BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 17 2007, pages 2265–2272
doi:10.1093/bioinformatics/btm337

Gene expression

Improving gene quantification by adjustable spot-image restoration

Antonis Daskalakis[1,*], Dionisis Cavouras[2], Panagiotis Bougioukos[1], Spiros Kostopoulos[1], Dimitris Glotsos[1], Ioannis Kalatzis[2], George C. Kagadis[1], Christos Argyropoulos[1] and George Nikiforidis[1]

[1]Medical Image Processing and Analysis (MIPA) Group, Laboratory of Medical Physics, School of Medicine, University of Patras, 265 04 Rio and [2]Medical Image and Signal Processing Laboratory, Department of Medical Instruments Technology, Technological Educational Institute of Athens, 122 10 Athens, Greece

Received on February 9, 2007; revised on June 16, 2007; accepted on June 19, 2007

Advance Access publication June 28, 2007

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: One of the major factors that complicate the task of microarray image analysis is that microarray images are distorted by various types of noise. In this study a robust framework is proposed, designed to take into account the effect of noise in microarray images in order to assist the demanding task of microarray image analysis. The proposed framework, incorporates in the microarray image processing pipeline a novel combination of spot adjustable image analysis and processing techniques and consists of the following stages: (1) gridding for facilitating spot identification, (2) clustering (unsupervised discrimination between spot and background pixels) applied to spot image for automatic local noise assessment, (3) modeling of local image restoration process for spot image conditioning (adjustable wiener restoration using an empirically determined degradation function), (4) automatic spot segmentation employing seeded-region-growing, (5) intensity extraction and (6) assessment of the reproducibility (real data) and the validity (simulated data) of the extracted gene expression levels.

Results: Both simulated and real microarray images were employed in order to assess the performance of the proposed framework against well-established methods implemented in publicly available software packages (Scanalyze and SPOT). Regarding simulated images, the novel combination of techniques, introduced in the proposed framework, rendered the detection of spot areas and the extraction of spot intensities more accurate. Furthermore, on real images the proposed framework proved of better stability across replicates. Results indicate that the proposed framework improves spots' segmentation and, consequently, quantification of gene expression levels.

Availability: All algorithms were implemented in Matlab™ (The Mathworks, Inc., Natick, MA, USA) environment. The codes that implement microarray gridding, adaptive spot restoration and segmentation/intensity extraction are available upon request. Supplementary results and the simulated microarray images used in this study are available for download from: ftp://users:bioinformatics@mipa.med.upatras.gr

Contact: daskalakis@med.upatras.gr

*To whom correspondence should be addressed.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Microarray technology provides a powerful approach for genomics research, since it assays large-scale gene sequences and assists gene expression analysis (Alizadeh *et al*., 1998). This unique technology allows for molecular biologists and bioinformaticians to identify simultaneously thousands of genes and predict their functionality within a larger system, such as the human organism (Schena *et al*., 1995).

In a typical microarray experiment, gene expression patterns between two samples (i.e. treatment and control samples) are compared. Initially, the samples are printed on a glass microscope slide by a robotic arrayer, thus, forming circular spots of known diameter (Schena, 2000). From each sample, the RNA (Ribonucleic acid) is extracted and is labeled with a fluorescent dye [Cy-3(green) for the control and Cy-5 (red) for the treatment sample]. Following labeling, RNA samples are mixed, are competitively hybridized at each spot of the microarray slide, and the slide is scanned, using suitable wavelengths to capture red and green dyes, resulting in two images, one for each dye (Jain, 2004). The relative fluorescence intensity between the two dyes (red/green) in each spot represents the expression level of the corresponding gene.

In order to extract those relative intensities from microarray images (Schena, 2002; Wang and Ghosh, 2001), a series of image analysis techniques have been proposed namely griding (Li *et al*., 2005; Rueda and Vidyadharan, 2006), spot segmentation (Angulo and Serra, 2003; Barra, 2006; Demirkaya *et al*., 2005; Nagarajan, 2003; Rahnenfuhrer, 2005; Stanley *et al*., 2002), and intensity extraction (Yang *et al*., 2002). Extracted mean intensities correspond to gene expression levels that are translated into biological conclusions by molecular biologists using data mining techniques (Eisen *et al*., 1998) for clustering genes with similar expression levels, for identifying differentially expressed genes, etc. (Chen and Liu, 2005).

One of the major factors that complicate the task of image analysis and data mining is that microarray images are contaminated by various types of noise (biological and experimental). Improper treatment of noise may result in erroneous biological conclusions (Nykter *et al*., 2006). Biological noise is intrinsic, it includes the stochastic internal noise of the cell and error sources related to sample preparation (Blake *et al*., 2003), and it induces image blurring (Nykter *et al*., 2006). Experimental noise can be subdivided into source noise and detector noise. Source noise is generated during the fabrication and target labeling, whereas detector noise is generated during the amplification and digitization stages. These types of noise produce microarray images, which are corrupted by irregularities in the shape, size and position of the spots, and are dominated by spatially inhomogeneous noise (Balagurunathan *et al*., 2004).

