

MORPHOLOGICAL MODEL-BASED MICROARRAY SPOT CLASSIFICATION AND SEGMENTATION IN POLAR COORDINATES

JESÚS ANGULO

Centre de Morphologie Mathématique, Mines de Paris, 35 rue Saint Honoré, 77300 Fontainebleau, France
e-mail: jesus.angulo@ensmp.fr
(Submitted)

ABSTRACT

Robust image analysis of spots in microarrays (quality control + spot segmentation + quantification) is a requirement for automated software which is of fundamental importance for a high-throughput analysis of genomics microarray-based data. This paper deals with the development of model-based image processing algorithms for qualifying/segmenting/quantifying adaptively each spot according its morphology. A series of morphological models for the spot intensities are introduced. The spot categories represent most of possible qualitative cases identified from a large database (different routines, techniques, etc.). Then based on these spots models, a classification framework has been developed. The spot feature extraction and classification (without segmenting) is based on converting the spot image to polar coordinates and, after computing the radial/angular projections, the calculation of granulometric curves and derived parameters from the projections. Spot contour segmentation can be solved by working in polar coordinates, and then calculating the up/down minimal path, easily obtained with the generalized distance function. With this model-based technique, the segmentation can be regularized by controlling different elements of the algorithm. According to the spot typology (e.g., doughnut-like or egg-like spots), several minimal paths can be computed to obtain a multi-region segmentation. Moreover this segmentation is more robust and sensible to weak spots, improving the previous approaches.

Keywords: mathematical morphology, cDNA microarray image, polar coordinates, spot modelling, spot segmentation, shortest path segmentation.

INTRODUCTION

DNA microarrays are an experimental biotechnology to identify and quantify levels of gene expression. The method consists in arrays of thousands of discrete DNA sequences (genes) that are printed as spots on a support. The aim is to compare the relative abundance of each of these gene sequences in two DNA samples (Brown and Botstein, 1999). Spot finding and signal intensity determination are performed with the help of image analysis software. Successful work on spot location and segmentation has already been done during the last years (Chen *et al.*, 1997) (Steinfath *et al.*, 2001) (Yang *et al.*, 2002). We have previously proposed an automatic spot segmentation based on advanced morphological operators (Angulo and Serra, 2003). This inner marker (spot center) and outer marker (bounding box from rectangular grid) watershed-based segmentation yields satisfactory results for “normal” spots. However, it is observed, on the one hand, problems of segmentation for low intensity spots or for spots on strong noisy background; and on the other hand, difficulties to define a right segmentation/quantification for structured spots (e.g., doughnut-like and egg-like spots). In addition, several typologies of abnormal or irregular spots can be related to different problems

of preparation of microarrays and consequently, a qualitative automatic evaluation of spots can be of help for flagging the suspect spots necessary for data analysis.

This paper is organised into two parts and it deals with the development of model-based image processing algorithms for qualifying/ segmenting/ quantifying adaptively each spot according its morphology. In a first part, we focuss on the morphological modelling and automated classifying of spots according to different typologies. Several models have been suggested for spot intensity distribution, including specially statistical models: a stochastic/geometric model (Balagurunathan and Dougherty, 2002), a scaled bivariate Gaussian density function (Steinfath *et al.*, 2001), a difference of two Gaussian densities or a cylinder (Wierling *et al.*, 2002), a polynomial-hyperbolic model (Ekstrom *et al.*, 2004), linear models based on PCA (Glasbey and Khondoker *et al.*, 2005). These models are typically used for image simulation or for estimation of model parameters. We prefer here to propose a morphological model with spot categories which represent most of possible qualitative cases identified from a large database (different routines, techniques, etc.). Then based on these spots models, a classification

framework has been developed. The spot feature extraction and classification (without segmenting) is based on converting the spot image to polar coordinates, and after computing the radial/angular projections, calculating granulometric curves and derived parameters from the projections.

