

An automated method for gridding and clustering-based segmentation of cDNA microarray images

Nikolaos Giannakeas^{a,b}, Dimitrios I. Fotiadis^{b,c,*}

^a Laboratory of Biological Chemistry, Medical School, University of Ioannina, Ioannina, Greece

^b Unit of Medical Technology and Intelligent Information Systems, Department of Computer Science, University of Ioannina, Ioannina, Greece

^c Biomedical Research Institute - FORTH, Ioannina, Greece

ARTICLE INFO

Article history:

Received 24 March 2008

Received in revised form

18 September 2008

Accepted 6 October 2008

Keywords:

Microarray image processing

Gridding

Segmentation

K-means

Fuzzy C means

ABSTRACT

Microarrays are widely used to quantify gene expression levels. Microarray image analysis is one of the tools, which are necessary when dealing with vast amounts of biological data. In this work we propose a new method for the automated analysis of microarray images. The proposed method consists of two stages: gridding and segmentation. Initially, the microarray images are preprocessed using template matching, and block and spot finding takes place. Then, the non-expressed spots are detected and a grid is fit on the image using a Voronoi diagram. In the segmentation stage, K-means and Fuzzy C means (FCM) clustering are employed. The proposed method was evaluated using images from the Stanford Microarray Database (SMD). The results that are presented in the segmentation stage show the efficiency of our Fuzzy C means-based work compared to the two already developed K-means-based methods. The proposed method can handle images with artefacts and it is fully automated.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

The analysis of gene expression has become simple and efficient using microarray technology [1]. The cycle of the microarray experiment begins with the biological question and ends with the data analysis results along with the biological conclusions, which might lead to a new question and so on. Microarrays can yield expression profiles for thousands of genes simultaneously in a single hybridization experiment.

During a biological experiment, two messenger Ribonucleic Acid (mRNA) samples are reverse transcribed into complementary Deoxyribonucleic Acid (cDNA). Most experiments compare a normal sample with a cancer sample in order to find, which genes are related to the current type of cancer. Deoxyribonucleic Acid (DNA) obtained from thousands of known genes of interest is printed on a glass microscope slide by a robotic arrayer. The cDNA samples, labeled with two different fluorescent dyes, are hybridized with the known genes on the slide at the same time. The most common dyes for tagging cDNA are the red fluorescent dye Cy5 (emission from 630 to 660 nm) and the green-fluorescent dye Cy3 (emission from 510 to 550 nm) [2].

The microarray images are generated using a scanner. The slide is scanned in the two wavelengths, producing two 16-bit images corresponding to the two fluorescent dyes. In terms of microarray image analysis the two channels are referred to as red and green channels. Thus, the image can be represented as an RGB image with blue being zero. The digitization of the image is modified in order to generate spots with 10 pixels diameter (2–5 mm). The images contain several blocks (or subgrids), which consist of several spots, placed in rows and columns. The level of intensity of each spot represents the amount of sample hybridized with the corresponding gene.

The processing of microarray images [3] provides the input for further analysis of the extracted microarray data [4]. It includes the following stages:

- Spot addressing and gridding, which are the processes of assigning the location of each spot and fit a grid on the image.
- Segmentation, which is the process of grouping the pixels with similar features (this step results in the separation of foreground and background pixels).
- Intensity extraction, which calculates red and green foreground fluorescence intensity pairs and background intensities.

An ideal microarray image must have the following properties [5]:

- All the blocks are of the same size.
- The spacing between the blocks is regular.

* Corresponding author at: Unit of Medical Technology and Intelligent Information Systems, Department of Computer Science, University of Ioannina, PO Box 1186, GR 451 10 Ioannina, Greece. Tel.: +30 26510 98803; fax: +30 26510 98889.

E-mail address: fotiadis@cs.uoi.gr (D.I. Fotiadis).

- The spots are centered on the intersections of rows and columns.
- The size and shape of the spots is perfectly circular and identical for all spots.
- The location of the blocks is fixed in the images for a given slide type.
- No dust or other contamination is present on the slide.
- The background intensity is minimal and uniform.

However, a scanned microarray image has none of the above properties. Thus, automated gridding and spot addressing is crucial for the microarray image processing.

The gridding methods proposed in the literature are manual, semi-automated and automated. Several software packages have been developed like Genepix [6] and ScanAlyze [7], which require inputs by the user. Towards automated methods, there exist several approaches, which process the vertical and horizontal projections of the image and generate lines in the image according to the valleys of the 1-D signal [8–10]. Morphological operators [11] and smoothing filtering [12] have also been used. Other methods preprocess the image using an initial segmentation and assign the centers of the objects in the image. Such techniques use histogram-based segmentation [13], template matching [14,16] combined with graph models [13], several types of transformations such as Affine [14], and Radon [15], or others, which take into account the symmetries of the microarray image [16]. The proposed gridding work attempts to engage the approaches, which are based on the template matching techniques with the approaches, where the projections of the image are processed. It consists of several steps to deal with the image rotation, the existence of low-intensity spots and the existence of artefacts and other types of noise. The current method is fully automated and does not require tuning of its parameters. In addition, our gridding approach is able to allocate effectively the non-expressed spots. The use of outlier detection can be considered as an advantage since artefacts existing in the image, are detected effectively.

The primary goal of microarray image processing is to extract the intensities of red and green channels. Thus, image segmentation must be applied, grouping the pixels of the image into foreground and background. The background intensities are used to adjust the foreground intensities for local noise, resulting in corrected red and green intensities for each spot [17].

The methods, which have been proposed for the segmentation of microarray images can be classified into four categories: (i) early approaches, which are based on fixed or adaptive circle segmentation. A circle is fit around each spot, characterizing the pixels in the circle as signal pixels and the pixels out of the circle as background pixels. Fixed and adaptive circle segmentation is used by ScanAlyze [7] and Dapple [18]. (ii) Histogram-based segmentation, which is used by QuantArray [19]. A threshold value is calculated [20], and pixels with intensity lower than the threshold are characterized as background pixels, whereas pixels with higher intensity as signal pixels. (iii) Adaptive shape segmentation methods, which are based on the watershed transform [21], seed region growing [22,23] and Markov Random Field (MRF) [24]. (iv) The most recent techniques, which employ clustering algorithms like K-means [25–29], Fuzzy C means (FCM) [25], Expectation–Maximization (EM) [9], Partitioning Around Medoids (PAM) [30] and model-based clustering [31]. Lehmußola et al. [32] compared most of the above segmentation approaches and they have shown that the most effective are those based on K-means clustering. The proposed segmentation method also employs clustering techniques, including an innovative set of features to deal with artefacts and donut spots. Instead, most of the already developed methods for the clustering-based segmentation of the microarray image use only the intensities of the two channels as features. On the other hand, a set of features is used by the pro-

posed method for each channel. In this way, the expression levels of each gene are accurately computed, according to the requirements of the biological experiment.

