

COMPARATIVE EVALUATION OF A GAUSSIAN MIXTURE MODELS AND A SEEDED REGION GROWING TECHNIQUES FOR THE SEGMENTATION OF MICROARRAY IMAGES

E. Athanasiadis^{*}, A. Daskalakis^{*}, P. Spyridonos^{*}, D. Glotsos^{*}, I. Kalatzis^{**}, D. Cavouras^{**} and G. Nikiforidis^{*}

^{*} Medical Image Processing and Analysis Group, Laboratory of Medical Physics, School of Medicine, University of Patras, 26500 Patras, Greece
e-mail: mathan@upatras.gr, web page: <http://mipa.med.upatras.gr>

^{**} Medical Image and Signal Processing Laboratory, Department of Medical Instruments Technology, Technological Educational Institute of Athens.
e-mail: cavouras@teiath.gr, web page: <http://medisp.bme.teiath.gr>

The purpose of the present study was to investigate and compare the segmentation ability of the Gaussian Mixture Models (GMM) against the Seeded Region Growing (SRG) methods in microarray spots segmentation. A simulated microarray image, each containing 200 spots, was produced. An automatic gridding process was developed in MATLAB and it was applied on the images for identifying the centers of spots and their surrounding borders (cells). The GMM, developed in MATLAB and the SRG algorithms, using MAGIC Tool software, were applied to each spot separately for discriminating foreground from background. The segmentation abilities of the GMM and SRG algorithms were evaluated by calculating the segmentation matching factor for each spot. Optimal segmentation results were obtained by the GMM, especially in cases where the spot's mean intensity value was close to the background. The GMM technique was found to be an accurate algorithm in delineating the boundary of microarray spots and, thus, in discriminating the spot from its surrounding background.

Introduction

In the field of bioinformatics, microarray imaging is used for the identification of thousands of genes simultaneously [1]. The identification of Gene is closely associated with the identification of spot. By locating the spot in a complementary DNA (cDNA) microarray, measurements, such as the mean or median fluorescence intensity values, are obtained. These measurements are related to the abundance of the messenger RNA (mRNA).

For the task of measuring spot intensity values, three major steps are taken [1, 2]: First, the *gridding step*,

where the exact location of each spot is located. Second, the *segmentation step*, in which foreground and background recognition of each spots is accomplished, and, finally, the *intensity extraction* step, where the mean fluorescence value is calculated for each spot.

In the past, several algorithms and software packets have been developed for the task of processing microarray images [3, 4, 5, 6, 7]. In the ScanAlyze [3] packet, a *fixed circle segmentation* method is used, where all spots are considered circular with a fixed predefined radius. In the GenePix [4] packet, an *adaptive circle segmentation* technique is employed and the radius of each spot is suitably determined. In the Spot [6] packet, an *adaptive shape segmentation* technique is followed. The most representative algorithms employed for that technique are watershed [8] and seeded region growing [9]. In the ImaGene [7] packet, a *histogram based segmentation* method is applied, in which the 80th and 95th percentiles contribute to the calculation of the mean intensity value. In all those techniques, the major disadvantages are either that spots are not circular, or a-priori knowledge of the precise position of centers is not accurately defined [10].

The purpose of the present study was to investigate and compare the segmentation ability of the *Gaussian Mixture Models* (GMM) [11] method against that of the *Seeded Region Growing* (SRG) [9] method in microarray spots segmentation. Evaluation of the results was achieved by calculating the segmentation matching factor [13]

Material and Methods

On complementary DNA (cDNA) Microarray experiments [12], two messenger RNA (mRNA) samples are first reverse transcribed into cDNA. Next, the two samples are labeled using two different fluorescence dyes, Cyanine Cy3 (red channel) and Cy5

(green channel) respectively. The two samples are hybridized [1] and after the scanning procedure, two colored fluorescence Tagged Image File Format (TIFF) images are produced for each channel [1]. The whole procedure is illustrated in Figure 1.

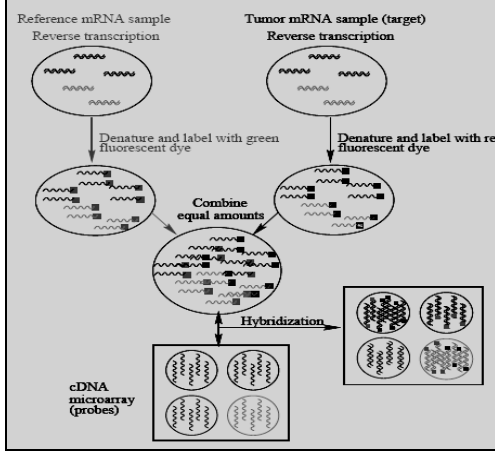


Figure 1. Hybridization of cDNA Microarray images. [13]

The fluorescence intensity value of each spot is related to the expression abundance of the corresponding DNA sequence. In the present study, a microarray image of 200 spots (16-bit colored TIFF image) was created using the *Microarray Scan Simulator* [14].

