An Unsupervised and Fully-Automated Image Analysis Method for cDNA Microarrays

E. Zacharia, D. Maroulis
Dept. of Informatics and telecommunication, University of Athens, Greece
rtsimage@di.uoa.gr

Abstract

Microarray gene expression image analysis is a labor-intensive task and requires human intervention since microarray images are contaminated with noise and artifacts while spots are often poorly contrasted and ill-defined. The analysis is divided into two main stages: Gridding and Spot-Segmentation. In this paper, an original, unsupervised and fully-automated approach to gridding and spot-segmenting microarray images, which is based on two genetic algorithms, is presented. The first genetic algorithm determines the optimal grid while the second one determines, in parallel, the boundaries of multiple spots. Experiments on 16-bit microarray images show that the proposed method is effective and achieves more accurate gridding and spot-segmentation results in comparison with existing methods.

1. Introduction

cDNA microarrays [1] is a powerful biotechnology tool for analyzing expressions levels of thousands of genes in normal and abnormal samples. The gene expression information is obtained by fluorescently scanning the hybridized glass slides and analyzing the scanned images. A typical microarray image is composed of blocks, each containing a number of spots of various fluorescence intensities, which correspond to the level of hybridization of the samples. A variety of software packages have been developed for the analysis of microarray images. In these packages, image analysis is divided into two main stages: Gridding and Spot Segmentation [2].

The analysis of the scanned images is not a straightforward process since the quality of microarray images suffers from the existence of noise (i.e. dust on the slide), artifacts (i.e. inner holes and scratches) and uneven background [3][4]. Moreover, lowly expressed genes in microarray experiments are depicted by spots that are poorly contrasted and ill-defined. Increasing the photometric gain during scanning is not ideal for studying these spots since it may cause some pixels of highly expressed genes to become saturated [5]. Given that the automatic image analysis is defective, all of the above problems lead to errors that propagate to all the following stages of the statistical analysis [6]. Therefore, image analysis tools require human intervention in order to recognize the boundaries of spots.

The first step in image analysis is the spot-gridding problem. Software packages such as ScanAlyze [7] and GenePix [8] are based either on a grid with uniform cells or on manual intervention in order to specify the grids properly. Amongst other well-known techniques, the axis projections method [9], the morphological method [10], the Markov random field [11], and the template matching and seeded region growing method [12] are most commonly used. However, all the aforementioned techniques are only semiautomatic as they require mandatory input parameters and, at times, manual intervention in order to locate the grid precisely.

Gridding is followed by spot-segmentation techniques. ScanAlyze software [7] uses a fixed circle segmentation method while GenePix [8] uses an adaptive circle segmentation method. Both of them are not ideal for non-circularly spots. QuantArray software [2] computes a

Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)
0-7695-2905-4/07 $20.00 © 2007 IEEE

threshold based on the nonparametric Mann–Whitney test. However, it does not provide reliability in an automated system given that it requires a background sample. Other methods are the seeded growing method [13] and the method based on a generalized hit-or-miss transformation [14]. The former relies on the selection of a seed point while the latter requires a training sample from which a typical spot shape can be acquired.

In this paper, an original, unsupervised and fully-automated approach for gridding and spot-segmenting microarray images, based on two genetic algorithms, is presented. The first genetic algorithm determines the optimal line segments which constitute the grid while the second one determines, simultaneously, the boundaries of multiple spots by optimally fitting them into diffusion models. Current experiments demonstrate that the proposed method gives very effective results and the accuracy of location reaches a very high value. To the best of our knowledge genetic algorithms have not been previously applied to microarray gridding and spot-segmenting.

The rest of this paper is structured in three sessions. Section 2 describes the proposed approach to gridding microarray images and segmenting cDNA spots. The results of its application in microarray images are presented in section 3, whereas the conclusions of this study are summarized in section 4.

2. Image analysis using genetic algorithms

Locating the optimal grid and the optimal boundaries of spots in microarray images is not a straightforward process, seeing that we know little for them. Therefore, we developed an original method for cDNA microarray image analysis based on genetic algorithms since they are stochastic, robust optimizers, suitable for solving problems for which there is little or no a priori knowledge of the underlying processes [15].

