International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:5, 2008

An Automatic Gridding and Contour Based Segmentation Approach Applied to DNA Microarray Image Analysis

Alexandra Oliveros, and Miguel Sotaquirá

***Abstract*—**DNA microarray technology is widely used by geneticists to diagnose or treat diseases through gene expression. This technology is based on the hybridization of a tissue's DNA sequence into a substrate and the further analysis of the image formed by the thousands of genes in the DNA as green, red or yellow spots. The process of DNA microarray image analysis involves finding the location of the spots and the quantification of the expression level of these. In this paper, a tool to perform DNA microarray image analysis is presented, including a spot addressing method based on the image projections, the spot segmentation through contour based segmentation and the extraction of relevant information due to gene expression.

***Keywords*—**Contour segmentation, DNA microarrays, edge detection, image processing, segmentation, spot addressing.

I. Introduction

DNA microarrays consist of large number of known DNA probe sequences, arranged over a bi-dimensional grid on a slide. Each sequence is a probe for capturing a given gene, then, every probe will hybridize with a specific sequence of complementary RNA. By comparing healthy and sick tissues, the array becomes a tool to measure how the genes are differentially expressed in two cell cultures. The genes extracted from the tissues to be studied are first labeled with a red dye (Cy5) for the sample, and a green dye (Cy3) for the control tissue, and hybridized with the sequences spotted on the array [1]. Finally, the array is scanned to obtain an image where each spot represents the differential level of hybridization of each gene in the two tissues. The more activated genes will be evident on the array as they will light up. So, the output of a microarray experiment is a digital image whose size is of tens of MB.

Manuscript received June 26th, 2008. This work was supported in part by the Internal Program for Young Researchers in the Strategy for Strengthening Groups and Research Institutes at Fundacion Universidad del Norte and the Department of Electronic Engineering from Universidad del Norte in Barranquilla, Colombia.

Alexandra Oliveros is with the Department of Electronic Engineering as a researcher in the group of Telecommunications and Signals in Universidad del Norte, Barranquilla, Colombia (phone: 57-5-3509270; e-mail: alexandra_oliveros@ ieee.org).

Miguel Sotaquirá, is with the Electronic Engineering Department, Universidad del Norte, Km.5 via Pto. Colombia. Barranquilla, Colombia (e-mail: msotaquira@uninorte.edu.co).

Besides allowing experts to measure gene expression in a single hybridization experiment, this technology is useful to study thousands of DNA nucleotide sequences on a single substrate. It has been used in numerous studies that include toxicological research and diagnosis in cancer, diabetes and genetic diseases [2]-[3]. For this reason, DNA microarrays may be used to diagnose a disease, create treatments or medicines [4].

In the last few years, commercial and research purpose software have become available to perform microarray image analysis. This process involves three basic tasks: *Spot addressing* which consists of assigning coordinates to each spot, *segmentation* that allows the identification of the foreground (spot signal) and background, and *quantification* of the information extracted from the red (Cy5), green (Cy3) fluorescence intensity and background intensities [5].

In this article, a tool to perform microarray image analysis is proposed, and it is outlined as follows: First, in Section II the gridding method based on the axis image projections and distance maps is described. In Section III, the segmentation approach that combines edge detection and threshold segmentation to determine the spot signal is presented. Then, in Section IV, the process to perform the quantification of the quality of spots and signal information are shown. Finally, some conclusions are discussed.

II. Spot Addressing

Finding the exact location of a spot is essential to obtain an accurate segmentation of spots. Furthermore, the accurateness of this process affects the quality of the gene expression information and impacts the confidence in any conclusions derived from this task [6]. Due to the imperfections in the construction of the microarrays, the resulting image contains artifacts such as highly fluorescent dust particles, unattached dye, and fibers or deposits from evaporated solvents, which makes the image processing task a computationally intensive process. Likewise, the coordinate system of the robot printing may be slightly rotated with respect to the microarray image coordinate system [6]-[7]. Therefore, this angle must be considered in the spot locating process, because if the image is rotated back to its original position, valuable information will get lost, leading to inaccurate results during the feature extraction stage.