One of the most undesirable effects of noise is that it causes inaccurate spots' segmentation (i.e. the boundaries of spots are erroneously estimated). The latter, as a direct effect, evokes wrong estimation of the relative mean spots' intensities and reduces the reproducibility and validity of the gene expression levels, derived from microarray images. Noise complicates all microarray image processing tasks (gridding, segmentation, intensity extraction), but mostly segmentation. For noiseless images, spot segmentation would have been a trivial task even by using simple segmentation methods, but this is not the case. It has been shown that 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 (Ahmed *et al*., 2004). The question then arises: which segmentation method is the most accurate and why different methods lead to different differentially expressed genes? The answer is not straightforward. The segmentation method, as an individual process, may give accurate spot boundary detection (this can only be objectively assessed using simulated data) but its combination with preceding gridding and subsequent data analysis does not necessarily guarantee that the end result—the gene expression quantification—will be more accurate. It turns out that different differentially expressed genes are obtained even by changing the gridding or the data analysis technique. Thus, it is not only important to assess the performance of each analysis stage independently, as it has been done in most previous studies (Yang *et al*., 2002) (i.e. whether gridding or spot boundary detection is accurate or not) but also the performance of all processing steps as a whole in terms of reproducibility and validity in computing gene expression levels.

From the above, it is evident that noise reduction is an essential process, which has to be incorporated into the microarray image analysis pipeline. One possible solution proposed in previous studies (Lukac and Smolka, 2003; Lukac *et al*., 2005; Mastriani and Giraldez, 2006; Wang *et al*., 2003) for addressing microarray image noise is image enhancement. Results of these studies have indicated a superior quality of the enhanced images, without however examining whether enhancement leads to more accurate spot segmentation or reduces the variability of the extracted gene expression levels. What is missing here is a complete framework of microarray image processing steps that will properly model and address the effects of noise in such a way that it will not only increase the accuracy of spot segmentation but also the reproducibility and validity of gene expression levels.

This article presents a robust framework for microarray image analysis, which is designed to take into account the effect of local spot-image noise in microarray images for improving spot segmentation and subsequently gene quantification. The proposed framework incorporates in the microarray image analysis pipeline a novel combination of image processing and analysis techniques originating from the comprehensive quantitative investigation of the impact of noise on spot segmentation and intensity extraction. In details, the proposed framework consists of the following stages: (1) gridding for facilitating spot identification, (2) clustering (unsupervised discrimination between spot and background pixels) applied to spot image for automatic local noise assessment, (3) modeling of local image restoration process for spot image conditioning (adjustable wiener restoration using an empirically determined degradation function), (4) automatic spot segmentation employing seeded-region-growing, (5) intensity extraction and (6) assessment of the reproducibility (real data) and the validity (simulated data) of the extracted gene expression levels.

The proposed method was comparatively evaluated against well-established publicly available software packages, Scanalyze (fixed circle) and SPOT (seeded region growing) (Eisen, 1999; Yang *et al*., 2002) and a recent study (Baek *et al*., 2007). Comparisons with available software were performed on both simulated and real microarray images in terms of valid and reproducible extraction of gene expression levels, which is the case of concern in microarray image processing task.

2 METHOD

Gene quantification is affected by various image degradation processes that reduce microarray image quality, resulting in erroneous delineation of the spots' boundaries. The image degradation process (Gonzalez and Woods, 1992) may be formulated in the spatial domain as shown in Equation (1):

$$g(x,y) = f(x,y) * h(x,y) + n(x,y) \quad (1)$$

where, $g(x, y)$ is the degraded microarray image, $f(x, y)$ is the original image, $h(x, y)$ is the degradation process, $n(x, y)$ is image noise, considered additive and the symbol '*' indicates convolution.

In the case of microarray images, the degradation process $h(x, y)$ may be considered approximately constant across the image and it reflects the end result of the degradations, caused by the cell-population effect $h_{\mathrm{CFP}}(x, y)$ (Lähdesmäki *et al*., 2003) and the image acquisition apparatus $h_{\mathrm{Apparatus}}(x, y)$, as shown in (2):

$$h(x,y) = h_{\mathrm{CPF}}(x,y) * h_{\mathrm{Apparatus}}(x,y) \quad (2)$$

Regarding the noise term of equation (1), it includes both biological errors and measurements errors, which can be presented in the compact form of Equation (3) (Nykter *et al*., 2006):

$$n(x,y) = m(x,y) + l(x,y) \quad (3)$$

where $m(x, y)$ is a non-linear function depending on the gene expression level of each spot of the microarray image and $l(x, y)$ is a signal independent error term. Thus, $n(x, y)$ may be considered to depend on the local properties of the microarray image and, in particular, of the spot-image. As a consequence, a solution to Equation (1) with respect to $f(x, y)$ should be given locally by processing each individual spot-image independently, i.e. in a spot-image adjustable manner. Such a measure would produce a restored version of each spot-image that would facilitate accurate spot boundary determination and, thus, improved gene quantification.

