

Genes expression level quantification using a spot-based algorithmic pipeline

Antonis Daskalakis, Dionisis Cavouras, Panagiotis Bougioukos, Spiros Kostopoulos, Pantelis Georgiadis, Ioannis Kalatzis, George Kagadis and George Nikiforidis

Abstract—An efficient spot-based (SB) algorithmic pipeline of clustering, enhancement, and segmentation techniques was developed to quantify gene expression levels in microarray images. The SB-pipeline employed i/a gridding procedure to locate spot-regions, ii/a clustering algorithm (enhanced fuzzy c-means or EnFCM) to roughly segment spots from background and estimate background noise and spot's center, iii/ an adaptive histogram modification technique to accentuate spot's boundaries, and iv/a segmentation algorithm (Seeded Region Growing or SRG), to extract microarray spots' intensities. Extracted intensities were comparatively evaluated in term of Mean Absolute Error (MAE) against the MAGIC TOOL's SRG employing a dataset of 7 replicated microarray images (6400 spots each). MAE box-plots mean values were 0.254 and 0.630 for the SB-pipeline and the MAGIC TOOL respectively. Total processing times for the dataset evaluated (7 images) were 2100 seconds and 3410 seconds for the SB-pipeline and MAGIC TOOL respectively.

I. INTRODUCTION

DNA microarray technologies are hybridization based methods that enable the simultaneous assessment of the expression levels of thousands of genes [1]. In this way, microarrays provide an easy way to compare gene expression profiles between biological samples, by detecting either their expression or differential expression. In a microarray slide, thousands of discrete DNA sequences are printed by a robotic arrayer, thus, forming circular spots of known diameter. To compare the relative abundance of each of these gene sequences in two DNA samples, the two samples are 1/ labeled one with red and the other with green fluorescent dye, 2/ mixed, and 3/ competitively hybridized to the microarray slide. The end product of the comparative hybridization experiment is scanned, using lasers that excite each dye at the appropriate wavelength. The relative fluorescence intensity between each dye on each spot,

which, in turn, represents the relative expression level of the corresponding gene in both samples, is recorded in 2 array images, one for each dye [2].

Relative intensities are extracted through an image analysis workflow, which includes gridding, spot segmentation, and intensity extraction techniques [3]. Initially, the coordinates of each rectangular region containing the spots are located (gridding). Region's pixels are then classified as either signal (spot's foreground) or surrounding area (spot's background) and, from the spot's foreground, the fluorescence signal's mean intensity is calculated. Intensities correspond to gene expression levels that, in turn, are translated into biological conclusions from molecular biologists, by employing data mining techniques.

Microarray experiments involve a number of error-prone steps (occurring during fabrication, target labeling, and hybridization), which induce noise on the resulting images [4]. Microarray images are also corrupted by irregularities in the shape, size, and position of the spot [4]. Unless these sources of error are addressed, they will propagate throughout the stages of the analysis, leading to inaccurate biological inferences.

One of the most undesirable effects of noise is that it contributes to inaccurate spot segmentation, which leads to wrong estimation of the mean spots' intensities and reduces the reproducibility of the gene expression levels, derived from microarray images [5]. Existing software tools, utilized for the analysis of microarray images, focus mainly on accurate spot localization and segmentation by various segmentation techniques [6-13]. Despite the variety of image preprocessing techniques [14] in addressing noise effects, only few studies [15-18] have examined their impact on cDNA images. Those studies have evaluated the benefits of pre-processing techniques in terms of image quality improvement without, however, evaluating their effect on facilitating spot segmentation and, thus, in increasing the accuracy of the extracted gene expression levels.

In the present study, an efficient spot-based (SB) algorithmic pipeline of clustering, enhancement, and segmentation techniques was developed to quantify gene expression levels in microarray images. The SB-pipeline utilizes i/a gridding procedure to locate spot-regions, ii/a clustering algorithm (the Enhanced fuzzy c-means [19]) to crudely segment spots from background and to estimate

Manuscript received March 30, 2007. This work was supported in part by the General Secretariat for Research and Technology, Greece (Grant PENED 2003/136 jointly funded with the European Union).

A. Daskalakis, P. Bougioukos, S. Kostopoulos, P. Georgiadis, G. Kagadis and G. Nikiforidis are with the Medical Image Processing and Analysis Group (M.I.P.A.), Department of Medical Physics, School of Medicine, University of Patras, Rio, GR-26503 Greece (correspondence author; phone: 2610-995012; e-mail: daskalakis@med.upatras.gr).

