EXTRACTION OF SPOTS IN DNA MICROARRAYS USING GENETIC ALGORITHM

A. Sreedevi[1] and Dakshayani. S. Jangamshetti[2]

[1]Department of E&E E, R.V. College of Engineering, Bangalore, India
[2]Department of E&E E, Basaveshwar Engineering College, Bagalkot, India

ABSTRACT

DNA microarray technology is an eminent tool for genomic studies. Accurate extraction of spots is a crucial issue since biological interpretations depend on it. The image analysis starts with the formation of grid, which is a laborious process requiring human intervention. This paper presents a method for optimal search of the spots using genetic algorithm without formation of grid. The information of every spot is extracted by obtaining a pixel belonging to that spot. The method developed selects pixels of high intensity in the image, thereby spot is recognized. The objective function, which is implemented, helps in identifying the exact pixel. The algorithm is applied to different sizes of sub images and features of the spots are obtained. It is found that there is a tradeoff between accuracy in the number of spots identified and time required for processing the image. Segmentation process is independent of shape, size and location of the spots. Background estimation is one step process as both foreground and complete spot are realized. Coding of the proposed algorithm is developed in MATLAB-7 and applied to cDNA microarray images. This approach provides reliable results for identification of even low intensity spots and elimination of spurious spots.

KEYWORDS

genetic algorithm, dna microarray, gridding, foreground, background estimation

1. INTRODUCTION

Genomic engineering is an emerging technology which relates biology, medicine and engineering. Genomic engineering plays a very important role in medical field, such as drug discovery, gene discovery, diagnosis of disease, effect of drug etc. [1]. The DNA microarray analysis targets to identify differentially expressed genes in control sample with respect to reference sample, which can be utilized to study the functions of genes and gene expression levels. Classical methods deal with analyzing a single gene (probe), whereas DNA microarrays contain tens of thousands of genes. DNA microarrays are available as single, double and multi-fluorescent images, depending on labeling of complimentary DNA (cDNA). Double fluorescent images are most common, in which cDNA are labeled with red (532nm) and green (632nm) intensities. The cDNA extracted from control and reference samples, labeled with different fluorescent dyes are hybridized and then spotted on glass slides with robotic means [2].

DOI : 10.5121/sipij.2013.4607

In order to arrive at meaningful biological conclusions regarding DNA expression, the initial steps involved require image processing tools. As the DNA expression is a measure of intensity of spot, intensities of all the pixels belonging to individual spots have to be measured accurately. Hence extensive image processing and analysis of microarray images is necessary for reliable biological interpretations.

Analysis of microarray images comprises of three major steps: gridding, segmentation and intensity extraction. The techniques for gridding that are most commonly used are only semi-automatic as they require mandatory input parameters such as number of horizontal and vertical lines and at times manual intervention, in order to locate the grid precisely [3]. Placing the grid is a challenge as the spots are not uniformly spaced. So most of the techniques are semi-automatic and may involve human intervention for choosing the position for the grid, since gridding forms unevenly spaced parallel and perpendicular lines. Also the non-uniformity in the position of the spots makes it almost impossible to get a grid, such that a single complete spot is positioned in an individual grid cell. Gridding is followed by spot segmentation techniques. ScanAlyze software uses a fixed circle segmentation method while GenePix uses adaptive circle segmentation method. Both the methods are not idyllic for measuring spot intensities of non circular spots.

Gene expression data derived from arrays may be used for gene clustering, cancer detection and other analysis. Then DNA microarrays provide a medium for matching known and unknown DNA samples based on base-pairing rules and automating the process of identifying the unknowns.

2. LITERATURE SURVEY

During the microarray slide preparation non-uniformity in spot positioning is caused by the equipment, since the pins in the spotting machine might bend over the time leading to irregularity in spacing between the spots. Under such conditions template based approaches for grid formation may lead to inaccurate results [4-6]. The limitation of microarray image analysis is due to cumulative errors from numerous sources; hence often require manual analysis of array image data to ensure accuracy [7].

There is no direct approach for both gridding and spot segmentation. It is difficult to obtain optimal location of the grid location due to non uniformity in spot location. Hence it is a complicated, time consuming process and requires human intervention.