Accordingly, a microarray griding procedure to identify and isolate individual spot-images must be initially applied on the microarray images. Such a procedure would produce a series of rectangular spot-images, each one consisting of a spot-region and a background-region (see Section 2.1).

Although, exact estimation of noise at each spot-image point may not be possible, estimation of the general noise statistics may be obtained from the spot-image's background-region, by means of the region's variance σ^2. Thus, the Fuzzy C-Means (Bezdek, 1981) unsupervised classification (clustering) method was employed to roughly separate the two regions (see Section 2.2). The background-region was used to assess noise (σ^2) while the spot-region provided an initial estimation of the spot's position and centroid for use as starting point by the seeded region growing (SRG) segmentation algorithm (see Section 2.4).

Assessment of spot-image noise may now provide an approximate estimate $\hat{f}(x, y)$ of spot-image $f(x, y)$ in Equation (1) (now considered to represent the degradation model of each spot-image) by Wiener restoration (Gonzalez and Woods, 1992) (see Section 2.3).

Restored spot-images $\hat{f}(x, y)$ were finally segmented using the SRG algorithm (Hojjatoleslami and Kittler, 1998) (see Section 2.4).

The spot-region's boundary, thus determined, was referred to the corresponding spot-image in the original microarray image and the spot-region's intensity was evaluated as the mean value of all pixels contained within the boundary. This was necessary, since intensities in the processed spot-images were altered by the restoration process.

2.1 Microarray image gridding

Since typical microarray images contain thousands of spots, the gridding method must be characterized by accuracy, automation and simplicity (Blekas *et al.*, 2005). In a recent study (Rueda and Vidyadharan, 2006), a highly accurate and simple gridding procedure has been proposed, which takes no assumptions of microarray slide details (i.e. number of spots, spots' size, etc.) and requires only the boundaries of each subgrid to be specified. A similar gridding procedure was employed by the proposed method for locating spot-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 (Cleveland, 1979), it ensured minimization of irregularities, introduced by the printing procedure, and, therefore, success of the gridding procedure. The bandwidth used for the smoothing process approximately equals the width of a typical spot. Spot sites, in terms of width and height, were finally estimated from the peak-valley distance in each line profile. Mathematical formulation of the aforementioned procedure is provided in Section S1.A of the Supplementary Material.

2.2 Clustering for local noise and spot position estimation

The Fuzzy C-Means unsupervised classification (clustering) algorithm searches iteratively for cluster centers (centroids) that minimize the dissimilarity function (Bezdek, 1981):

$$J = \sum_{i=1}^{M} J_i = \sum_{i=1}^{M} \sum_{j=1}^{N} u_{i,j}{}^{m} d_{i,j}{}^{2}(\mathbf{x}_j, \mathbf{c}_i) \tag{4}$$

where: $\mathbf{x}_j, j = 1, 2, \ldots, N$, are the pixels of the spot-image, $\mathbf{c}_i$, $i = 1,2, \ldots, M$, are the cluster centers, $d_{i,j}$ is the Euclidean distance between centroid c_i and data point $\mathbf{x}_j$, and u_{ij} is the element of a fuzzy membership function matrix $U = [u_{ij}]$ with values $0 \leq u_{ij} \leq 1$ and m is a weighting exponent ($m = 2$). The output of the iterative procedure is two clusters containing the pixels belonging to spot-region and background-region (see Section S1.B of the Supplementary Material).

2.3 Spot image restoration

Considering the discrete fourier transform (DFT), of Equation (1) we obtain Equation (5):

$$G(u, v) = F(u, v) \cdot H(u, v) + N(u, v) \tag{5}$$

where $G(u, v)$, $F(u, v)$, $H(u, v)$, and $N(u, v)$ are the DFTs of $g(x, y)$, $f(x, y)$, $h(x, y)$, and $n(x, y)$ respectively and u, v are spatial frequencies.

An estimation $\hat{\mathrm{F}}(x, y)$ of the original image $F(u, v)$ may be provided by the Wiener restoration algorithm (Gonzalez and Woods, 1992):

$$\hat{F}(u,v) = \left[\frac{\left|\mathrm{H}(u,v)\right|^2}{\left|H(u,v)\right|^2 + K} \right] \frac{G(u,v)}{H(u,v)} \tag{6}$$

where K is a constant that can be approximated by $K = 2 \times \sigma^2$ (Gonzalez and Woods, 1992) where σ^2 is the spot's background-region variance.