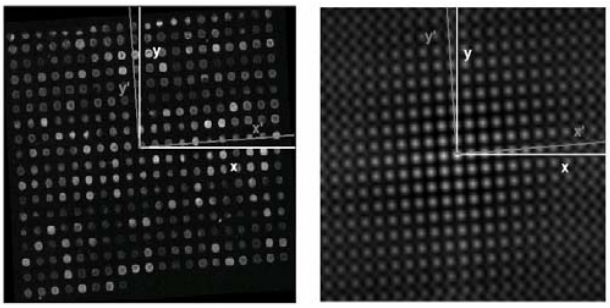
Some of the available software for gene expression analysis

include gridding methods that require manual intervention (e.g., specify dimension of a grid and a radius of each spot) [8]-[9], semi-automated methods (e.g., combine manual specification of grid dimensions and an automated search of grid lines) [10] and fully automated methods [7]-[11]. However, the last one could fail due to variations in microarray fabrication process and imperfections.

The approach of our method to find the grid lines is based on a simple technique, which is the analysis of 1-D projections on both axes of the image [11].

A. Main Axes and Image Orientation

The first step in the gridding process is to compute the image orientation. In order to accomplish this, we have used the autocorrelation image (see Fig. 1), which reveals important details of the DNA microarray digital image [12]. For instance, based on the autocorrelation image we identify the main axes (*x',y)'* of the image with an iterative algorithm. Then, we find the mean of the difference of the correlation values with respect to the maximum, and through trigonometric operations we find the angle with respect to the conventional axes (x, y).



(a) (b)

Fig. 1 a) Original DNA microarray image b) Autocorrelation image of a). In this figure, x and y are the conventional axis and x' and y' are the axis of the rotated image

B. Distance Maps Computation

Once the main axes of the image are found, we create two binary images. The first one has a marker with a line parallel to the *x'* axis and the other has one line parallel to the *y'* axis. Then, a distance transform is applied to each binary image using the City block metric [13]. The result of this process is a gray level image, where each pixel represents the distance between every 0 labeled pixel and the marker. The distance of the pixels is calculated according to (1), where x_i, y_i are the coordinates of each pixel in the binary image.

$$d_{\min} = |x_1 - y_1| + |x_2 - y_2| \quad (1)$$

After finding the distance transform, a normalization process is done in order to differentiate one line from the other, so each line will be identified by an integer value. The gray level image obtained by applying the distance transform is known as distance map. The distance maps are a group of parallel lines, which have the same integer value.

C. Projection

In order to find the lines that determine the rows and columns of the array to locate the spots, we found the projection onto the x' and y' axes. The algorithm designed for this task, reads each line of the horizontal and vertical distance map finding the mean of the gray levels of every row of the oriented image and then these values are stored in a vector. By applying a local minima detection algorithm to the vectors created, the positions for the lines of the arrays can be specified. The projection onto the *x'* and *y'* axes of a DNA microarray image is shown in Fig. 2.

