Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

An Automated Method for Gridding in Microarray Images

Nikolaos Giannakeas, Dimitrios I. Fotiadis, *Member, IEEE* and Anastasia S. Politou

***Abstract*—Microarray technology is a powerful tool for analyzing the expression of a large number of genes in parallel. A typical microarray image consists of a few thousands of spots which determine the level of gene expression in the sample. In this paper we propose a method which automatically addresses each spot area in the image. Initially, a preliminary segmentation of the image is produced using a template matching algorithm. Next, grid and spot finding are realized. The position of non-expressed spots is located and finally a Voronoi diagram is employed to fit the grid on the image. Our method has been evaluated in a set of five images consisting of 45960 spots, from the Stanford Microarray Database and the reported accuracy for spot detection was 93 %.**

I. Introduction

DNA microarray technology provides simple and effective gene expression analysis [1]. During the biological experiment, the two mRNA samples to be compared are reverse transcribed into cDNA. Thousands of known genes are printed on a glass slide by a robotic printer. The cDNA samples, labeled with two difference fluorescent dyes, are hybridized with the known genes at the same time. The most common dyes for tagging mRNA are the red fluorescent dye Cy5 (emission in the 630-660 nm) and the green-fluorescent dye Cy3 (emission in the 510-550 nm) [2].

The DNA microarray is scanned in both wave lengths in order to generate two grayscale images. The level of intensity of every spot represents the amount of sample hybridized with the corresponding gene. A typical microarray image is shown in Fig.1.

Due to the abundance of experimental data, techniques for automated processing and analysis of microarray images are required. The processing of microarray images [3], usually consists of the following steps: (i) gridding and spot finding, which is the process of assigning the location of each spot in the image, (ii) segmentation, which is the process of grouping the pixels with similar features (this step results in the separation of foreground and background pixels), (iii) intensity extraction, which calculates red and green foreground intensity pairs and background intensities.

Manuscript received April 3, 2006.

N. Giannakeas is with Laboratory of Biological Chemistry, Medical School, University of Ioannina, Ioannina, Greece, GR 45110. (e-mail: me01310@cc.uoi.gr).

D. I. Fotiadis is with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Ioannina, Greece, GR 45110 (0030-26510-98803; fax: 0030-26510-97092; e-mail: fotiadis@cs.uoi.gr).

A. S. Politou is with Laboratory of Biological Chemistry, Medical School, University of Ioannina, Ioannina, Greece, GR 45110.

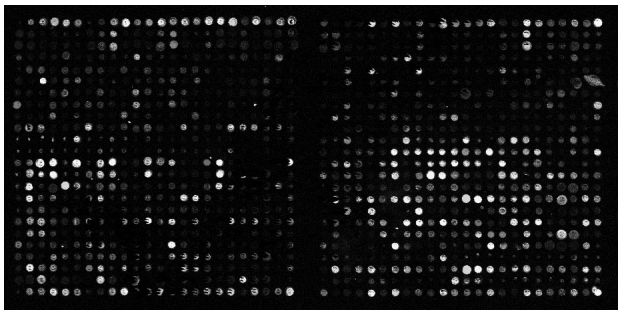

Fig.1: Two blocks of a typical microarray image consist of 24x24 spots each one.

Gridding and spot finding is crucial for making accurate expression measurements, due to their importance into all subsequent steps. Common problems that have to be addressed are (i) the variations of block and spot positions, (ii) the spot shape, which is not perfectly circular (iii) the existence of non-expressed spots that have zero intensity (iv) dust or other contamination on the slide generates artefacts in the image.

According to the user interference, gridding methods can be divided into three main categories: (i) manual, (ii) semi-automated, and (iii) automated. Various techniques have been developed at the end of the previous decade like Genepix [4] and Scanalyse [5] which require inputs from the user. Due to the need for automated processing several approaches have been proposed. Some of them process the vertical and horizontal projections of the image and generate lines in the image according to the valleys of the 1-D signal [6-8]. Morphological operators [9] and smoothing filtering [10] have been used for this purpose. Other methods preprocess the image with an initial segmentation and assign the centers of the objects in the image. Towards the last approach, histogram based segmentation methods [11] and template matching techniques [12-14] are combined with graph models [7] or other techniques taking into account the symmetries of the microarray image.

