Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Improved Microarray Spot Segmentation by Combining two Information Channels

Th. Margaritis, K. Marias, *Member, IEEE,* and D. Kafetzopoulos.

Abstract— **High-throughput gene expression is an important aspect of modern post-genomic research. Microarray technology is the driving force of this revolution, a technology that allows the simultaneous monitoring of expression for thousands of genes. The need for accurate and reproducible research has driven the development of robust analysis frameworks for maximizing the information content of biological data. In microarray imaging technologies, several non-linearities in the experimental process render the measured expression values prone to variability and often, to poor reproducibility. Accurate segmentation of the true signal is a very important task, not least because a single value per spot needs to be derived for further knowledge discovery analysis. In this paper, we present a fully automatic segmentation method for improving the spot segmentation result. The method doesn't make any assumptions concerning the number of classes present in each image spot, and it isn't driven only by the most intense features, since it takes into account the underlying "hybridization ground truth" derived from both information channels of the spotted arrays. Our method is compared to widely used, state-of-the-art segmentation methods in microarray image analysis in a study of a metabolic disorder in yeast, where replicates of reporters are present. Initial results indicate that our method yields more reproducible log ratio measurements across replicates.**

I. INTRODUCTION

IN microarrays, an array of DNA reporters is hybridized with labeled samples to study differential expression or patterns of gene expression. The expression of each gene results in increased concentration of the corresponding mRNA. DNA microarrays are used for estimating the concentration of mRNAs of living cells using reporters, that each matches a particular mRNA in the cells. The extracted mRNA is converted to cDNA and then every sample is labeled. For expression analysis, there are many technologies of microarrays production but the field has been dominated by two major technologies, the Affymetrix GeneChips and the spotted microarrays.

The first technology is patented and the manufacturing of probes is done by direct synthesis, photolithographically. The Affymetrix chips can handle only one fluorochrome, so two chips are required for the comparison between two samples. The segmentation of the square probes is straightforward; the signal is the 75th percentile of all but the edge pixels.

The second technology uses manual deposition of probes and the manufacturing can be done even with home-made robots [1]. The user can label multiple samples with different fluorescent dyes. The two most commonly used dyes – also referred as fluorochromes – are the cyanine dyes, Cy3, that is green and Cy5, that is red. After mixing, the labeled samples (usually two) are hybridized to the reporters on the glass slides. Then, the unhybridized material is washed away and the slide is scanned. The scanned area is divided into equally sized pixels and the scanner produces for each dye a digital map (image channel) of the fluorescence intensities for each pixel. Within each spot, pixel intensities represent the relative amount of fluorescent dyes which, in turn, is proportional to the reporter quantity. The spotting and drying procedure introduces spatial variability of the reporter quantity across each spot. However, this quantity distribution is equal in both channels (e.g. Cy3, Cy5), and therefore, the ratio of the corresponding pixel intensities should be constant. This fact is exploited in our segmentation method.

Accurate spot segmentation is an essential analysis step in spotted arrays technologies (a comprehensive review on the subject can be found in [2-3]). The aim is to reduce the image to single gene-expression values per spot, i.e. the log ratio of the fluorescent intensities. Background pixels can underestimate the true expression value of each channel, leading to potentially false negative calls in differential expression. On the other hand, outlier pixels, representing hybridization defects, nearby spots, dust, etc., may overestimate the expression value and create potential false positive calls. The first broadly used method, used a fixed circle segmentation algorithm, included in the ScanAlyze Software (Michael Eisern, Univeristy of Berkeley, California) [4]. This algorithm only works when spots are

Thanasis Margaritis is with the Institute of Molecular Biology, IMBB-FORTH, Vassilika Vouton, P.O.Box 1385, GR 711 10 Heraklion, Greece (email: thama@imbb.forth.gr) and the University of Crete, Biology Department, Vassilika Vouton, P.O.Box 2208, GR 714 09 Heraklion, Greece.

Kostas Marias is with the Institute of Computer Science, ICS-FORTH, Vassilika Vouton, P.O.Box 1385, GR 711 10 Heraklion, Greece (e-mail: kmarias@ics.forth.gr).