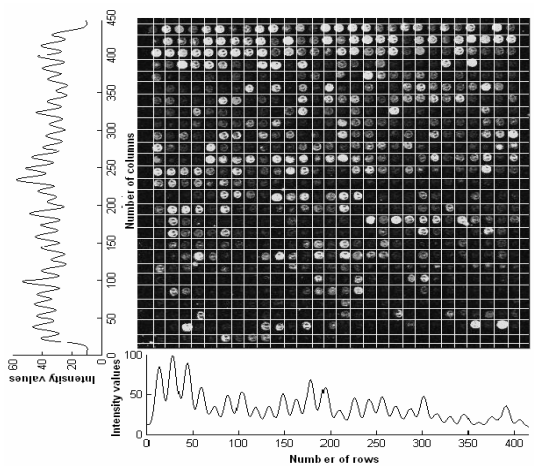


Fig. 2 DNA microarray image projections using distance maps

Due to imperfections of the probe placement, some of the lines of the array cross the spots (see Fig. 2). For this reason, it is necessary to do a refinement of the algorithm. It consists of guessing the initial position of the spot and then move the line from left to right (for columns) and top to bottom (for rows) by specifying a threshold for the minimum gray level along the edges of the spot analyzed. When the threshold condition is met, the coordinates for the cell, in which the spot is inscribed, are stored in a matrix. In Fig. 3 the gridding of the DNA microarray image of Fig. 2 is improved with the algorithm described before.

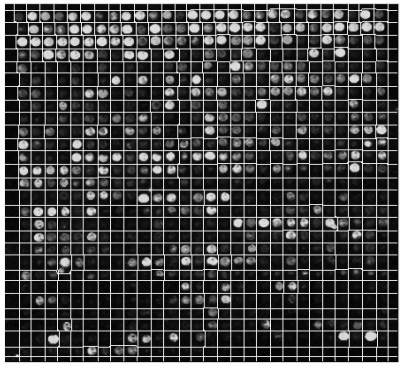


Fig. 3 Improved gridding of Fig. 2

III. Segmentation

The segmentation process performs the identification of the foreground (i.e. the spot signal) and the background. This task allows the further quantification of the fluorescence intensity of every probe from the DNA sequence. Accurate spot segmentation is an essential step in microarrays image analysis, because it aims to reduce the image into single gene-expression values [14].

Some of the methods used to perform microarray segmentation includes: Fixed circle segmentation algorithm, included in ScanAlyze Software [8], which generates a binary masked based on a fixed ratio circle placed on each grid cell, but it only works well for well rounded and equal size spots. Adaptive circle segmentation, in which the ratio is adapted to the spot ratio [5], k-means clustering, seeded region growing algorithms, shape based and histogram based segmentation [5].

Choosing an inappropriate method to separate the foreground pixels from the background pixels can modify the true expression value of each gene, leading to potentially false negative calls in the quantification process. On the other hand, outlier pixels, representing hybridization defects, nearby spots, dust, etc., may overestimate the expression value and create potential false positive calls [15].

The segmentation method proposed in this paper is based on a edge detection algorithm combined with a histogram based approach to eliminate outlier pixels. In Figs. 1 and 2, it can be seen that there are bright regions inside the spots, which will cause faulty binarization of the image, but applying a Sobel filter and morphological operations of opening and closing the problem is solved by smoothing the intensity level of the spot signal. Fig. 4 illustrates the comparison of the grayscale version of part of a microarray image before and after applying the algorithm mentioned before.

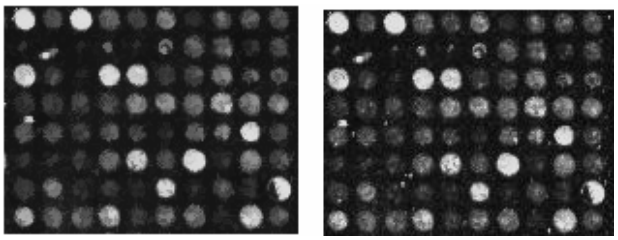


Fig. 4 Comparison of smoothed image on the left and original grayscale microarray image on the right

Once the grid of the subarray is obtained and the coordinates of the grid cells, the standard deviation of the gray level in the pixels, p_{ij}, of each cell, C, is calculated in order to determine if the cell contains only noise or significant spot signal. The mean of the standard deviation in the pixels gray level of each row in the cell as well as in each column are obtained as in (3) to (7). Then, the average of these values is calculated. If the standard deviation of the pixels gray level in the cell is more than 4 ($\sigma_{cell} \geq 4$) [16], the grid cell is considered to contain a potential good quality spot. If not, it is just noise or the spot signal will not produce reliable information.

