Microarray Image Analysis based on an Evolutionary Approach

Eleni Zacharia and Dimitirs Maroulis
Dept. Of Informatics and Telecommunications, University of Athens, Greece
eezacharia@gmail.com, dmaroulis@di.uoa.gr, rtsimage@di.uoa.gr

Abstract

Biological conclusions reached during microarray experiments can be greatly affected by human intervention, which is currently required in microarray image analysis. Therefore, accurate and automatic analysis of cDNA microarray images becomes crucial. In this paper, an automatic approach to microarray image analysis is presented. The proposed approach is based on the concept of evolution in order to process the microarray images. Conducted experiments in a set of real microarray images confirm the effectiveness of the proposed approach.

1. Introduction

It is well known that microarray technology is being applied increasingly in biological and medical research as it enables scientists to simultaneously measure the expression levels of thousands of genes [1]. The end product of a microarray experiment is a digital image containing thousands of spots. Given that the expression level of a particular gene is proportional to the intensity of the corresponding spot, the gene expression information is obtained by processing the digital images.

The first two main stages in the image-analysis procedure are the following: I) Gridding, which is the process of segmenting the microarray image into numerous compartments, each containing one only spot and background, and II) Spot Segmentation, which is the process of segmenting each determined compartment into one spot and its background area [2].

This work was supported by the Greek General Secretariat of Research and Technology (25%), the European Social Fund (75%), and private sector, under the framework of Measure 8.3 of E.P. Antagonistikotita – 3rd European Support Framework – PENED 2003 (grant no. 03-ED-332).

There is a considerable number of techniques and software systems that have been proposed and developed in order to implement the afore-mentioned two stages [3]. However, both of these stages remain challenging tasks. The reason for this lies in the nature of the microarray images [4]. More precisely, the quality of the images is often degraded due to noise and artifacts. Moreover, microarray images may contain low-intensity spots which are not clearly visible. Last but not least, the spots may not have a circular shape and may not be located in an ideal rectangular grid. As a result, the existing techniques or software packages require human intervention in order to specify mandatory parameters or to correct their results. This lack of automation is detrimental as it can affect the final gene-expression results, as it is reported in [5].

In this paper, an original and automatic approach to analyzing microarray images is presented. The gridding and spot segmentation procedures are expressed as optimization problems which are subsequently solved using genetic algorithms (GAs). Specifically, the gridding stage constitutes an improved version of an earlier approach of ours described in [6], while the spot-segmentation stage is an original method. It should be noted that GAs are powerful optimization tools based on the principles of natural selection and evolution. Compared to traditional search and optimization tools, GAs demonstrate superior performance, given that they are suitable for solving problems for which there is little or no a priori knowledge of the underlying processes [7]. Current experiments in real microarray images demonstrate that the proposed method is very effective. The results showed that the proposed method is noise-resistant and it can be applied to images containing low-intensity spots. Moreover, the gridding stage can effectively cope with rotations, misalignments and local deformations of the ideal rectangular grid.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

The rest of this paper is structured in four sections. Section 2 describes the proposed approach to gridding microarray images, while section 3 describes the proposed approach to spot-segmentation. The results of their application in microarray images are presented in section 4, whereas the conclusions of this study are summarized in section 5.

2. Microarray Gridding

In an earlier approach of ours [6], the gridding procedure was expressed as an optimization problem which was performed by a specific GA that determined parallel and equidistant line-segments constituting the borders between two adjacent blocks (depicted as rectangulars, fig. 1,2), or spots (depicted as circles, fig.1,2). The GA was executed twice: Firstly, it determined the "vertical" line-segments of the grid structure (L_1, L_2, and L_3, Fig.1) and subsequently it determined the "horizontal" line-segments of the grid structure (L_4, L_5, and L_6, Fig.1).

However, the determined line-segments may slightly vary from the optimal ones as misalignments and local deformations of the ideal rectangular grid may exist. In order to tackle this problem, the following refinement procedure is applied during which each determined line-segment is examined and, if needed, it is replaced by another one (Fig.2).

