Buletinul Ştiinţific al Universităţii "Politehnica" din Timişoara

Seria ELECTRONICĂ şi TELECOMUNICAŢII

TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Tom 54(68), Fascicola 2, 2009

Microarray Image Segmentation Using Marked Point Processes

Barna Keresztes [1], * Bogdan Belean [2], † Monica Borda [2], Olivier Lavialle [1]

Abstract

This paper presents a new method dedicated to unsupervised segmentation of spots in cDNA type microarray images. The framework relies on a marked point process algorithm. We shall create random circular objects to fit the spot distribution in the image. The interaction rules between the objects complete the model.

Using a Markov Chain Monte Carlo (MCMC) method, the algorithm converges to a configuration which is close to the spot distribution in the images. At each step, the configuration is evaluated considering its proximity to the target distribution. In order to achieve this task, we propose a data model using a Gaussian gray level distribution and gradient detection to evaluate the likelihood of the current configuration.

Finally, the results on the microarray images illustrate the efficiency of the segmentation and suggest that the marked point processes can be a promising tool for spot detection.

1 Introduction

The most common technique used in molecular biology and medicine to measure the gene expression levels for thousands of genes in parallel is the cDNA (complementary DNA) microarray experiment. By gene expression we understand the transformation of genes information into proteins. The informational pathway in gene expression is as follows: DNA mRNA protein. The protein coding information is transmitted by an intermediate molecule called messenger ribonucleic acid. This molecule passes from nucleus to cytoplasm carrying the information to build up proteins. This mRNA acid is a single stranded molecule from the original DNA and is subject to degradation, so it is transformed into stable complementary DNA for further examination.

Specific single stranded cDNA probes are arrayed on a cDNA microarray glass slide or microchip. Usually, samples from two sources are labeled with two different fluorescent markers and hybridized on the same array (glass slide). By hybridization we understand the tendency of two single stranded DNA molecules to bind together. After the hybridization, the array is scanned using two light sources with different lengths (red and green) to determine the amount of labeled sample bound to each spot through hybridization process. The light sources induce fluorescence in the spots, which is captured by a scanner and a composite image is produced. The most common use of cDNA microarrays is for the determination of patterns of differential gene expression, comparing differences in mRNA expression levels between identical cells subjected to different stimuli or between different cellular phenotypes or development stages [2].

Further on, image processing techniques are used to quantify gene expression levels present in the captured microarray image, in order to identify the differential gene expression between normal and abnormal cells, labeled with the two different fluorescent markers. The flow of processing a microarray image [13] is generally separated in the following tasks: addressing, segmentation, intensity extrac-

*[1] IMS Bordeaux, *barna.keresztes@laps.ims-bordeaux.fr*
†[2] Universitatea Tehnică din Cluj-Napoca

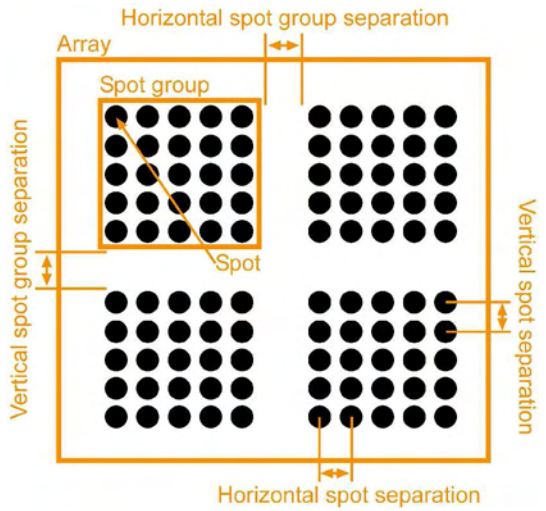

Figure 1: The structure of a microarray image

tion and preprocessing to improve image quality and enhance weakly expressed spots. The first step associates an address to each spot of the image. In the second one, pixels are classified either as foreground, representing the DNA spots, or as background. The last step calculates the intensities of each spot and also estimates background intensity values. The major tasks of microarray image processing, which contributes in fulfilling the last mentioned steps, are to identify the array format including the array layout, spot size and shape, spot intensities and distances between spots. The main parameters taken into consideration in image processing are accuracy and time. There is hardware implementation for spot detection based on vertical and horizontal projections. This paper proposes an original method for spot segmentation using high level approach which aims to model the spot distribution using circular objects.

It is time consuming to analyze in each pixel of the image space each possible circular spot, so a stochastic algorithm is proposed, specifically the spatial marked point process algorithm, in order to achieve a fast convergence towards an optimal distribution of the objects. The marked point processes were first used in image segmentation by Baddeley and Van Lieshout in [1].