2. Material and methods

The proposed method consists of two stages. Fig. 1 demonstrates the flowchart of the proposed work. In the first stage, a gridding algorithm [34] is implemented to identify regions around each spot. Next, clustering techniques are used for the image segmentation in the above regions, which result in the identification of the signal and background pixels and artefacts. The first stage consists of five steps. Initially, the raw microarray image is preprocessed using a template matching technique and the image is rotated using least square fitting. In the second step, the blocks of the image are detected, using 1-D projections of the image. In the third step, the coordinates of the centers of each spot are addressed and in the fourth step the non-expressed spots are detected. Finally, in the last step a grid is fit on the image using a Voronoi diagram. In the second stage (segmentation), K-means and Fuzzy C means clustering is employed for grouping the pixels of the images into foreground, background pixels and artefacts. Note, that the above procedure is applied in parallel to both red and green channels.

2.1. Gridding

2.1.1. Preprocessing

The preprocessing of the image begins with the background removal. This kind of noise appears in most of the microarray images due to the emission of the slide, which affects the entire image. Since the number of background pixels is larger than that of the signal pixels, the median value of all pixel intensities is very close to the value of the global background. According to this assumption, the median value is used as the threshold for the removal of the noise, which is generated from the emission of the slide. The intensities of the pixels, which have intensities lower than the median intensity of the grayscale image, are set equal to zero [16]. All the other values of the image (i.e. the values, which have intensities greater the median value) are not changed during the background removal procedure.

In order to locate the objects of the image, which are similar to a spot, we perform an initial segmentation using template matching. The appropriate template is selected and placed at each pixel in the image. The size of the template is approximately equal to the size of the theoretical spot and the template follows a 2-D Gaussian distribution as it is shown in Fig. 2. Unfortunately, the size of the spots of the image varies due to the level of the hybridization. The aim of this procedure is to measure the similarity between the template and the image at each pixel. For this reason, we calculate the correlation coefficient of the values of the template with the corresponding values of the image. If all the spots have 11×11 size and their pixels follows the Gaussian distribution, the correlation coefficient of the center pixel of the spot will be equal to 1. As the correlation increases, the probability of a pixel to belong to a spot increases. Using a threshold value for the correlation coefficient, the image is converted to binary. This threshold was set to $r=0.3$, which has been obtained heuristically after several experiments.

Since the microarray is often not parallel to the image boundaries, the estimation of the rotation angle is necessary. For this reason, we compute and afterwards eliminate the rotation between the microarray and the boundaries of the image. The slope between the four boundaries of the microarray and the corresponding boundaries of the image is found. The median value of these four slopes is calculated and the array is rotated accordingly. In some

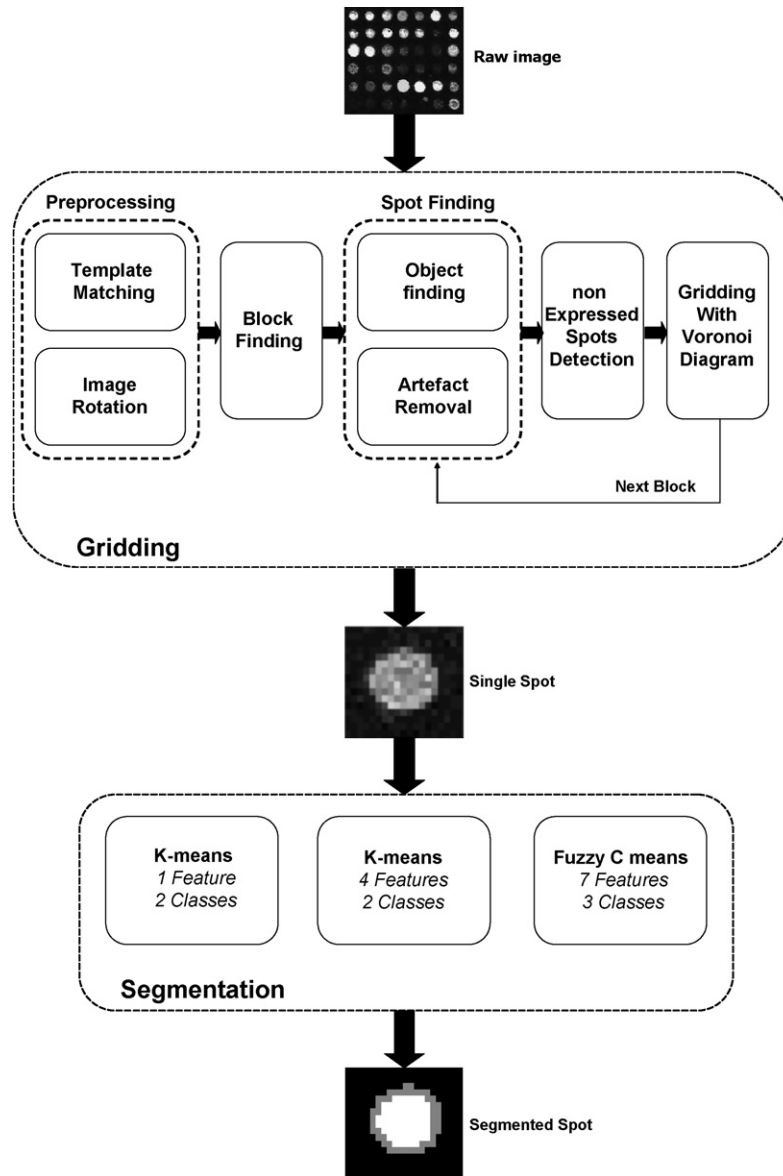


Fig. 1. The flowchart of the proposed method.

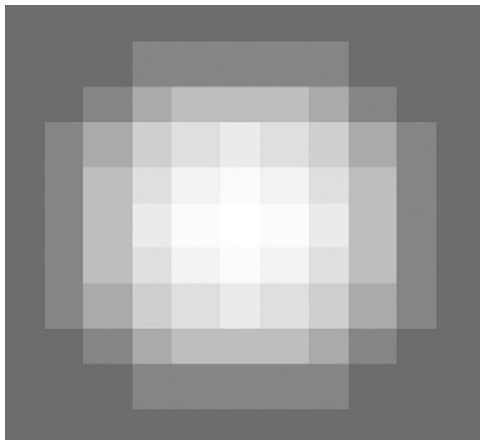


Fig. 2. 2-D 11×11 Gaussian template.

cases, due to the large number of artefacts the computed slope is large. This fact makes the use of median value of the four slopes more convenient than the mean value. For each boundary of the microarray the first pixels with non-zero intensities in each row or column of pixel are detected. For instance, for the upper boundary the first non-zero intensity pixels of each column of pixels in the image are detected. Having the above set of points, we fit a line using least squares fitting and we calculate the angle of this line with the corresponding boundary of the image. Following image rotation, we calculate the center of mass for all objects that are characterized as spots by the template matching procedure. In the calculation of centers of mass the intensity distribution of each spot is taken into account:

$$X_C = \frac{\sum_{i \in D} x_i I_i}{\sum_{i \in D} I_i}, \quad Y_C = \frac{\sum_{i \in D} y_i I_i}{\sum_{i \in D} I_i}, \quad (1)$$

where X_C and Y_C are the coordinates of the center of mass, I_i is the intensity of the i th pixel, and D is the area of the spot.

