MICROARRAY IMAGE GRIDDING VIA AN EVOLUTIONARY ALGORITHM

Eleni Zacharia, and Dimitris Maroulis, Member IEEE

Dept. Of Informatics and Telecommunications, University of Athens, Greece,
Emails: eezacharia@gmail.com, dmaroulis@di.uoa.gr, rtsimage@di.uoa.gr

ABSTRACT

Gridding is the first, essential stage of processing cDNA microarray images. The existing tools for allocating the grid structure in a microarray image often require human intervention which causes variations to the gene expression results. In this paper, an original and fully-automatic approach to gridding microarray images is presented. The proposed approach is based on a Genetic Algorithm which determines parallel and equidistant line-segments constituting the grid structure. Thereafter, a refinement procedure follows which further improves the existing grid structure, by slightly modifying the line-segments. Experiments on 16-bit microarray images have shown that the proposed method is effective as well as noise-resistant. Additionally, it achieves an accuracy of more than 95% and it outperforms existing methods.

Index Terms—Microarrays, Image, Gridding

1. INTRODUCTION

cDNA microarrays is a fundamental biotechnological tool which has been utilized in a variety of biomedical application areas, such as cancer research [1]. The reason for its popularity in the scientific community is the fact that scientists can gain insight into the expression of thousands of genes in a single experiment. The end product of a microarray experiment is a digital image which contains one or more distinct blocks, each one containing equal number of spots. A typical microarray image is depicted in Figure 1.

In order for scientists to monitor the expression levels of genes, it is necessary to analyze the digital image. The first important stage in the microarray image analysis is gridding; that is the process of segmenting a microarray image into numerous compartments, each containing one individual spot and background. Gridding however, is far from being a straightforward procedure, as images are contaminated with noise and artifacts, while some spots are poorly contrasted and ill-defined [2]. Additionally, there may be rotations, misalignments and local deformations of the ideal rectangular grid [3].

The available commercial or experimental software packages require human intervention in order to specify the grids properly. For instance, ScanAlyze [4], and ImaGene [5] software programs as well as the morphological method [6], and the Markov random field [7] require human intervention in order to define mandatory input parameters as well as to locate properly the grid structure. The absence of automation in the gridding procedure leads to significant discrepancies in the results of the gene expression levels, even for the same microarray slide as it is reported in [8].

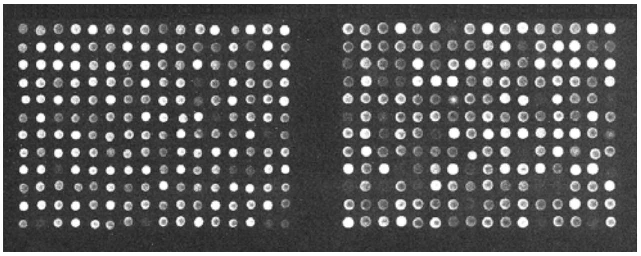

Fig. 1. A typical microarray image

Moreover, techniques which have been proposed to solve the rotation and misalignment problems are not always adequately effective, as they address only limited aspects of these challenging problems. For example, Ho et al [9] can cope with rotation effectively only when the grids are smoothly distorted.

In this paper, an original, fully-automatic approach to gridding microarray images is presented. The proposed approach improves the one reported in [10]. It is based on a genetic algorithm which determines parallel and equidistant line-segments constituting the grid structure. Thereafter, a refinement procedure follows which further improves the existing grid structure, by slightly rotating or transposing the line-segments. The proposed method is noise-resistant and it can effectively cope with rotations, misalignments and local deformations of the ideal rectangular grid.

The rest of this paper is structured in three sections as follows: In Section 2, the details of the proposed gridding method are presented. Experimental results are discussed in section 3 and concluding remarks are apposed in section 4.

2. PROPOSED APPROACH

The proposed approach to gridding microarray images is divided into the following two main stages: I) the microarray image is segmented into blocks and II) each block of the microarray image is segmented into spots. The segmentation of stage I is equivalent to the determination of a set S_B of line-segments whose members constitute the borders of adjacent blocks, while the segmentation of stage II is equivalent to the determination of a set S_S of line-segments whose members constitute the borders of adjacent spots. Let G be a microarray image or block. Each of the S_B or S_S sets can be divided into the following two sub-sets: 1) a sub-set S_V of line-segments whose members are defined by the two vertical sides of G, and 2) a sub-set S_H of line-segments whose members are defined by the two horizontal sides of G.

2.1. A Genetic Algorithm

The determination of line-segments which are included in either the S_V or the S_H sub-sets can be viewed as an optimization problem which is tackled by using the proposed Genetic Algorithm which determines the exact values of the variables of all the line-segments included in both subsets, one sub-set at a time.