Regarding the degradation function, the authors of a previous study (Nykter *et al.*, 2006) have proposed a 9-point kernel $\{10E-8,\ 10E-4,\ 0.152,\ 0.312,\ 0.362,\ 0.162,\ 0.12,\ 10E-4,\ 10E-8\}$ in the spatial domain. We found that the spectral response of that kernel could be adequately represented (0.0025 in terms of root mean square error) by the spectral response of a low-pass Butterworth filter, shown in (7):

$$Fh^{\mathrm{LP}}(v) = \frac{1}{1 + 0.414(v/f_{\mathrm{co}})^{2n}} \tag{7}$$

where n is the degree of the filter, v is the spatial frequency, f_{co} the cut-off frequency.

Subsequently, the 2D $H(u, v)$ was modeled as in (8) (Gonzalez and Woods, 1992):

$$H(u,v) = Fh^{LP}(\sqrt{u^2 + v^2}) \tag{8}$$

$$\sqrt{u^2 + v^2} <= N \tag{9}$$

where, N is the maximum dimension of the spot-image (which was zero-padded in the case of non-square spot-image).

The restored spot-image was transferred into the spatial domain by the 2D Inverse DFT (2d-IDFT) of (5) as:

$$\hat{f}(x,y) = 2\mathrm{d} - \mathrm{IDFT}(\hat{F}(u,v)) \tag{10}$$

2.4 Spot image segmentation and intensity extraction

Restored spot-images were segmented using the SRG algorithm (Hojjatoleslami and Kittler, 1998). SRG initially segmented each spot-image into spot-regions of pixels starting from the spot's center, as determined by the Fuzzy C-Means rough 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 spot-image continued until all pixels of the spot-image were assigned to either the spot-region or its background. Mathematical description is provided in Section S1.C of the Supplementary Material.

3 PERFORMANCE EVALUATION

Since performance evaluation of microarray segmentation is not a straightforward task to consider (Lehmussola *et al.*, 2006), we used as test images a set of customized synthetic microarray images (with no artifacts), produced by a microarray simulator (Martin and Horton, 2004). In each synthetic/simulated image, pixels were pre-assigned as spot or background.

3.1 Simulated experiments

Initially, a pair of microarray grayscale TIFF images, representing the red and green channels of a two-color experiment, containing 200 different spots was produced by the microarray image simulator. In this pair of images, spots' background was initial set to be zero. Therefore, spots' boundaries were known a priori. Based on this gold standard pair of images (reference images), a series of customized test images were further produced. Initially, blurring introduced from biological noise was modeled by convolving the image in the frequency domain with a first order low-pass Butterworth filter using cut off frequencies in the range of $0.1 \times \mathrm{N}$ to $0.9 \times \mathrm{N}$ (nine pairs of images) where N is the dimension of the image (non-square images were zero-padded). Furthermore, on the blurred images, experimental noise, modeled as additive, signal dependent, random noise for four different noise percentage levels (10, 30, 50 and 70%), was introduced (36 pairs of images). Resulting images (overall 45 different images of 200 spots each) contained spots of various shapes and sizes, aiming at complicating the spots' segmentation and consequently the intensities extraction procedure. Data are available for downloading from: ftp://users:bioinformatics@mipa.med.upatras.gr.

For assessing the pixel-based segmentation accuracy of the proposed method, we selected two traditional measures namely the discrepancy which was based on the number of mis-segmented pixels and the discrepancy which was based on the position of mis-segmented pixels (Zhang, 1996). These methods provided information not only for the number of erroneously segmented pixels but also for their spatial location in order to ensure that different segmented images provided the same discrepancy measure values.

The discrepancy that was based on the number of mis-segmented pixels was assessed using the probability of error (PE), defined as (Lee *et al.*, 1990):

$$\mathrm{PE} = \mathrm{P(O)} \times \mathrm{P(B|O)} + \mathrm{P(B)} \times \mathrm{P(O|B)} \tag{11}$$

where P(B|O) is the probability of error in classifying objects as background, P(O|B) is the probability of error in classifying background as objects, P(O) and P(B) are a priori probabilities of objects and background in images. For our case spot is considered to be the object that must be discriminated from the background.

The discrepancy that was based on the position of mis-segmented pixels was defined as (Yasnoff *et al.*, 1977):

$$D = \frac{\sqrt{\sum_{i=1}^{N} d^2(i)}}{A} \tag{12}$$

where N is the number of mis-segmented pixels, d(i) the Euclidean distance between the ith mis-segmented pixel and the nearest pixel of its true class and A is the number of pixels in the image.

Although, pixel-based segmentation performance is the best way to objectively characterize segmentation schemes, publicly available software packages that were used in this study do not provide such information, i.e. boundaries of spots' and background regions. So, in order to present comparable results and ensure their validity we had to further calculate the pairwise differences between the extracted spots' intensities (for each one of the 45 evaluated images) and the original synthetic image spots' intensities using the mean absolute error (MAE) (Lehmussola *et al.*, 2006).

