

# Effective Quantification of Gene Expression Levels in Microarray Images Using a Spot-Adaptive Compound Clustering-Enhancement-Segmentation Scheme

Antonis Daskalakis<sup>1</sup>, Dionisis Cavouras<sup>2</sup>, Panagiotis Bougioukos<sup>1</sup>, Spiros Kostopoulos<sup>1</sup>, Pantelis Georgiadis<sup>1</sup>, Ioannis Kalatzis<sup>2</sup>, George Kagadis<sup>1</sup>, and George Nikiforidis<sup>1</sup>

<sup>1</sup> Department of Medical Physics, School of Medicine, University of Patras, Rio, GR-26500 Greece

daskalakis@med.upatras.gr

<sup>2</sup> Medical Signal and Image Processing Lab, Department of Medical Instruments Technology, Technological Educational Institute of Athens, Ag. Spyridonos Street, Aigaleo, 122 10, Athens, Greece

**Abstract.** A spot-adaptive compound clustering-enhancement-segmentation (CES) scheme was developed for the quantification of gene expression levels in microarray images. The CES-scheme employed 1/gridding, for locating spot-regions, 2/Fuzzy C-means clustering, for segmenting spots from background, 3/ background noise estimation and spot's center localization, 4/emphasizing of spot's outline by the CLAHE image enhancement technique, 5/segmentation by the SRG algorithm, using information from step 3, and 6/microarray spot intensity extraction. Extracted intensities by the CES-Scheme were compared against those obtained by the MAGIC TOOL's SRG. Kullback-Liebler metric's values for the CES-Scheme were on average double than MAGIC TOOL's, with differences ranging from 1.45bits to 2.77bits in 7 cDNA images. Coefficient-of-Variation results showed significantly higher reproducibility ( $p < 0.001$ ) for the CES-Scheme in quantifying gene expression levels. Processing times for 1024x1024 16-bit microarray images containing 6400 spots were 300 and 487 seconds for the CES-Scheme and MAGIC TOOL respectively.

**Keywords:** DNA, microarray image analysis, microarray gridding, CLAHE, SRG.

## 1 Introduction

DNA microarrays are an experimental technology for exploring the genome. This technology provides to biomedical investigators a simple tool for monitoring the expression levels of thousands of genes, under the same experimental conditions, and thus an easy way to identify and quantify gene expression levels for all genes in an organism [1-3].

Microarrays are arrays of glass microscope slides, in which 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 with red and green fluorescent dye, respectively, 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 on 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 [4].

In order to extract those relative intensities [5, 6], a series of image analysis techniques have been employed namely gridding, spot segmentation, and intensity extraction [7, 8]. Gridding is the process of identifying and locating the coordinates of each cell containing the spot; the cell is a rectangular region containing the pixels of both the spot and its background. Segmentation refers to the classification of cell-pixels as either signal (spot's foreground) or surrounding area (spot's background). Spots' intensity extraction refers to the calculation of the fluorescence signal's mean intensity from the spot's foreground. Extracted mean intensities correspond to gene expression levels that, in turn, are translated into biological conclusions from molecular biologists, by employing data mining techniques.

A major factor that complicates the task of image analysis and data mining is that microarray experiments involve a number of error-prone steps (occurring during fabrication, target labeling, and hybridization), which induce noise on the resulting images [9, 10]. Microarray images are also corrupted by irregularities in the shape, size, and position of the spot [10, 11]. Improper treatment of noise may result to erroneous biological conclusions.

One of the most undesirable effects of noise is that it contributes to inaccurate spot segmentation (i.e. the boundaries of spots are incorrectly estimated), which leads to wrong estimation of the mean spots' intensities and reduces the reproducibility of the gene expression levels, derived from microarray images [12]. Although a variety of image pre-processing techniques have been suggested for correcting these sources of error [13], existing software tools, utilized for the analysis of microarray images, focus mainly on accurate spot localization and segmentation by various segmentation techniques [14-19]. Only few studies [20-24] have examined the impact of image pre-processing on cDNA image quality, however, without evaluating the effect of image enhancement on facilitating spot segmentation and, thus, on increasing the repeatability of the extracted gene expression levels.