Manual and semi automated methods require user's intervention. However, most of the proposed automated methods require a large number of tuning parameters. In the current work, an automated method for microarray gridding is proposed. Our method requires only a few user defined parameters. In addition, our approach is able to detect effectively the non-expressed spots. Furthermore, the use of outlier detection can be considered as an advantage since, artefacts existing in the image, are removed effectively.

II. Materials and Method

The proposed method consists of five steps (Fig.2). Initially, the raw microarray image is preprocessed with a template matching technique. In the second step, the blocks of the image are located. The third step is the spot finding in each block. In this step outlier detection is applied on each row and column of spots in order to remove the artefacts. The next step is the detection of the non-expressed spots. Finally a grid is fit on the image using a Voronoi diagram.

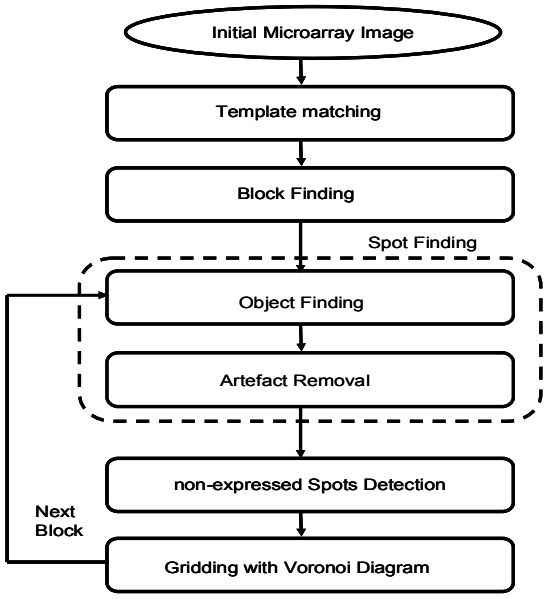

Fig.2: The flowchart of the proposed method

A. Preprocessing

The preprocessing of the image begins with the removal of the background. We set the intensities of the pixels which are under the median intensity value of the greyscale image equal to zero [13].

Next, we locate the objects of the image that are similar to a spot. For this purpose, we apply the template matching technique. We choose the appropriate template and place it at each location in the image. Then, we calculate the similarity between the template and the image on each point, by comparing the values of the template with the corresponding values of the image. The correlation coefficient was used as the measure of similarity. Having the template X_{mn} and the image Y_{mn}, the correlation coefficient is given as:

$$r = \frac{\sum_m \sum_n (X_{mn} - \bar{X})(Y_{mn} - \bar{Y})}{\sqrt{(\sum_m \sum_n (X_{mn} - \bar{X})^2)(\sum_m \sum_n (Y_{mn} - \bar{Y})^2)}}, \quad (1)$$

where $\bar{X}$ is the mean value of the template matrix, $\bar{Y}$ is the mean value of the intensities of the image, m is the number of rows and n the number of columns of the template. The size of the template was approximately equal to the size of the theoretical spot, following a 2-D Gaussian distribution. A pixel is characterized as a signal pixel if the correlation is higher than a threshold value. After several experiments this threshold was set to $r = 0.3$

Finally, we find the locations of all the objects that are characterized as spots. For this reason, we calculate the centre of mass for all the objects. The coordinates of the centre of mass (X_C, Y_C) for a spot are:

$$X_C = \frac{\sum_{i \in D} x_i I_i}{\sum_{i \in D} I_i}, \quad (2)$$

$$Y_C = \frac{\sum_{i \in D} y_i I_i}{\sum_{i \in D} I_i}, \quad (3)$$

where x_i and y_i are the coordinates of the pixel, I_i is the intensity of the pixel in the domain of spot, and D is the domain defining the area of the spot.

B. Block Finding

Due to the deviations of the block positions, it is essential to process each block separately. In order to detect the positions of each block, the preprocessed image is converted into binary and the elements of each row and column of the image are summed. This results in two 1-D signals, with the projections of the image in vertical and horizontal direction. A median filter is then applied to smooth the signals. The above procedure is shown in Fig.3.