The paper is organized as follows: In the next section we introduce the marked point process and our object model, in section 3 we discuss the bayesian interface of the process, and in section 4 the Monte Carlo chain used for the convergence of the process will be presented.

2 Marked point process

2.1 Notations

Let I be the actual image, $I = [0, W] \times [0, H]$ (the value of W and H is around 1500-2500 pixels). A configuration of objects in the image I will be noted Y. Using a marked point process X we try to approximate the observed configuration Y.

A marked point process $X = P \times K$ is a random configuration of points P in the image space, where a mark K is assigned to each point. This mark is a collection of parameters which define an object.

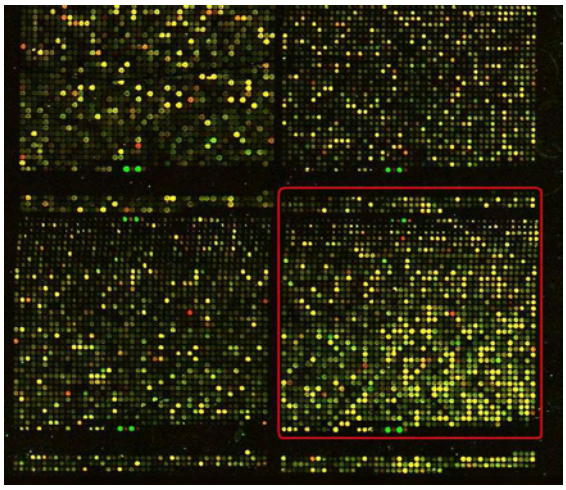

Figure 2: Scanned microarray image

2.2 The object model

As presented in the introduction, the genomic images contain luminous spot arrays representing gene samples arranged in groups (figure 2).

The spots are circular patches in the image, so we'll use a circle defined by the center P and radius R to describe an object. The object used to describe the spots will be a simple object with 3 parameters, the point process built to describe the system will be defined in a limited subset of $\mathbb{R}^3$.

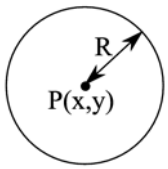

Figure 3: Object model

3 The probability density function of the process

Let $f(X)$ be the density of a configuration X of objects, in the given image I. According to Bayes formula, the expression of this density can be expressed as:

$$f(X) = f(X|I) \propto f_p(X) f(I|X) \tag{1}$$

The $f_p(X)$ contains all the a priori knowledge about the configuration, and the $f(I|X)$ the likelihood between the image and the current configuration; this will be further noted as $\mathcal{L}(I|X)$.

3.1 The *a priori* term

We can make some restrictions on the object configuration based on the a priori knowledge about the shape and distribution of the objects. The first restriction we can make is to limit the R (radius) parameter between $[R_{min}, R_{max}]$ where the two limits were determined experimentally the radius of the spots is fairly constant. The x and y parameters are limited by the size of the image: $[0, W]$ and $[0, H]$.

The a priori term can be described using the following formula:

$$f_p(X) \propto \alpha h(X) \tag{2}$$

The α function defines the probability density of the process. In this application the fibers are considered to have a homogenous Poisson distribution:

$$\alpha = \beta^{n}(X)$$

where β is the density parameter of the process and $n(x)$ represents the number of objects in the configuration.

The h function defines the interaction between the different objects. Since the spots cannot intersect, a repulsive Strauss pairwise interaction model [12] will be used, which penalizes the overlapping object configurations.

Two objects are overlapping if their silhouettes touch. This interaction will be noted $\sim_o$ and defined by:

$$x_i \sim_o x_j \Leftrightarrow S_i \cup S_j \neq \emptyset \tag{3}$$

The spots are arranged along a rectangular grid, in groups. The grid is rectangular, and in the same group the spots are equidistant. As the plates are scanned, the grid isn't perfectly horizontal, it has an inclination α.

If the distance between two objects are smaller than a given threshold d_c, they are considered close objects (relation $\sim_c$):

$$x_i \sim_c x_j \Leftrightarrow d_{i,j} < d_c \tag{4}$$

Two objects are neighbors ($\sim_n$)if the distance between them is equal to the grid step and the angle between them is equal to the inclination of the grid.

$$x_i \sim_n x_j \Leftrightarrow \begin{cases} \alpha_{i,j} = \alpha_{grid} \\ d_{i,j} = d_{grid} \end{cases} \quad \text{and} \tag{5}$$

We consider that the objects which are close to each other should be neighbors. The closeness relationship between two objects is a repulsive relationship, and as it gets closer to the neighboring relationship, its force decreases (no interaction force) using a Gaussian function. If the neighborhood relationship would be an attractive one, the probability of the inactive spots would increase, leading to their detection.

$$h_c(x_i, x_j) = \frac{1}{2\pi\sigma_d\sigma_\alpha} e^{-\frac{(d_{i,j}-d_{grid})^2}{2\sigma_d^2} - \frac{(\alpha_{i,j}-\alpha_{grid})^2}{2\sigma_\alpha^2}} \tag{6}$$

where $x_i \sim_c x_j$.

