# Improving Microarray Spots Segmentation by K-Means driven Adaptive Image Restoration.

Antonis Daskalakis, Dionisis Cavouras, Panagiotis Bougioukos, Spiros Kostopoulos, Christos Argyropoulos and George Nikiforidis

***Abstract*—Complementary DNA microarray experiments are used to study human genome. However, microarray images are corrupted by spatially inhomogeneous noise that deteriorates image and consequently gene expression. An adaptive microarray image restoration technique is developed by suitably combining unsupervised clustering with the restoration filters for boosting the performance of microarray spots segmentation and for improving the accuracy of subsequent gene expression. Microarray images comprised a publicly available dataset of seven images, obtained from the database of the MicroArray Genome Imaging & Clustering Tool website. Each image contained 6400 spots investigating the diauxic shift of Saccharomyces cerevisiae. The adaptive microarray image restoration technique combined 1/a griding algorithm for locating individual cell images, 2/a clustering algorithm, for assessing local noise from the spot's background, and 3/a wiener restoration filter, for enhancing individual spots. The effect of the proposed technique quantified using a well-known boundary detection algorithm (Gradient Vector Flow snake) and the information theoretic metric of Jeffrey's divergence. The proposed technique increased the Jeffrey's metric from 0.0194 bits to 0.0314 bits, while boosted the performance of the employed boundary detection algorithm. Application of the proposed technique on cDNA microarray images resulted in noise suppression and facilitated spot edge detection.**

## I. Introduction

MICROARRAY technology provides a useful tool to assay large-scale gene sequence and gene expression analysis [1], [2]. Molecular biologists and bioinformaticians are using microarrays for identifying genes in biological sequences and predict genes function within a larger system, such as the human organism [3]. Microarray analysis involves three basic stages namely experimental design, image processing, and gene quantification [4].

Initially, the DNA obtained from the genes of interest (targets) is printed on a glass microscope slide by a robotic arrayer, thus, forming circular spots of known diameter. Each spot serves as a highly specific and sensitive detector (probe) of the corresponding gene [5], [6]. In order to create a genome expression profile of a biological system with microarrays, the messenger RNA (mRNA) from a particular sample is isolated, is labelled using Cy3 (green) and Cy5 (red) fluorescent dyes, and it is hybridized on the microarray. Following hybridization, the arrays are scanned by activation using lasers that excite each dye on the appropriate wavelength. The relative fluorescence between each dye on each spot is then recorded using methods contingent upon the nature of the labelling reaction i.e. confocal laser scanners, and Charged Couple Devices [7], [8].

The data output of such systems are two sample 12 to 16-bit TIFF images, one for each fluorescent channel. The relative intensities of each channel represent the relative abundance of the DNA product in each of the two samples. Image processing and analysis plays a crucial role in the extraction and quantitative analysis of the relative abundance of the DNA product, since it affects the following steps that lead to gene expression and quantification. The basic stages in a microarray image processing workflow are: griding, spot segmentation, and intensity extraction [4], [9]-[11]. Griding is the process of assigning coordinates to each cell; the latter is a square ROI containing the pixels of both the spot and its background. Segmentation, classifies cell-pixels as foreground (spot-pixels) or background. Intensity extraction calculates ratios of red to green fluorescence intensities for the foreground and background respectively.

Data mining techniques [12] are employed to group genes so that molecular biologists may extract meaningful biological information or make assumptions regarding unknown genes. Gene quantification is confounded by a number of technical factors, which operate during the fabrication, target labelling, and hybridization stages. Microarray images are corrupted by spatially inhomogeneous noise and by irregularities in the shape, size, and position of the spot [13], [14]. Another source of degradation is due to noise and MTF of the confocal laser scanner, employed as "reading" method. These sources of error may propagate and thus affect biological expression.

In spite of the potential importance of image pre-processing in correcting these error sources, existing software tools [15]-[20] pay little attention to pre-processing and focus mainly on spot localization and microarray image

Manuscript received June 30, 2006. 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, C. Argyropoulos 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 is with the Medical Signal and Image Processing Lab, Department of Medical Instrumentation Technology,Technological Education Institution of Athens, Ag. Spyridonos Street, Aigaleo, 122 10, Athens, Greece (e-mail: cavouras@teiath.gr).

segmentation. Few studies [21]-[24] have examined the impact of image pre-processing upon spot enhancement, however, we have found no studies to actually quantify the benefit of image enhancement in facilitating segmentation and consequently gene quantification.