$$C = \begin{bmatrix} p_{11} \cdots & p_{1m} \\ \vdots & \vdots \\ p_{n1} \cdots & p_{nm} \end{bmatrix}_{n \times m} \quad (2)$$

$$\sigma_{row} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(p_{ij} - m_{row} \right)^2}, \; j = 1,2,\ldots n \quad (3)$$

$$m_{row} = \frac{1}{m} \sum_{j=1}^{m} p_{ij}, \; j = 1,2,\ldots n \quad (4)$$

$$\sigma_{col} = \sqrt{\frac{1}{n} \sum_{k=1}^{n} \left(p_{kl} - m_{col} \right)^2}, l = 1,2,\ldots m \quad (5)$$

$$m_{col} = \frac{1}{n} \sum_{k=1}^{n} p_{kl}, \; l = 1,2,\ldots m \quad (6)$$

$$\sigma_{cell} = \frac{\sigma_{row} + \sigma_{col}}{2} \quad (7)$$

After assuring the grid cell contains valuable information, the histogram of the same is calculated and a threshold value equal to the median of the histogram is used to remove background and noisy pixels. Then, the spot contour is identified using a Canny filter approach for edge detection. This method was chosen, because it is less likely to be fooled by noise and it is based on looking for the local maxima of the gradient in the spot cell using the derivative of a Gaussian filter [17]. This way, we obtained a binary mask that contains the spot pixels and that is not overestimated due to outlier pixels as it is shown in the Results section.

IV. Quantification

When the binary masks of the spots are obtained, the pixels corresponding to spot signal and background are identified. Consequently, the fluorescence intensity and quality measures from each spot in the subgrid can be specified. The hybridization process in DNA microarray produces a spot signal formed by the background and the red (Cy5) and green (Cy3) channel intensities. Hence, the real red and green fluorescence intensity will be obtained after removing the background from the spot signal, so the fluorescence intensity will be obtained as in (2) and (3).

$$Cy5\_intensity = Cy5\_spot\_signal - Cy5\_background \quad (2)$$

$$Cy3\_intensity = Cy3\_spot\_signal - Cy3\_background \quad (3)$$

A. Background Removal

The process of removing the background pixels intensity from the spot signal consists on taking each grid cell and

eliminating the spot signal pixels. The morphological operation of dilation is applied to the image of this cell, and each background pixel intensity is stored in a vector to calculate the background mean intensity. The image is dilated with the purpose of not considering the pixels immediately surrounding the spots. If these pixels are included, the background estimation could be sensitive to outliers and artifacts in the grid cell, deviating the intensity value from the real one.

B. Foreground Intensity

After finding the mean background intensity for each channel, this value is subtracted from every spot signal pixel from the red and green channel as shown in (2) and (3). Later, the red and green fluorescence intensity is stored in a vector for the calculation of the mean intensity in both color planes.

In general, the most significant measure extracted for each grid cell is the median of the background and the median fluorescence intensity, because this parameter is not affected by noise or outlier pixels. Some of the information extracted for each spot by this DNA microarray image analysis tool, includes: Background intensity mean, background intensity median, red fluorescence intensity mean, red fluorescence intensity median, green fluorescence intensity mean, green fluorescence intensity median, green to red fluorescence intensity ratio (Cy3/Cy5), red to green fluorescence intensity ratio (Cy5/Cy3), diameter of the spot and number of spot pixels.

These parameters are then normalized, in order to avoid systematic errors and bias introduced by the microarray imperfections [9]. Also, the normalization process taking the logarithm to based 2 of the parameters obtained is done to ensure the data is of high quality and suitable for analysis.

V. Results

In order to evaluate the performance of the proposed framework, subgrids from 50 microarray images from the Stanford microarray database [18] were analyzed. The characteristics of the images used included: Different rotation angles and imperfections such as: noise contamination, misaligned spots and defects on the images like those in Fig. 1 and Fig. 2. Although some of the microarray image acquisition systems produce good quality images, this kind of images were chosen in order to test the robustness of our method to noise and outlier pixels.