The value of the *a priori* density for a given configuration X is:

$$f_p(X) = \beta^{n(X)} \prod_{x_i \sim_o x_j} \gamma_o \prod_{x_i \sim_c x_j} h_c(i,j) \tag{7}$$

where γ_o is the repulsion force between the overlapping objects; $0 < \gamma_o < 1$.

The spot distance and the angle are recalculated at each step as the mean distance and angle between neighboring objects, and the variation of these parameters.

$$\begin{gathered} d_{grid} = mean(d_{i,j}) \quad \sigma_d = var(d_{i,j}) \\ \alpha_{grid} = mean(\alpha_{i,j}) \quad \sigma_\alpha = var(\alpha_{i,j}) \\ d_c = 1.5\, d_{grid} \end{gathered} \tag{8}$$

where x_i and x_j are neighboring objects.

3.2 The data term

In this step we have to determine the probability of the existence of a configuration based on the likelihood function $\mathcal{L}(I|X)$.

As a first approach we used the classical luminosity-based likelihood detection [10]. Two classes will be defined, the object class and the background class. A pixel belongs to the object class if it is a part of the silhouette of an object; otherwise it is considered background. The likelihood of a pixel with a class is determined using a Gaussian distribution function of the luminosities:

$$\mathcal{L}(p|X) = \frac{1}{\sqrt{2\pi}\sigma_\varphi} e^{-\frac{(y_p-\nu_\varphi)^2}{2\sigma_\varphi^2}} \tag{9}$$

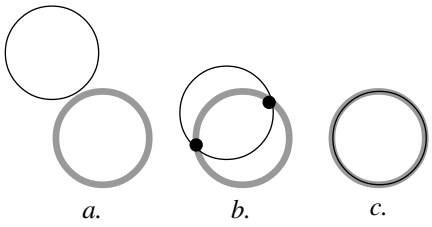

Figure 4: Gradient-object intersection: a. no intersection b. two points c. all points

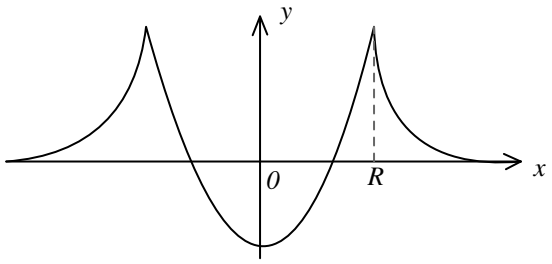

Figure 5: 2D section of the gradient estimation function

where φ denotes the class (object or background), μ_φ and σ_φ are the mean and variation of luminosities of the current class.

The likelihood of the image is the product of the pixel likelihood through the image:

$$\mathcal{L}(I|X) = \prod_p \mathcal{L}(p|X) \tag{10}$$

To get a more robust detection scheme, a gradient-based approach was considered, too [8]. This method will help the objects to reach the real boundaries of each spot. Using the marked point process it is difficult to define a convergent function toward the optimal solution.

In the gradient-based likelihood detection we can't determine the likelihood of the image conditioned by the current configuration of objects. The method used in this case is based on external fields energy, the probability of the individual objects is calculated and the final data term will be determined based on these values.

There are three cases of intersection between the contour of the object and the contour of a spot in the figure 4:

Based on the intersection points it is impossible to achieve a convergence towards the solution; therefore another function will be proposed with the following properties:

- its maxima are at the contour of the detected object (maximum likelihood)
- it is monotonically increasing on each side towards these values outside the object the values are all positives (we accept the possibility of fiber contours outside our contour)
- close to the center of the object the values are negative (we penalize the inside contours)

There are a lot of functions which fulfill these conditions. We can use for example a truncated gaussian function, a paraboloid combined with an exponential function, a two dimensional function rotated around its axis, a.s.o.

For example, $g(x, y)$ can be described for a circular object in the following way:

$$g(x, y) = -abs(th - e^{\frac{x^2+y^2}{2\sigma^2}}) + th \tag{11}$$

where th is a truncation threshold and σ a fixed parameter. The function is scaled to match the size of the object.