3.2 Real experiments

Microarrays used in this study comprised a publicly available dataset of seven images obtained from the database of the MicroArray Genome Imaging & Clustering Tool (MAGIC) website (Heyer). Each image contained 6400 spots investigating the diauxic shift of *Saccharomyces cerevisiae*. Images included spots of various shapes as well as artifacts (scratches and dust). The particular dataset was selected because the authors (DeRisi *et al.*, 1997) used a common reference messenger

RNA pool (green, Cy-3) to control for biological variability (Churchill, 2002).

Thus, exploiting the benefits of the replicated common reference channel (Cy-3), we quantitatively assessed the performance of the proposed method in terms of extracted genes expression reproducibility using the coefficient of variation metric (CV) [Equation (13)], since each spot in the common reference channel should have the same intensity throughout the replicated experiments.

$$CV = \frac{\sigma}{\mu} \tag{13}$$

where σ is the SD and μ is the mean value for each spot evaluated for all the replications (seven replications totally). CV allows for the comparison of variability estimates regardless of the magnitude of the measurement (Reed *et al.*, 2002). Additionally, in order to quantify the efficiency and robustness of the proposed method, we calculated the pairwise MAE between the replicates (altogether 21 pairwise MAE values) for the common reference channel.

Extracted intensities, for the same series of microarray images, were comparatively evaluated against the intensities obtained from both commercial software used in the current study (Scanalyze and SPOT) and the recent study of Baek *et al.* (Baek *et al.*, 2007). All extracted intensities were normalized using global normalization (Schuchhardt *et al.*, 2000).

4 RESULTS

The degradation function $H(u, v)$ in Equation (6) was optimally designed with respect to simulated data segmentation accuracy, by a first degree ($n = 1$) low-pass Butterworth filter using $f_{co} = 0.6 \times N$, with N being the spot-image dimension, and it was modeled according to Equations (7–9). The other parameter in equation (6) that needed to be specified was $K = 2 \times \sigma^2$, an estimate of the spot-image's background noise. That was computed as the SD of the spot-image's background region. The latter was automatically determined by the Fuzzy C-Means clustering algorithm.

Regarding the segmentation accuracy of the proposed method, the mean value of the probability of error segmentation metric (concerning the 200 segmented spots), for the 45 evaluated images, ranged between 0.055–0.130 and 0.037–0.097 with mean value 0.084 and 0.067 for the red and green channels, respectively. Additionally, to depict the improvement on the segmentation procedure stage due to the intermediate step of image restoration, we compared the results of the proposed method with an implementation of the same procedure but without the step of image restoration. Results for the segmentation metric of the discrepancy, based on misclassified pixels positions, were 0.022–0.027 and 0.024–0.029 with mean value 0.022 and 0.024 with and without the restoration step, respectively.

Following the segmentation procedure, the extracted intensities were compared with the results obtained from both commercial software used in this study by measuring the pairwise MAE as explained in Section 3.1. Boxplots of Figure 1 illustrate the MAE values for the series of the customized simulated images.

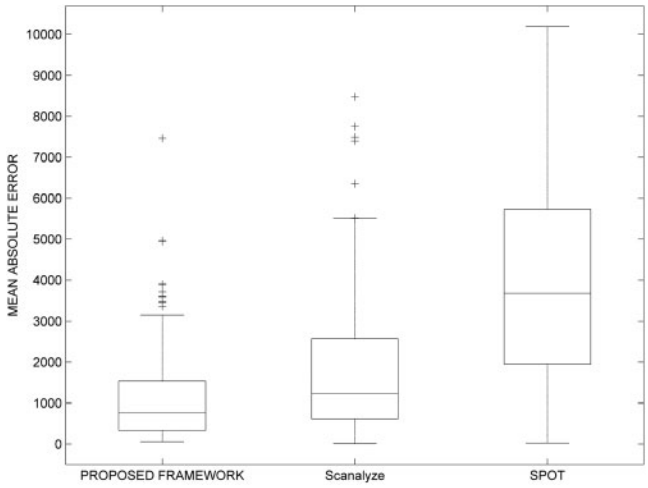

Fig. 1. MAE for the simulated data. Obelisks are MAE values characterized as outliers.

Table 1. Mean values (in terms of intensity) of the MAE boxplots of Figure 1

	Proposed method	Scanalyze	SPOT
Red channel	1110.8	1779.1	3908.1
Green channel	1169.9	1877.5	5370.2

Table 1 provides the mean values of the MAE boxplots (Fig. 1), in terms of intensity, for the evaluated 45 images and for both channels.