Automatic gridding techniques are available like hill climbing approach [8], and gridding using genetic algorithms (GA) [9]. Since last decade, GA is implemented to search for optimum value in a given search space. GA is more suitable when the search space is large because of its parallel searching capability. GA search gives global optimal solution. Genetic algorithms are stochastic, robust optimizers, suitable for solving problems, where there is little or no prior knowledge is available.

In our earlier work, an approach based on GA consists, a single step wherein the spots are located without the formation of gridding [10]. This paper presents a method for identification of foreground and background pixels of spots and spot centers along with its mean values.

3. GENETIC ALGORITHM AND ITS IMPLEMENTATION

The work concentrates on detailed way in which using GA, spots in the microarray are identified and the estimation of foreground and background has been computed.

The following steps are involved in Image Analysis of DNA microarray images

1. Preprocessing
2. Identification of spots
3. Segmentation
4. Finding the center and average intensity
5. Removal of spurious spots
6. Background estimation

3.1 Preprocessing

An ideal spot representing DNA should contain pixels of uniform intensity. But hybridization and slide preparation leads to non uniform intensities of the pixels belonging to an individual spot. So preprocessing is necessary to remove non uniformity and the noise.

Microarray image is preprocessed with Median filter which is an order stochastic filter and non linear in nature. Median filter replaces the value of a pixel by the median of gray levels in the neighborhood of that pixel. Median filters provide excellent noise reduction capability for random and salt and pepper noises with considerably less blurring than linear smoothing filters of same size. Hence edges are preserved and noise is eliminated. Also isolated clusters of pixels with an area less than $n^2/2$ are eliminated by an n x n median filter, hence even spurious spots can be removed to a larger extent.

3.2. Identification of spots

Identification of spots is carried out using GA approach without formation of grids. Genetic algorithm searches fitness function which is the combination of objective function and constraint function. Finding a pixel belonging to a spot is a maximization problem in GA. A part of DNA Microarray image of size n x m pixels is considered. Certain number of pixels having total fitness greater than the threshold value is encoded as a chromosome.

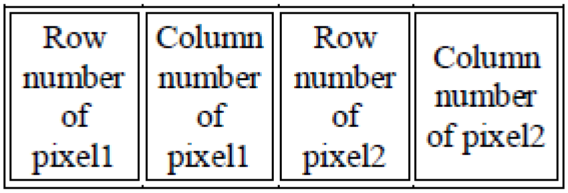
3.2.1. Chromosome:

The "chromosomes" encode a group of linked features. In our work chromosome is encoded as an array of binary digits. Coding consists of number of strings, each string representing either row or column of a pixel. Number of strings is decided by the number of pixels considered for a chromosome.

Figure 1 shows sample coding when two pixels are considered for a chromosome. Row and column numbers of each pixel is coded in binary. Number of bits of each string and hence length of chromosome depends on the size of the sub image chosen for processing.

For example a sub image of size 31 x 31 and two pixels per chromosome are considered for processing.

Length of each substring = 5 bits
Length of each string (chromosome) = 20 bits



Row number of pixel1	Column number of pixel1	Row number of pixel2	Column number of pixel2

Figure1. Coding of one chromosome

3.2.2. Population:

Set of chromosomes form the population. Information of set of pixels in a sub image is taken as a chromosome. Initial population is randomly selected. Three genetic operators with elitism technique are realized. Genetic operators Figure 2 gives flow chart of procedure to obtain spot using GA.

Selecting number of copies of an existing chromosome is done by Roulette Wheel selection method to obtain parent population for the purpose of crossover in order to obtain next generation of chromosomes with better fitness. This process is repeated till the maximum number of iterations, giving the result as a pixel within a spot. Genetic operators like mutation and crossover are applied to get best results. Number of chromosomes in a population is maintained constant for all the generations.

After the crossover is performed, mutation takes place, which prevents falling all solutions in population into a local optimum. Mutation changes the new offspring randomly. In case of binary encoding, few randomly chosen bits are switched from 1 to 0 or from 0 to 1.