2.1 Refinement procedure

Each line-segment L_i which is determined by this GA has a probability $P(L_i)$ to be part of the grid which is defined as follows:

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

R_{Li} denotes the region of the microarray image or block 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} which belong to the background, while the real-valued function $f_S^{R_{Li}}(L_i)$ expresses the percentage of pixels of the region R_{Li} which belong to areas containing spots. It should be noted that a pixel belongs to background if its intensity value is less or equal to I_B otherwise it belongs to spot. I_B is the intensity value that is present in most pixels of the microarray image or block.

Each of the determined line-segments L_i is replaced with a new one L_i' so long as the following are valid: I) the line-segment L_i' is located near the L_i ($L_i' \in R_{Li}$) and, II) $P(L_i')$ is higher than $P(L_i)$ by more than a threshold T_p ($P(L_i') - P(L_i) > T_p$).

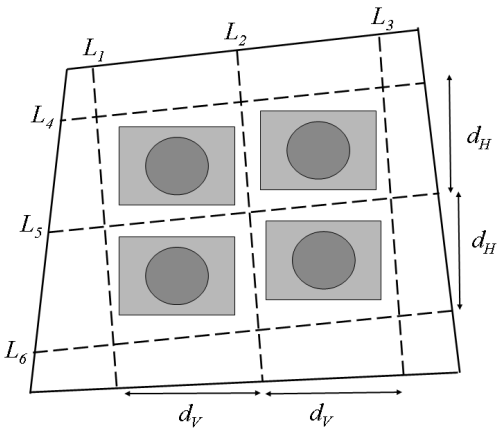


Figure 1. Vertical and horizontal line-segments constituting the grid structure in a microarray image or block. L_1, L_2 and L_3 are parallel and equidistant as well as L_4, L_5 and L_6.

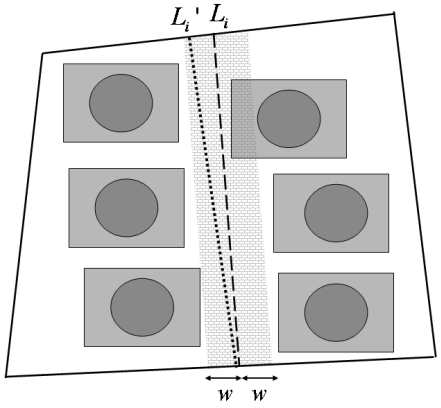


Figure 2. An example of the refinement procedure. The line-segment L_i is examined and it is replaced by the L_i'.

3. Spot Segmentation

Microarray spots can be modeled using a mathematical function. This is because spots share with each other some common characteristics such as an approximately elliptical shape, a "plateaus" or "volcano" shape and an isotropic distribution [8]. An example of a real microarray spot is depicted in Fig.3.

As a result of the afore-mentioned observation, the segmentation of a spot is equivalent to its modeling. Indeed, the contour of a spot can be depicted by drawing the cross-section between the image-plane and its spot-model.

In the proposed method, the modeling of a microarray spot is expressed as an optimization problem which is solved using a specific GA. The GA determines the optimal values of the parameters of the mathematical model. It should be noted that the mathematical function used for the modeling of the microarray spots is the diffusion model proposed by Bettens et al [9]. According to this model a spot can be determined using the following seven parameters: x_o, y_o, (the coordinates of the substance on the plane), B (the background intensity), C_o (the initial concentration of the substance), a' (the area of the disc containing the substance), and D_x', D_y' (the diffusion constants in the two main directions of diffusion).

More precisely, the GA searches for the optimal parameters of the model as follows: Firstly, it creates an initial population (*Pop1*) of chromosomes. Each chromosome represents a spot-model, thus encoding the values of the seven parameters of the diffusion model.

Subsequently, the chromosomes constituting the *Pop1* are evaluated using the fitness function which is described in §3.1. Thereafter, the GA makes the population *Pop1* evolve into a new population *Pop2:* The P_r% of the best chromosomes of *Pop1* is maintained in *Pop2*. The rest are reproduced by applying: 1) the joint application of the BLX-a crossover and of the Dynamic Heuristic one [10] and 2) the Wavelet mutation [11]. This Evolutionary Cycle from one population to the next (*Pop1* to *Pop2*, *Pop2* to *Pop3* and so forth) continues until a maximum number of populations is reached, for which the best fitness value has remained unchanged.