Furthermore, spot segmentation can be approached in a more flexible and understandable way when working in polar coordinates. But the same weaknesses of the watershed on the low or noisy gradients are still underlying. The spot contour in polar coordinates is equivalent to calculating the left/right markers watershed-based transformation. This well-posed problem of segmentation can be also solved by calculating the up/down minimal path (easily obtained with the generalized distance function). The aim of the second part of the paper is just to introduce an innovative model-based spot segmentation according to this paradigm, where the type of segmentation is adapted to the spot typology. Several issues must be addressed, mainly the way for “filtering” the image on which the distance is computed and the manner to obtain a “circular” segmentation (circular shortest path). The shortest path segmentation can be “regularised” by controlling different elements of the algorithm. The segmentation of microarray spots in polar coordinates has been also addressed by (Appleton and Talbot, 2005), as an example of application of globally optimal geodesic active contours, but without considering the different typologies of spots. Another recent work has proposed a model-based spot segmentation by means of clustering algorithms (Li *et al.*, 2005).

Notation and basic definitions: In the framework of digital grids, a grey tone image associated to a scanned microarray can be represented by a function $f(\mathbf{x}) : E \rightarrow \mathcal{T} = \{t_{min}, t_{min} + 1, \dots, t_{max}\}$, where E is a discrete space ($E \subset \mathbb{Z}^2$), domain of definition of the function f , and \mathcal{T} is an ordered set of discrete grey-levels, i.e. a subset of \mathbb{Z} . Typically, $t_{min} = 0$ and $t_{max} = 2^{16} - 1 = 65535$ for a 16-bits image file. $f(\mathbf{x})$ is the intensity value of the image at point $\mathbf{x} = (x, y)$. The spots are structures placed on the microarray image. Let the image zone $Z_i \subset E$ be defined as the cell influence region (or bounding box region since the spots are usually place into an orthogonal array structure) around the spot i , i.e., pixels of the zone where their distance to the center of spot i is lower than the distance to the other spot centers. Ideally, we can suppose that $Z_i \cap Z_j = \emptyset, \forall i, j \setminus i \neq j$ (i.e., overlapping between neighbouring spots is not possible). The image signal intensity in the cell associated to the spot i at pixel position \mathbf{x} is denoted by $f_i(\mathbf{x}) : Z_i \rightarrow \mathcal{T}$, where obviously $f_i(\mathbf{x}) = f(\mathbf{x})$, that is, the function f_i

is a restriction of the function f to the set of support Z_i . In order to consider individually each spot but establishing spot models, we refer by $s_i(\mathbf{x} - \mathbf{x}_i^c) = f_i(\mathbf{x})$ the function $s_i(\mathbf{y}), \mathbf{y} \in E$, translated at \mathbf{x}_i^c , the central point of the spot i .

The polar transformation converts the cartesian image function $f(x, y) : E \rightarrow \mathcal{T}$ into another polar image function $f^\circ(\rho, \theta) : E_{\rho, \theta} \rightarrow \mathcal{T}$, where the angular coordinates are placed on the vertical axis and the radial coordinates are placed on the horizontal one. More precisely, with respect to a central point $\mathbf{x}^c = (x^c, y^c)$: $\rho = \sqrt{(x - x^c)^2 + (y - y^c)^2}$; $0 \leq \rho \leq R$; $\theta = \arctan\left(\frac{y - y^c}{x - x^c}\right)$; $0 \leq \theta < 2\pi$. The support is the space $E_{\rho, \theta}, (\rho, \theta) \in (\mathbb{Z} \times \mathbb{Z}_p)$ (discrete period of p pixels equivalent to 2π). A relation is established where the points at the top of the image ($\theta = 0$) are neighbors to the ones on the bottom ($\theta = p - 1$). The application of mathematical operators to images in (log-)polar coordinates has been recently studied by Luengo-Oroz *et al.* (2005).

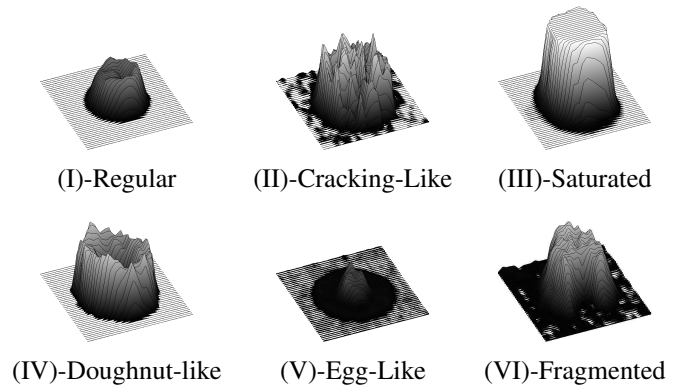


Figure 1. Examples of spot categories.