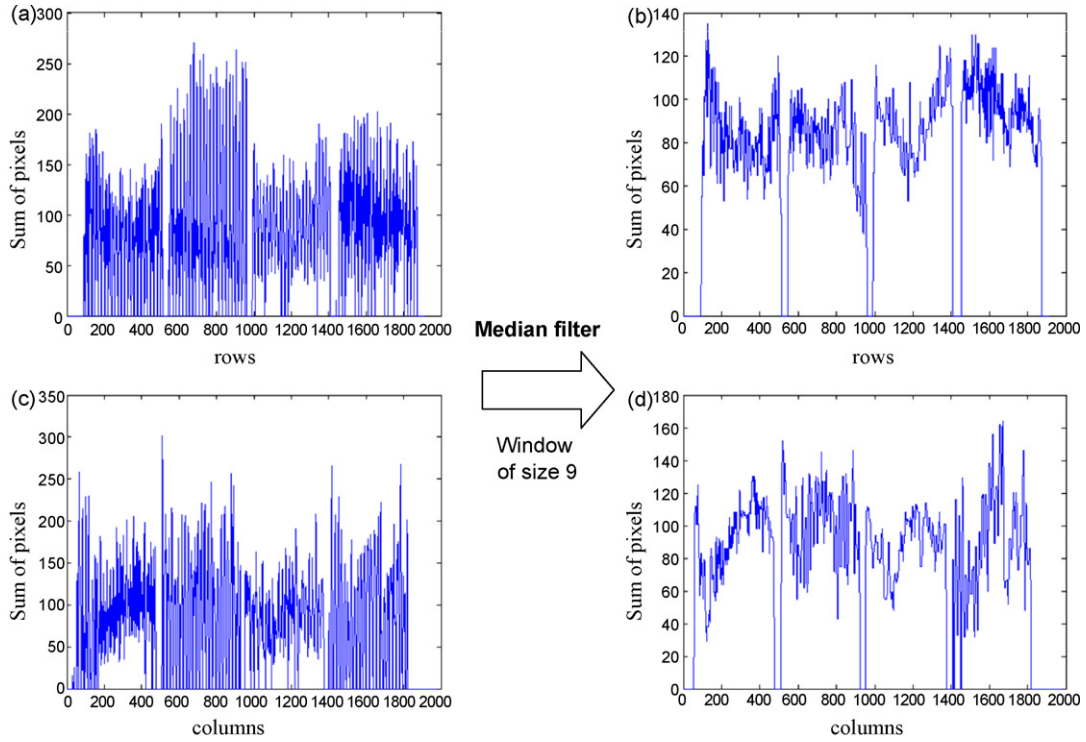


Fig. 3. (a) The projection of the Microarray image into the vertical direction, (b) the same projection smoothed using a median filter with window of size 9, (c) the projection of the microarray image into the horizontal direction, (d) the same projection smoothed using a 3×3 median filter.

2.1.2. Block detection

It is essential to process each block separately since some blocks are not of the same size and the spacing between the blocks is not the same. For this reason, we determine the area of each block in the image, in order to split the image into blocks. Initially, the elements of each row and column of the image are summed. This results in two 1-D signals, which are the projections of the image in the vertical and the horizontal direction. Finally, the signals are smoothed using a 3×3 median filter, as it is shown in Fig. 3.

Next, we detect the local minima (valleys) in both 1-D-signals, since during the smoothing procedure, the local minima are set equal to zero and the rest of the signal is smoothed to non-zero values. These valleys of the two 1-D signals assign the coordinates of the points that are used to split the image into individual blocks in the 2-D image. v_i and h_j are the coordinates of the valley, of the vertical and horizontal projections, respectively, $i = 1, 2, \dots, N_v$, $j = 1, 2, \dots, N_h$ and N_v and N_h are the total number of valleys in the vertical and horizontal direction, respectively. The vectors V and H contain the points, which are used to split the image into individual blocks:

$$V = [1, v_1, v_2, \dots, v_{N_v}, D_v], \tag{2}$$

$$H = [1, h_1, h_2, \dots, h_{N_h}, D_h], \tag{3}$$

where D_v and D_h are the dimensions of the image in the vertical and the horizontal direction, respectively.

2.1.3. Spot detection

In this step, the noise of the microarray (small and high intensity objects characterized by the template matching above as spots) is removed to obtain a well-defined grid, which specifies the regions of each spot. All the spots of a row should have similar vertical (y -coordinate) coordinates, while all the spot of a column should have approximately the same horizontal (x -coordinate) coordinates in the image. Other objects, whose coordinates differ

from the mean values of vertical or horizontal coordinates of the row or column, respectively, are probably artefacts and should be removed. For this reason, the coordinates of the centers of each spot are compared with the mean value of the coordinates of the centers of the spots, which belong to the same row or column with the corresponding spot. If the coordinates of the center of a spot satisfy the following criteria the spot is characterized as an outlier [34,35] and it is removed:

$$X_e(p) < \text{mean}(\text{COL}_x) - 2 \times \text{stddev}(\text{COL}_x), \tag{4}$$

$$X_e(p) > \text{mean}(\text{COL}_x) + 2 \times \text{stddev}(\text{COL}_x), \tag{5}$$

$$Y_e(p) < \text{mean}(\text{ROW}_y) - 2 \times \text{stddev}(\text{ROW}_y), \tag{6}$$

$$Y_e(p) > \text{mean}(\text{ROW}_y) + 2 \times \text{stddev}(\text{ROW}_y), \tag{7}$$

where X_e, Y_e are the coordinates of the center of the spot p , COL_x is the vector with x -coordinates of the column where spot p belongs, and ROW_y is the vector with y -coordinate of the row where spot p belongs. Suppose a row of a spots and the mean value of the y -coordinates of its members. If the y -coordinate of a spot satisfies the criterion of the inequality 7 this spot is located in a position lower than the other spots of the current row of the image, and thus this object is removed.

2.1.4. Non-expressed spot detection

The detection of the non-expressed spots is necessary in order to reduce the number of missing values in the microarray data. Initially, we detect the coordinates of the four corners of each block. The center of the upper-left spot is expected to have, x -coordinate approximately equal to the x -coordinates of spots that belong to the first column of the block, and y -coordinate approximately equal to the y -coordinates of the spots, which belong to the first row of the block. If no spot is detected close to these coordinates, the corner spot is replaced with a spot having the row and column coordinates, which are the mean coordinates of the corresponding row and column.

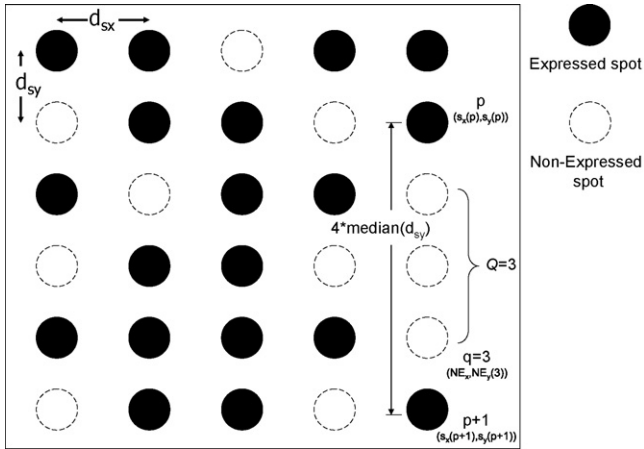


Fig. 4. Detection of the non-expressed spots.