3.1 Fitness Function

The fitness of a chromosome m as a solution to the particular optimization problem is defined by the following equation:

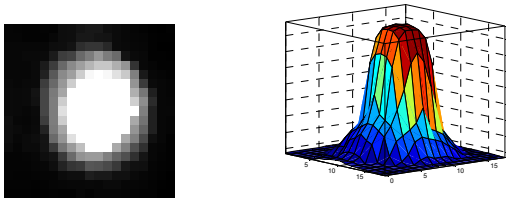
$$F(m) = \sum_{p \in I} \begin{cases} D_{MAX} - D(p), & if \quad D(p) < Th_{AC} \\ -D(p), & otherwise \end{cases} \quad (2)$$

where

$$D(p) = Min(|I(p) - M(p)|, D_{MAX}) \quad (3)$$

$I(p)$ denotes the intensity value of a pixel p of the image I, while $M(p)$ denotes the intensity value of the pixel p of the model M - encoded by the chromosome m. $D(p)$ denotes the error between a pixel p of the model M and the image I. If the error $D(p)$ is more than D_{Max}, then $D(p)$ is set equal to D_{Max}.

It should be noted that the fitness function $F(m)$ equals to D_{max}-$D(p)$ or $-D(p)$ according to the value of $D(p)$. If the error $D(p)$ of a pixel p is less than a threshold Th_{AC}, it means that the model M resembles the image I on that pixel (1st case). In this case, the less the error $D(p)$ is, the higher is the value added in the Fitness Function $F(m)$. On the other hand if the error $D(p)$ of a pixel p is more than the threshold Th_{AC}, it means that the model M does not resemble the image I on that pixel (2nd case). In this case, the higher the error $D(p)$ is, the higher is the value deducted from the fitness function $F(m)$. According to the fitness function $F(m)$, the higher the fitness value of the chromosome m, the higher the resemblance of the model M (encoded by the chromosome m) to the image I.



(a) (b)

Figure 3. A real microarray spot in 2D and in 3D dimensions.

4. Results

Several experiments were performed in order to evaluate the efficiency of the proposed approach. In this respect, a set of real microarray images - from the Stanford Microarray Database (SMD) - was used, which contained thousands of spots [12].

The parameters of the GA have been experimentally adjusted once and for all. Thus, the values of the parameters remained stable in all the experiments which were performed. The value of the margin (w) was set to 8 when the GA was searching for line-segments constituting the borders between two adjacent blocks. Respectively, when the GA was searching for line-segments constituting the borders between two adjacent spots, the margin (w) was set to 2. The T_p threshold in the refinement procedure was set to 0.2. The D_{Max} and Th_{AC} thresholds were set to 30.0 and 3.0 respectively.

The accuracy of the proposed method was analyzed by means of a statistical analysis. In detail, using the proposed gridding approach, 95.1% of spots were perfectly placed inside a compartment, 4.3% were very nearly gridded while only 0.6% of spots were gridded incorrectly. These results are higher than the ones mentioned in [6]. Indeed, using the earlier approach of ours [6], 94.6% of spots were perfectly placed inside a

compartment, 4.8% were very nearly gridded and only 0.6% of spots were gridded incorrectly. Using the proposed spot-segmentation approach, 85.3% of spots were very nearly segmented. Moreover, the afore-mentioned stages of the proposed approach are automatic as they do not require any human intervention.

Examples of the gridding and spot-segmentation stages are presented in Fig.4. It becomes obvious that the proposed approach has optimally detected the grid structure. Moreover, it segmented all the spots, even those which are not clearly visible.