The proposed approach consists of the following three main steps: a) Preprocessing of input images by applying wavelet-based noise reduction and Box-Cox transformation adjustment, b) Gridding the preprocessed images using a genetic algorithm which determines the optimal lines which constitute the grid, c) Spot-segmenting together with model-based quantificating of individual spots using another genetic algorithm, which determines, concurrently, the parameters of multiple diffusion models that optimally fit the characteristics of possible spots.

2.1. Preprocessing cDNA microarray images

The quality of microarray images suffers due to the deficiencies of the equipments used in microarray experiments [3][4]. Therefore, stationary wavelet-based filtering of the input images [4] and Box-Cox transformation of the de-noised images [5] are employed as a pre-processing step with the purpose of improving the quality of the images.

2.2. Gridding cDNA microarray images

Gridding of microarray images is divided in the following two main stages: a) Detection of the borders between blocks and b) Identification of the borders between spots. In each stage, the genetic algorithm is used twice. Firstly, it performs a search in order to determine the optimal parameters of the line segments which constitute the borders between the blocks (1st stage) or spots (2nd stage) and are defined by the two vertical sides of the image or block respectively. Then the image or block is rotated by 90^{o} and the genetic algorithm is used again.

2.2.1. Chromosome: Let G be an M_1xM_2 preprocessed microarray image or block. The chromosome encodes the parameters of a line segment which is defined by a point $A(a_1,a_2)$ in the left vertical side (i.e. a_2=0) of G and $B(b_1,b_2)$ a point in the right vertical side (i.e. b_2=M_2-1) of G. Since a_2, b_2 are known, only the parameters a_1, b_1 are encoded into a single real-valued chromosome m_G.

2.2.2. Genetic Operators and Termination criterion: The initial population of randomly generated chromosomes is evolved by the subsequent use of a) the elitist reproduction, b) the BLX-a crossover and the Dynamic Heuristic one [16] and c) the Wavelet mutation [17].

The genetic algorithm is executed until a maximum number of generations have been reached from which no chromosome has larger fitness than a threshold T_f. It should be noted that during the evolution, chromosomes, with fitness value above the given threshold T_f, are considered solutions, i.e. line segments of the grid, and they are depicted in image G.

2.2.3. Fitness Function: The fitness of a chromosome m_G as a solution to the particular optimization problem is defined by the following equation:

$$F_G(m_G) = f_p(m_G) - f_n(m_G) \quad (1)$$

where the real valued functions $f_p(m_G)$ and $f_n(m_G)$ are named as positive term and negative term of the fitness respectively.

The positive term $f_p(m_G)$ expresses the percentage of pixels of the image G whose a) intensity is smaller than a value I_B and b) distance from the line encoded by the chromosome m_G, is less than a constant D. I_B is the largest intensity value that is present in a percentage of pixels, k% less than the maximum percentage of pixels depicting equal intensity value. D is a constant which controls the width of the margin existing between blocks or spots.

Respectively, the negative term expresses the percentage of pixels whose a) intensity is larger than I_B and b) distance from the line encoded by the chromosome m_G is less than D.

2.3. Segmenting cDNA spots

cDNA spots have some common characteristics, such as an approximately "volcano" or "plateaus" elliptical shape and an isotropic distribution. These characteristics can be captured by tuning the parameters of a mathematical model so that it fits to the image region containing a spot.

Thus, the proposed genetic algorithm for spot-segmenting performs a parallel search for determining the optimal model parameters of the cDNA spots which belong to an N_1xN_2 window of adjacent regions of the grid. The boundaries of cDNA spots are the cross section between the models and the image plane.

2.3.1. Chromosome: Since cDNA strands are hybridized by a diffusion process [18], the diffusion model proposed by Bettens et al [19] can be used in order to model cDNA spots. According to this model, a spot can be defined by the set of the following parameters: x_o, y_o, B, C_o, a', D_x', D_y', where x_0, y_0 stands for the coordinates of the substance on the plane. B is the background intensity. C_0 is the initial concentration of the substance. a' is the area of the disc containing the substance. D_x' and D_y' represent the diffusion constants in the two main directions of diffusion.