In the present study, an efficient spot-adaptive compound clustering-enhancement-segmentation (CES) scheme is proposed for the quantification of gene expression levels in microarray images. The CES-scheme suitably incorporates the benefits of image enhancement into the spot segmentation procedure. Accordingly, the CES-scheme combines 1/a gridding algorithm for locating individual spots 2/the Fuzzy C-means clustering algorithm [25], for automatically differentiating the spot's foreground and background, 3/a procedure for the assessment of local noise from the spot's background and for the determination of the spot's center, 4/the contrast limited adaptive histogram equalization (CLAHE) technique, for enhancing individual spot boundaries [13], 5/the seeded region growing (SRG) [26] segmentation technique, for segmenting microarray spots, employing information from step 3, and 6/a procedure for quantification of individual spot intensities. The proposed CES-scheme was implemented as part of a custom software that we have developed in our

lab [24]. The efficiency of the proposed CES-scheme was assessed employing both the Kullback-Liebler divergence metric [27], for goodness of segmentation, and the Coefficient of Variation metric, for assessing the reproducibility of extracted genes expression. Results were compared against those obtained by the SRG technique employed within the MAGIC TOOL's software [28], using a set of publicly available microarray cDNA images [29].

## 2 Material and Methods

Microarrays used in the current study comprised a public available dataset of 7 images, downloaded from the MicroArray Genome Imaging & Clustering Tool (MAGIC) website [29]. Each image of the dataset contained 6400 spots investigating the diauxic shift of *Saccharomyces cerevisiae*. In the particular dataset, the authors [30] have used a common reference messenger RNA pool (green, Cy-3) to control for biological variability [2, 31, 32]. Such a design provides an adequate degree of replication (7 images), required for the quantitative assessment of image segmentation and subsequent gene quantification.

### 2.1 CES-Scheme

The CES-Scheme was initialized by applying a gridding procedure on the images. Ideally, spots are located at certain positions on the rectangular grid. By summing up the intensities across the pixels in each row and each column of the grid (line profiles), each spot center was represented by a peak-valley pattern, where peaks corresponded to spot centers and valleys to spot sites edges. Smoothing the line profiles by the Lowess filter [33], it ensured minimization of irregularities, introduced by the printing procedure, and, therefore, success of the gridding procedure. Spot sites, in terms of width and height, were finally estimated from the peak-valley distance in each line profile. Thus, rectangular spot-containing regions (cell-regions) were formed.

Within each cell-region, the Fuzzy C-Means unsupervised algorithm [25] was employed to discriminate spots from surrounding background. The Fuzzy C-Means searched iteratively for cluster centers (centroids) that minimize the dissimilarity function (eq.1)

$$J(U, c_1, c_2, \dots, c_c) = \sum_{i=1}^c J_i = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m d_{ij}^2 \quad (1)$$

where  $c_i$  is the centroid of cluster  $i$ ,  $d_{ij}$  is the Euclidean distance between  $i$ -th centroid ( $c_i$ ) and  $j$ -th data point,  $U$  is the membership function matrix, and  $m$  is a weighting exponent ( $m=2$ ). Subsequently, from the segmented spot and background, the location of the spot's center and local noise (standard deviation of background) respectively were assessed.

Following Fuzzy C-Means rough segmentation, spots in the individual cell-regions were enhanced using the CLAHE method [34]. CLAHE functioned adaptively on each cell-region aiming to maximize the contrast of each cell-spot relative to its background. The basic steps for its implementation were: a) division of the cell-region into a number of non-overlapping contextual regions of equal sizes (in our case regions were experimentally set to be approximately 40 pixels), b) computation of the histogram of each contextual region, c) enhancement of the contrast of each contextual region, by clipping its respective histogram under a certain threshold ( $t=0.001$ ), d) redistribution of the histogram in such a way that its height did not exceed the clip limit, e) bilinear interpolation of the neighboring contextual regions and modification of the cells'-region gray-levels according to the Cumulative Distribution Function (CDF) (eq.2) of each contextual region.

$$f(n) = \frac{(N-1)}{M} \sum_{k=0}^n h(k) \quad (2)$$

where  $M$  and  $N$  are the numbers of pixels and gray-levels in the cell-region respectively and  $h(k)$  for  $k=0,1,\dots,N-1$  is the histogram of each cell-region.