Dimitris Kafetzopoulos is with Institute of Molecular Biology, IMBB-FORTH, Vassilika Vouton, P.O.Box 1385, GR 711 10 Heraklion, Greece (email: kafetzo@imbb.forth.gr).

1-4244-0033-3/06/$20.00 ©2006 IEEE.

circular and roughly equal in size, which is far from the truth in both aspects. To cope with this problem, adaptive circle segmentation was developed, usually based on histogram thresholding between background and spot [5]. This method was further enhanced with spatial constraints to eliminate "outlier" pixels, (e.g. the constrained region growing algorithm in ImaGene, one of the most widely-used commercial analysis software [6]). In [3], the existing algorithms (at that time) are compared, and the Spot software is proposed, which uses a seeded region growing (SRG) algorithm for spot segmentation. The latest version of the software includes a second segmentation option, using globally optimal geodesic active contour (GOGAC) [7]. In [8], a model based approach for segmentation is presented, that clusters each spot into 1 (no spot present) to 3 (background, spot and outliers) classes, using the appropriate Gaussian mixture model that maximizes the Bayesian Information Criterion (BIC). This is incorporated into 'spotSegmentation', an open-source software package.

We argue that the gene-expression signal in each microarray spot can't be considered as a single homogenous entity. Therefore, no assumptions should be made concerning the number of classes that might be present within each spot, since this can vary according to the reporter quantity distribution, the background variability, and the presence of artifacts. Additionally, narrowing the initial estimate of classes to a fixed value (as in [8]), means that a number of potential outliers can affect the value of the signal, especially if they are spatially connected to true spot signal regions.

In this paper, we propose a two-channel segmentation framework that aims to provide a more robust and intuitive segmentation. We first introduce a pre-processing step that removes high intensity outliers outside the expected spot area (explained in Section II), thus partially normalizing the dynamic range of values in the Cy3, Cy5 channels. Then, the number of clusters is determined by applying the Bayesian Information Criterion, in the log product of Cy3 and Cy5. The rationale behind this step is that the reporter area in each spot is the same in both channels. Therefore, even weak signals will be significantly amplified with respect to the background, where isolated high intensity noise pixels are random for each channel, and therefore, the log multiplication minimizes their influence.

We also introduce an optimization step in the segmented log ratio image of each spot, where we remove pixels, which differ significantly from the others. This is inspired from the fact that each spot image has the same inhomogeneity in its intensity values emanating from the distribution of the reporter's quantity. As a result, the true log ratio in the pixels of each spot is constant. This in turn also suggests that the log ratio of replicate spots should be the same (even if the reporter's concentration differs). This observation is used for comparing our segmentation results to well know methods, as is described in section 3.

The proposed method is described in the next section.

II. Methods

A. Spot Addressing

In order to automatically extract significant information from microarray images, it is imperative to address each spot separately, in order to be able to compare it to its local background, segment it and compute the log ratio across the channels. A detailed account of our approach for microarray spot addressing is out of the scope of this paper (the authors aim to publish a more extensive account of their image analysis framework later). We summarize the steps that comprise our approach:

a) Channel Registration: This is a potential problem for microarray imaging, (e.g. when individual channels are scanned sequentially, some motion can occur). Since we combine the pixel information from both channels, accurate registration at pixel level is necessary. In order to align the image data, a plethora of 'classic' algorithms is available. We have used an image similarity approach based on previously introduced measures (e.g. mutual information, cross-correlation), in order to geometrically align microarray images.

b) Artificial grid spot addressing: Based on the image resolution and spot spacing in a given experiment, we construct an artificial grid where the spots are perfect circular regions with diameters equal to the average spot diameter of the brighter spots (e.g. automatically defined by thresholding and labeling). Initially, the grid is registered to the registered channels (as in step a), as shown in Fig 1c. The grid is then affine transformed, in order to account for possible skew in the array. This is done by rotating first horizontally and then vertically and calculating the rotation angles for which the alignment signal (sum of rows, columns respectively), is maximized. The calculated binary grid defines the theoretical spot area (hereafter denoted as TSA), as shown in Fig 1d. However, we need to stress that this is not the solution of the segmentation problem, since within each TSA there is inherent variability of pixel values corresponding to different classes. For this reason, in the next step we utilize a square image segment around each TSA.