An automatic gridding procedure was developed, as described in K. Blekas et. al. [15], according to which, line profiles of the horizontal (x) and vertical (y) axes of the hybridized image were calculated. Then, a low-pass filter was applied to smooth the two signals. Sub-array boundaries were obtained by finding the local minima of each signal and the procedure was repeated within each sub-array separately. Identification of the centers and the boundaries of each spot were calculated by finding the local maxima and minima of each signal respectively.

The K-means clustering [15, 16] algorithm was firstly employed and two classes ($K=2$) were produced, the foreground (FG) and the background (BG) class corresponding to the spot and surrounding areas respectively. For each class j , the Probability Density Function (PDF) was estimated. Classification based on Bayes' rule was then performed, according to Equation 1.

$$P(C_j | X) = p(X | C_j) \frac{P(C_j)}{p(X)} \quad \text{Eq. 1}$$

where, $P(X|C_j)$ is the PDF of class j evaluated at X , $P(C_j)$ is the prior probability of for class j , and $p(X)$ is the overall PDF evaluated at X .

GMM estimates $P(X|C_j)$ as a weighted average of multiple Gaussians. The estimation of the PDF of each class j is calculated by Equation 2.

$$p(X | C_j) = \sum_{k=1}^{N_c} \pi_k G_k \quad \text{Eq. 2}$$

where π_k is a weighting factor of the k -th Gaussian G_k ,

described in Equation 3. N_c is considered to be the number of features belonging to class j .

For the initialization of the GMM algorithm, an estimation of the mean value μ_k , the covariance matrix V_k , both estimated for each class k produced by k-means, and finally π_k , that initially is set to 0.5 for both classes was performed.

$$G_k = \frac{1}{(2\pi)^{n/2} |V_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T V_k^{-1} (x-\mu_k)} \quad \text{Eq. 3}$$

where T is the transpose matrix and n is the number of features used.

The algorithm was next divided into two steps, the Estimation Step (*E-Step*) and the Maximization Step (*M-Step*). In the E-Step, an estimation of the prior probability of every pixel value (pattern X_p) in cluster i was made, according to Equation 4. In the M-Step, the model's parameters θ (*i.e.* π , μ , V) were recomputed, according to Equations 5, 6 and 7.

$$t_{ip} = P(G_i | X_p) = \frac{\pi_i p(X_p | \theta_i, C_i)}{p(X_p)} = \frac{\pi_i p(X_p | \theta_i, C_i)}{\sum_{j=1}^{N_c} \pi_j p(X_p | \theta_j, C_j)} \quad \text{Eq. 4}$$

$$\pi_i(t+1) = \frac{1}{N_c} \sum_{p=1}^{N_c} t_{ip}(t) \quad \text{Eq. 5}$$

$$\mu_i(t+1) = \frac{\sum_{p=1}^{N_c} t_{ip}(t) X_p}{\sum_{p=1}^{N_c} t_{ip}(t)} \quad \text{Eq. 6}$$

$$V_i(t+1) = \frac{1}{N_c \pi_i(t)} \sum_{p=1}^{N_c} t_{ip}(t) [(X_p - \mu_i(t))(X_p - \mu_i(t))^T] \quad \text{Eq. 7}$$

Steps E and M were repeated for each iteration $t+1$, until $V_i(t+1) = V_i(t)$ or the number of iterations reached a specific predefined value. For each pattern X_p , the confidence (Equation 8) for each class j was computed and the pattern was classified to the class with the highest confidence.

$$P(C_j) p(X | \theta_j, C_j) \quad \text{Eq. 8}$$

The output of the algorithm was a binary image; white color for the BG and black color for the FG class.

The accuracy of segmentation was numerically calculated by using the segmentation matching factor (Equation 9) [13].

$$accuracy = \frac{A_{calc} \cap A_{real}}{A_{calc} \cup A_{real}} \times 100 \quad \text{Eq. 9}$$

where A_{calc} is the area of the spot as determined by the proposed algorithm and A_{real} is the actual spot area. A perfect match is indicated by a 100% score, any score higher than 50% indicates reasonable segmentation [13] while a score of less than 50% indicates poor segmentation [13].

Results and discussion

According to our results, the GMM algorithm was more successful than the SRG method in delineating spots. The SRG algorithm employed was that incorporated in the Magic Tool [17] software. It is clear from the calculated matching factors (see Table 1), that in cases where the FG value was close to the BG value, the proposed methodology managed to delineate spots in a more accurate way.