The aim of the present study is to evaluate the impact that image pre-processing techniques may have on improving the accuracy of spot segmentation and gene quantification. Consequently, this study proposes an adaptive microarray image restoration (A.μA.I.R) technique, which combines 1/a griding algorithm for locating individual cell images, 2/a clustering algorithm, for assessing local noise, and 3/a restoration filter, for enhancing individual cell images, in order to facilitate accurate spot detection and gene quantification. Objective quantification of restoration results was based on information theoretic measures.

## II. Materials and methods

Microarrays used in this study comprised a publicly available dataset of images obtained from the database of the MicroArray Genome Imaging & Clustering Tool (MAGIC) website [25]. Each image contained 6400 spots investigating the diauxic shift of *Saccharomyces cerevisiae.* The particular dataset was selected because the original authors [26] used a common reference messenger RNA pool (green, Cy-3) to control for biological variability [27]-[29]. This particular design affords an adequate degree of replication required for the quantitative statistical assessment of the effects of pre-processing on the image segmentation and subsequent gene quantification.

### A. *Griding*

The adaptive microarray image restoration technique, developed in the current study, initially applied an image griding procedure [30] on the images in order to locate spot sites (cell images).

### B. *Clustering for local noise estimation*

Next, individual spots were crudely segmented from surrounding background by unsupervised segmentation, using the K-Means algorithm [31]. The latter is a least-squares partitioning method that divides a collection of objects into K groups according to their pixel intensities, based on an iterative procedure, which minimizes a Euclidean distance. Subsequently, from each segmented cell-image local noise $(2\times\sigma^2)$ [32] was assessed from the spots background. This parameter was used to restore each cell image of the microarray image by employing the wiener image restoration technique.

### C. *Cell image restoration*

The latter incorporates both the degradation function and statistical characteristics of noise into the restoration process as in (1):

$$\hat{F}(u,\upsilon)=\left[\frac{\left|MTF(u,\upsilon)\right|^2}{\left|MTF(u,\upsilon)\right|^2+2\times\sigma^2}\right]\frac{G(u,\upsilon)}{MTF(u,\upsilon)} \qquad \textbf{(1)}$$

where $MTF$ is the Fourier transform of the degradation function (Point Spread Function), considered constant across the image, $G$ is the Fourier transform of the degraded image. Subsequently, the restored image in the spatial domain is obtained by the inverse Fourier transform of (1).

An estimation of the degradation function $MTF$ in (1) was modeled as a low pass Butterworth filter:

$$Fh^{LP}(v)=\frac{1}{1+0.414\left(\frac{v}{f_{co}}\right)^{2n}} \qquad \textbf{(2)}$$

where $n$ is the degree of the filter, v is the frequency, and $f_{co}$ the cut-off frequency [32]. $MTF$ was then obtained by (3)

$$MTF(u,v)=Fh^{LP}(\sqrt{u^2+v^2}) \qquad \textbf{(3)}$$

and

$$\sqrt{u^2+v^2}<=N \qquad \textbf{(4)}$$

where, N is the dimension of the cell-image.

All algorithms were implemented using Matlab custom-made code.

### D. *Evaluation*

The effect of the A.μA.I.R was tested by 1/applying a Gradient Vector Flow (snake) [33] boundary detection algorithm on both the original and the processed microarray images and 2/applying a metric to both images, the Jeffrey's (J) measure of divergence [34], [35], for estimating the ‘goodness’ of segmentation in each cell image. In turn, this is an indication as to the accuracy of spot detection and gene quantification.

### E. *Cell-image segmentation*

Twenty randomly selected cell-images from the original microarray image and the corresponding processed cell-images were segmented by the snake algorithm, which determined boundary points separating the spot from its background in each cell image. All boundary points were referred to the original cell-images, since intensities in the processed cell images were altered by the restoration process.

### F. *Quantification of the benefit*

Following segmentation, foreground (spot) and background intensity values for the common reference channel (green, Cy-3) were extracted. Those values were used to form two density distributions employing a non-parametric kernel density estimation method [36]. The distance between those two distributions was determined employing the Jeffrey's (J) measure of divergence, shown in (5):

$$J(S,B) = \sum_i \left( p_{B,i} - p_{S,i} \right) \log \frac{p_{B,i}}{p_{S,i}} \qquad (5)$$

Where $p_{S,i}$ *and* $p_{B,i}$ are the spot and background density distributions respectively. Higher values of *J* correspond to more distant distributions and consequently to more accurate segmentation, considering that intensities are evaluated on the original image alone. Those J values were further tested, using Wilcoxon non-parametric statistical test, in order to provide the statistical significance of the proposed methodology.