2.1.1. Chromosome

The Chromosome m represents all line-segments L_i, i=1,…,$N(m)$ belonging to the S_V or S_H sub-set, where $N(m)$ is the number of the line-segments belonging to the respective sub-set. Before explaining how the Chromosome can represent all the line-segments, it is worth making the following two observations:

1) Any line-segment can be represented on a Cartesian plane and it is defined by its end-points. In this case, for each of the S_V or S_H sub-sets, let the x-axis of the plane be the one of the two sides of G, that does not intersect with the corresponding line-segments. It is obvious that the end-points of line-segments L_i belonging to the S_V or S_H sub-sets are located on the sides of the quadrilateral in shape G. As a result, any line-segment L_i belonging to the S_V or S_H sub-sets can be defined under the condition that the y-coordinates of its end-points are known.

2) Due to the alignment of blocks inside the microarray image and the arrangement of spots inside the blocks, the line-segments belonging to the S_V or the S_H sub-sets are ideally parallel and equidistant. As a result, the distance d between two adjacent line-segments, belonging to the same sub-set, is considered as constant.

According to the above observations, the Chromosome m has been encoded as a string of three real values; the two y-coordinates of the end-points of one line-segment and the distance d between two adjacent line-segments. In the case when the Genetic Algorithm searches for the exact values of the variables of the optimal line-segments belonging to the S_V sub-set, its Chromosome will encode the y-coordinates of the end-points of "$line_{V1}$" and "d_V" (Figure 2). In the case when the Genetic Algorithm searches for the exact values of the variables of the optimal line-segments belonging to the S_H sub-set its Chromosome will encode the y-coordinates of the end-points of "$line_{H1}$" and "d_H".

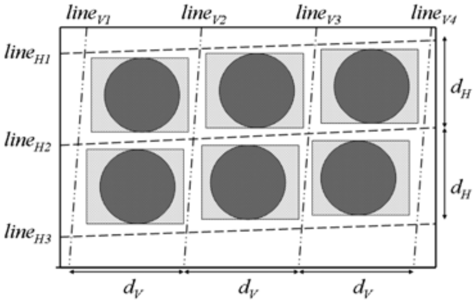

Fig. 2. Line-segments constituting the grid structure in a microarray image or block.

2.1.2. Fitness Function

A line-segment which is part of the grid is located in an area empty of spots. The pixels of this area are part of the background and their intensities are generally lower than the intensities of the pixels constituting spots. As a result, we define the probability $P(L_i)$ of a line-segment L_i to be part of the grid by the following equation:

$$P(L_i) = f_B^{R_{Li}}(L_i) - f_S^{R_{Li}}(L_i) \qquad (1)$$

R_{Li} denotes the region of G which contains those pixels whose distance from the line-segment L_i is less than a margin w. The real-valued function $f_B^{R_{Li}}(L_i)$ expresses the percentage of pixels of the region R_{Li} whose intensity is lower than a value I_B, while the real-valued function $f_S^{R_{Li}}(L_i)$ expresses the percentage of pixels of the region R_{Li} whose intensity is higher than a value I_B. I_B is an intensity value which is defined as the value which is present in most pixels of G. Any pixel below this intensity value I_B belongs to the background.

The Fitness Function $F(m)$ of a Chromosome m that encodes a possible solution to the particular optimization problem is defined by the following equation:

$$F(m) = \begin{cases} S_p(m) \cdot N(m), \; if \; f_{LS}(m) \le f_{Max} \\ S_p(m), \; otherwise \end{cases} \qquad (2)$$

The real-valued function $S_p(m)$ denotes a total sum of the probabilities $P(L_i)$ of the line-segments L_i, i=1,…,$N(m)$, that are represented by the Chromosome m, and have a high probability $P(L_i)$ to be part of the grid. The real-valued function $f_{LS}(m)$ denotes the percentage of the line-segments L_i, i=1,…,$N(m)$, that are represented by the Chromosome m, and have a low probability $P(L_i)$ to be part of the grid. A high probability $P(L_i)$ is the one which is higher than a threshold P_{MAX} while a low one is the one which is lower than a threshold P_{LOW}, where $P_{LOW} < P_{MA}$. $N(m)$ denotes the total number of the line-segments L_i which are represented by the Chromosome m.

2.1.3. Genetic Operators and Termination criterion

The initial population of randomly generated chromosomes evolves because of the subsequent use of: 1) the elitist reproduction, 2) the BLX-a crossover and the Dynamic Heuristic one [11] and 3) the Wavelet mutation [12].

New Populations are produced until the following criterion is met: the Genetic Algorithm is executed up to a maximum number of Populations G_{Fit} for which the best Fitness Value has remained unchanged.