MODELS FOR SPOTS IN CDNA MICROARRAY IMAGES

Based on empirical observations of spots, we consider that the image intensity distribution for any spot i is given by the following expression:

$$f_i(\mathbf{x}) = a_i s_i(\mathbf{x} - \mathbf{x}_i^c) + n_i(\mathbf{x}) \quad (1)$$

where $s_i(\mathbf{y})$ corresponds to the morphological shape distribution for the spot i . For our purposes of classification and segmentation, we assume that s_i is represented by a cylindrical model. Moreover, a_i is the height of the “cylindrical” peak for the spot i , \mathbf{x}_i^c are the coordinates of the center position of the peak for the spot i , and $n_i(\mathbf{x})$ is a function that describes the noise.

Background noise

Two different sources of background noise can be distinguished:

$$n_i(\mathbf{x}) = n^g(\mathbf{x}) + n_i^l(\mathbf{x}) \quad (2)$$

$n^g(\mathbf{x})$ is the global background at point \mathbf{x} : typically, this function can be described by a randomly Gaussian distributed noise for the whole image, i.e. $n^g \sim N(\mu_n, \sigma_n^2)$. This part of the noise can be considered as associated to the acquisition system (photon-electronic scanner, CCD camera, etc.)

$n_i^l(\mathbf{x})$ is the local background noise (regionalised variable). It can be associated to different local phenomena: inhomogenous illumination, artefacts and inhomogeneities on the surface of support, errors in the preparation, etc.

Morphological spot typologies

The intensity distribution for the spot i is a cylindrical peak with a variable radius and height:

$$s_i(\mathbf{y}) = r_i(\theta)t_i(\mathbf{y}) \quad (3)$$

$r_i(\theta)$ is a ‘‘shape’’ function in polar coordinates describing the contour of the spot i . It defines a closed boundary such that

$$s_i(\mathbf{y}) = \begin{cases} t_i(\mathbf{y}) & \text{if } \|\mathbf{x} - \mathbf{x}_i^c\| \leq r_i(\theta) \\ 0 & \text{if } \|\mathbf{x} - \mathbf{x}_i^c\| > r_i(\theta) \end{cases}$$

$t_i(\mathbf{y})$ is a ‘‘texture’’ function, that is, any spatial variable (more or less regular) function of intensity. Note that this structural variation of intensity at point \mathbf{x} , associated to the spot (biochemistry, hybridisation, washing and fixing, etc.), is different from the background noise.

According to the particular distributions of $r_i(\theta)$ and $t_i(\mathbf{y})$ in this model, it is possible to identify six types of spots, see the examples of Figure 1.

(I) Regular spot: In the case of a typical regular spot, the cDNA deposition on the spot is considered to be circular with an homogenous intensity distribution. The radius can be modeled by a normal distribution having mean μ_r and variance σ_r^2 : $r \sim N(\mu_r, \sigma_r^2)$. Typically, the radius mean is random over a small range within the array and it can be considered as a uniform distribution, $\sigma_r \sim U(r_{min}, r_{max})$. The global variation of intensity, $a_i t_i(\mathbf{y})$ can be modeled as a normal distribution function, where the texture is a normal distribution with mean $\mu_t = 1$ and variance σ_t^2 : $t \sim N(1, \sigma_t^2)$. The coefficient a_i is considered as the ground truth expression signal, modeled as another uniform distribution, $a \sim U(t_{min}, t_{max})$.

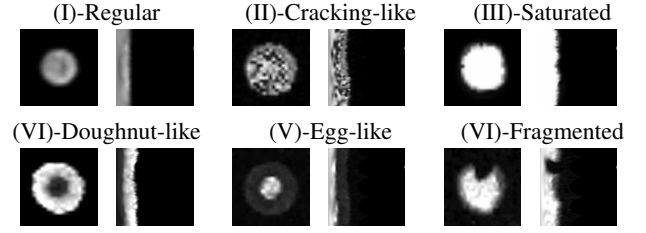


Figure 2. Examples of spot categories in cartesian and polar coordinates.