## III. Results

The image degradation function was optimally designed by a first degree (n=1) low-pass Butterworth filter using $f_{co}$=0.3xN, with N being the dimension of the square cell image; non-square cell-images were zero padded. Figure 1, shows the result of the A.μA.I.R.

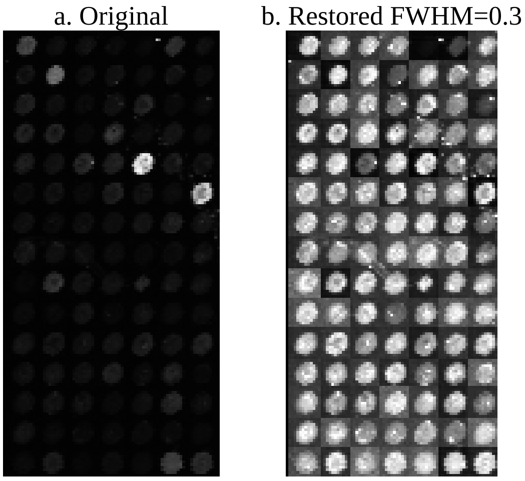

**Fig. 1.** Original and Adaptive Wiener restored sections of microarray images for the optimally designed degradation function according to the Butterworth Filter.

Figure 2 depicts the results of the snake boundary detection algorithm for two different spots using as initialization points the spots' sites located by the griding procedure.

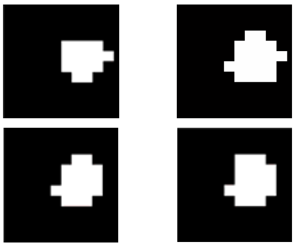

**Fig. 2.** GVF segmentation results based on the original (left column) and A.μA.I.R. (right column) restored images for two different spots.

Table 1 shows the values of the Jeffrey's divergence between spot and background log intensity distributions for the snake segmentation procedure.

Differences between J distances were found to differ at a statistical significant level of p<0.001, employing 20 randomly selected cell images.

TABLE I
Jeffrey's measure of divergence

| | **Original Spot** | **Restored Spot** |
|---|---|---|
| Spot 1 | 0.0449 | 0.0520 |
| Spot 2 | 0.0182 | 0.0298 |
| Spot 3 | 0.0217 | 0.0282 |
| Spot 4 | 0.0126 | 0.0166 |
| Spot 5 | 0.0136 | 0.0162 |
| Spot 6 | 0.0262 | 0.0326 |
| Spot 7 | 0.0209 | 0.0329 |
| Spot 8 | 0.0194 | 0.0314 |

Typical values (measured in bits) of divergence between signal and background intensities for original and adaptive A.μA.I.R restored spots respectively

## IV. Discussion

In the present work, the performance of an adaptive microarray image restoration technique was explored for evaluating the impact that image pre-processing techniques may have on improving the accuracy of spot segmentation and gene quantification. The proposed technique combined 1/a griding algorithm for locating individual cell images, 2/a clustering algorithm, for assessing local noise, and 3/a restoration filter, for enhancing individual cell images, in order to facilitate accurate spot detection and gene quantification.

By the visual inspection of the original and the corresponding A.μA.I.R restored sections, shown in Figure 1, it can be observed that the proposed technique removed noise components while preserved the sharpness of spot edges. The adaptive wiener filter has been previously employed in enhancing microarray images [21] and it has been found to perform worse than the stationary wavelet transform. This is, however, expected, since noise has been modelled as an additive quantity that it is locally assessed, taking no provision of existing structural noise. In contrast, in the present work, noise was considered spatially inhomogeneous, and thus it was locally assessed, and structural noise was not incorporated by focusing on the cell-image background. This rendered the result of the A.μA.I.R successful enough to boost the performance of the employed boundary detection algorithm.

Objective quantification for both the original and the enhanced images, based on the Jeffrey's divergence, confirmed the influence of the proposed technique in the subsequent steps of the microarray pipeline (e.g. segmentation, gene quantification) as Table 1 shows. The implemented technique performed as anticipated by increasing the divergence (J) between the distributions of signal and background intensity distributions.

## V. Conclusion

In the present work an adaptive microarray image restoration technique has been presented. Application of the proposed technique on cDNA microarray images resulted in noise suppression while boosted spot edge detection. Improved accurate spot detection was objectively quantified by employing information theoretic metrics.