Dionisis Cavouras, I. Kalatzis is with the Medical Signal and Image Processing Lab, Department of Medical Instruments Technology, Technological Educational Institute of Athens, Ag. Spyridonos Street, Aigaleo, 122 10, Athens, Greece (e-mail: cavouras@teiath.gr).

background noise and spot's center, iii/ an adaptive histogram modification technique [14] to accentuate spot's boundaries, and iv/a segmentation algorithm (Seeded Region Growing or SRG [20]) in order to extract microarray spots' intensities and quantify individual genes' expression levels. The efficiency of the proposed SB-pipeline was comparatively evaluated against the MAGIC TOOL software [21], using a set of publicly available microarray cDNA images [22].

II. MATERIAL AND METHODS

A publicly available dataset of 7 microarray images, downloaded from the MicroArray Genome Imaging & Clustering Tool (MAGIC) website [22], was employed to quantify the efficiency of the proposed SB-pipeline. Each image of the dataset contained 6400 spots investigating the diauxic shift of *Saccharomyces cerevisiae*. In the particular dataset, the authors [23] have used a common reference messenger RNA pool (green, Cy-3) to control for biological variability [24]. Such a design provides an adequate degree of replication (7 images), required for the quantitative assessment of image segmentation and subsequent gene quantification.

The procedure followed in the proposed algorithmic pipeline was influenced by the fact that each spot has unique characteristics (biological and signal dependent noise) that directly affect its expression level [25]. Therefore, it is necessary for each spot to be localized and processed independently. Moreover, in order to improve contrast between spot and background in each spot-image and, thus, facilitate the segmentation procedure, an intermediate step of adaptive image enhancement was adopted.

A. SB Algorithmic Pipeline

The SB-pipeline was initialized by applying a gridding procedure on the images, following the method proposed in a previous study [26] for dividing the image into rectangular spot-containing regions (spot-images). Within each spot-image, the enhanced fuzzy c-means clustering algorithm (EnFCM) [19] was utilized to discriminate spots from surrounding background. The EnFCM algorithm segmented each spot by minimizing the objective function of equation 1:

$$J_s = \sum_{i=1}^C \sum_{l=1}^q \gamma_l u_{il}^m (\xi_l - v_i)^2 \quad (1)$$

where v_i represents the prototype of the i -th cluster, $u_{i,l}$ represents the fuzzy membership of gray value l with respect to cluster i , $m \in (1, \infty)$ is a weighting exponent on each fuzzy membership ($m=2$ in our case), which controls the degree of fuzziness, q denotes the number of gray levels

of the given image, and γ_l is the number of the pixels having gray value equal to l , with $l=1, \dots, q$.

In equation (1), ξ is a linearly-weighted sum image, formed from the original image and its local neighbor average image as in (eq.2):

$$\xi_k = \frac{1}{1+a} \left(x_k + \frac{a}{N_R} \sum_{j \in N_k} x_j \right) \quad (2)$$

where ξ_k denotes the gray value of the k -th pixel of the image ξ , x_j represents the neighbors of x_k , N_k stands for the set of neighbors falling into a window around x_k ,

$\frac{\sum_{j \in N_k} x_j}{N_R}$ is a neighbor average gray value around x_k , and a is a parameter used to control the effect of the neighboring terms (experimentally determined as $a = 10$).

Subsequently, from the segmented spot and background, the location of the spot's center and local noise (standard deviation of background) respectively were assessed.

Additionally, spot-images were enhanced using an adaptive histogram modification technique [27], which maximized the contrast between spot and background. The steps followed for this procedure were:

Step 1: Each spot-image was divided into a number of non-overlapping contextual regions of equal sizes, experimentally set to be 2x2 pixels.

Step 2: The histogram of each contextual region was calculated.

Step 3: A clip limit, for clipping histograms, was set ($t=0.001$). The clip limit was a threshold parameter by which the contrast of the spot-image could be effectively altered; a higher clip limit increased the spot's-image contrast.

Step 4: Each histogram was redistributed in such a way that its height did not exceed the clip limit.

