\#of\ correctly\ identified\ background\ pixels}{total\ \#\ of\ background\ pixels}, \quad (7)$$

$$Acc = \frac{\#of\ correctly\ identified\ pixels}{total\ \#\ of\ pixels}. \quad (8)$$

The results for each block for the pixel by pixel classification are presented in Table II.

Table II: Results obtained when evaluating our method in a set of 9 image blocks, acquired from the SMD.

Block	Se (%)	Sp (%)	Acc (%)
lc4b069rex2B1	87.0	74.8	83.1
lc4b069rex2B2	82.8	74.8	80.3
lc4b069rex2B3	88.7	72.5	83.5
lc4b069rex2B5	84.2	62.0	77.0
lc4b069rex2B6	88.7	76.1	84.6
lc4b069rex2B7	89.9	74.4	84.9
lc4b069rex2B8	92.1	69.4	84.7
lc4b069rex2B9	87.7	74.4	83.4
lc4b069rex2B15	82.4	73.6	79.6
Overall	**87.1**	**72.4**	**82.3**

IV. DISCUSSION

In this paper, a supervised classification-based method for the segmentation of microarray images is presented. A set of features for each pixel is extracted and the Bayes classifier classifies the pixels of the image as signal and background. The two channels are processed simultaneously, using features from both the grayscale images, providing a multichannel approach for microarray image segmentation.

The proposed method detects efficiently the signal pixels of high expressed spots as it shown in Fig. 4. However, several low-expressed spots are segmented incorrectly, such as the first spot of the first row in the image. As it is shown in Table II, better results are reported for specificity. The main reason for this is the imbalanced dataset, i.e. it contains a large number of background pixels compared to the signal ones. Unfortunately, none of the already developed segmentation methods present pixel by pixel accuracy results, since there exist no databases or programs to provide pixel by pixel annotation. Thus, a comparison of our method with existing methods is not possible.

The absence of a database that contains annotation in a pixel by pixel (i.e. signal, background or artefact) basis limits the improvements of the proposed work. The existence of such a database would be beneficial for the application of classification based segmentation techniques in microarray imaging. This constrain could be overcome using a dataset of simulated Microarray images, where the annotation of the artefacts would be extracted during the simulation. Finally, more advanced classification methods would be employed.

V. REFERENCES

[1] M.B. Eisen, and P.O. Brown, "DNA Arrays for Analysis of Gene Expression," *Methods Enzymol,* vol. 303, pp.179-205, 1999.

[2] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative motoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, pp. 467-470, 1995.

[3] M.B. Eisen, ScanAlyse, http://rana.Stanford.EDU/software/, 1999.

[4] J. Buhler, T. Ideker, and D. Haynor, "Dapple: improved techniques for finding spots on DNA microarrays," *UWCSE Tech Rep. UWTR Dept. of Computer Science and Eng.,* University of Washington, 2000.

[5] QuantArray Analysis Software, http://lifesciences.perkinelmer.com.

[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.2(4), pp.364–374, 1997.

[7] K.I. Siddiqui, A. Hero, and M. Siddiqui, "Mathematical Morphology applied to Spot Segmentation and Quantification of Gene Microarray Images," *in Proceedings of Asilomar Conference on Signals and Systems*, 2002.

[8] M.J. Buckley, Spot User's Guide, *CSIRO Mathematical and Information Sciences*, Sydney, Australia, 2000.

[9] X. Wang, S. Ghosh, and S.W. Guo, "Quantitative quality control in microarray image processing and data acquisition," *Nucleic Acids Research,* vol. 29(15), pp. E75-E82, 2001.

[10] O. Demirkaya, M.H. Asyali, M.M. Shoukri, and K.S. Abu-Khabar, "Segmentation of microarray cDNA spots using MRF-based method," *in Proc. of the 25th Anual Int. Conf. of the IEEE EMBS*, vol.1, pp.674-677, 2003.

[11] D. Bozinov, and J. Rahnenfuhrer, "Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering," *Bioinformatics,* vol. 18, pp. 747–756, 2002.

[12] E. Ergüt, Y. Yardimci, E. Mumcuoglu, O. Konu, "Analysis of microarray imagesusing FCM and K-means clustering algorithm," in Proc IJCI, pp.116-121, 2003.

[13] K. Blekas, N. Galatsanos, A. Likas, and I.E. Lagaris, "Mixture Model Analysis of DNA Microarray Images," *IEEE Transactions on Medical Imaging,* vol. 24(7), pp. 901-909, 2005.

[14] R. Nagarajan, "Intensity-Based Segmentation of Microarray Images," *IEEE Trans. Med. Imag*., vol.22, pp. 882–889, 2003.

[15] 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, pp.2875–2882, 2005.

[16] N. Giannakeas, D.I. Fotiadis, and A.S. Politou, "An Automated Method for Gridding in Microarray Images," *in Proc. EBMS – IEEE*, pp.5876-5879, 2006.

[17] R.C. Gonzalez, R.E. Woods, and S.L. Eddins, "Digital image processing using MATLAB*", Prentice Hall,* Upper Saddle River, NJ, 2004.

[18] J. Gollub, C. A. Ball, G. Binkley, K. Demeter, D. B. Finkelstein, J. M. Hebert, T. Hernandez-Boussard, H. Jin, M. Kaplper, J. C. Matese, M. Schroeder, P. O. Brown, D. Botstein, and G. Sherlock, "The Stanford Microarray Database: data access and quality assessment tools," *Nucleic Acids Res*., vol. 31, pp. 94-96, 2003.