## VI. Acknowledgement

This work was supported by a grant form 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. 373-9, 1998.

[2] M. Taniguchi, K. Miura, H. Iwao, and S. Yamanaka, "Quantitative assessment of DNA microarrays--comparison with Northern blot analyses," *Genomics*, vol. 71, pp. 34-9, 2001.

[3] M. Chee, R. Yang, E. Hubbell, A. Berno, X. C. Huang, D. Stern, J. Winkler, D. J. Lockhart, M. S. Morris, and S. P. Fodor, "Accessing genetic information with high-density DNA arrays," *Science*, vol. 274, pp. 610-4, 1996.

[4] J. Angulo and J. Serra, "Automatic analysis of DNA microarray images using mathematical morphology," *Bioinformatics*, vol. 19, pp. 553-62, 2003.

[5] E. Southern, K. Mir, and M. Shchepinov, "Molecular interactions on microarrays," *Nat Genet*, vol. 21, pp. 5-9, 1999.

[6] Mark Schena, *Microarray Biochip Technology*, 1st ed: Eaton Publishing Company, 2000.

[7] K. K. Jain, "Current status of fluorescent in-situ hybridisation," *Med Device Technol*, vol. 15, pp. 14-7, 2004.

[8] P. A. t Hoen, F. de Kort, G. J. van Ommen, and J. T. den Dunnen, "Fluorescent labelling of cRNA for microarray applications," *Nucleic Acids Res*, vol. 31, pp. e20, 2003.

[9] V. Barra, "Robust segmentation and analysis of DNA microarray spots using an adaptative split and merge algorithm," *Comput Methods Programs Biomed*, vol. 81, pp. 174-80, 2006.

[10] 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. 325-32, 2002.

[11] Y. H. Yang, M. J. Buckley, and T. P. Speed, "Analysis of cDNA microarray images," *Brief Bioinform*, vol. 2, pp. 341-9, 2001.

[12] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc Natl Acad Sci U S A*, vol. 95, pp. 14863-8, 1998.

[13] Y. Balagurunathan, E. R. Dougherty, Y. Chen, M. L. Bittner, and J. M. Trent, "Simulation of cDNA microarrays via a parameterized random signal model," *J Biomed Opt*, vol. 7, pp. 507-23, 2002.

[14] 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. 663-78, 2004.

[15] "GenePix4000A User's Guide," 1999.

[16] 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. 634-41, 2001.

[17] 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. 3578-9, 2005.

[18] 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. SOFTWARE0001,1-0001,9, 2002.

[19] O. s. M. QuantArray Analysis Software, 1999.

[20] M. B. Eisen, "ScanAlyze."

[21] 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. 184-9, 2003.

[22] R. Lukac, P. K.N., S. B., 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*, 2005.

[23] 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.

[24] 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.

[25] http://www.bio.davidson.edu/projects/MAGIC/MAGIC.html.

[26] 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. 680-6, 1997.

[27] G. A. Churchill, "Fundamentals of experimental design for cDNA microarrays," *Nat Genet*, vol. 32 Suppl, pp. 490-5, 2002.

[28] E. Sterrenburg, R. Turk, J. M. Boer, G. B. van Ommen, and J. T. den Dunnen, "A common reference for cDNA microarray hybridizations," *Nucleic Acids Res*, vol. 30, pp. e116, 2002.

[29] Y. H. Yang and T. Speed, "Design issues for cDNA microarray experiments," *Nat Rev Genet*, vol. 3, pp. 579-88, 2002.

[30] K. Blekas, N. Galatsanos, A. Likas, and I. Lagaris, "Mixture model analysis of DNA microarray images.," *IEEE Trans Med Imaging*, vol. 24, pp. 901-9, 2005.

[31] S. Theodoridis and K. Koutroubas, *Pattern Recognition*: Academic Press, 1999.

[32] R. C. Gonzalez and R. E. Woods, *Digital Image Processing* 1992.

[33] http://iacl.ece.jhu.edu/projects/gvf/, "Gradient Flow Vector active contours."

[34] S. M. Ali and S. D. Silvey, "A general class of Coefficients of Divergence of One Distribution from another," *J.R.Stat.Soc.B.*, vol. 28, pp. 131-142, 1966.

[35] S. Kullback, *Information Theory and Statistics*, 2nd ed: Dover Publications, 1968.

[36] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, 1997.