The likelihood of an object p will be the correlation between the function g and the gradient image:

$$\mathcal{L}(x) = \int_{-n}^{n}\int_{-m}^{m} g(u+x, v+y) I_G(u+x, v+y) dxdy \tag{12}$$

The final likelihood value used to determine the validity of the object is determined using the likelihood values obtained fusing the the grayscale values and the gradients in the image [7]:

$$\mathcal{L}(X) = \mathcal{L}_1(I|X) \star \prod_{x \in X} \mathcal{L}_2(x) \tag{13}$$

where $\star$ is a fuzzy fusion operator, such as the symmetrical sum.

4 The MCMC simulation

After the model has been defined, an algorithm that assures the convergence of the process towards the minimal energy of the system is created. Here, the energy is related to the density of a point process, so the optimal configuration is the one that maximizes this density.

$$X_{MAP} = \underset{X}{\operatorname{argmax}}\, f(X) \tag{14}$$

In the case of marked point processes the most common method for this is the Monte Carlo Markov chain (MCMC) coupled with simulated annealing.

To simulate the MCMC, tested two algorithms: the Metropolis-Hastings-Green (MHG) algorithm

1. given the configuration x_t, we generate y using the translation kernel q.
2. we calculate the ratio between the probability of the current configuration and the proposed one:
$$r = \frac{f(y)q(y, x_t)}{f(x)q(x_t, y)}$$
3. with the probability $\alpha = min(1, r)$ we accept $x_{t+1} = y$

Algorithm 1: Metropolis-Hastings-Green algorithm

1. Generate a random number of new objects
2. Sort all the objects based on their probability
3. Propose to delete each object calculating the acceptance ratio of the transition and accepting the new configuration with the probability $\alpha = min(1, r)$

Algorithm 2: Birth and death algorithm

[6], which was adapted by Geyer and Moller to point processes [5] and the birth and death algorithm [11].

The MHG algorithm consists in proposing a new, random state y for the current state x_t. The transition kernel, noted with $q(\cdot, \cdot)$ consists in some allowed movements between the two states. The allowed transitions are:

- birth (adding an object to the configuration)
- death (deleting an object)
- birth and death of a neighboring object
- translation
- dilation

The algorithm can be described in the following way:

The initial configuration x_0 is considered the empty configuration. The third step of the MHG algorithm ensures that the chain wont be struck in a local minimum of energy. The disadvantage of this approach is that the process will take a longer time to converge towards the maximum a posteriori configuration.

To optimize the chain, a simulated annealing is introduced; $f(X)$ term will be replaced by $f^{1/T}(X)$, where T is the temperature of the system, and it is a parameter with a decreasing value towards 0.

Another disadvantage of the algorithm is that the birth of a high number of objects require a high number of steps in the MCMC chain. To optimize the convergence, a continuous time algorithm is considered: the birth and death algorithm. This algorithm has only two transition kernels: adding or deleting an object. As the algorithm is a continuous time algorithm, these transitions occur at a given random moment, with a Poisson distribution in time.

To have a fast convergence, the birth and death algorithm is sampled using a large step size [4], at each step an important number of objects being added and deleted from the configuration:

In order to further optimize the process, a birth map is precalculated based on the greyscale values in the microarray image: applying a gaussian blur and considering the resulting image as the probability map that an object in born in each point, the probability of correct births will be much higher.

The simulated annealing is used in the case of the birth and death algorithm, too. The number of objects to be added doesn't change, but the acceptance ratio is getting more selective as the temperature of the system decreases.

5 Results

We tested the model several microarray images, with a resolution of between 1 megapixels (1000*1000) and 5 megapixels, 8 bits/pixel, greyscale. The results are shown in figure 6.

The Markov chain used to simulate the marked point process converges in around 500 000 steps using the MHG algorithm and around 20 000 steps using the birth and death method; the birth and death algorithm was around three times faster than the MHG algorithm.

The results show a good estimation of spot positions, without the necessity of user intervention which is needed by existing platforms to confirm the spot positioning. Also the gradient-based approach and the inclusion of the a priori knowledge minimized the number of false detections.