Contrast enhanced cell-regions were segmented using the SRG algorithm. SRG initially segmented each cell-region into regions of pixels starting from the spot's center, as determined by the Fuzzy C-Means segmentation. Pixel regions were iteratively augmented by assigning neighboring pixels that satisfied a homogeneity criterion: the neighboring pixels should be 1/of higher intensity than local noise, as it was calculated during the rough Fuzzy C-Means segmentation stage, and 2/of intensity close to the mean intensity of the so far seeded region. This iterative procedure of growing pixel regions within each cell-region continued until all pixels of the cell-region were assigned to either the spot or its background.

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

## 2.2 Evaluation of Extracted Genes Expression Levels

Foreground (spot) and background intensity values for the common reference channel (green) were subsequently extracted and, considering all segmented spots for each image, two density distributions were produced, employing a non-parametric kernel density estimation method [35]. The distance between those two distributions was determined employing the Kullback-Liebler (K-L) measure of divergence, shown in (3):

$$K - L(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (3)$$

where  $P$  and  $Q$  are the spot and background density distributions respectively. Higher values of divergence correspond to more distant distributions and, consequently, to

more accurate segmentation, considering that intensities are evaluated on the original image alone.

Exploiting the benefits of the replicated common reference channel (Cy-3), provided through the design of the particular dataset, we quantitatively assessed the performance of the CES-Scheme in terms of extracted genes expression reproducibility using the Coefficient of Variation metric (*CV*) (eq.4).

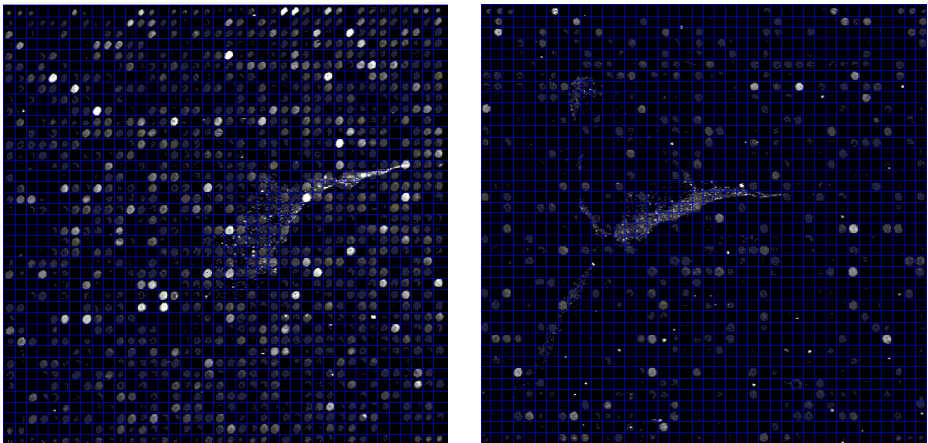
$$CV = \frac{\sigma}{\mu} \quad (4)$$

where  $\sigma$  is the standard deviation and  $\mu$  is the mean value for each spot evaluated for the all the replications (7 replications totally).

Extracted intensities employing the CES-Scheme were comparatively evaluated, in terms of both *CV* and *K-L* divergence, against the intensities obtained by a recently published commercial software (MAGIC TOOL [28]). 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.