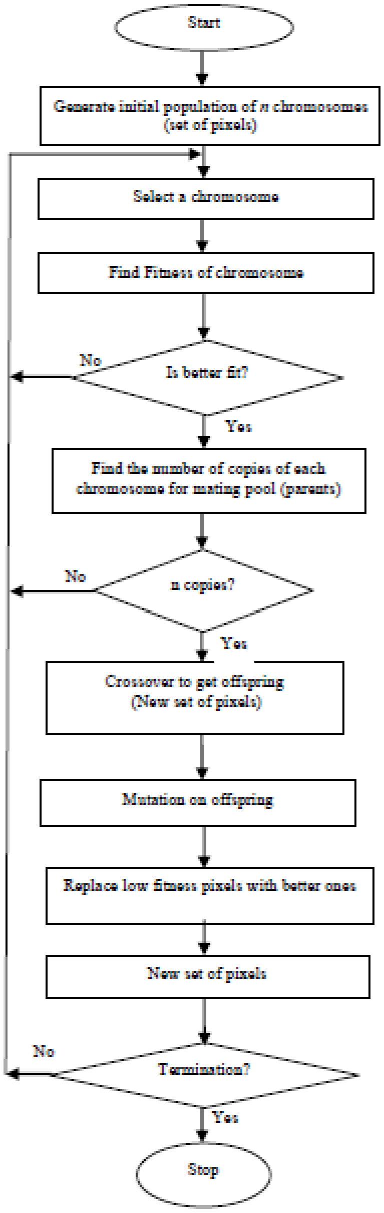

Start
Generate initial population of n chromosomes (set of pixels)
Select a chromosome
Find Fitness of chromosome
Is better fit?
No
Yes
Find the number of copies of each chromosome for mating pool (parents)
n copies?
No
Yes
Crossover to get offspring (New set of pixels)
Mutation on offspring
Replace low fitness pixels with better ones
New set of pixels
Termination?
No
Yes
Stop

Figure 2. Flowchart to identify spots (step2)

3.2.3 Objective function:

Objective function is combination of fitness function and constraint.

Fitness function:

The fitness f(i) of a particular chromosome which is derived for the solution of optimization problem, is defined by the equation (1)

$$f(i)=\sum_{j=1}^{k} ins(j) \qquad \text{-----} \qquad (1)$$

$$TotalFitness=\sum_{i=1}^{m} f(i) \qquad \text{-----} \qquad (2)$$

$$Probability(i)=\frac{f(i)}{TotalFitness} \qquad \text{-----} \qquad (3)$$

ins (j) = Intensity of j^{th} pixel in the chromosome
Total Fitness = Fitness of population
Probability (i) = Selection probability of an i^{th} chromosome

Constraint:

$d > d_{min}$ and $ins\ (j_l) \mathrel{!=} ins\ (j_m)$

Where

ins (j) = intensity of pixel
d = distance between two spots
d_{min} = minimum distance between spots

GA searches for best value of objective function Objective function with violated row and column values will be penalized to discourage the solution. Genetic Algorithm searches for optimal locations of a pixel belonging to a spot by maximizing the fitness function.

The values assigned for variables used in the GA method are given below. The termination of the program is decided by maximum number of generations and the size of the population gives the information of number of search points. The application of genetic operators is dependent on the respective probabilities. Scaling is used to obtain appropriate fitness value.

Maximum number of generations	=	60 to 100
Maximum population size	=	40
Crossover probability	=	0.08
Mutation probability	=	0.03
Scaling	=	1.1

3.2.4. Segmentation:

Individual spot is identified by a pixel, which is obtained by applying method developed, as described in step2. Once the complete image is processed by the GA method, all the spots are identified.

Segmentation process extracts the pixels representing the spot. The pixels belonging to a spot are considered as foreground. All these pixels in a spot have the same intensity in an ideal case.

To identify the foreground pixels of a spot, a specific local threshold value is selected depending on the intensity level of spot. Both minimum and maximum values of intensity are derived to distinguish between neighboring spots. The algorithm scans for the pixels lying between two threshold values and distinguishes between foreground and background. Once all the foreground pixels are identified mean or median of spot intensity is calculated.

Threshold values are calculated by following equations

Minimum = inty_p - 0.4*inty_p
Maximum = inty_p + 0.2*inty_p
Where inty_p is the intensity of the pixel obtained from GA.