B. Segmentation

The segmentation algorithm we propose is summarized below:

i. High intensity clusters outside the TSA are removed in both channels by amplifying structures that are brighter than their surroundings. This is based on morphological reconstruction as described in [9]. Then, such clusters are labeled and removed, if their center is outside the TSA. Note that we don't completely remove background (until step v.), so as to capture it's statistics and identify similar regions inside the spot. The effect of this step is illustrated in Fig.3b, where background clusters (visible in Fig3.a), have been removed.

ii. The A, M images are computed from the two channels: Ch1 is denoted as R(i, j) and Ch2 as G(i, j). Then A and M can be calculated for each pixel according to (1,2)

$$A(i,j) = 0.5 \cdot \log_2[R(i,j) \cdot G(i,j)] \tag{1}$$

$$M(i,j) = \log_2[R(i,j) / G(i,j)] \tag{2}$$

iii. Then the Bayesian Information Criterion (BIC) is computed on the log product image A (see [8] for details on this application of BIC), for 1:20 classes segmentation using the well known fuzzy clustering algorithm. The model that corresponds to the highest BIC is chosen to segment image A. In Fig. 2 (last row), the correct number of classes (as is indicated by the BIC calculated in A), is 7. If the number of classes is constrained (e.g. to 3 as in [8]), the segmentation result is incorrect (the doughnut in Fig 2, middle row, is included in the segmentation result although its corresponding M values are clearly different from the rest of the spot), and the selected class contains noisy features mainly on the edges of the spot.

iv. The largest clustered class (in number of pixels) is chosen, provided that more than half of its pixels lie within the TSA. This will ensure that a doughnut won't be selected as the spot signal in low quality printed spots, since it values are clustered together wih the background. Based on this class, Ch1 and Ch2 are segmented. For eliminating any residual background pixels in the segmented class, step i. is repeated.

v. Occasionally, impurities (usually of high intensities), inside the TSA are also segmented as true spot. This affects significantly the measured mean intensity of each channel, and consequently the resulting log ratio of the spot. To address this, we compute the Median Average Deviation (MAD) of the segmented spot M values and eliminate pixels with significantly different log ratios (2 MADN from median, where MADN is MAD/0.6745). This is illustrated in Fig3. d, resulting in a reduction of the log ratio range and removal of the outliers. Fig. 3c shows that the ranges of values in the corrected segmented images are normalized when excluding the M outliers. This has a significant effect in the comparative results presented in the next section.

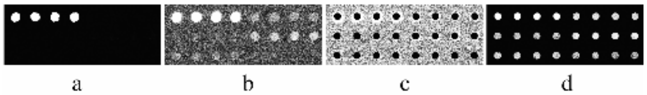

Fig. 1. a: Original image, b: Original image histogram equalized, c) Registration grid (black circles), aligned with original image, d) Corresponding regions of the original image based on the grid determining the theoretical spot area (TSA).