The Euclidean distances for each pair of neighbor spots are calculated in both directions. The median values of these distances in both directions are larger than the real distances between two expressed spots, since there exist many non-expressed spots. The proposed approach is applied in both rows and columns and it is shown schematically in Fig. 4. For example, in each row we add non-expressed spots in positions where the Euclidean distances are larger than the median distance in each direction. The number of non-expressed spots Q , between two neighbor expressed spots (p and $p+1$), is computed as

$$Q = \left\lceil \frac{(S_x(p+1) - S_x(p))}{\text{median}(d_{sxp})} - 1 \right\rceil \quad (8)$$

where S_x is the x -coordinate vector of the expressed spots, p and $p+1$ are two neighbor expressed spots, and d_{sxp} is a vector with all the Euclidean distances in the horizontal direction. The coordinates of the non-expressed spots NE_x and NE_y are defined as

$$NE_x(q) = S_x(p) + q \times (\text{median}(d_{sxp})), \quad (9)$$

and

$$NE_y(q) = \text{mean}(\text{ROW}_y), \quad (10)$$

where ROW_y is the y -coordinate vector of the expressed spots in the row where spots p and $p+1$ belong. Following the above procedure the outlier detection (Eqs. (4)–(7)) is applied for both rows and columns to identify spots, which are located incorrectly. The coordinates of these spots are replaced with the mean values of the corresponding row and column coordinates.

2.1.5. Gridding using a Voronoi diagram

An area (cell) around each spot is marked, applying a Voronoi diagram [36] on each block using as vertices the coordinates of the center of each spot. The Voronoi cell of a center of spot is defined as the set of points, which are closer to this center than any other center. Accordingly, the Voronoi diagram provides an effective spatial solution to deal with the variations of the spot positions in a single block of microarray images. To restrict the Voronoi diagram to the edges of the block, two rows (above and below the block) and two columns (left and right of the block) of fictitious centers were added, as is shown in Fig. 5.

2.2. Segmentation

Since Voronoi cells around each spot are determined, K-means and FCM algorithms are employed, to group all the pixels in the

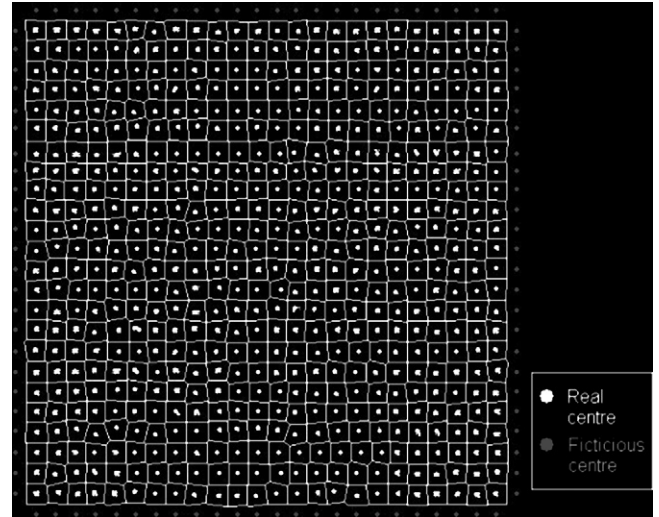


Fig. 5. The result of Voronoi diagram application in a block of the image.

Voronoi cell into two or three clusters. Those clustering techniques are fed with pixel features, which belong to three categories: (i) features, which are related to the intensity of each pixel, (ii) spatial features, and (iii) features, which are related to the shape of a theoretical circular spot (see Table 1). We compared our work with two previous methods [25,26], which are based on K-means. Ergüt et al. [25] employed K-means using only the intensity of the pixel as a feature. Wu and Yan [26] used all three intensity-based features (see Table 1) as well as the Euclidean distance between the pixel and the center of the spot, as the fourth feature. Both methods [25,26] were implemented and tested with the same dataset. The number of clusters is set to $K=2$ (signal pixels and background pixels), in both cases. The K-means [37] employs a square-error criterion, which is calculated for each of the two clusters. The square error criterion for the two clusters is given as

$$E^2 = \sum_{k=1}^2 e_k^2, \quad (11)$$

where

$$e_k^2 = \sum_{i=1}^{n_k} (x_{ik} - M_k)^2, \quad k = 1, 2, \quad (12)$$

and x_{ik} is the feature vector of the i th pixel, n_k is the number of pixels, which belong to the k th cluster and M_k are the centroids of the k th cluster, respectively:

$$M_k = \left(\frac{1}{n_k} \right) \sum_{i=1}^{n_k} x_{ik}. \quad (13)$$

Table 1
Pixel features used in the clustering.

1	Intensity features	Intensity of the pixel
2		Mean intensity of the 3×3 neighborhood of the pixel
3		Standard deviation of the intensity of the 3×3 neighborhood of the pixel
4	Spatial features	x -Coordinate of the pixel
5		y -Coordinate of the pixel
6		Euclidean distance between the pixel and the centre of the spot
7	Shape features	Correlation coefficient of the neighborhood of the pixel and the Gaussian template

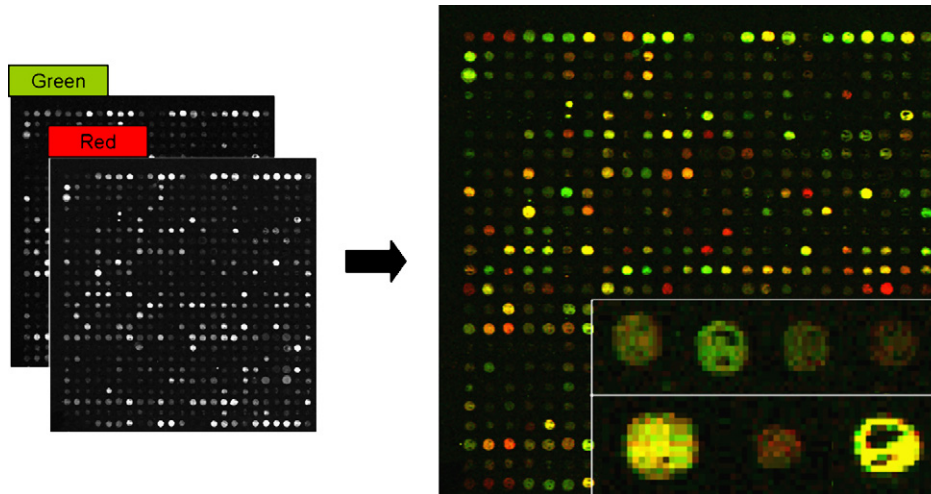


Fig. 6. A block of a microarray image from the SMD which consists of 576 spots (24×24). In the left the two greyscale channels are shown and in the right side the colour microarray image. The zoomed spots present the variations of the spot location in a same row, and the variations of the spot size in the image. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

The dimension of the feature vector and the centroid is 1×1 or 4×1 depending on the number of features (one in the first case, four in the second case).