Regarding real microarray images, $H(u, v)$ was empirically determined with respect to the minimization of CV. $H(u, v)$ parameters for optimal performance were ($n = 1, f_{co} = 0.6 \times N$). Figure 2 illustrates the actual distribution of CV values of the extracted gene expression levels from the set of seven 1024×1024 16bit replicated images (Cy-3), as they were calculated by the proposed method, the Scanalyze, the SPOT and the Baek's method, respectively. Accordingly, the calculated CV values were 0.211 for the proposed method, 0.228 for the SPOT 0.288 for the Scanalyze software and 0.299 for Baek's *et. al.*'s procedure.

Figure 3 shows the calculated pairwise MAE between the expression ratios of all possible pairs of the common reference channel for the dataset of the seven replicated real images.

Table 2 provides the mean values of the pairwise MAE (Fig. 3) as they calculated for the seven replicates of the common reference channel.

5 DISCUSSION

Microarray technology has transformed the field of genomic research by allowing the simultaneous profiling of thousands of genes. The microarray process is based entirely on the accurate extraction of quantitative information from images. In the present study, a robust framework for microarray image analysis was developed, proposing a novel combination of

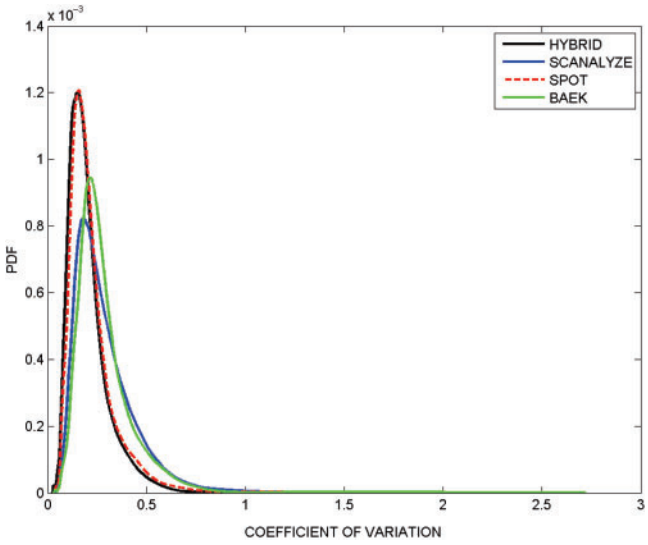

Fig. 2. Probability density functions (PDF's) of the coefficient of variation for all the spots as calculated from the seven replications of the common reference channel. Black line corresponds to the results obtained using the proposed method. Blue, red and green line correspond to the Scanalyze, SPOT and Baek's approach, respectively.

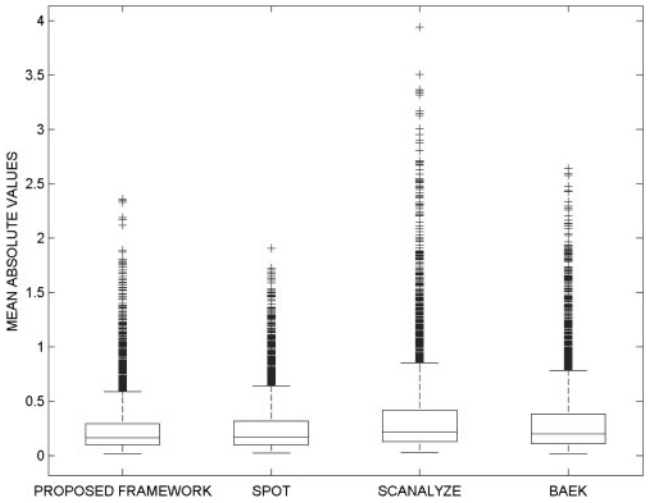

Fig. 3. Boxplots illustrating the pairwise MAE between all replicates (totally 21 MAE values from which the mean value for each spot is illustrated here). Obelisks are MAE values characterized as outliers.

image processing and analysis techniques. The proposed framework was derived following the quantitative investigation of the impact of noise on spot segmentation and intensity extraction and consists of the following stages: (1) grid creation for facilitating spot identification (gridding), (2) noise parameters assessment for noise modeling (noise estimation), (3) application of image restoration process for noise reduction (adaptive wiener restoration using an empirically determined degradation function), (4) segmentation for spots identification on the restored images and (5) intensity extraction.

Table 2. Mean values of the calculated 21 pairwise MAE for the common reference channel

	Proposed method	Scanalyze	SPOT	Baek *et al.*
Common reference channel (Green)	0.254	0.362	0.262	0.323

The proposed method was comparatively evaluated against the well-established methods of Scanalyze (fixed circle) and SPOT (seeded region growing) (Eisen, 1999; Yang *et al.*, 2002), employing both simulated and real microarray images, and against a recent study (Baek *et al.*, 2007).