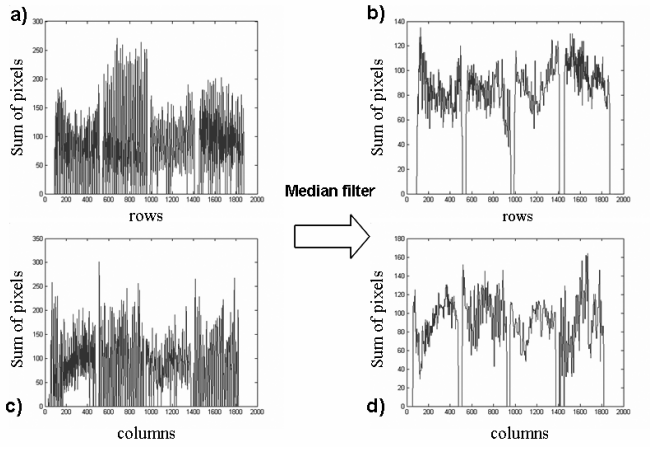

Fig.3: a) The projection in vertical direction, b) the projection in vertical direction smoothed by median filter, c) the projection in horizontal direction, d) the projection in horizontal direction smoothed by median filter.

We detect the valleys in both the projection diagrams and assign the coordinates of the points that are used to split the image into individual blocks. Having v_i and h_i the coordinates of the valleys for the vertical and horizontal

projections respectively, these points are used to split the image into blocks.

C. Spot Finding

The purpose of this step is to find and remove the objects that are generated from contaminations on the microarray slide, in order to obtain the best grid. For this task, each spot is ranked in rows and columns depending on its coordinates. The coordinates of each spot are compared with the mean value of the row and the column coordinate where it belongs. A spot is characterized as an outlier [15], and thus it is removed, if the coordinates of its centre satisfy the following criteria:

$$\begin{matrix} X_e < mean(X_i) - 2stddev(X_i) \\ X_e > mean(X_i) + 2stddev(X_i) \end{matrix}, \quad (4)$$

$$\begin{matrix} Y_e < mean(Y_i) - 2stddev(Y_i) \\ Y_e > mean(Y_i) + 2stddev(Y_i) \end{matrix}, \quad (5)$$

where X_e, Y_e are the centre coordinates to be examined and X_i, Y_i are the coordinates of the centres in the row and column where X_e, Y_e belong.

D. Non-Expressed Spot Finding

An important step in microarray gridding is to find the positions of non-expressed spots. This is required for the reduction of the number of missing values of the microarray data. In order to deal with the non-expressed spots, we first assign the four corners of the block. If a corner-spot is a non-expressed spot, we replace it with a spot, having as row and column coordinates the mean values of row and column coordinates where it belongs.

Then, the distances of neighbour spots are calculated in both directions. The median values of the distances in both directions are also calculated. We prefer the median value and not the mean value since it provides a better approximation of the distance between two expressed spots. After that, the method finds the distances that are higher than the median distance and interpolates the appropriate number of missing spots *N*, according to:

$$N = round\left(\frac{(x_{i+1} - x_i)}{median(d_{x_i})}\right) - 1, \quad (6)$$

where x_{i+1} and x_i are two neighbour spots and d_{x_i} the distance of neighbour spots. The coordinates of the non-expressed spots are:

$$x_n = x_i + n\left(median\left(d_{x_i}\right)\right), \quad (7)$$

and

$$y_n = y_i + n\left(median\left(d_{y_i}\right)\right), \quad (8)$$

where *n=1,2,...,N*. The above procedure is shown schematically in Fig. 4.

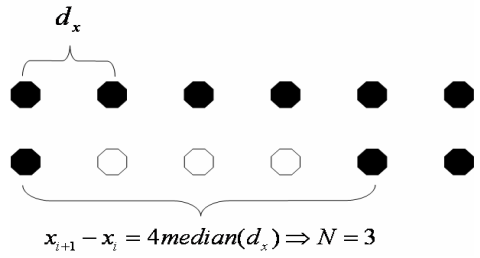

Fig.4: Interpolation of non-expressed spots

This process is realized in both rows and columns. We repeat the process of outlier detection in order to remove spots not correctly located. The coordinates of these spots are replaced with the mean values of the row and the column coordinate where they belong.