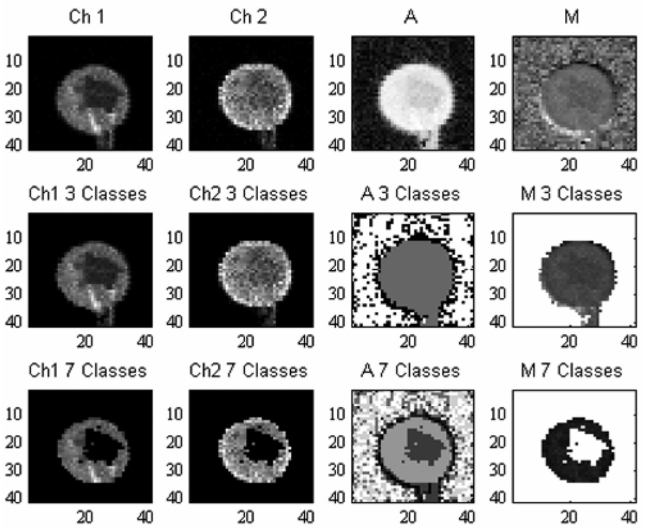

Fig. 2. Top row, left to right: Channel 1 (Ch1) and Ch2 are combined to form the A, M images. Middle row, left to right: Corresponding results when using a constrained, to three classes, segmentation. Last row, left to right: Segmentation results using our method with no constrains on the number of classes (BIC indicates 7 distinct classes in A).

III. Results

For assessing the presented method, 100 spots (41 duplicates and 6 triplicates) were used from a gene expression study of a metabolic disorder in yeast that was performed in our laboratory. In order to define an unbiased dataset, the expert selected half of the cases to be "clearly defined" spots in all replicates (i.e. relatively easy to segment) and the other half "subtle" (i.e. difficult to segment). Our method was compared to the following widely used methods: ImaGene [6], Spot Software with both SRG [3] and GOGAC [7] segmentation algorithms, and spotSegmentation [8]. The log ratio range in replicates was used as a measure for comparing the above methods. Fig. 4 shows the boxplots of the results for all the methods. The calculated mean and standard deviation for each method were: SegA: 0.35±0.22, SegAM: 0.27±0.17, ImaGene: 0.33±0.29, GOGAC: 0.55±0.37, SRG: 0.73±0.54, spotSegmentation: 0.43±0.32 (the software failed to return spot results in 50% of the 'subtle' cases, in at least one of the replicates analyzed). Note that in Fig. 4, we have reported our results with (SegAM), and without (SegA), the optimization step (Section IIB v), in order to assess its added value.

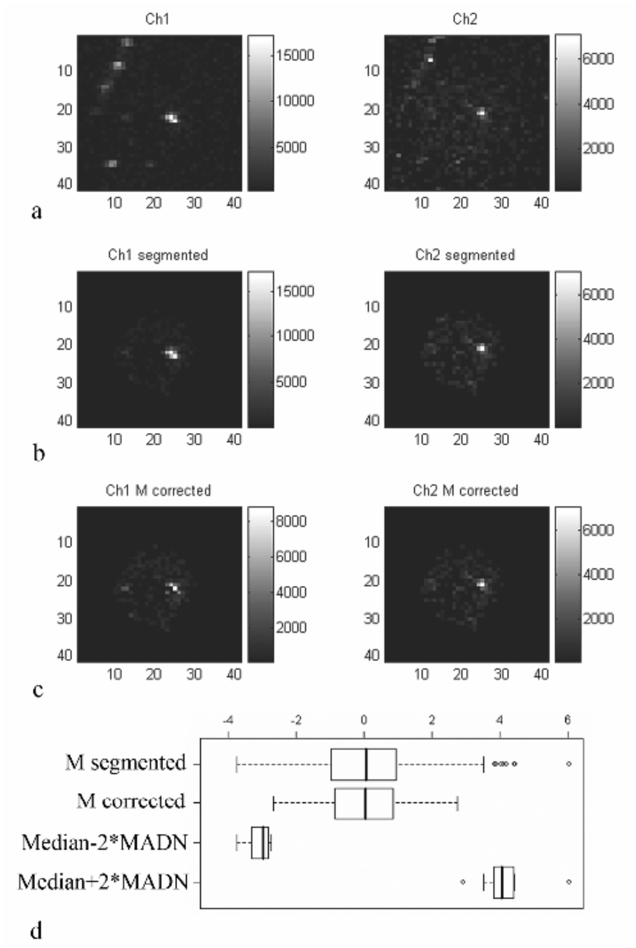

Fig. 3. a: The original spot channels, b: Corresponding segmented regions, c: Optimization of the segmentation result by robustly removing M outliers, as is shown in the corresponding boxplots in d. Note the gradual reduction of the gray-scale dynamic range from a to c in the noisy Ch1.