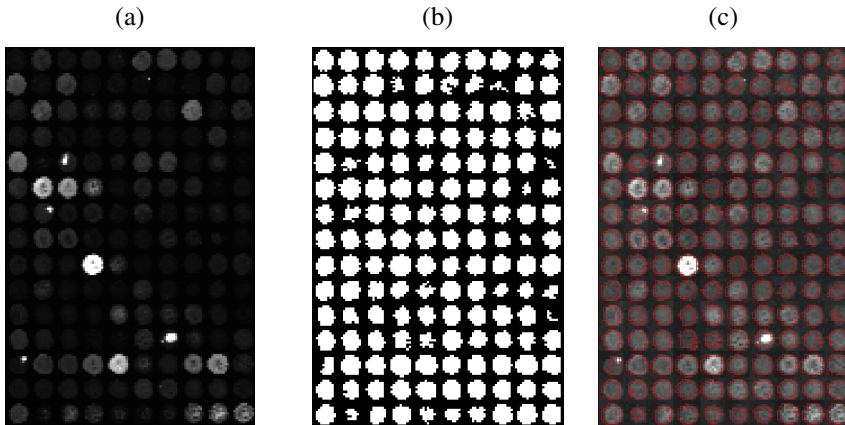
### 3 Results

Figure 1 shows the result of the successive gridding procedure for two selected sub-blocks of the microarray images used in the present study.



**Fig. 1.** Results of the gridding procedure

Figure 2 shows the result of the Fuzzy C-Means (Fig. 2b), for rough estimation of spot and background, and the resulting segmented image (Fig. 2c), with spots' boundaries outlined, employing the proposed CES-Scheme.



**Fig. 2.** a/Original image, b/Fuzzy C-means segmented image, and c/enhanced image, result of the CES-Scheme for microarray boundary determination

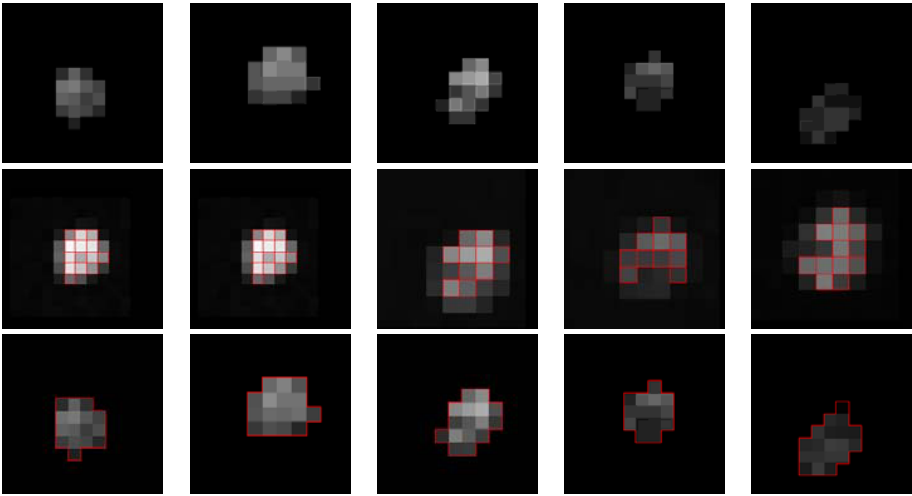
Table 1 presents the results of the Kullback-Liebler divergence between spot and background intensity distributions (of the common reference channel Cy-3) for the proposed methodology and the MAGIC TOOL respectively.

**Table 1.** Kullback-Liebler divergence metric values (in bits) between spot (signal) and background intensity values for the green (common reference sample) channel, for the seven evaluated cDNA images

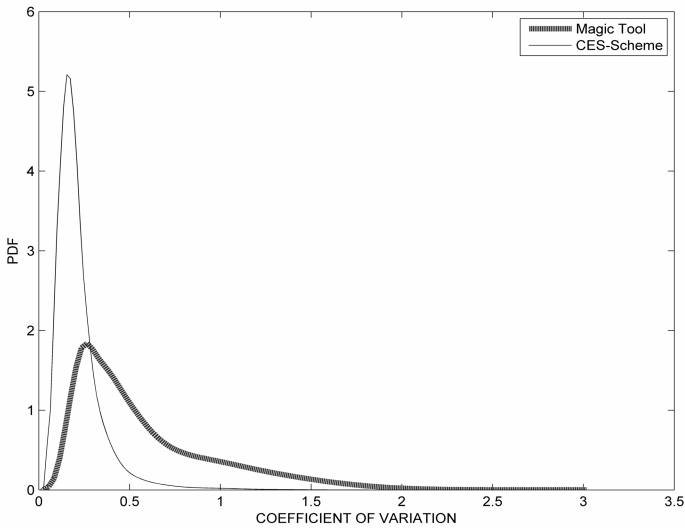
Images	CES-Scheme	MAGIC TOOL
1302_OD370	10.620	4.807
1303_OD014	8.075	2.917
1309_OD690	10.620	7.284
1310_OD046	7.546	4.858
1311_OD080	6.404	4.405
1312_OD180	10.505	4.680
1313_OD370	8.253	5.127