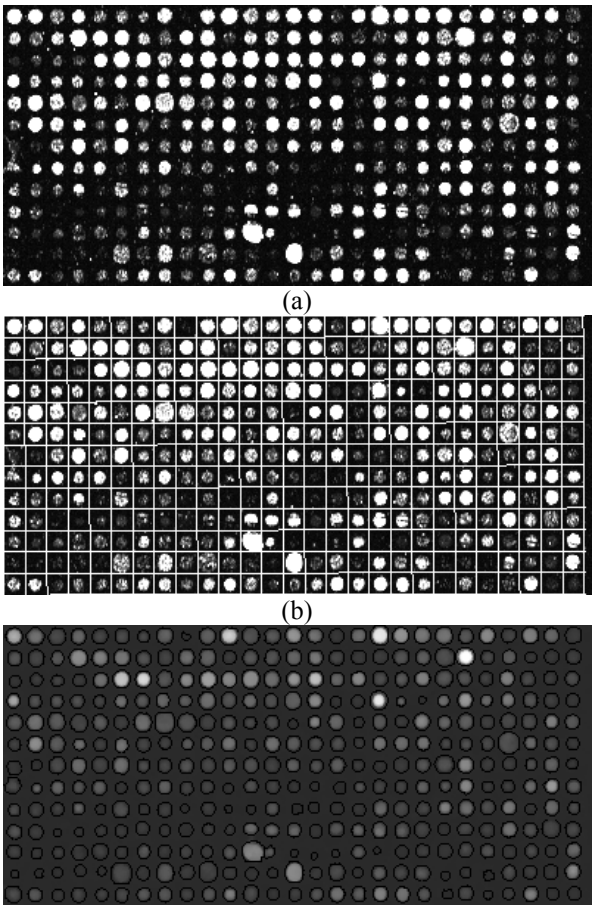


Figure 4. Microarray image analysis using the proposed approach. (a) an area of the microarray image containing low intensity and distorted spots, (b) gridding result , (c) spot Segmentation result

5. Conclusions

In this paper, an original method for gridding and spot-segmenting is presented. The proposed approach is based on the concept of evolution in order to process microarray images. Two GAs are used. The first one determines the line-segments constituting the grid structure, while the second one determines the models which optimally represent each spot.

The proposed approach is automatic as it does not require any human intervention. The experimental results over real images demonstrate that the proposed method is efficient and it outperforms competing state-of-the-art gridding techniques. It can also be applied to distorted microarray images as well as to images missing spots or containing spots of various intensities.

References

[1] D. Stekel. *Microarray Bioinformatics*. Cambridge University Press, UK, 2003.

[2] Y.H. Yang, M. Buckley, S. Dudoit, T. Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, 11(1):108–136, March 2002.

[3] P. Bajcky. An overview of DNA microarray grid alignment and foreground separation approaches. *Journal on Applied Signal Processing*, 2006:1-13, 2006.

[4] M. Ceccarelli, G. Antoniol. A deformable Grid-Matching Approach for Microarray Images. *IEEE Trans. on Image Processing*, 15(10):3178-3188, Oct 2006.

[5] W. Zhang, I. Shmulevich, J. Astola. *Microarray Quality Control*, JohnWiley & Sons, Inc., Hoboken, New Jersey, 2004.

[6] E. Zacharia, D. Maroulis. An original Genetic Approach to the Fully-Automatic Gridding of Microarray Images. *IEEE Trans. on Med. Imaging*, 27(6):805-813, Jun 2008.

[7] D. E. Goldberg. *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison-Wesley, Boston, 1989.

[8] C.T. Ekstrom, S. Bak, C. Kristensen, M.Rudemo. Spot shape modeling and data transformations for microarrays, *Bioinformatics*, 20(14):2270-2278, 2004.

[9] E. Bettens, P. Scheunders, D.V. Dyck, L. Moens P.V. Osta. Computer analysis of two dimensional electrophoresis gels: A new segmentation and modelling algorithm. *Eletrophoresis*, 18 (5): 792-798, May 1997.

[10] F. Herrera, M. Lozano, A. M. Sanchez. Hybrid crossover operators for real-coded genetic algorithms: An experimental study. *Soft Computing*, 9(4):280-298, 2005.

[11] S. H. Ling, F. H. F. Leung. An improved genetic algorithm with average-bound crossover and wavelet mutation operations. *Soft Computing*, 11(1):7-31, 2007.

[12] Standford microarray database. [Online]. Available: http://genomewww5.stanford.edu/