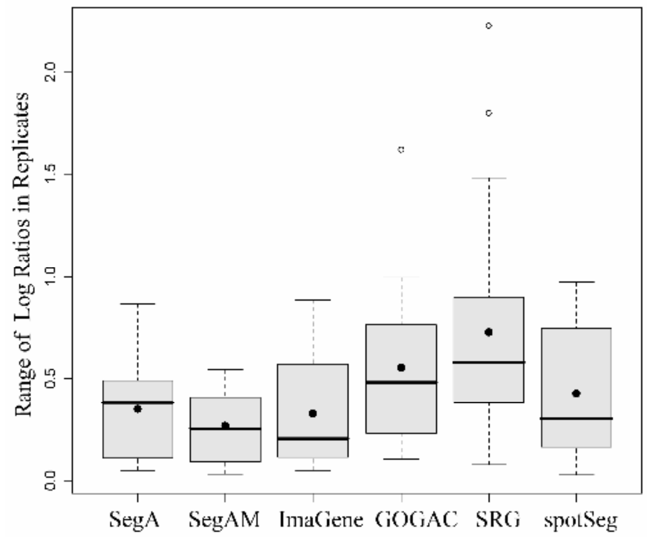

Fig. 4. Comparison of different segmentation methods by using the range of log ratios in replicate spots as a measure of reproducibility. The range should be equal to zero in the ideal case.

IV. Discussion

A novel method for spot segmentation is presented. The method uses both information channels without any assumptions for the number of classes present, which may compromise the result (see Fig. 2). Since both segmented image channels have the same spatial extent, the log ratio M can be estimated as an array, and this allows us to optimize the segmentation result on the basis of further removing image values that drive the log ratio outside a 2 MADN from median interval. This robust estimation of outliers renders the final segmentation result more homogenous with respect to the range of corresponding M values. This is in line with the theoretical concept that the segmented channel values should follow the reporter quantity distribution within each spot, while log ratios remain constant.

The initial results indicate that our method can achieve the smallest range of estimated log ratios in replicate spots, which is a measure of increased reproducibility. We aim to validate this method in a large number of replicate spots, and when possible in the same data used in previous publications, for direct comparison. We believe that reduction of variability between replicate spots increases the ability to detect differential expression, but this will be carefully validated in our future work. We also argue that the range of log ratios in replicates used in this paper for assessing different segmentation methods could become a standard quality control measure also providing uncertainty weights for individual differential expression values.

Acknowledgment

This work was supported in part by the E.U. Integrated Project ‘ACGT’, FP6-2005-IST-026996.

References

[1] M.B. Eisen, P.O. Brown., “DNA arrays for analysis of gene expression,” *Methods Enzymol.* 303:179-205,1999.

[2] L. Qin, L. Rueda, A. Ali, A.Ngom, “Spot detection and image segmentation in DNA microarray data,” *Appl Bioinformatics*, 4(1):1-11, 2005.

[3] Y.H. Yang, M.J. Buckley, T.P.Speed, “Analysis of cDNA microarray images,” *Brief Bioinform.*, 2(4):341-9Dec. 2001.

[4] M.Eisen, 1999. Scanalyze. http://rana.lbl.gov/EisenSoftware.htm

[5] Y. Chen, E.R. Dougherty, and M. L. Bittner, “Ratio-based decisions and the quantitative analysiss of cdna microarray images.” *Journal of Biomedical Optics*, 2:364-374, 1997.

[6] http://www.biodiscovery.com/index/imagene

[7] B. Appleton, H Talbot, “Globally optimal Geodesic Active Contours,” *Journal of Mathematical Imaging and Vision* 23: 67–86, 2005.

[8] Q. Li, C. Fraley, R.E. Bumgarner, K.Y. Yeung, A.E.Raftery, “Donuts, scratches and blanks: robust model-based segmentation of microarray images,” *Bioinformatics*, Jun 15;21(12):2875-82, 2005.

[9] P. Soille, *Morphological Image Analysis: Principles and Applications*, Springer, 1999, pp. 164-165.