In the current work we proposed an FCM-based method [38] using the entire set of features (11-features) and the number of clusters $K=3$. The third cluster contains artefacts or low-intensity inner holes in the spot (donut spot) [9]. More specifically, FCM is based on the minimization of the following objective function:

$$J_m = \sum_i \sum_{k=1}^3 (u_{ik})^m \|x_{ik} - M_k\|^2, \quad (14)$$

where x_{ik} is the feature vector (11×1) of the i th pixel, M_k is the 11×1 centroid of each cluster, $u_{i\mu}$ is the degree of membership of x_{ik} in each cluster, $\|x_{ik} - M_k\|^2$ is the Euclidean distance between x_{ik} and M_k , n_k is the number of the pixels belonging to the k th cluster. The parameter m is the weighting exponent for u_{ik} , which controls the fuzziness of the resulting clusters (experimentally we set $m=2$). Each pixel is classified in the cluster with the maximum calculated membership, J_m (Eq. (14)).

Once the clustering is realized, the signal and background clusters should be recognized. Thus, the cluster with the maximum mean intensity value is characterized as signal cluster and the cluster with the minimum mean intensity value is characterized as background class. Due to the high-level intensity of several artefacts, the third cluster is sometimes characterized as signal cluster. For this reason, the Euclidean distance between the spatial component of each centroid and the center of the spot is used as a second criterion to discriminate between signal, background and artefacts. The signal cluster is assumed to be the one with the smaller Euclidean distance.

In order to quantify the expression levels of genes, a set of measurements is used. Variables R and G refer to the red and green channels, while the indexes b and f refer to background and foreground pixels, respectively. The mean values of the intensities of the foreground pixels R_f and G_f and the median values of the intensities of background pixels R_b and G_b for each spot are calculated for the intensity extraction procedure, in both red and green channels of the image [39]. The background-corrected intensities are R and G :

$$R = R_f - R_b, \quad (15)$$

and

$$G = G_f - G_b. \quad (16)$$

The ratio of the background-corrected intensities, RAT, the log-differential expression ratio, M , and the log-intensity, A , of each spot can be computed as

$$\text{RAT} = \frac{R}{G}, \quad (17)$$

$$M = \log_2 \left(\frac{R}{G} \right), \quad (18)$$

$$A = \frac{1}{2} \log_2(R \times G). \quad (19)$$

2.3. Datasets

The evaluation of the proposed method is realized for both the gridding and the segmentation stage. For this purpose, images from the Stanford Microarray Database (SMD) [33] are employed. Each microarray experiment in SMD provides two grayscale images (green and red channel), each one consisting of 16 blocks. The size of the images is 2100×2050 pixels, while the size of each block is approximately 500×500 pixels. Each block consists of 576 spots, i.e. 24×24 rows and columns of spots. Fig. 6 presents a block of a microarray image provided by SMD. Some of the critical problems of the microarray image, such as the variation of the spot size, the variation of the location of spots in the same row and the dark inner holes of donut spots, are illustrated in the zoomed images. 70 blocks were used for the evaluation of the proposed method (total 40,320 spots).

The annotation of the microarray is also available, providing information about the intensity and the position of each spot in the image, for both channels. For the evaluation of the gridding stage the spatial information (i.e. the location of each spot in the image) of each spot is needed. More specifically, the annotation contains the coordinates of the top, bottom, left and right edges of a square area around of each spot and not directly the coordinates of the center of each spot. Given the square area around of a spot, the center of the spot that the annotation provides is assumed to be the center of this square area.

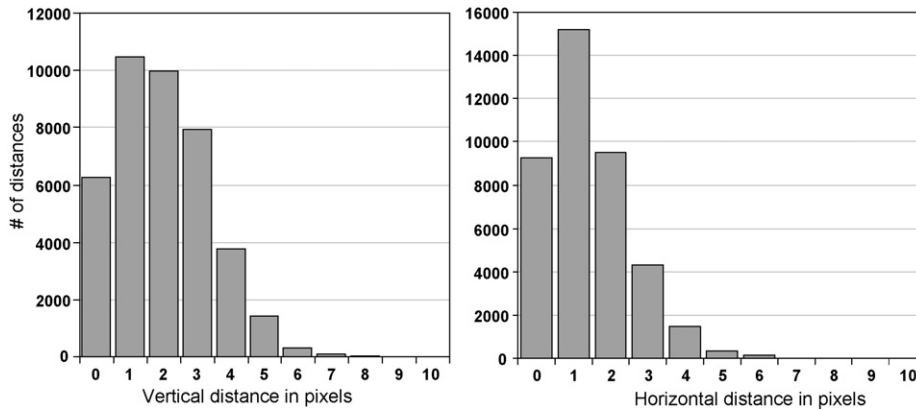


Fig. 7. Histograms of horizontal and vertical distances between the calculated and the annotated spot centres, in the whole dataset.

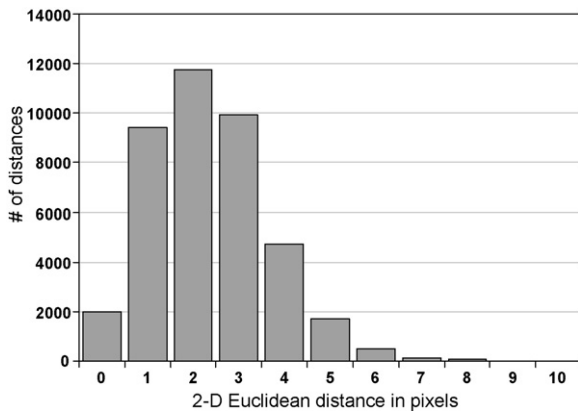


Fig. 8. Histograms of 2-D distances between the calculated and the annotation spot centres, in the whole dataset.

Finally, the ratio (Eq. (17)) of the green and red values, which is also provided in the annotation, is used for the evaluation of the segmentation stage. Unfortunately, pixel-by-pixel information is not available, inhibiting the evaluation of the proposed method in a pixel-by-pixel level.

Table 2

Mean (MEAN) and standard deviation (S.D.) of the gridding stage in both channels.