2.2. The Refinement procedure

As mentioned before, due to the alignment of blocks inside the microarray image and the arrangement of spots inside the blocks, the line-segments - having the same direction and constituting the borders of blocks (or spots) - are ideally equidistant. However, this observation may not come true when rotations, misalignments and local deformations of the ideal rectangular grid exist. As a result, the determined line-segments may slightly vary from the optimal ones.

In order to tackle this problem, each line-segment L_i belonging to the S_V or S_H sub-sets is replaced with a new one, L_i', under the following two conditions: 1) the line-segment L_i' is located inside the region R_{Li} of G, 2) the probability $P(L_i')$, of the line-segment L_i', to be part of the grid, is higher than the equivalent probability of L_i ($P(L_i)$), by more than a threshold T_p. An example of the refinement procedure is depicted in figure 3.

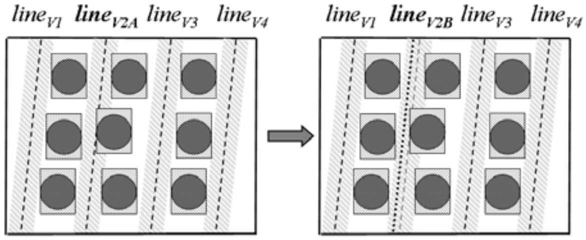

Fig. 3. The line-segment $line_{V2A}$ is replaced with the line-segment $line_{V2B}$. The high-lighted areas on either sides of the line-segments denote the regions R_{Li}.

3. RESULTS

Several experiments were performed in order to evaluate the efficiency of our proposed approach. The set of microarray images, which were used for the evaluation, were obtained from the Stanford Microarray Database (SMD) which is publicly available. This set contained 25 real microarray images, which were digitized at ~5000x 2000 pixels at 16-bit grey level depth and they were stored in TIFF format. Each microarray image contained thousands of spots. It should be noted that due to the fact that the microarray images contained low-intensity spots, the Box-Cox transformation was applied as a pre-processing step, prior to gridding, in order to adjust microarray spot intensities [13].

The parameters of the Genetic Algorithm have been experimentally adjusted once and for all. Thus, the values of the parameters remained stable during the gridding procedure. The population size of the Genetic Algorithm was set to 100. The percentage of each Population which was reproduced was set to 10%. In accordance with [14][15], both the Crossover and the Mutation probabilities were chosen to be 80%. The Termination criterion was satisfied when G_{Fit} equaled 200. The value of the margin (w) was set to 8 when the Genetic Algorithm was searching for line-segments constituting the borders between two adjacent blocks. Respectively, when the Genetic Algorithm was searching for line-segments constituting the borders between two adjacent spots, the margin (w) was set to 2. The values P_{MAX}=0.7, P_{LOW}=0.5, f_{Max}=0.2, and T_p=0.1 were adopted as the most appropriate ones.

Using the proposed approach, 95.1% of spots were perfectly placed inside a compartment, 4.3% were very nearly gridded while only 0.6% were gridded incorrectly, while using a previous version of the approach [10], 94.6% of spots were perfectly placed inside a compartment. It should be noted that the spot areas, used as a reference, were the ones annotated in the SMD. Three gridding results are presented in figures 4, 5 and 6. In the first example, it is obvious that the proposed method has efficiently located the grid structure even though the block is contaminated with noise. In the second example, despite the existence of a local deformation, the proposed method has efficiently located the grid structure, while in the third example the proposed method has efficiently located the grid structure in a rotated sub-image.

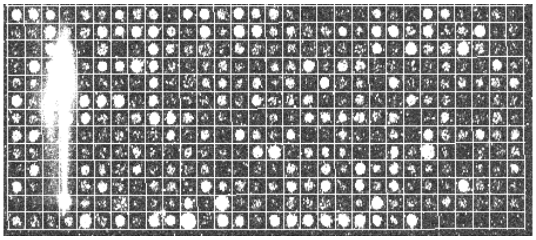

Fig. 4. Gridding results in a microarray sub-image contaminated with noise.

4. CONCLUSIONS

In this paper, an original method for the fully-automatic determination of the grid structure in a microarray image has been presented. The experimental results over real images demonstrate that the proposed method is efficient even when the image is contaminated with noise, or artifacts. Moreover, it can efficiently cope with various kinds of perturbations such as arbitrary rotations or local deformations. It should be noted that following its application to several images, the proposed method achieved an accuracy of more than 95%. To our knowledge, this percentage is much higher than the ones obtained from state-of-the-art gridding techniques.