Figure 3 depicts a randomly selected spot as appeared in 5 out of the 7 replicated images for the common reference channel. The middle row shows the result of the SRG algorithm on the latter mentioned spot according to the MAGIC TOOL software, and the bottom row presents the segmentation results based on the CES-Scheme.

Figure 4 illustrates the actual distribution of  $CV$  values of extracted gene expression levels from the set of seven 1024x1024 16-bit replicated images (Cy-3), each containing 6400 spots, as they were calculated by Magic Tool and the proposed CES-Scheme.



**Fig. 3.** Spot-containing cell region as appeared in 5 replica images and segmentation results according to the MAGIC TOOL's SRG (middle row) and the CES-Scheme (bottom row)



**Fig. 4.** Probability Density Functions (PDF's) of the Coefficient of Variation for all the spots as evaluated from the 7 replicated images of the common reference channel. Thick line corresponds to the results obtained using the MAGIC TOOL's SRG and thin line to the results obtained using the CES-Scheme.

## 4 Discussion

DNA microarrays represent a technological intersection between biology and computer science, which enables the gene expression analysis on a genome-wide scale. A

unique capability provided by the microarray technology is the collection of the necessary data for genomic analysis from just one experiment.

The major challenges of this approach appear in the image processing stage in which the spotted DNA sequences have to be extracted from the produced microarray images and quality measures have to be calculated. The accuracy of this stage has a straightforward impact on the accuracy and the effectiveness of the subsequent gene expression and identification analysis.

The task of microarray image processing is complicated, since the microarray experiments involve a number of error-prone steps, which induce noise on the produced images [9, 10]. Noise contributes to inaccurate spot segmentation (i.e. the boundaries of spots are incorrectly estimated) and, thus, indirectly it evokes the wrong estimation of the extracted genes expression levels [12].

Despite the variety of existing image pre-processing techniques that address noise effects [13], existing software tools focus mainly on spot recognition (spot localization) and extraction of genes' intensities employing various segmentation techniques [14-19]. Nevertheless, as a previous study has indicated [12], different segmentation methods, while accurate in simulated microarray images, lead to a different number of differentially expressed genes, when applied to identical real microarray images. Thus, a solution to the problem of accurate gene quantification might be provided by an intermediate pre-processing step.

Only few studies have examined the impact of image preprocessing techniques on cDNA image quality [20-23]. However, most of them were applied globally on the microarray images without taking into account the specific characteristics of individual spots [36]. Moreover, to the best of the authors' knowledge, there are no studies evaluating the effect of image enhancement in facilitating spot segmentation and, thus, in increasing the repeatability of the extracted gene expression levels.

In the present study, a compound Clustering-Enhancement-Segmentation (CES) scheme was developed to efficiently quantify genes expression levels in microarray images. The proposed compound CES-Scheme comprised 1/a gridding algorithm for locating individual spot containing cell-regions, 2/the Fuzzy C-Means algorithm for automatically segmenting each cell-region into spot and background, 3/a procedure for assessing local noise and the spot's center, 4/the contrast limited adaptive histogram equalization (CLAHE) technique, for enhancing individual spot boundaries, 5/the seeded region growing (SRG) segmentation technique, for outlining the spot's boundary, and 6/a routine for quantifying the intensities of individual microarray spots on the original image.

Observing the original and the corresponding segmented images, as Figure 2 depicts, supports our initial hypothesis that an intermediate step of spot-based image enhancement improves the display of spots and emphasizes the depiction of spot edges even for those spots that are not well-defined (low intensity spots). Such an intermediate step resulted in more accurate and reproducible segmentation results, since it facilitated the spot's edge-detection. This conclusion is supported by the results depicted in Figure 3, where the same spot was segmented for 5 replicated images of the common reference channel using the MAGIC TOOL's SRG (middle row) and the proposed CES-Scheme (end row) .