However, the diffusion process of cDNA strands differs from the one proposed in [19] because it is isotropic. Therefore, cDNA spots can be modeled on the aforementioned diffusion model only if it is properly modified from an anisotropic to an isotropic one. Consequently, we assume that D_x' cannot differ from the D_y' by more than a threshold T_D.

The parameters of the diffusion models that correspond to the cDNA spots, which belong to an N_1xN_2 window of adjacent regions of the grid, are encoded into a single three-dimensional chromosome m_s which consists of m_{ijz} cells, i=1,2,3,…N_1, j=1,2,3,…,N_2, z=1,2,…7. The chromosome also consists of segments s_{ij}, i=1,2,3,…N_1, j=1,2,3,…,N_2. Each segment s_{ij} encodes the diffusion-model parameters of the spot which belongs to the equivalent position (i,j) of the N_1xN_2 window. The B, C_o, a', D_x', D_y' parameters of a spot model are encoded as real value numbers while the x_o, y_o parameters are encoded as integers.

2.3.2. Genetic Operators and Termination criterion: The initial population is evolved in a similar manner to the one used in the gridding approach. However, since the chromosome is three-dimensional, the parents of the crossover can be either different chromosomes or columns of the same chromosome.

The genetic algorithm is executed until all the regions of the grid have been tested as possible positions of spots.

2.3.3. Fitness Function: The fitness of a chromosome m_s, as a solution to the particular optimization problem, is defined by the following equation:

$$F_S(m_s) = \sum_{i=1}^{i=N_1} \sum_{j=1}^{j=N_2} f_L(s_{ij}) \tag{2}$$

where the real valued function $f_L(s_{ij})$ is defined as the local fitness of a chromosome segment s_{ij}, i=1,2,3,…,N_1 , j=1,2,3,…,N_2 and is computed by the following equation:

$$f_L(s_{ij}) = \begin{cases} \dfrac{d_T(s_{ij}) + d_I(s_{ij})}{E(s_{ij})}, & \text{if } \dfrac{d_T(s_{ij}) + d_I(s_{ij})}{E(s_{ij})} < T_s \\ d_T(s_{ij}) + d_I(s_{ij}), & otherwise \end{cases} \tag{3}$$

where $d_T(s_{ij})$ expresses the number of pixels p belonging to the model spot encoded by the segment s_{ij}, and whose intensities do not differ from the intensities of the equivalent pixels in the image by more than a value $|I_D(p)|$. $I_D(p)$ is defined as: $I_D(p)=w \cdot I_p$, where $0<w<1$ is a constant and I_p is the intensity of the pixel p. $d_I(s_{ij})$ specifies the number of pixels p belonging to the inside of the model spot encoded by the s_{ij}, and whose intensities do not differ from the intensities of the equivalent pixels in the image by more than the value $|I_D(p)|$. $E(s_{ij})$ defines the number of pixels belonging to the spot model encoded by s_{ij} . T_s is a threshold for the local fitness.

3. Results

Several experiments were executed so as to evaluate the performance of the proposed algorithm on a set of microarray images. Each image of the set contains thousands of spots at 16-bit grey level depth. Figure 1 illustrates a partial of image "Array1.tif" (Array1.tif analysis: 1916 x 1872) used in [20].

In both genetic algorithms, a population of 100 chromosomes was used. In each generation of the genetic algorithm, 10% of the best chromosomes were maintained in the population, whereas the rest were reproduced by crossover and mutation operations. In accordance with [21][22], a high crossover probability of 0.8 and a high mutation probability of 0.8 were chosen. The genetic algorithm for gridding was applied using the following values of the thresholds. T_f=0.8, D=100 when the genetic algorithm is performed for detecting the boundaries of blocks and D=10 when the genetic algorithm is performed for detecting the boundaries of spots. The genetic algorithm for spot-segmenting was performed using a window with N_1=3 and N_2=3. Moreover, T_D=0.25, and T_s = 1.4.