A. Spot Addressing

The accuracy of the gridding algorithm was measured in terms of the number of successful spots located, the number of non-addressed spots and the amount of spots non-located correctly because of imperfections in the image and due to a confusion of the algorithm with the spot and the background.

According to the images tested, the results show that from a total of 29.867 spots, there is a 98% of spots located successfully and a 1.64% of spots misallocated. The spots included in this group are those whose area crossed by the grid line in more than 40%. The 1.22% of this group of spots was not addressed correctly due to imperfections in the image, such as overlapping spots or residues of spot material from the fabrication process. At the same time 0.41% of the spots were not properly located because the gray level threshold considered in the algorithm to do the gridding refinement process was very similar to the background in the spot cell, which tends to confuse the spot signal with the background.

Additionally, we also tested the need to consider the rotation of the image. From the 50 images analyzed, 8 presented rotation angles from 0.5° to 3.7°. We compared the grid tracing considering the image orientation and obviating the rotation angle. We found that 20.69% of the spots are crossed significantly by the grid lines, whereas 79.31% of the spots are located correctly. For an image rotation angle smaller than 0.5°, the spots are not significantly crossed by the grid line, but this one would be slightly displaced from the correct path. In Fig. 5 a microarray image with of 3.7° angle rotation is shown, here it can be seen that when the rotation angle is not obtained prior to finding the projections, some of the grid lines are not traced whereas others are placed in the wrong path. A wrong spot addressing process would then lead to an inaccurate spot segmentation avoiding the extraction of the truth fluorescence intensity in each spot.

This method is considered automatic in finding the gridding on the image, because it does not demand any type of information about the spots size as other softwares like ScanAlyze [8]. The performance of this algorithm provides good results compared to similar algorithms that do not consider the image orientation [6]. Additionally, the approach used here is simpler than others that are based on finding all expected rotational angles [19], in analyzing the four edges on the image with the disadvantage of introducing pixel distortions [20] or in applying the radon transform, which is computationally expensive [21].

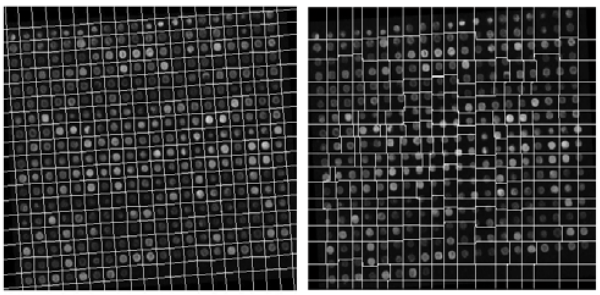


Fig. 5 Gridding of microarray image with a 3.7° rotation. The figure on the left considers the angle when finding the grid, whereas the one on the right does not

B. Segmentation

The segmentation algorithm evaluation was done in the same images used to test the performance of the gridding algorithm. According to the criteria established in Section II, to define if a grid cell contains or not a spot, from the total numbers of spots (29.867), in only 26.518 of the grid cells, spots with significant information were found. The rest of the grid cells (3.349), did not contain spots or had spots with unreliable information (i.e. spots with very low fluorescence intensity or very similar to the background).

From the total of grid cells segmented, 99.99% were segmented correctly. The remaining 0.11%, were not segmented properly due to spot size smaller than the minimum (10 pixel radius) and because of outlier pixels present in the grid cell due to imperfections in the microarray fabrication process. Furthermore, from the total of spots segmented, 0.3% had artifacts as part of the binary mask obtained for the microarray image. Fig. 6 shows some of the spots segmented, including round shape spots, non-round spots and artifacts that alter the adequate spot segmentation.