To quantify the effectiveness of the proposed CES-Scheme, the information theoretic metric of the Kullback-Liebler divergence (Table 1) was employed and our

results were comparatively evaluated against those obtained using publicly available software (MAGIC TOOL). Results, according to the proposed scheme, were obtained by superimposing spot-outlines from the processed cell-region images on the original cell-spots. In this way, a higher Kullback-Liebler divergence, achieved by a particular method between the actual spots and its surrounding background, would eventual lead to better spot-boundary detection result. Table 1 confirms that the proposed scheme performed as anticipated, by increasing the divergence (K-L) between signal and background intensity distributions, as compared to corresponding distributions obtained using the MAGIC TOOL software.

The dataset used in the current study has been designed to control for biological variability (green channel was a common reference channel) [31]. Hence, an adequate degree of replication has been provided to quantitatively assess the reproducibility of the intensities extracted by the proposed CES-Scheme. Normally, replicate experiments have been used for reducing experimental variation [37]. Due to the replication provided, each spot should have the same intensity throughout the replicated experiments, and therefore the coefficient of variation between replicated experiments should be minimal (as close as possible to zero). Figure 4 shows the probability density functions (PDF) of the coefficient of variation for the common reference channel for all the images in the dataset using the CES-Scheme (thin line) and the MAGIC TOOL's SRG (thick line). CES-Scheme's PDF is narrow and sharp with a peak-value close to zero in contrast to MAGIC TOOL's PDF, which is more spread and far from zero. Distributions were found significantly different ( $p < 0.001$ ), employing Matlab's non-parametric statistical test (signtest).

Regarding processing time, the CES-Scheme took 300 seconds against MAGIC TOOL's 487 secs for the same  $1024 \times 1024$ , 16-bit cDNA image, containing 6400 microarray spots, and on the same computer.

**Acknowledgments.** 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. Alizadeh, A., Eisen, M., Botstein, D., Brown, P.O., Staudt, L.M.: Probing lymphocyte biology by genomic-scale gene expression analysis. *J Clin Immunol.* 18, 373–379 (1998)
2. Churchill, G.A.: Fundamentals of experimental design for cDNA microarrays. *Nat Genet.* 32, 490–495 (2002)
3. Taniguchi, M., Miura, K., Iwao, H., Yamanaka, S.: Quantitative assessment of DNA microarrays—comparison with northern blot analyses. *Genomics.* 71, 34–39 (2001)
4. Schena, M.: *Microarray biochip technology*, 1st edn. Eaton Publishing Company (2000)
5. Chen, Y., Dougherty, E., Bittner, M.: Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* 2, 364–374 (1997)
6. Schena, M.: *Microarray analysis*, 1st edn. New York (2002)
7. Jain, A.N., Tokuyasu, T.A., Snijders, A.M., Segraves, R., Albertson, D.G., Pinkel, D.: Fully automatic quantification of microarray image data. *Genome Res.* 12, 325–332 (2002)
8. Yang, Y.H., Buckley, M.J., Speed, T.P.: Analysis of cDNA microarray images. *Brief Bioinform.* 2, 341–349 (2001)

9. Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., Herzel, H.: Normalization strategies for cDNA microarrays. *Nucleic Acids Res.* 28, 47 (2000)
10. Balagurunathan, Y., Wang, N., Dougherty, E.R., Nguyen, D., Chen, Y., Bittner, M.L., Trent, J., Carroll, R.: Noise factor analysis for cDNA microarrays. *J Biomed Opt.* 9, 663–678 (2004)
11. Balagurunathan, Y., Dougherty, E.R., Chen, Y., Bittner, M.L., Trent, J.M.: Simulation of cDNA microarrays via a parameterized random signal model. *J Biomed Opt.* 7, 507–523 (2002)
12. Ahmed, A.A., Vias, M., Iyer, N.G., Caldas, C., Brenton, J.D.: Microarray segmentation methods significantly influence data precision. *Nucleic Acids Res.* 32, e50 (2004)
13. Gonzalez, R.C., Woods, R.E.: *Digital image processing*, 1st edn (1992)
14. Axon Instruments. *Genepix4000a user's guide* (1999)
15. Steinfath, M., Wruck, W., Seidel, H., Lehrach, H., Radelof, U., O'Brien, J.: Automated image analysis for array hybridization experiments. *Bioinformatics.* 17, 634–641 (2001)
16. White, A.M., Daly, D.S., Willse, A.R., Protic, M., Chandler, D.P.: Automated microarray image analysis toolbox for matlab. *Bioinformatics.* 21, 3578–3579 (2005)
17. Zapala, M.A., Lockhart, D.J., Pankratz, D.G., Garcia, A.J., Barlow, C., Lockhart, D.J.: Software and methods for oligonucleotide and cDNA array data analysis. *Genome Biol.* 3, SOFTWARE0001.0001-0001.0009 (2002)
18. QuantArray Analysis Software, O.s.M. Available: via the INTERNET. Accessed
19. Eisen, M.B.S.: (Accessed 06/12/2006) via the INTERNET, Available: <http://rana.stanford.edu/software>
20. Wang, X.H., Istepanian, R.S., Song, Y.H.: Microarray image enhancement by denoising using stationary wavelet transform. *IEEE Trans Nanobioscience.* 2, 184–189 (2003)
21. Lukac, R., Plataniotis, K.N., Smolka, B., Venetsanopoulos, A.N.: 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)
22. Mastriani, M., Giraldez, A.E.: Microarrays denoising via smoothing of coefficients in wavelet domain. *International Journal of Biomedical Sciences.* 1, 1306–1316 (2006)
23. Lukac, R., Smolka, B.: Application of the adaptive center-weighted vector median framework for the enhancement of cDNA microarray. *Int. J. Appl. Math. Comput. Sci.*, 13, 369–383 (2003)
24. Daskalakis, A., Cavouras, D., Bougioukos, P., Kostopoulos, S., Argyropoulos, C., Niki-foridis, G.C.: Improving microarray spots segmentation by k-means driven adaptive image restoration. In: *Proceedings of the ITAB Ioannina, Greece* (2006)
25. Jain, A.K.: *Fundamentals of digital image processing*. Prentice-Hall, Englewood Cliffs (1989)
26. Adams, R., Bischof, L.: Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 16, 641–647 (1994)
27. Kullback, S.: *Information theory and statistics*, 2nd edn. Dover Publications, Mineola (1968)
28. Heyer, L.J., Moskowitz, D.Z., Abele, J.A., Karnik, P., Choi, D., Campbell, A.M., Oldham, E.E., Akin, B.K.: Magic tool: Integrated microarray data analysis. *Bioinformatics.* 21, 2114–2115 (2005)
29. (Accessed 06/12/2006) Available: via the INTERNET. <http://www.bio.davidson.edu/projects/MAGIC/MAGIC.html>
30. DeRisi, J.L., Iyer, V.R., Brown, P.O.: Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science.* 278, 680–686 (1997)

31. Sterrenburg, E., Turk, R., Boer, J.M., van Ommen, G.B., den Dunnen, J.T.: A common reference for cDNA microarray hybridizations. *Nucleic Acids Res.* 30, e116 (2002)
32. Yang, Y.H., Speed, T.: Design issues for cDNA microarray experiments. *Nat Rev Genet.* 3, 579–588 (2002)
33. Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 829–836 (1979)
34. Pizer, S.M., Amburn, E.P.: Adaptive histogram equalization and its variations. *Graphics, and Image Processing* 39, 355–368 (1987)
35. Bowman, A.W., Azzalini, A.: *Applied smoothing techniques for data analysis*. Oxford University Press, Oxford (1997)
36. Nykter, M., Aho, T., Ahdesmaki, M., Ruusuvuori, P., Lehmustola, A., Yli-Harja, O.: Simulation of microarray data with realistic characteristics. *BMC Bioinformatics* 7, 349 (2006)
37. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 97, 77 (2002)