Using the proposed approach, the accuracy of locating the spot reaches the high value of 84.6% while MicroZip software program achieves only 77%. Moreover we have observed

that the three dimensional chromosome permits the genetic algorithm to accelerate its convergence.

Examples of gridding and spot-segmenting results are illustrated in Figure 1. This figure contains 192 real microarray spots. As we can observe, the proposed method detects all the lines of the grid (Fig.1c) while MicroZip software failed to detect the last raw (Fig. 1b). Moreover, the proposed approach segmented all the 192 real spots. On the other hand, the MicroZip software did not detect 24 real microarray spots and failed to segment 29 more spots.

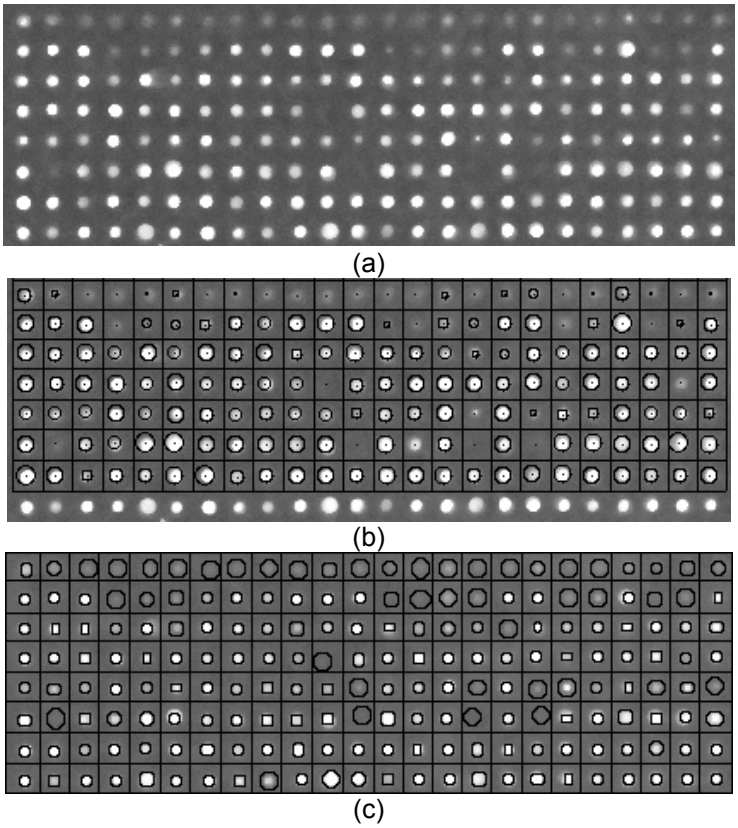

Figure 1: Microarray image analysis results: (a) input microarray sub-image, (b) output of the MicroZip software program, (c) output of the proposed method

4. Conclusions

In this paper, an original method to detect the grid of and segment microarray spots in microarray images based on two genetic algorithms has been presented. The first genetic algorithm conducts a search in order to determine the optimal parameters of the line segments which constitute the borders between the blocks or spots. The second genetic algorithm performs a parallel search for determining the optimal model parameters for the cDNA spots which belong to a window of adjacent regions of the grid.

The proposed method has the following advantages: a) it is an unsupervised technique as it does not require a training phase, b) it is capable of segmenting spots which have volcano or plateaus shape, c) it is capable of segmenting spots distorted by imperfect diffusion and d) its spot-segmentation rate is higher than that of other software packages.

Future development of this project includes further experimentation, optimization and parallelization of the proposed method, and its integration into a complete user-friendly software application.

5. Acknowledgments

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. 03-ED-332).

6. References

[1] Y.F. Leung and D. Cavalieri, "Fundamentals of cDNA microarray data analysis", Trends in genetics, Elsevier, vol. 19, no. 11, Nov. 2003, pp. 649-659.