Regarding pixel-based segmentation accuracy on the simulated images, the proposed methodology achieved high segmentation results. Even though the image quality of the evaluated images varied significantly, the accuracy of the proposed methodology, in terms of mean probability error for the 200 spots, remained high. The success is mostly due to the intermediate step of adaptive spot restoration. According to the results provided by the metric of the mean discrepancy error based on misclassified pixels position, the intermediate step of adaptive image restoration facilitated the segmentation procedure, since segmentation accuracy without this intermediate step was lower. Thus, the initial Fuzzy C-Means segmentation procedure is of major importance, since it provides the necessary information to estimate the noise parameter which, in turn, is used to restore spot-images.

To obtain comparable results with existing software and considering that available software does not provide information about pixel-based segmentation performance, we calculated the pairwise MAE for the extracted intensities by the proposed method, the SPOT and the Scanalyze software. Figure 1 shows the MAE boxplots as calculated for 200 spots in 45 customized test microarray images and Table 1 illustrates the mean values of those boxplots. The goal was to minimize MAE, since such a result proves the validity of the extracted intensities. As the results clearly support, the proposed framework outperformed commercial software providing intensities closer to those of the simulated images.

Regarding real images we had to assess the performance of the proposed method against the SPOT and Scanalyze software in terms of providing reproducible gene expression levels, since the actual spot boundaries (and subsequent spot intensity levels) on the real images were not available. For this reason, we selected to evaluate a dataset (DeRisi *et al.*, 1997), which was designed to control the biological variability and reduce the experimental variation in a microarray experiment. Accordingly, for the common reference channel (Cy-3, Green channel), an adequate degree of replication was provided to quantitatively assess the reproducibility of the extracted intensities. Due to the replication, 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 2 shows the PDF of the coefficient of variation for the

common reference channel for all the images in the dataset using the proposed method (black line), the SRG method implemented in SPOT (red line), the fixed circle method used in Scanalyze (blue line) and gamma-t mixture model (green line) employed in Baek *et al*. (Baek *et al*., 2007). The proposed method's PDF is narrow and sharp with a peak-value close to zero in contrast to Scanalyze's PDF and Baek's method, which is more spread and far from zero. Regarding SPOT's PDF curve, while narrow and sharp is further away from zero as compared with the proposed method's curve. This may be seen by comparing the corresponding CV values, 0.211, 0.228, 0.288, 0.299, for the proposed method, SPOT, Scanalyze and Baek's method, respectively. Since the plots of Figure 2 represent PDF's, a highly peaked and narrow curve close to zero represents a microarray image processing methodology, which results in more reproducible extracted intensities and, thus, in more repeatable computation of gene's expression levels.

Exploiting the benefits of the provided replication in real images, we explored the validity of the extracted gene expression levels by measuring the 'sameness' of replicates using their pairwise MAE (totally 21 pairwise MAE values). Figure 3 illustrates the boxplots of MAE as they were calculated for the common reference channel of the seven replicated microarray images and Table 2 depicts the mean values of those boxplots. Lower MAE are indicative of higher segmentation performance and, thus, of more accurate (*valid*) extraction of gene expression levels. Again, as shown in Table 2, the proposed method achieved better results than the publicly available software and Baek's method. This may be due to the employment by our method of the automatic local restoration step, which incorporated in the procedure valuable structural information from the spot's background, as estimated by the Fuzzy C-Means clustering.

Regarding processing time, the proposed method took $\sim$300 s to extract the intensities from a 1024×1024, 16bit cDNA image, containing 6400 microarray spots. This may seem computationally intensive and time consuming as compared to commercial software used in the present study, since the code has not been optimized for speed, as yet. On the contrary, the proposed method proved to be more robust and efficient, since it provided more accurate and reproducible results, which is the case of concern in microarray image processing tasks.

6 CONCLUSION

The findings of the present study revealed that by applying local spot-image restoration and by incorporating structural information from the spot-image, spot-image segmentation and, consequently, quantification of gene expression is improved. This is a step that publicly available and commercial software should take into account.

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.

Conflict of Interest: none declared.

REFERENCES

Ahmed,A.A. *et al*. (2004) Microarray segmentation methods significantly influence data precision. *Nucleic Acids Res*., **32**, e50.

Alizadeh,A. *et al*. (1998) Probing lymphocyte biology by genomic-scale gene expression analysis. *J. Clin. Immunol*., **18**, 373–379.

Angulo,J. and Serra,J. (2003) Automatic analysis of DNA microarray images using mathematical morphology. *Bioinformatics*, **19**, 553–562.

Baek,J. *et al*. (2007) Segmentation and intensity estimation of microarray images using a gamma-t mixture model. *Bioinformatics*, **23**, 458–465.

Balagurunathan,Y. *et al*. (2004) Noise factor analysis for cDNA microarrays. *J. Biomed. Opt*., **9**, 663–678.