Step 5: All histograms were modified by the transformation function

$$T(r_k) = \sum_{j=0}^k p_r(r_j) \quad (3)$$

where

$$p_r(r_j) = \frac{n_j}{n} \quad (3)$$

is the probability density function of the input image grayscale value j , n is the total number of pixels in the input image, and n_j is the input pixel number of grayscale value j .

Step 6: The neighboring tiles were combined using bilinear interpolation and the spot-image grayscale values

were altered according to the modified histograms.

In the last stage of the SB-pipeline, the contrast enhanced spot-images were segmented using the SRG algorithm. SRG initially segmented each spot-image into regions of pixels starting from the spot's center, as determined by the EnFCM segmentation procedure. Pixel regions were iteratively augmented by assigning neighboring pixels that satisfied a homogeneity criterion: the neighboring pixels should be 1/10th higher intensity than the mean of background pixels, as determined in the EnFCM segmentation stage and 2/10th intensity close to the mean intensity of the so far seeded region. This iterative procedure of growing pixel regions within each spot-image continued until all pixels of the spot-image were assigned to either the spot or its background.

The spot's boundary, thus determined, was referred to the corresponding spot-image on the original image and the spot's mean intensity was evaluated. This was necessary, since intensities in the processed spot-images were altered by the enhancement process.

B. Evaluation of the SB-pipeline

To the best of the authors' knowledge, a standard cDNA image data set with known expression levels (e.g confirmed by quantitative PCR [28]) is not available. Thus, exploiting the benefits of the replicated common reference channel (Cy-3), provided with the particular dataset, we quantitatively assessed the performance of the SB algorithmic pipeline using the pairwise mean absolute error (MAE) between the replicates [29] (altogether 21 pairwise MAE values). Extracted intensities by the proposed SB method were comparatively evaluated (in terms of MAE) against the intensities obtained by a recently published commercial software (MAGIC TOOL [21]). For evaluation purposes, the same microarray images were introduced to both methods and the SRG segmentation option of the MAGIC TOOL was chosen, for fairness of comparison.

III. RESULTS

Figure 1 shows part of the original image and the resulting processed image, with spots boundaries outlined, employing the proposed SB algorithmic pipeline.

Figure 2 depicts a randomly selected spot as appeared in replicated images for the common reference channel (1st row). The middle row shows the segmentation result of the SB-pipeline and the 3rd row the result of the MAGIC TOOL's SRG algorithm.

Figure 3 illustrates the calculated pairwise MAE between the expression ratios of all possible pairs of the common reference channel for the dataset of the 7 replicated real images (6400 spots per image). Table 1 provides the mean values of the pairwise MAE of Figure 3.

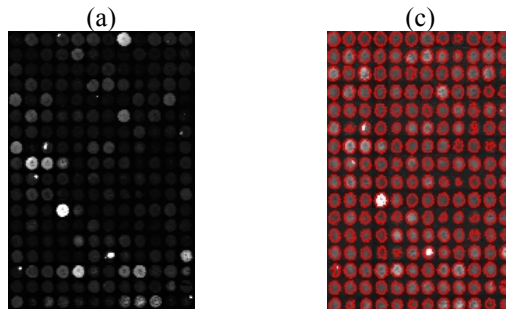


Fig. 1. a/Original image, b/ enhanced image, with spot's outlines superimposed (as determined by the SB-pipeline).

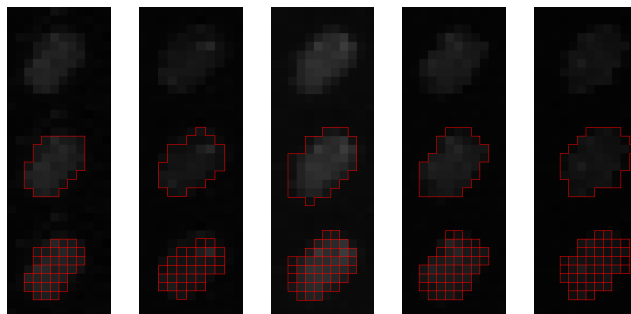


Fig. 2. Segmentation results as obtained by the SB-pipeline (2nd row) and by the MAGIC TOOL's SRG (3rd row) for 5 replica spot-images (1st row).