E. Voronoi Diagram

After the detection of the centres of all spots in each block, we mark an area (cell) around each spot. For this task, we apply a Voronoi diagram [16] on each block using as vertices the centres of each spot (Fig.5). The Voronoi cell of a centre is the set of points that are closer to this centre than any other center of block. In this way, the Voronoi diagram provides a solution for the variations of the spot positions. In order to restrict the Voronoi diagram to the correct borders, two rows and columns were added, to be considered as the borders of the block.

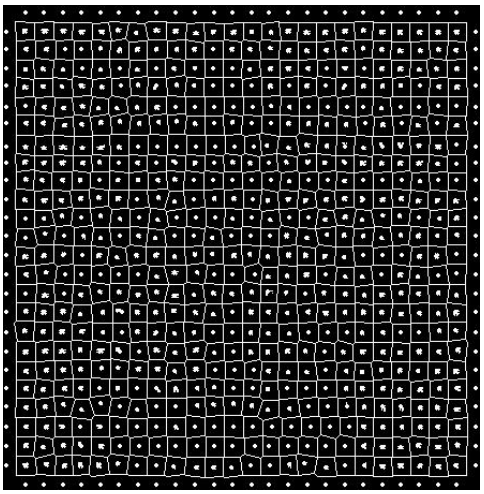

Fig. 5: Voronoi diagram

III. Dataset and Results

In order to evaluate our method, we randomly choose 5 images from the Stanford Microarray Database (SMD) [17]. Every image has 9192 spots ranked in 16 blocks. Scanalyse [5] has been used for the annotation of the images. We compared the centres of the spots that were allocated with the ones of the Scanalyze annotation. The distances between the centres of the annotated and detected spots were calculated in both vertical and horizontal direction. The

histograms of the absolute distances are shown in fig.6.

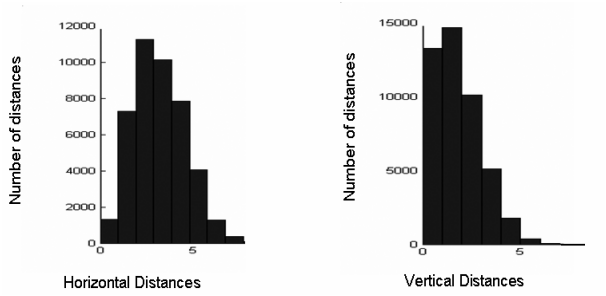

Fig. 6: Histograms of horizontal and vertical absolute distances between the centres of the annotated and detected spots.

In order to calculate the accuracy of our method we define the 2-D distance between the centres of the annotated and detected spots, to be the maximum of the horizontal and the vertical distance:

$$d_{2d} = \max\left(d_x, d_y\right), \tag{9}$$

where d_x and d_y are the distances in the horizontal and vertical direction for a spot. The mean and standard deviation of the distribution of 2-D distances are calculated for each image. We consider a spot detection as correct, if its distance from the annotation is less than 5 pixels. This threshold was selected in order to consider a correct spot detection only if the centre detected by our method is inside the borders of the theoretical spot, which has a diameter of 10 pixels. With the above consideration, we can calculate the accuracy of our method, defined as follows:

$$Acc = \frac{\#\ of\ \ correctly\ detected\ spots}{total\ \#\ of\ \ spots\ in\ the\ image}. \tag{10}$$

Table I presents the mean and standard deviation of the calculated distances as well as the accuracy for every image under consideration. The overall values were also computed.

Table I: The experimental results obtained when evaluating our method in a set of five images, acquired from the SMD.

Image	Mean	Std	Accuracy (%)
lc4b013rex2G	3.6388	1.3033	92.63
lc4b013rex2R	3.9217	1.2949	91.53
lc4b040rex2R	2.3338	2.0931	93.74
lc4b057rex2R	3.4369	1.3151	92.63
lc4b013rex2G	3.4794	1.2289	93.09
overall	**3.3498**	**1.5897**	**92.73**

IV. Discussion

In the current work we presented a method for microarray image gridding. Our method consists of five steps. In the first step the image is preprocessed using template matching. In the second step, the location of the blocks is addressed. In the third step, spot finding takes place. The non-expressed spots are detected in the fourth step and finally, in the fifth step gridding of the image is realized using a Voronoi diagram.