3.2.5. To find the center and average intensity:

Mean intensity of each spot is calculated from the output of foreground estimation. Centre is determined from extreme rows and columns. Average spot intensities along with their respective centre co-ordinates are shown in figures 4, 5 and 6 for the different sizes of sub-images considered.

3.2.6. Removal of spurious spots:

The spots with very low intensity and small size are considered as spurious spots. The spurious spots are eliminated and the actual spots are considered.

3.2.7. Background estimation:

Intensity of the spots depends on signals from scanner devices originating due to fluorescent molecules attached to hybridize DNA, signals due to coating of glass and contamination in hybridization and washing processes. Even limitation of scanner bandwidth plays an important role in background estimation [12]. Hence measured spot intensity is true intensity of the spot plus its local background.

Basically there are two methods to estimate the background. In one of method histogram is used, in which range of pixel intensities in the image is obtained. Hence global background can be calculated. Where as in the other, pixels close to spot are assumed to be the local background. In the process of estimation of background, normally local background level is assumed to be same as that of the intensity of the pixels in the proximity of the corresponding spot.

In this approach the background of each spot is obtained from the foreground pixel data and the complete spot. The mean (or median) intensities for foreground and background are calculated. The actual intensity of the spot is the difference in foreground and background mean (or median) values.

4. RESULTS

In this paper a part of microarray image is considered for which the results are evaluated and presented. Figure 3 shows the part of DNA microarray image of size 200x90.

GA algorithm is implemented on sub image of size n x n to identify the spots. The process is repeated for entire image to extract all the spots. This algorithm is applied to images of 3 different sizes of the sub images and the results are discussed. Figures 4, 5 and 6 show the average intensity of spots plotted at their respective centre co-ordinates, for sub image of size 31x31, 15x15 and 7x7 respectively.

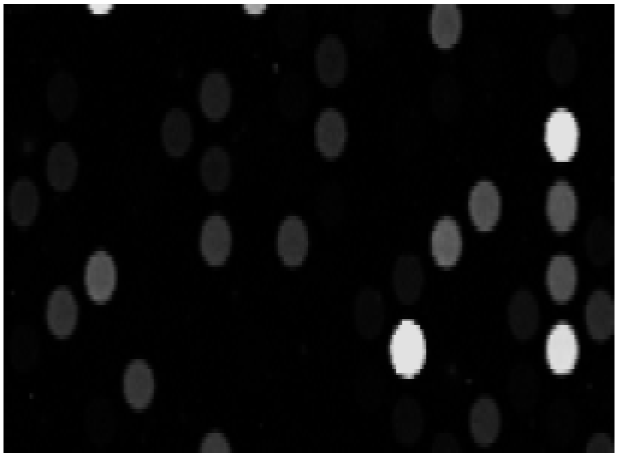


Figure 3. Part of microarray image
(Original image courtesy of Prof. Paturu Kondaiah,
Dept of M.R.D.G, IISc, Bangalore)