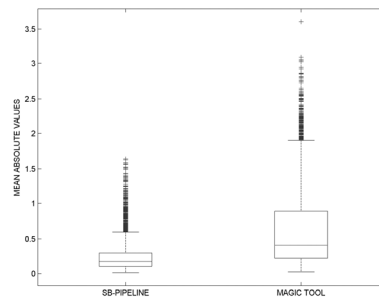


Fig. 3. Boxplots illustrating the pairwise MAE between all replicates (totally 21 MAE values from which the mean value for each spot is illustrated here). Obelisks are MAE values characterized as outliers.

TABLE I
MEAN VALUES OF THE CALCULATED 21 PAIRWISE MAE FOR THE COMMON REFERENCE CHANNEL

	SB-Pipeline	MAGIC TOOL
Common Reference Channel	0.254	0.630

IV. DISCUSSION

In the present study, an efficient spot-based (SB) algorithmic pipeline of clustering enhancement and segmentation techniques was developed to quantify gene expression levels in microarray images. The SB-pipeline utilizes i/a gridding procedure to locate spot-regions, ii/ a clustering algorithm (the Enhanced fuzzy c-means) to roughly segment spots from background and to estimate background noise and spot's center, iii/ an adaptive histogram modification technique to accentuate spot's boundaries, and iv/the Seeded Region Growing (SRG) technique, which extracted microarray spots' intensities and quantified individual genes' expression levels by incorporating information from step 2.

As it may be observed from Figure 1, the intermediate step of spot-based image enhancement improves the display of spots and emphasizes spot's boundaries even for those spots that are not well-defined (low intensity spots). For comparison reasons with a public available software, Figure 2 shows the same spot (for 5 replicated images of the common reference channel) segmented using the MAGIC TOOL's SRG and the proposed SB-pipeline. Visual inspection may reveal the improvements achieved by the SB-pipeline.

Figure 3 illustrates the boxplots of MAE as they were calculated for the common reference channel of the 7 replicated microarray images and Table 1 depicts the mean values of those boxplots. Lower MAE values are indicative of higher segmentation performance and, thus, of more accurate (*valid*) extraction of gene expression levels. As shown in Table 1, the proposed method achieved better results than the MAGIC TOOL's SRG.

Regarding processing time, the SB-pipeline took 2100 seconds against MAGIC TOOL's 3410 seconds for the evaluated dataset of the 7 real cDNA microarray images (1024x1024, 16-bit tiff format), each containing 6400 spots. Comparisons were performed on the same computer (Intel Core 2 Duo at 2.66 GHz, 2GB RAM).

ACKNOWLEDGMENT

This work was supported by a grant from the General Secretariat for Research and Technology, Ministry of Development of Greece (136/PENED03) to B.A.