	MEAN	S.D.
Red	2.7306	1.0985
Green	2.3402	1.0693

3. Results

To evaluate the gridding stage, we compared the centers of the spots found with the annotation of SMD. We use both channels of the microarray image. Given the coordinates of a square area around each spot, the coordinates of the center of the corresponding spot are extracted from the annotation. Next, the center of the same spot is calculated by the proposed method, in order to compare the coordinates of the two centers. The distances in pixels between the centers of the annotated and the detected spots are used in both vertical and horizontal directions. The histograms of the absolute distances are shown in Fig. 7. As we can see, most of the distances are less than 5 pixels. Thus, the centers detected by our method are located inside the borders of the theoretical spot. Taking into account measures such as the mean and standard deviation, use-

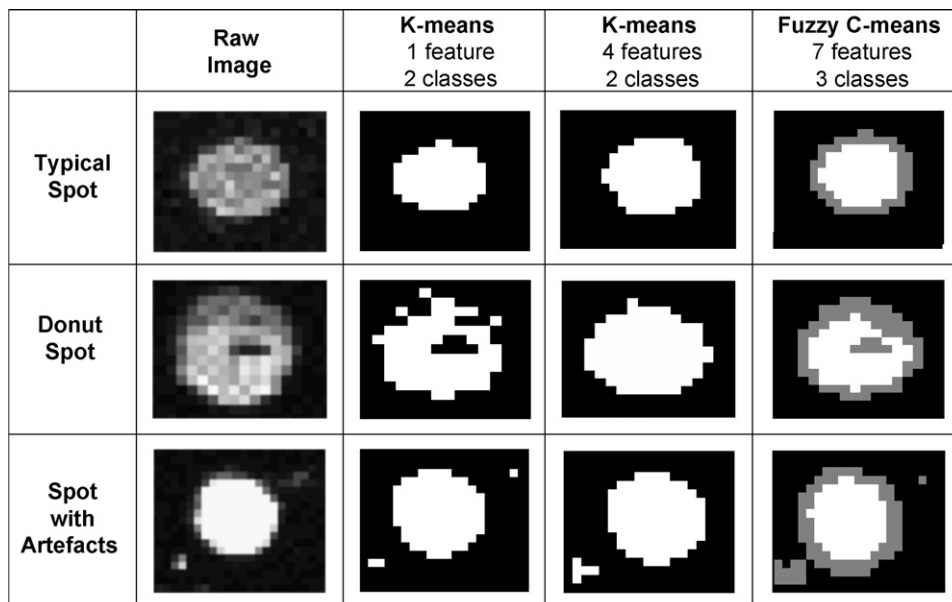


Fig. 9. Segmentation results in a typical spot, a donut spot and a spot with artefacts using K-means with one feature ($K=2$), K-means with four features ($K=2$), FCM with 7 features ($C=3$).

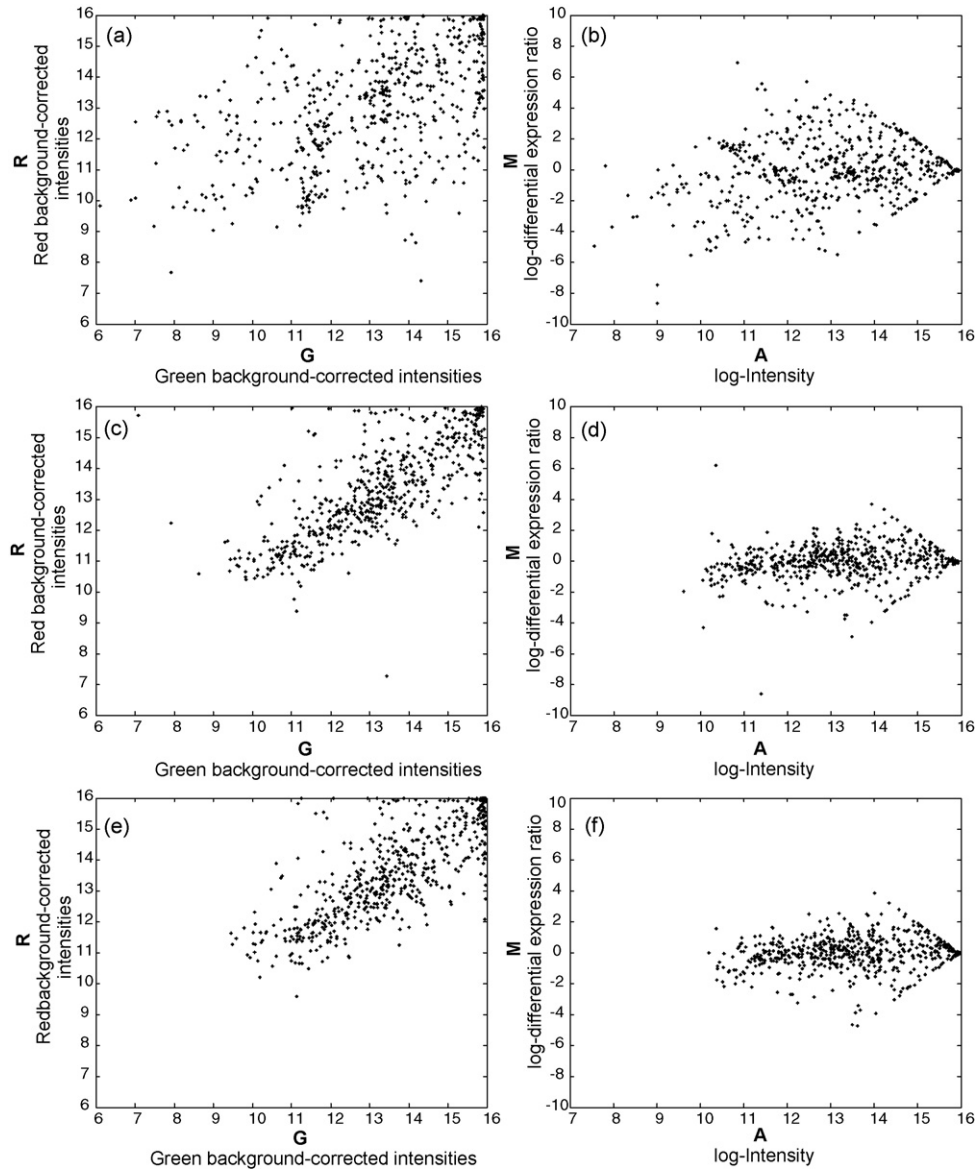


Fig. 10. (a) Scatter plot of K-means with 1 feature and ($K=2$), (b) $M-A$ plot of K-means with 1 feature ($K=2$) (c) Scatter plot of K-means with 4 features ($K=2$), (d) $M-A$ plot of K-means with 4 features ($K=2$), (e) Scatter plot of FCM with 7 feature ($C=3$), (f) $M-A$ plot of Fuzzy C means with 7 features ($C=3$).

ful conclusions can be drawn about the efficiency of the proposed method.

It is convenient to calculate a 2-D distance, d_{2-D} , to represent the distance between the centers of the annotated and detected spots in both directions. For this reason, the Euclidean distance is selected. To generate the histogram for the 2-D distance the Euclidean distances are rounded (Fig. 8).

The mean and the standard deviation of the distribution of the above-defined 2-D Euclidean distances are computed. Table 2 presents the mean and standard deviation of the calculated distances and the obtained for each channel in the dataset.

Fig. 9 presents the results of the segmentation stage for a typical spot, a spot with artefacts and a donut spot. The results of the clustering obtained by K-means with 1 feature and K-means with 4 features and $K=2$ are shown as binary images. The white pixels represent the pixels that K-means characterizes as signal pixels and the black pixels represent the background pixels. The results of FCM with 7 features and $C=3$ are represented as a three level grayscale image. White pixels correspond to signal pixels, black pixels are

the background pixels and gray pixels correspond to artefacts, low-intensity pixels in the area of a donut spot, and also pixels belonging to the contour of a spot.