The results presented here evidence the accurateness of the method implemented, combining histogram based segmentation and contour segmentation, which allows a reliable identification of the spot signal and in the information extracted later.

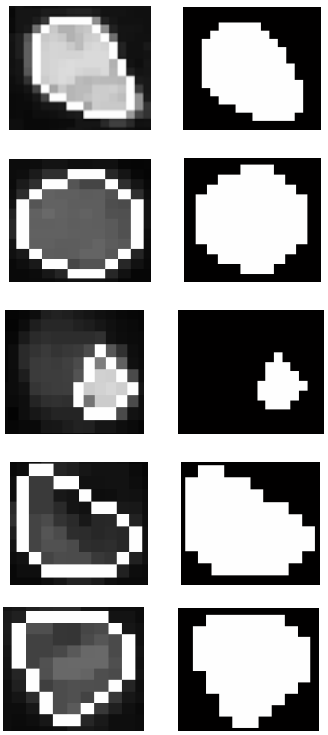


Fig. 6 Spots of a microarray gray level image with its corresponding binary mask obtained after segmentation

VI. Conclusion

In this article, a novel, robust and efficient tool for spot addressing, segmentation and information quantification in DNA microarray images was described.

The gridding method is robust to common properties of microarray images such as: axes rotation, misaligned and overlaping spots as well as noise. Using the distance transform and the distance maps to associate a magnitude to each pixel of the rotated image the 1D projection onto the image axis can be obtained. Likewise, applying filtering techniques, such as a median filter and morphological operations, most of the noise present in the image can be eliminated to get a uniform background and to perform an accurate addressing of the spots, and then the segmentation.

On the other hand, at first, the segmentation method implemented did not present useful results due to the noise pixels in the background of some images or severe artifacts in others. So, an initial guess of the background pixels was done by applying histogram based segmentation, using the histogram median as a threshold. This way, some of the noise pixels were discarded obtaining a highly accurate segmentation. Adding contour based segmentation, an adaptive method to different spot shapes is achieved producing reliable information regarding spot quality and gene expression.

For the spot signal information extraction, background correction must be considered in order to avoid deviations from the real value due to microarray fabrication. Also, because the fluorescence intensity in the spot is proportional to the gene expression, a normalization process needs to be done to obtain avoid biased information.

With the methods proposed in this article for spot addressing and segmentation, spots have been located automatically, considering orientation of the image. The results presented above, evidence the robustness of the method. Additionally to the advantages of the techniques applied here with respect to other algorithms that do not consider image orientation in the gridding finding process, or that do not consider the features and imperfections in the image.

References

[1] M. B. Eisen and P. O. Brouwn, "DNA arrays for analysis of gene expression," in *Methods in Enzymology.*, vol. 303, 1999, pp. 179–205.
[2] R.S.H. Istepanian, "Microarray image processing: Current status and future directions", *IEEE Transactions on Nanobioscience,* vol 2, issue 4, pp. 173–175, 2003.
[3] X.Y. Zhang, *et al*, "Signal processing techniques in genomic engineering", *Proceedings of the IEEE,* vol. 90 issue 2, pp. 1822–1833, 2002.
[4] M. J. Marton, *et al*, "Drug Target Validation and Identification of Secondary Drug Target Effects Using DNA Microarrays," *Nature Medicine*, vol 4, 1998, pp.1293–1301.
[5] Yee Hwa Yang, Michael J. Buckley, Sandrine Dudoit, and Terence P. Speed. "Comparison of Methods for Image Analysis on cDNA Microarray Data". *Journal of Computational and Graphical Statistics,* vol 11, no 1, 2000, pp 1–29.
[6] Peter Bajcsi. "An overview of DNA grid alignment and foreground separation approaches". *EURASIP Journal on applied signal processing*, Article ID 80163 , 2006, p.1-13.
[7] M. Katzer, F. Kummert, and G. Sagerer, "Methods for automatic microarray image segmentation," *IEEE Transactions on Nanobioscience*, vol. 2, no. 4, pp. 202–212, 2003.
[8] M. Eisen, "ScanAlyze," Product Description. Available: http://rana.lbl.gov/EisenSoftware.htm.
[9] J. Buhler, T. Ideker, and D. Haynor, "Dapple: improved techniques for finding spots on DNA microarrays," Technical Report. UWTR 2000-08-05, UV CSE, Seattle,Washington, USA.
[10] S. Draghici, *Data Analysis Tools for DNA Microarrays*, CRC Mathematical Biology and Medicine Series, Chapman & Hall, London, UK, 2003.
[11] A.M. Machado, M.F. Campos, A.M. Siqueira and O.S. Carvalho. "An iterative algorithm for segmenting lanes and gel electrophoresis images". *Proceedings of the X Brazilian Symposium on Computer Graphics and Image Processing*. Brazil, pp. 140-146, 1997.
[12] J.C. Russ. The Image Processing Handbook. CRC Press & IEEE Press. Florida, 2002.