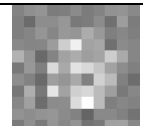


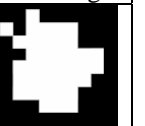

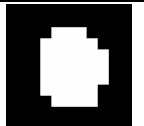

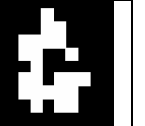




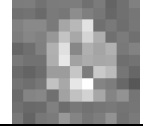
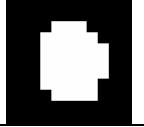
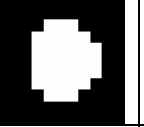
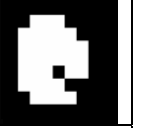
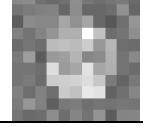
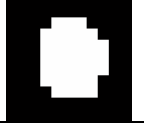
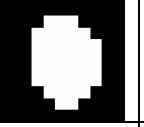
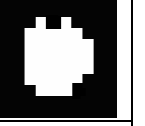
Original Cell	Real Boundaries	GMM	Seeded Region Growing
			
spot 1	Accuracy:	97,5207	91,7355
			
spot 2	Accuracy:	95,8678	88,4297
			
spot 3	Accuracy:	94,2149	90,9091
			
spot 4	Accuracy:	99,1736	97,5207
			
spot 5	Accuracy:	97,5207	97,5207

Table 1: Comparative results for 5 identical spots (G channel). The first column indicates the simulated spot, the second column indicates the actual boundaries of spot and finally third and fourth columns present the results of GMM and Seeded Region Growing (SRG) algorithms (using the MAGIC Tool), as well as the calculated corresponding matching factors, respectively

Conclusion

The GMM technique was found to be an accurate algorithm in delineating the boundary of microarray spots and, thus, in discriminating the spot from its surrounding background. Moreover, the processing time was less than 20 sec. on a typical desktop PC (Pentium 4, 2.3GHz, 512MB RAM) for the 200 spots.

Acknowledgements

We would like to thank the Greek State Scholarships Foundation (I.K.Y.) for funding the above work.

References

- [1] Y.H. Yang, M. J. Buckley, S. Duboit and T.P.Speed (2002), "Comparison of methods for Image Analysis on cDNA Microarray Data", Journal of Computational and Graphical Statistics, vol. 11, pp 108-136.
- [2] Y.H. Yang, M. J. Buckley, S. Duboit and T.P.Speed (2001), "Analysis of cDNA Microarray images", Briefing in Bioinformatics, vol.2, No.4, pp. 341-349.
- [3] M.B. Eisen, ScanAlyze (1999). <http://rana.lbl.gov/EisenSoftware.htm>
- [4] Axon Instruments, Inc. (1999): GenPix 4000A User's guide
- [5] GeneSifter data center: <http://www.genesifter.net/web/dataCenter.html>
- [6] M.J. Buckley (2000), The Spot user's guide. CSIRO Mathematical and Information Science. <http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm>
- [7] ImaGene, ImaGene 6.1 User Manual <http://www.biodiscovery.com/index/papps-webfiles-action>.
- [8] S. Beucher, F. Meyer (1993). "The morphological approach to segmentation: The watershed transformation", Optical Engineering, Vol. 34, pp. 433-481.
- [9] R. Adams and L. Bischof (1994). "Seeded Region Growing", IEEE Trans. Pattern Anal. Machine Intell, vol 16, pp 641-647.
- [10] D. Bozinov and J. Rahenfuhrer (2002), "Unsupervised technique for robust target separation and analysis of DNA Microarray spots through adaptive pixel clustering", Journal of Bioinformatics, vol 18, pp 747-756.
- [11] K. Blekas, N.P. Galatsanos and I. Georgiou (2003), "An unsupervised Artifact Correction Approach for the Analysis of DNA Microarray Images", in Proc. IEEE International Conf. on Image Processing (ICIP), vol 2, pp 165-168.
- [12] M. Schena, D. Shalon, R.W. Davis and P. O. Brown (1995), "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", Science 270, pp. 467-470.
- [13] Sandrine Dudoit, Jane Fridlyand and Terry Speed (2002), "Comparison of discrimination methods for the classification of tumours using gene expression data". Journal of the American Statistical Association, 97(457):77-87

- [14] Bill Martin and Robert M. Horton (2004), “A Java Program to Create Simulated Microarray Images”, Proceedings on the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004).
- [15] K. Blekas, N. Galatsanos, A. Likas, and I.E. Lagaris, (2005) “*Mixture Model Analysis of DNA Microarray Images*”, IEEE Transactions on Medical Imaging, vol 24. pp. 901-907.
- [16] R. Nagarajan, C. Peterson, (2002) “*Identifying Spots in Microarray Images*”, IEEE Transactions on nanobioscience, Vol. 1, No.2, pp. 78-84.
- [17] L. J. Heyer, D. Z. Moskowitz, J. A. Abele, P. Karnik, D. Choi, A. M. Campbell, E. E. Oldham and B. K. Akin, (2005) “MAGIC Tool: integrated microarray data analysis”, Bioinformatics, Vol 21, no. 9, pp. 2114 – 2115.