Scatter and $M-A$ plots of the processed spots are shown in Fig. 10. Scatter plot is the diagram of logarithmic values of background-corrected red intensities vs. log values of background-corrected green intensities. $M-A$ plot is the diagram of M values vs. A values of each spot. Both of these plots provide an overview of the biological experiment. Most of the genes have similar expression levels in the two samples. Thus, most of the points of scatter plot should be close to the diagonal line due to the fact that the ratio of red and green intensities is close to one. Therefore, when the points are concentrated along the diagonal line, the proposed method is more efficient. As it is shown in Fig. 10, the points for K-means with 4 features and FCM are much more confined around the diagonal line than K-means with 1 feature. The same information can be obtained from $M-A$ plots.

Finally, the ratio of red and green background-corrected intensities, RAT, is used as a measure for the evaluation of the segmentation

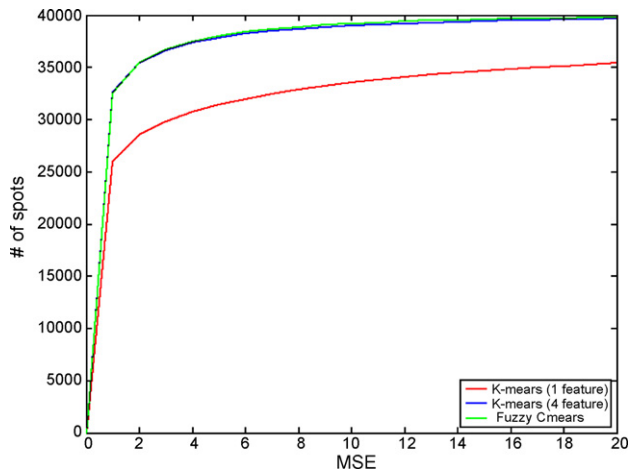


Fig. 11. Number of spots vs. MSE.

stage. We compare the ratio, RAT, of the annotation with the obtained by the proposed method. The Mean Square Error (MSE) is employed for each clustering case:

$$\text{MSE} = \frac{1}{\# \text{ of spots}} \sum_{k=1}^{\# \text{ of spots}} (\text{RAT}_{\text{scanalyse}}(k) - \text{RAT}_{\text{proposed method}}(k))^2. \quad (21)$$

In all cases, there exist several spots (mainly low expressed spots) whose square errors are very high. The number of spots vs. MSE of the proposed method is shown in Fig. 11. Variations of the MSE value allow the discovery of spots, which have a lower MSE value than the selected one. These are the spots that are correctly segmented for the given MSE. Green, blue and red curves correspond to K-means with 1 feature, K-means with 4 features, and FCM, respectively. Fig. 11 shows that K-means with one feature does not perform so well as the other two approaches. The red curve converges to a lower value than the yellow and the green one. This means that a large number of spots (approximately 5000) are incorrectly segmented. The green curve, which corresponds to FCM is placed higher than the other two approaches, indicating higher efficiency.

4. Discussion

In the current work we have presented a method for microarray image analysis. The proposed method is implemented in two stages: gridding and segmentation. In the first stage template

matching is performed to address block and spot detection. During spot detection, outlier analysis is applied to exclude small artefacts between the spots. Next, the non-expressed spots are detected and the gridding of the image is realized using a Voronoi diagram. In the second stage, clustering techniques are employed for image segmentation, using a set of informative features. K-means and FCM are fed with the pixel's features to group the pixels of the image into two or three clusters. More specifically, 1 feature K-means, 4 features K-means and 7 features FCM are applied. Finally, the extracted clusters are characterized as signal, background or artefacts.

In the gridding stage, the template matching technique recognizes efficiently the spots taking into account the similarity of an object in the image with the theoretical spot. Thus, several objects are correctly excluded, while other preprocessing methods, such as thresholding, characterize all the high intensity objects as spots. Another advantage of the proposed work is the effective detection of spots, including the non-expressed ones and the removal of small artefacts during the spot addressing procedure. Outlier detection is employed to remove the artefacts in the image. The spatial information of the detected spots is also used to extract the coordinates of the centers of non-expressed spots. To isolate each spot in an individual area, the Voronoi diagram is employed providing an effective tool for the determination of these areas. A disadvantage of the proposed method is that template matching fails to deal with oblong objects near the edge of the image, which are generated due to emission of noise in SMD images. In this case, template matching produces a sequence of small objects like spots when this noise is processed. Thus, the whole gridding procedure recognizes these objects like a row or column of spots.

The use of three clusters ($C=3$) combined with the informative set of features is the main advantage of the proposed work computed to the already developed segmentation methods. The third cluster contains pixels of artefacts, low-intensity pixels in the area of a donut spot, and pixels belonging to the contour of the spot. These types of pixels cannot be characterized neither as signal pixels nor as background pixels. The mean value of signal pixels for a spot is decreased, if the low-intensity pixels of donut spot or contour pixels are included. On the other hand, the median value of background pixels is often increased, if artefacts and contour pixels are taken into account. Due to the fact that the spatial components of centroids for signal and background clusters are very close (both are very close to the center of the spot), the coordinates of each pixel are not used as features in the two clusters case. When $C=3$, the spatial component of the centroid of the third cluster is far from the center of the spot if artefacts exist in the spot area. For this reason, the pixel's coordinates are included in the feature vector only when $C=3$.

Table 3
Comparison of our method with other works for spot segmentation.

Authors	Segmentation Category	Description	Dataset (# of spots)	Metrics
Buhler et al. 2000 [18]	Adaptive circle	Circular mask with independently estimated radius for each spot	4608	z-Score
Chen et al. 1997 [20]	Histogram-based	Computing a threshold with Mann–Whitney test	1368, National human genome research institute	RAT several genes
Yang et al. 2002 [3]	Adaptive shape	Seed region growing algorithm	122, Apolipo-protein AI	Standard deviation of log-intensities
Demirkaya et al. 2005 [24]	Adaptive shape	Markov random field modeling of pixels	King Faisal Specialist Hospital	–
Rahnenführer et al. 2004 [28]	Clustering	Hybrid K-means	864, University of Nebraska Medical Center	Median stability
Li et al. 2005 [31]	Clustering	Model-based clustering of pixels	1536	Sum of square distances (SSD)
Blekas et al. 2005 [9]	Clustering	EM clustering of pixels	Artificial & real spots	Classification error & mean square error
Proposed method	Clustering	FCM clustering of pixels	40320 SMD	Mean square error

The proposed method for spot segmentation was directly compared with other segmentation approaches based on the K-means algorithm [25,26]. Other segmentation approaches that are based on clustering techniques [9,27,28,31], or fixed and adaptive shape [18], or histogram techniques [21,22,19], have been also proposed. However, the performance of these methods cannot be compared directly with the proposed work since different measures have been used for the evaluation. Moreover, none of these works employ the same set of images with the current approach. In Table 3 we present a qualitative comparison with previous works.

5. Conclusions

The proposed work addresses automated gridding and clustering-based segmentation for microarray image processing. Pixels of the image, which belong to artefacts and inner holes of donut spots are excluded from the intensity extraction procedure. Thus, the gene expression levels are efficiently quantified with respect to the biological experiment. For further improvement of the proposed method, filtering techniques could be employed as a preprocessing step, in order to enhance the initial image and to obtain better results. Apart from the background noise which occurs due to the emission of the slide, filtering techniques could deal with the small artefacts which appear in the image as a result of the dust or other contamination on the slide.