Barra,V. (2006) Robust segmentation and analysis of DNA microarray spots using an adaptative split and merge algorithm. *Comput. Methods Programs Biomed*., **81**, 174–180.

Bezdek,J.C. (1981) *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York.

Blake,W.J. *et al*. (2003) Noise in eukaryotic gene expression. *Nature*, **422**, 633–637.

Blekas,K. *et al*. (2005) Mixture model analysis of DNA microarray images. *IEEE Trans. Med. Imaging*, **24**, 901–909.

Chen,Z. and Liu,L. (2005) RealSpot: software validating results from DNA microarray data analysis with spot images. *Physiol. Genomics*, **21**, 284–291.

Churchill,G.A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat. Genet*., **32** (Suppl.), 490–495.

Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc*., **74**, 829–836.

Demirkaya,O. *et al*. (2005) Segmentation of cDNA microarray spots using Markov Random Field Modeling. *Bioinformatics*, **21**, 2994–3000.

DeRisi,J.L. *et al*. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eisen,M.B. (1999) Scanalyze. http://rana.stanford.edu/software

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Gonzalez,R.C. and Woods,R.E. (2002) *Digital Image Processing*, Addison-Wesley Longman Publishing Co., Boston.

Heyer,L. (2000) Magic Tool Database. http://www.bio.davidson.edu/projects/MAGIC/MAGIC.html.

Hojjatoleslami,S.A. and Kittler,J.R.I.T. (1998) Region growing: a new approach. *IEEE Trans. Image Process*., **7**, 1079–1084.

Jain,K.K. (2004) Current status of fluorescent in-situ hybridisation. *Med. Device Technol*., **15**, 14–17.

Lähdesmäki,H. *et al*. (2003) Estimation and inversion of the effects of cell population asynchrony in gene expression time-series. *Signal Processing*, **83**, 835–858.

Lee,S.U. *et al*. (1990) A comparative performance study of several global thresholding techniques for segmentation. *CVGIP*, **52**, 171–190.

Lehmussola,A. *et al*. (2006) Evaluating the performance of microarray segmentation algorithms. *Bioinformatics*, **22**, 2910–2917.

Li,Q. *et al*. (2005) Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics*, **22**, 2875–2882.

Lukac,R. and Smolka,B. (2003) Application of the adaptive center-weighted vector median framework for the enhancement of cDNA microarray. *Int. J. Appl. Math. Comput. Sci*., **13**, 369–383.

Lukac,R. *et al*. (2005) cDNA microarray image processing using fuzzy vector filtering framework. *J. Fuzzy Sets Syst*., **152**, 17–35.

Martin,B. and Horton,R.M. (2004) A Java Program to Create Simulated Microarray Images. *IEEE Computational Systems Bioinformatics Conference (CSB 2004)*, .

Mastriani,M. and Giraldez,A.E. (2006) Microarrays denoising via smoothing of coefficients in wavelet domain. *Int. J. Biomed. Sci*., **1**, 1306–1216.

Nagarajan,R. (2003) Intensity-based segmentation of microarray images. *IEEE Trans. Med. Imaging*, **22**, 882–889.

Nykter,M. *et al*. (2006) Simulation of microarray data with realistic characteristics,. *BMC Bioinformatics*, **7**, 349.

Rahnenfuhrer,J. (2005) Image analysis for cDNA microarrays. *Methods Inf. Med*, **44**, 405–407.

Reed,G.F. *et al*. (2002) Use of coefficient of variation in assessing variability of quantitative assays. *Clin. Diagn. Lab. Immunol*., **9**, 1235–1239.

Rueda,L. and Vidyadharan,V. (2006) A hill-climbing approach for automatic gridding of cDNA microarray images. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **3**, 72–83.

Schena,M. (2000) *Microarray Biochip Technology*. Eaton Publishing Company, USA.

Schena,M. (2002) *Microarray Analysis*. Wiley-Liss, New York.

Schena,M. *et al*. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schuchhardt,J. *et al*. (2000) Normalization strategies for cDNA microarrays. *Nucleic Acids Res*., **28**, E47.

Stanley,R.J. *et al*. (2002) Microarray image spot segmentation using the method of projections. *Biomed. Sci. Instrum*., **38**, 387–392.

Wang,X. and Ghosh,S. (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res*., **29**, E75.

Wang,X.H. *et al*. (2003) Microarray image enhancement by denoising using stationary wavelet transform. *IEEE Trans. Nanobioscience*., **2**, 184–189.

Yang,Y.H. *et al*. (2002) Comparison of methods for image analysis on cDNA microarray data. *J. Comput. Graph Stat*., **11**, 108–136.

Yasnoff,W.A. *et al*. (1977) Error measures for scene segmentation. *Pattern Recognit*., **9**, 217–231.

Zhang,Y.J. (1996) A survey on evaluation methods for image segmentation. *Pattern Recogni*t., **29**, 1335–1346.