[2] Y. Chen, E.R. Dougherty and M.L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images", Journal of Biomedical Optics, SPIE, vol.2, no. 4, Oct. 1997, pp. 364–367.

[3] W.B. Chen, C. Zhang and W.L. Liu, "An Automated Gridding and Segmentation Method for cDNA Microarray Image Analysis", Proc. of the 19th IEEE Symposium on Computer based Medical Systems, Salt Lake City, UT, June 2006, pp. 893-898.

[4] X.H. Wang, R.S.H. Istepanian, Y.H. Song, "Microarray Image Enhancement by denoising using stationary wavelet transform, IEEE Transaction on nanobioscience, vol.2, no. 4, Dec. 2003, pp.184-189.

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

[6] K. Blekas, N.P. Galatsanos and I. Georgiou, "An unsupervised artefact correction approach for the analysis of DNA microarray images", IEEE International Conference Image Processing, Barcelona, Spain, Sep. 2003, pp. 165-168.

[7] M.B. Eisen, ScanAlyze documentation, http://rana.lbl.gov/EisenSoftware.htm, 1999.

[8] Axon Instruments, GenePix Pro documentation, http://www.axon.com, 2002.

[9] N. Deng and H. Duan, "The Automatic Gridding Algorithm based on projection for Microarray Image", Proc. of the International conference on Intelligent Mechatronics and Automation, Chendu, China, August 2004, pp.254-257.

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

[11] M. Katzer, F. Kummert and G. Sagerer, "A Markov Random Field Model of microarray gridding", Proc. of the 2003 ACM Symposium on Applied computing, ACM, Melbroun, Florida, March 2003, pp. 72-77.

[12] Y.H. Yang, M.J. Buckley, S. Duboit and T.P. Speed, "Comparison of methods for image analysis on cDNA microarray data", Journal of Computational and Graphical Statistics, vol. 11, no.1, 2002, pp. 108–136.

[13] O. Demirkaya, M.H.Ayali and M.M.Shoukri, "Segmentation of cDNA microarray spots using markov random field modelling", Bioinformatics, Oxford University Press, vol.21, no. 13, July 2005, pp. 2994-3000.

[14] P. Vesanen, "Calibration-free methods in segmentation of cDNA microarray images", Proc. IS&T/SPIE 12th Symp. Electronic Imaging Science and Technology, May 2002, pp. 291–302.

[15] D.E. Goldberg, Genetic Algorithms in Search, Optimization & Machine Learning, Addison-Wesley, Reading, 1989.

[16] F. Herrera, M. Lozano and A.M. Sanchez, "Hybrid crossover operators for real-coded genetic algorithms: An experimental study", Soft Computing - A Fusion of Foundations, Methodologies and Applications, Spinger, vol. 9, no.4, April 2005, pp. 280-298.

[17] S.H.Ling, F.H.F. Leung, "An improved genetic algorithm with average-bound crossover and wavelet mutation operations", Soft Computing - A Fusion of Foundations, Methodologies and Applications, Spinger, vol. 11, no.1, Aug. 2006, pp. 7-31.

[18] C. Gadgil, A. Yeckel, J.J. Derby and W.S. Hu, "A diffusion-reaction model for DNA microarray assays", Journal of Biotechnology, Elsevier, vol. 114, no. 1-2, Oct. 2004, pp. 31-45.

[19] E. Bettens, P. Scheunders, D.V. Dyck, L. Moens and P.V. Osta, "Computer analysis of two dimensional electrophoresis gels: A new segmentation and modelling algorithm", Eletrophoresis, vol. 18 no. 5, May 1997, pp. 792-798.

[20] S. Leonardi and Y. Luo, "Gridding and Compression of Microarray images", Proc. of the 2004 IEEE Computational Systems Bioinformatics Conference, Standford, USA, vol. 00, Aug. 2004, pp. 122-130.

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

[22] M. T. Miller, A. K. Jerebko, J. D. Malley, and R. M. Summers, "Feature selection for computer-aided polyp detection using genetic algorithms", Proceedings of SPIE, vol. 5031, 2003, pp. 102-110.