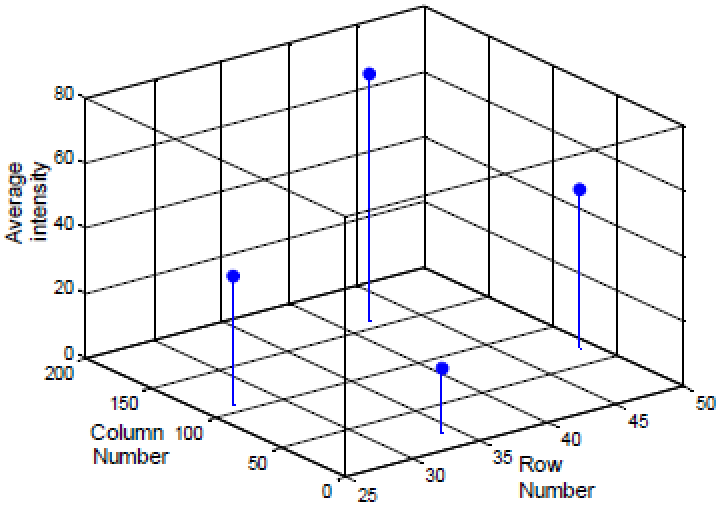


Figure .4. Average intensities of spots for sub image of size 31 x 31

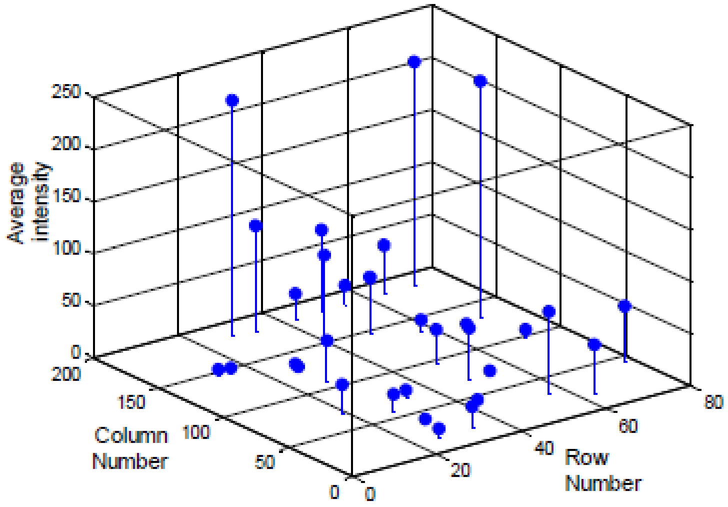


Figure 5. Average intensities of spots for sub image of size 15 x 15

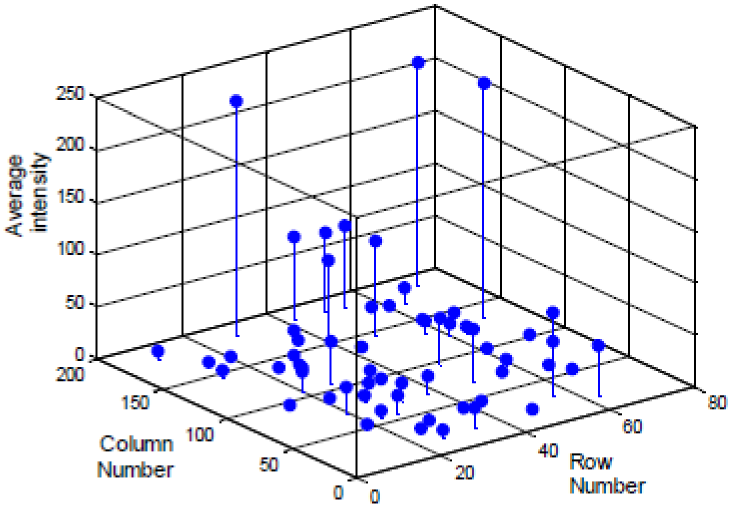


Figure 6. Average intensities of spots for sub image of size 7 x 7

Figure 7 shows the complete spot extracted by knowing the pixel features belonging to that spot from G.A. Once the spot is identified, foreground of the spot is estimated by implementing automatic segmentation technique. Figure 8 shows the foreground and figure 9 shows the corresponding background.