Acknowledgement

This work is part funded by the European Commission (MATCH project, IST-2005-027266).

References

- [1] Eisen MB, Brown PO. DNA arrays for analysis of gene expression. *Methods in Enzymology* 1999;303:179–205.
- [2] Schena M, Shalon D, Davis RW, Brown PO. Quantitative motoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;270:467–70.
- [3] Yang YH, Buckley MI, Dudoit S, Speed TP. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics* 2002;11(1):108–36.
- [4] Ahmed AA, Vias M, Iyer NG, Caldas C, Brenton JD. Microarray segmentation methods significantly influence data precision. *Nucleic Acids Research* 2004;32:e50.
- [5] Schena M. *Microarray biochip technology*. Eaton Publishing; 2000.
- [6] Axon Instruments. *GenePix A User's Guide*; 1999.
- [7] Eisen MB. *ScanAlyze*. Available: <http://rana.Stanford.EDU/software/>; 1999.
- [8] Bajcsy P. Gridline: automatic grid alignment in DNA microarray scans. *IEEE Transactions on Image Processing* 2004;13(1):15–25.
- [9] Blekas K, Galatsanos N, Likas A, Lagaris IE. Mixture model analysis of DNA microarray images. *IEEE Transactions on Medical Imaging* 2005;24(7):901–9.
- [10] Jain N, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, Pinkel D. Fully automatic quantification of microarray image data. *Genome Research* 2002;12:325–32.
- [11] Hirata JR, Barrera J, Hashimoto RF, Dantas DO. Microarray gridding by mathematical morphology. In: *Proceeding of International Symposium on Computer Graphics, Image Processing, and Vision*. 2001. p. 112–9.
- [12] Lonardi S, Yu L. Gridding and compression of microarray images. In: *Proceedings of IEEE Computational Systems Bioinformatics Conference-Workshops (CSBW'05)*. 2004. p. 122–30.
- [13] Jung HY, Cho HG. An automatic block and spot indexing with k-nearest neighbors graph for microarray image analysis. *Bioinformatics* 2002;2:141–51.
- [14] Galinsky VL. Automatic registration of microarray images. I. Rectangular grid. *Bioinformatics* 2003;19:1824–31.
- [15] Ceccarelli M, Antoniol G. A deformable grid-matching approach for microarray images. *IEEE Transaction on Image Processing* 2006;15(10):3178–88.
- [16] Steinfath M, Wruck W, Seidel H, Lehrach H, Radelof U, O'Brien J. Automated image analysis for array hybridization experiments. *Bioinformatics* 2001;17:634–41.
- [17] Smyth GK, Yang YH, Speed T. Statistical issues in cDNA microarray data analysis. *Functional genomics: methods and protocols*. In: Brownstein MJ, Khodursky AB, editors. *Methods in molecular biology series*. Totowa, NJ: Humana Press; 2002.
- [18] Buhler J, Ideker T, Haynor D. *Dapple: improved techniques for finding spots on DNA microarrays*. UWCSE Tech Report, UWTR Department of Computer Science and Engineering, University of Washington; 2000.
- [19] GSI Lumonics. *QuantArray Analysis Software. Operator's Manual*; 1999.
- [20] Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* 1997;2(4):364–74.
- [21] Siddiqui KI, Hero A, Siddiqui M. Mathematical morphology applied to spot segmentation and quantification of gene microarray images. In: *Proceedings of Asilomar Conference on Signals and Systems*. 2002.
- [22] Buckley MJ. *Spot user's guide*. Sydney, Australia: CSIRO Mathematical and Information Sciences; 2000.
- [23] Wang X, Ghosh S, Guo SW. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research* 2001;29(15):e75.
- [24] Demirkaya O, Asyali MH, Shoukri MM, Abu-Khabar KS. Segmentation of microarray cDNA spots using MRF-based method. In: *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1. 2003. p. 674–7.
- [25] Ergüt E, Yardimci Y, Mumcuoglu E, Konu O. Analysis of microarray images using FCM and K-means clustering algorithm. In: *Proceedings of International Conference on Signal Processing*. 2003. p. 116–21.
- [26] Wu H, Yan H. Microarray image processing based on clustering and morphological analysis. In: *Proceedings of 1st Asia Pacific Bioinformatics Conference*, vol. 2. 2003. p. 111–8.
- [27] Bozinov D, Rahnenführer J. Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering. *Bioinformatics* 2002;18:747–56.
- [28] Rahnenführer J, Bozinov D. Hybrid clustering for microarray image analysis combining intensity and shape features. *BMC Bioinformatics* 2004;5:1–11.
- [29] Abbaspour M, Abugharbieh R, Podder M, Tripp BW, Tebbutt SJ. Hybrid spot segmentation in four-channel microarray genotyping image data. In: *IEEE International Symposium on Signal Processing and Information Technology*. 2006. p. 11–6.
- [30] Nagarajan R. Intensity-based segmentation of microarray images. *IEEE Transactions on Medical Imaging* 2003;22:882–9.
- [31] Li Q, Fraley C, Bumgarner RE, Yeung KY, Raftery AE. Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics* 2005;21:2875–82.
- [32] Lehmussola A, Ruusuvoori P, Yli-Harja O. Evaluating the performance of microarray segmentation algorithms. *Bioinformatics* 2006;22:29100–32917.
- [33] Gollub J, Ball CA, Binkley G, Demeter K, Finkelstein DB, Hebert JM, et al. The Stanford microarray database: data access and quality assessment tools. *Nucleic Acids Research* 2003;31:94–6.
- [34] Giannakeas N, Fotiadis DI, Politou AS. An automated method for gridding in microarray images. In: *Proceedings of 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2006. p. 5876–9.
- [35] Kantardzic M. *Data mining concepts, models, methods, and algorithms*. IEEE Wiley Press; 2003.
- [36] Aurenhammer F. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys* 1991;23:345–405.
- [37] MacQueen JB. Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley: University of California Press; 1967. p. 281–97.
- [38] Bezdek JC. *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press; 1981.
- [39] Smyth GK, Yang YH, Speed T. Statistical issues in cDNA microarray data analysis. *Methods in Molecular Biology Series* 2003;224:111–36.

Nikolaos Giannakeas was born in Athens, Greece in 1980. He graduated from the Physics Department of the University of Ioannina, Greece in 2003. He is currently working towards the PhD degree in the Medical School of the University of Ioannina. His research interests include microarray image and data analysis, biomedical engineering and bioinformatics.

Dimitrios I. Fotiadis was born in Ioannina, Greece in 1961. He received the Diploma degree in Chemical Engineering from the National Technical University of Athens and the PhD degree in Chemical Engineering from the University of Minnesota, USA. Since 1995, he has been in the Dept. of Computer Science, University of Ioannina, Greece, where he is currently an Associate Professor. He is the director of the Unit of Medical Technology and Intelligent Information Systems. His research interests include biomedical technology, biomechanics, scientific computing and intelligent information systems.