REFERENCES

- [1] A. Alizadeh, M. Eisen, D. Botstein, P. O. Brown and L. M. Staudt, "Probing lymphocyte biology by genomic-scale gene expression analysis", *J Clin Immunol*, vol. 18, pp. 373-379, Nov 1998.
- [2] Mark Schena: *Microarray Biochip Technology*. Eaton Publishing Company, 2000.
- [3] A. N. Jain, T. A. Tokuyasu, A. M. Snijders, R. Segraves, D. G. Albertson and D. Pinkel, "Fully automatic quantification of microarray image data", *Genome Res*, vol. 12, pp. 325-332, Feb 2002.
- [4] Y. Balagurunathan, N. Wang, E. R. Dougherty, D. Nguyen, Y. Chen, M. L. Bittner, J. Trent and R. Carroll, "Noise factor analysis for cDNA microarrays", *J Biomed Opt*, vol. 9, pp. 663-678, Jul-Aug 2004.
- [5] A. A. Ahmed, M. Vias, N. G. Iyer, C. Caldas and J. D. Brenton, "Microarray segmentation methods significantly influence data precision", *Nucleic Acids Res*, vol. 32, pp. e50, 2004.
- [6] Axon Instruments, "GenePix4000A User's Guide", vol. pp. 1999.
- [7] M. Steinfath, W. Wruck, H. Seidel, H. Lehrach, U. Radelof and J. O'Brien, "Automated image analysis for array hybridization experiments", *Bioinformatics*, vol. 17, pp. 634-641, Jul 2001.
- [8] A. M. White, D. S. Daly, A. R. Willse, M. Protic and D. P. Chandler, "Automated Microarray Image Analysis Toolbox for MATLAB", *Bioinformatics*, vol. 21, pp. 3578-3579, Sep 1 2005.
- [9] M. A. Zapala, D. J. Lockhart, D. G. Pankratz, A. J. Garcia, C. Barlow and D. J. Lockhart, "Software and methods for oligonucleotide and cDNA array data analysis", *Genome Biol*, vol. 3, pp. SOFTWARE0001,0001-0001,0009, 2002.
- [10] M. B. Eisen, Scanalyze. Available: <http://rana.stanford.edu/software>.
- [11] J. Lee and D. Lee, "Dynamic characterization of cluster structures for robust and inductive support vector clustering", *IEEE Trans Pattern Anal Mach Intell*, vol. 28, pp. 1869-1874, Nov 2006.
- [12] V. S. Tseng and C. P. Kao, "Efficiently mining gene expression data via a novel parameterless clustering method", *IEEE/ACM Trans Comput Biol Bioinform*, vol. 2, pp. 355-365, Oct-Dec 2005.
- [13] J. Lee and D. Lee, "An improved cluster labeling method for support vector clustering", *IEEE Trans Pattern Anal Mach Intell*, vol. 27, pp. 461-464, Mar 2005.
- [14] R. C. Gonzalez and R. E. Woods: *Digital Image Processing* 1992.
- [15] X. H. Wang, R. S. Istepanian and Y. H. Song, "Microarray image enhancement by denoising using stationary wavelet transform", *IEEE Trans Nanobioscience*, vol. 2, pp. 184-189, Dec 2003.
- [16] R. Lukac, K. N. Plataniotis, B. Smolka and A. N. Venetsanopoulos, "cDNA Microarray Image Processing Using Fuzzy Vector Filtering Framework", *Journal of Fuzzy Sets and Systems: Special Issue on Fuzzy Sets and Systems in Bioinformatics*, vol. pp. 2005.
- [17] M. Mastriani and A. E. Giraldez, "Microarrays Denoising via Smoothing of Coefficients in Wavelet Domain", *International Journal of Biomedical Sciences*, vol. 1, pp. 1306-1216, 2006.
- [18] R. Lukac and B. Smolka, "Application of the adaptive center-weighted vector median framework for the enhancement of cDNA microarray", *Int. J. Appl. Math. Comput. Sci.*, vol. 13, pp. 369-383, 2003.
- [19] W. Cai, S. Chen and D. Zhang, "Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation", *Pattern Recognition*, vol. 40, pp. 825-838 2007.
- [20] R. Adams and L. Bischof, "Seeded region growing", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 641-647, 1994.
- [21] L. J. Heyer, D. Z. Moskowitz, J. A. Abele, D. Choi, A. M. Campbell, E. E. Oldham and B. K. Akin, "MAGIC Tool: integrated microarray data analysis", *Bioinformatics*, vol. 21, pp. 2114-2115, May 1 2005.
- [22] L. Heyer, Magic Tool Database. Available: <http://www.bio.davidson.edu/projects/MAGIC/MAGIC.html>.
- [23] J. L. DeRisi, V. R. Iyer and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale", *Science*, vol. 278, pp. 680-686, Oct 24 1997.
- [24] G. A. Churchill, "Fundamentals of experimental design for cDNA microarrays", *Nat Genet*, vol. 32 Suppl, pp. 490-495, Dec 2002.
- [25] M. Nykter, T. Aho, M. Ahdesmaki, P. Ruusuvaori, A. Lehmussola and O. Yli-Harja, "Simulation of microarray data with realistic characteristics", *BMC Bioinformatics*, vol. 7, pp. 349, 2006.
- [26] K. Blekas, N. Galatsanos, A. Likas and I. Lagaris, "Mixture model analysis of DNA microarray images." *IEEE Trans Med Imaging*, vol. 24, pp. 901-909, July 2005
- [27] S. M. Pizer and E. P. Amburn, "Adaptive histogram equalization and its variations", *Computer Vision, Graphics, and Image Processing*, vol. 39, pp. 355-368, 1987.
- [28] S. Poggeler, M. Nowrousian, C. Ringelberg, J. J. Loros, J. C. Dunlap and U. Kuck, "Microarray and real-time PCR analyses reveal mating type-dependent gene expression in a homothallic fungus", *Mol Genet Genomics*, vol. pp. Feb 16 2006.
- [29] A. Lehmussola, P. Ruusuvaori and O. Yli-Harja, "Evaluating the performance of microarray segmentation algorithms", *Bioinformatics*, vol. 22, pp. 2910-2917, Dec 1 2006.