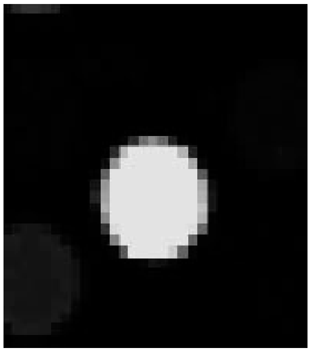


Figure 7. Complete spot

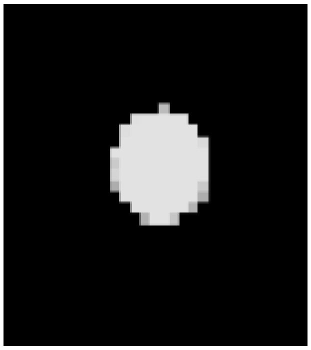


Figure 8. Foreground

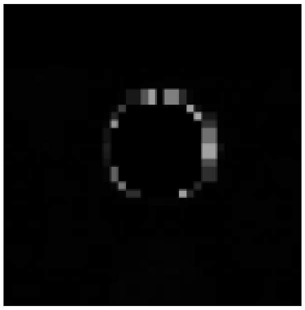


Figure 9. Background

5. CONCLUSION

GA with elitism technique is implemented on the cDNA Microarray image. Elitism removes pixels with low fitness from the off springs and is replaced by pixels with larger fitness (parents) to obtain next generation. This technique results in extraction of spots accurately. From the results it is evident that sub image of size 7x7 identifies all the spots but requires more time for processing. Appreciable results are obtained even with size 15x15. When sub images of size 31x31 are processed, only certain spots are realized as shown in Figure.4. Figures 7, 8 and 9 show one complete spot, its foreground and background respectively. The local background gives reliable results compared with global background.

REFERENCES

[1] Xin-Yuzhang, Fei Chen, Yuan-Ting Zhang, Shannon C. Agner, Metin Akay, Zu-Hong Lu, Mary Miu Yee Waye, and Stephen Kwok-Wing Tsui, “Signal Processing Techniques in Genomic Engineering”, Proceedings of IEEE, 0018-9219/02©2002 IEEE

[2] Mohammed Khabzaoui, Clarisse Dhanens, and El-Ghazali Talbi, “A Multicriteria Algorithm to analyze DNA Microarray data”, 0-7803-8515-2/04©2004 IEEE

[3] Peter Bajcsy, “Automatic grid alignment in DNA Microarray Scans”, IEEE Transactions on image processing, Vol 13, 1057-7149/04©2004 IEEE.

[4] Buhler J., T. Ideker , and D. Haynor, “dapple: Improved Techniques for finding spots on DNA Microarrays”, UV CSE Technical Report UWTR 2000-08-05.

[5] Scanalytics Inc, “Microarray Suite”, Product description at http://www.scanalytics.com/product/hts/microarray. html

[6] Peter Bajcsy, “An Over view of DNA Microarray Image requirements for Automated Processing”, Proceedings of the 2005 IEEE computer society conference on Computer vision and pattern Recognition.

[7] Christian Uehara and Ioannis Kakadiaris, “Towards Automatic Analysis of DNA Microarrays”, Proceedings of Sixth IEEE workshop on Applications of Computer Vision 2002.

[8] Luis Rueda and Vidya Vidyardharan, “A Hill Climbing Approach For Automatic Gridding Of cDNA Microarray Images”, IEEE/ACM transactions on computational Biology and Bioinformatics, Vol.3, No.1, January-March 2006.

[9] E.Zacharia and D. Maroulis, ”An Unsupervised and Fully- Automated image Analysis method for cDNA microarrays”, Twentieth IEEE International Symposium on Computer Based Medical Systems 2007.

[10] A.Sreedevi and D.S.Jangamashetti, "Automatically locating Spots in DNA microarray Image Using Genetic Algorithm without Gridding", IACSIT spring conference, 978-0-7695-3653-8/09 © 2009 IEEE .

[11] Anders Bengtsson and Henrik Bengtsson, "Microarray image analysis: background estimation using quantile and morphological filters", BMC Bioinformatics 2006, 7:96

AUTHORS

Dr. A.Sreedevi, born in Andhra Pradesh, India in 1961. She received her Electrical Engineering degree from Bangalore University during 1983, Her M.E degree from Pune University during1993 and Ph.D. from Visveswaraya Technological University, India. Since 1984 she is in teaching, presently working as Associate professor in the Department of Electrical and Electronics Engineering, RV College of Engineering, Bangalore. Her ares of research are signal and image processing.

Dr. D.S. Jangamshetti, born in Karnataka, India in 1964. She received Electrical Engineering degree from Karnatak University during 1985, M.Tech. degree from I.I.T. Khargpur, and Ph.D. from I.I.T. Bombay in 2003. She is presently working as professor in the Department of Electrical and Electronics Engineering, Basaveshwar Engineering Collage, Bagalkot, Karnataka.

She has several publications in IEEE Transactions. Her area of interest is signal processing. She has attended a number of international and national conferences. She has organized quite a few short term training programs for teachers and students. She is senior member of IEEE, member of B.M.E., I.S.T.E. and Institution of Engineers,India.