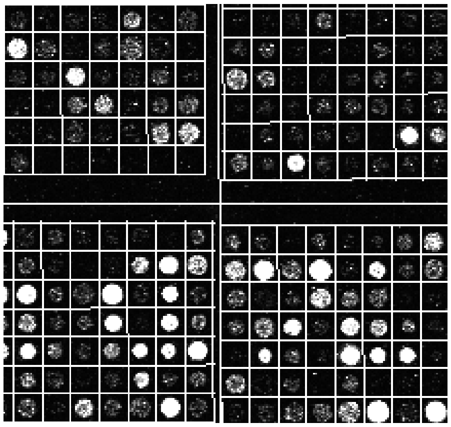

Fig. 5. Gridding results in a rotated microarray sub-image.

5. ACKNOWLEDGMENT

This work was supported by the Greek general Secretariat of Research and Technology and the European Social Fund, through the PENED 2003 program (grant no, 03ED332).

6. REFERENCES

[1] Y. H. Yang, M. J. Buckley, S. Duboit, and T. P. Speed, "Comparison of methods for image analysis on cDNA microarray data," J. Computat. Graph. Statist., vol. 11, pp. 108–136, 2002.

[2] W. B. Chen, C. Zhang, and W. L. Liu, "An Automated Gridding and Segmentation Method for cDNA Microarray Image Analysis," in Proc. 19th IEEE Symp. Computer-Based Medical Systems, Salt Lake City, 2006, pp. 893-898.

[3] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," in Proc. National Academy of sciences, USA, 2002, pp.14031-14036.

[4] M. B. Eisen. (1999). ScanAlyze. [Online]. Available: http://rana.lbl.gov/EisenSoftware.htm

[5] Biodiscovery Inc. (2005). ImaGene. [Online]. Available: http://www.biodiscovery.com/imagene.asp

[6] J. Angulo and J. Serra, "Automatic analysis of DNA microarray images using mathematical morphology," Bioinformatics, vol. 19, no. 5, pp. 553–562, 2003.

[7] O. Demirkaya, M. H. Asyali, and M. M. Shoukri, "Segmentation of cDNA microarray spots using Markov random field modeling," Bioinformatics, vol. 21, no. 13, pp. 2994–3000, 2005.

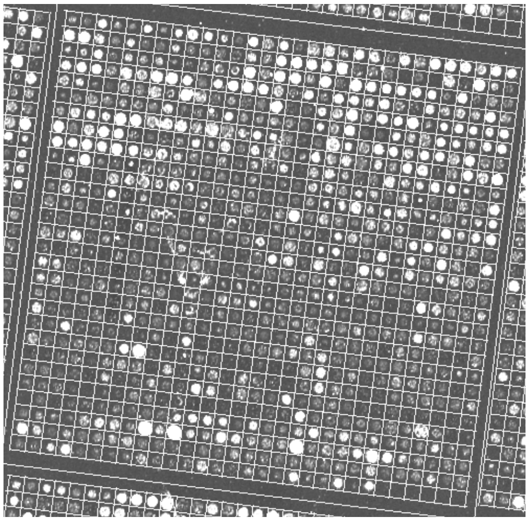

Fig. 6. Gridding results in a rotated microarray sub-image.

[8] N. D. Lawrence, M. Milo, M. Niranjan, P. Rashbass, and S. Soullier, "Reducing the variability in cDNA microarray image processing by Bayesian inference," Bioinformatics, vol. 20, no. 4, pp. 518–526, Mar. 2004.

[9] J. Ho, W. L. Hwang, H. H. S. Lu, D. T. Lee, "Gridding Spot Centers of smoothly distorted microarray images," IEEE Trans. Image Processing, vol. 15, no. 2, pp. 342-353, Feb. 2006.

[10] E. Zacharia, D. Maroulis, "An original Genetic Approach to the Fully-Automatic Gridding of Microarray Images," IEEE Trans. on Medical Imaging, vol. 27, no. 6, pp. 805-813, Jun. 2008.

[11] F. Herrera, M. Lozano, and A. M. Sanchez, "Hybrid crossover operators for real-coded genetic algorithms: An experimental study," Soft Computing, vol. 9, no. 4, pp. 280-298, Apr. 2005.

[12] S. H. Ling, and F. H. F. Leung, "An improved genetic algorithm with average-bound crossover and wavelet mutation operations," Soft Computing, vol. 11, no. 1, pp. 7-31, 2007.

[13] C. T. Ekstrom, S. Bak, C. Kristensen and M. Rudemo, "Spot shape modelling and data transformations for microarrays," Bioinformatics, vol. 20, no. 14, pp. 2270-2278, Sep. 2004.

[14] M. T. Miller, A. K. Jerebko, J. D. Malley, and R. M. Summers, "Feature selection for computer-aided polyp detection using genetic algorithms," in Proc. of SPIE, Santa Clara, 2003, pp. 102-110.

[15] C. Z. Janikow, Z. Michalewicz, "An experimental comparison of binary and floating point representations in genetic algorithms," in Proc. 4th Int. Conf. Genetic Algorithms, San Diego, 1991, pp. 31–6.