[13] F. de A. Zampirolli and R. de A. Lotuffo. "Classification of the distance transformation algorithms under the mathematical morphology approach". *Proceedings of the 13th Brazilian Symposium on Computer Graphics and Image processing*. Brazil, pp. 292-299, 2000.
[14] L. Qin, L. Rueda, A. Ali, A.Ngom, "Spot detection and image segmentation in DNA microarray data," *Applied Bioinformatics*, 4(1):1-11, 2005.
[15] Th. Margaritis, K. Marias and D. Kafetzopoulos. "Improved Microarray Spot Segmentation by Combining two Information Channels". *28th IEEE International Conference in Medicine and Biology,* Aug. 2006, pp 5850 – 5853.
[16] Iiris Hovatta *et al*, *DNA microarray data analysis*, CSC Scientific Computing. Helsinki, 2005.
[17] Rafael Gonzalez and Robert Woods, *Digital Image Processing*. Prentice Hall, pp 954 2008 N.Brandle, H. Bischof and H. Lapp. "Robust DNA microarray image analysis". *Machine Vision and Applications*, vol 15, no 1, pp. 11-28, 2003.
[18] Stanford DNA Microarry Database. Available: http://genomewww5.stanford.edu
[19] P. Bajcsy. "Gridline: automatic grid alignment in DNA microarray scans". *IEEE Transactions on Image Processing*, vol. 13, no 1, 2004, pp. 15-25.
[20] M. Steinfath, W. Wruck, H. Seidel, H. Lehbrach, U. Radelof and J. O'Brien. "Automated image analysis for array hybridization experiments". *Bioinformatics*, vol 17, no 7 , 2001 pp. 634-641.
[21] N.Brandle, H. Bischof and H. Lapp. "Robust DNA microarray image analysis". *Machine Vision and Applications*, vol 15, no 1, pp. 11-28, 2003.

Alexandra Oliveros was born in Barranquilla, Colombia on May 6th, 1984. She received the Bachelor degree in Electronic Engineering from Universidad del Norte, Barranquilla, Colombia, in 2007.

She participated in a research internship in the University of Girona, Spain and in the University of South Florida. She currently works as a researcher in the group of Telecommunications and Signals in Universidad del Norte, on DNA microarray image processing. Her research interests are modeling and design of drug delivery systems, biosensors and BioMEMs design.

Miguel Sotaquirá was born in Bogotá, Colombia on January 1st, 1979. He received a Masters degree in Electronic Engineering in 2005, and Electronic Engineer degree in 2001 form Industrial University of Santander, Bucaramanga, Colombia. He was Teaching Assistant from 2002 to 2005 in the same institution. He became assistant professor for the Electronic Engineering Department at the Universidad del Norte in 2005. He is a member of the Telecommunication and Signals Research Group since 2005, Universidad del Norte. His research interests include, medical images processing (e.g. DNA microrarray images) and BioMEMs design.