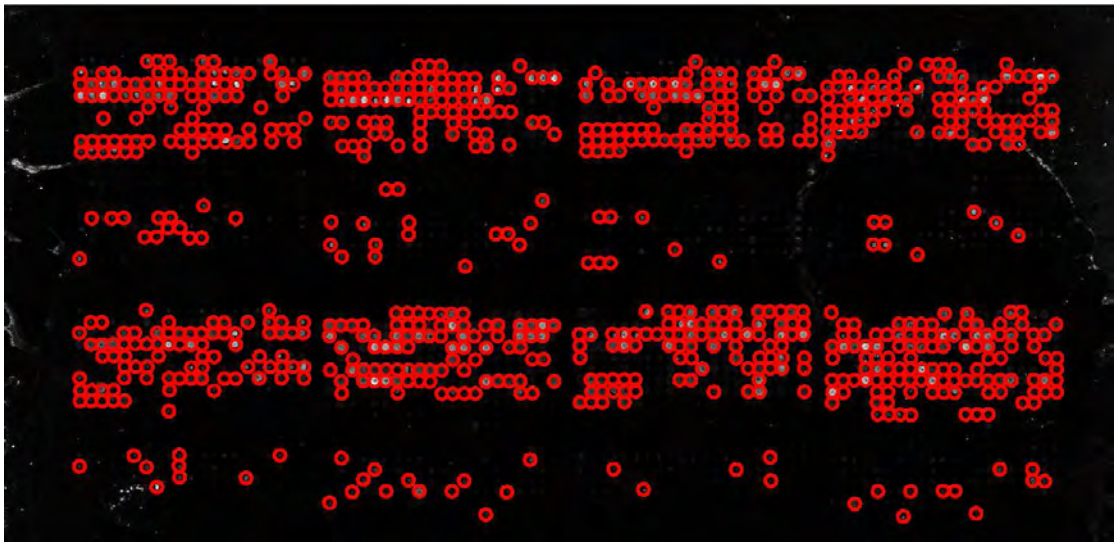

Figure 6: Segmented microarray image

6 Conclusions and perspectives

In this paper we presented a new kind of approach for microarray image processing and spot detection using an object-based model based on the marked point process method.

We created some novel methods to describe the object interaction in the image and the likelihood function, as the existing applications using marked point processes use only a simple approach using the pixel luminosity or homogeneity of the object silhouette. We proposed a new approach based on the image gradients, and a new decision system was created to determine the data term of a configuration.

Two different algorithms were used for the simulation of the Markov chain, one discrete and a continuous time algorithm, and the performance of the new algorithm was evaluated compared to the more commonly used MHG algorithm in the case of the detection of a large number of objects.

One of the main drawbacks of this method is the computing time which is much higher the methods using specialized hardware for spot detection [3]. To optimize the process, a point process detecting the spot groups instead the individual spots can be constructed. In this case the object can be a rectangular shape [9]. As the number of objects to detect is much lower in this case, this process will converge much faster. The informations about the spot groups, as the position and angle can be used later as *a priori* information for the spot detection algorithms based either on the point processes, or other non-stochastic methods.

Acknowledgments: This project was founded by National Grant CNCSIS PN2 IDEI no. 909/2007

References

[1] A. Baddeley and M.N.M. van Lieshout. Stochastic geometry models in high-level vision. *Statistics and Images*, 1993.

[2] P. Bajcsy. An overview of dna microarray image requirements for automated processing. *IEEE Transactions on Image Processing*, 2004.

[3] B. Belean, M. Borda, and A Fazakas. Adaptive microarray image acquisition system and microarray image processing using fpga technology. *KES2008*, 2008.

[4] X. Descombes, R. Minlos, and E. Zhizhina. Object extraction using a stochastic birth-and-death dynamics in continuum. *Journal of Mathematical Imaging and Vision*, 33:347–359, 2009.

[5] C.J. Geyer and Moller J. Simulation and likelihood inference for spatial point process. *Scandinavian Journal of Statistics*, 1994.

[6] P.J. Green. Reversible jump mcmc computation and bayesian model determination. *Biometrika*, 1995.

[7] B. Keresztes, O. Lavialle, S. Pop, and M. Borda. Seismic fault detection based on a curvilinear support. In *IEEE IGARSS*.

[8] B. Keresztes, O. Lavialle, S. Pop, and M. Borda. Fiber segmentation in composite materials using marked point processes. *Acta Tehnica Napocensis*, 2009.

[9] M. Ortner, X. Descombes, and J. Zerubia. A marked point process of rectangles and segments for automatic analysis of digital elevation models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2008.

[10] G. Perrin, X. Descombes, and J. Zerubia. Tree crown extraction using marked point processes. In *EUSIPCO*.

[11] B.D. Ripley and P.F. Kelly. Markov point processes. *Journal of the London Mathematical Society*, 1977.

[12] D.J. Strauss. A model for clustering. *Biometrika*, 1975.

[13] Y. Yang, M. Buckley, S. Dudoit, and T. Speed. Comparison of methods for image analysis on cdna microarray data. Technical report, Department of statistics - University of California Berkeley, 2001.