It should be mentioned that gridding of the microarray images is a relatively easy task. However, it is time consuming and many different methods have been proposed to assist the biologists in this direction. The proposed method is advantageous since it is fully automated and requires only a few user defined parameters. Also, it detects effectively the non-expressed spots and is able to deal with noisy images. The use of outlier detection in this approach is advantageous, since artefacts which exist in the image are removed effectively.

Future work should focus on the evaluation of our method in a larger set of images in order to obtain a better estimation of its performance. The segmentation step, which follows the gridding of the image should be considered in order to develop an integrated system for microarray image analysis.

References

[1] M.B. Eisen, and P.O. Brown, "DNA Arrays for Analysis of Gene Expressio," in *Methods Enzymol* vol. 303, pp.179-205, 1999.

[2] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative motoring of gene expression patterns with a complementary DNA microarray," in *Science*, ;vol. 270, pp. 467-470, 1995.

[3] Y.H. Yang., M.I. Buckley, S. Dudoit, and T.P. Speed, "Comparison of methods for image analysis on cDNA microarray data," in *J. Computational and Graphical Stat*, vol. 11, no. 1, pp. 108-136, 2002.

[4] Axon Instruments, *GenePix* A *User's Guide*, 1999.

[5] M.B. Eisen, *ScanAlyze*, http://rana.Stanford.EDU/software/,1999.

[6] P. Bajcsy, "Gridline: Automatic Grid Alignment in DNA Microarray Scans," in *IEEE T. Image Proc.*, vol. 13, pp. 15-25, Issue 1, 2004.

[7] K. Blekas, N. Galatsanos, A. Likas, and I.E. Lagaris, "Mixture Model Analysis of DNA Microarray Images," in *IEEE Transactions on Medical Imaging*, Vol. 24(7), pp. 901-909, 2005.

[8] A.N. Jain, T.A. Tokuyasu, A.M. Snijders, R. Segraves, D.G. Albertson, and D. Pinkel, "Fully automatic quantification of microarray image data," in *Genome Res.*, vol. 12, pp. 325-332, 2002.

[9] J.R. Hirata, J. Barrera, R.F. Hashimoto, and D.O. Dantas, "Microarray Gridding by Mathematical Morphology," in *Int. symp. on comp. graphics, image proc. vision*, pp. 112-119, 2001.

[10] S. Lonardi, and L. Yu, "Gridding and Compression of Microarray Images," in *CSB*, pp. 122-130, 2004.

[11] H.Y. Jung, and H.G. Cho, "An automatic block and spot indexing with k-nearest neighbors graph for microarray image analysis," in *Bioinformatics*, Vol. 2, pp. 141-151, 2002.

[12] V.L. Galinsky, "Automatic registration of microarray images. I. Rectangular grid," in *Bioinformatics*, vol. 19, pp. 1824-1831, 2003.

[13] M. Steinfath, W. Wruck, H. Seidel, H. Lehrach, U. Radelof, and J. O'Brien, "Automated image analysis for array hybridization experiments," in *Bioinformatics* Vol. 17: , pp. 634-641, 2001.

[14] J. Buhler, T. Ideker, and D. Haynor, "Dapple: improved techniques for finding spots on DNA microarrays," *UWCSE Tech Rep. UWTR* Dept. of Comp. Science and Eng., University of Washington, 2000.

[15] M. Kantardzic, *Data Mining Concepts, Models, Methods, and Algorithms*, IEEE Wiley Press, 2003.

[16] F. Aurenhammer, "Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure," in *ACM Comp. Surveys*, vol. 23, pp. 345-405, 1991.

[17] J. Gollub, C.A. Ball, G. Binkley, K. Demeter, D.B. Finkelstein, J.M. Hebert, T. Hernandez-Boussard, H. Jin, M. Kaplper, J.C. Matese, M. Schroeder, P.O. Brown, D. Botstein, and G. Sherlock, "The Stanford Microarray Database: data access and quality assessment tools," *Nucleic Acids Res*., vol. 31, pp. 94-96, 2003.