(II) Cracking-like spot: The spot has an aspect of cracked or ripped intensity: some dark tortous lines or strips cross the spot surface. These zones typically result in low intensities levels. The radius shape function for $r_i(\theta)$ has the same normal distribution as for a typical spot. The texture function can be given by the equation $t_i(\mathbf{y}) = \tilde{t}_i(\mathbf{y}) - \chi_i(\mathbf{y})$, where $\tilde{t}_i(\mathbf{y})$ has the same model as the typical spot and where the cracking function $\chi_i(\mathbf{y}) > 0$ if $\mathbf{y} \in \text{Crack Zone}$. The distribution of $\chi_i(\mathbf{y})$, the morphology of the strips (number, length, etc.) and spatial position are difficult to modeled but typically, the strip thickness is significant smaller than the spot radius r .

(III) Saturated spot: The fluorescence saturated spots are characterised by a saturated intensity, i.e., $a_i = t_{max}$, with no variation of texture $t_i(\mathbf{y}) = 1$, and a circular shape, i.e., $r_i(\theta)$ has the same normal distribution as for a typical spot.

(IV) Doughnut-like spot: The spot presents a circular ‘‘hole’’ in its center. The intensity distribution is the combination of two radial-defined texture functions:

$$t_i(\mathbf{y}) = \begin{cases} t_i^{low}(\mathbf{y}) & \text{if } \|\mathbf{x} - \mathbf{x}_i^c\| \leq r_i^{in}(\theta) \\ t_i^{high}(\mathbf{y}) & \text{if } r_i^{in}(\theta) \leq \|\mathbf{x} - \mathbf{x}_i^c\| \leq r_i^{ou}(\theta) \end{cases}$$

where $t_i^{low}(\mathbf{y})$ and $t_i^{high}(\mathbf{y})$ are the texture functions associated to the central part and to the peripheral part respectively; and $r_i^{in}(\theta)$ and $r_i^{ou}(\theta)$ are the radius functions of the center and of the spot contour respectively. We suppose that the inner and outer radius shape functions $r_i^{in}(\theta)$ and $r_i^{ou}(\theta)$ have the same normal distribution as for a typical spot (with mean μ_r^{in} and μ_r^{ou}). In a similar way, the texture functions $t_i^{low}(\mathbf{y})$ and $t_i^{high}(\mathbf{y})$ have a normal distribution. Moreover, usually, the mean for $t_i^{low}(\mathbf{y})$ tends to 0 and the mean for $t_i^{high}(\mathbf{y})$ tends to 1.

(V) Egg-like spot: Dual to the precedent, this spot has also two superposed intensity levels. More precisely, a circular peak of intensity t_i^{high} centered at

position \mathbf{x}_i^{ci} (but not necessary with $\mathbf{x}_i^{ci} = \mathbf{x}_i^c$) which is added to a pedestal of intensity t_i^{low} , i.e.,

$$t_i(\mathbf{y}) = \begin{cases} t_i^{high}(\mathbf{y}) & \text{if } \|\mathbf{x} - \mathbf{x}_i^{ci}\| \leq r_i^{in}(\theta) \\ t_i^{low}(\mathbf{y}) & \text{if } (r_i^{in}(\theta) \leq \|\mathbf{x} - \mathbf{x}_i^{ci}\|) \\ & \text{and } (\|\mathbf{x} - \mathbf{x}_i^c\| \leq r_i^{ou}(\theta)) \end{cases}$$

The inner and outer radius shape functions $r_i^{in}(\theta)$ and $r_i^{ou}(\theta)$, and the texture functions $t_k^{low}(\mathbf{x})$ and $t_k^{high}(\mathbf{x})$ have typically normal distributions, where here typically $\mu_{i^{high}}$ tends to 1 and $\mu_{i^{low}} > 0$.

(VI) Fragmented spot: A fragmented spot is characterised by a degenerated or irregular shape function $r_i(\theta)$, having also a size (surface area) lower than the typical spot within the array. The standard deviation σ_r is relatively important with respect to the mean. The texture function $t_i(\mathbf{y})$ can be still modeled as a normal distribution.

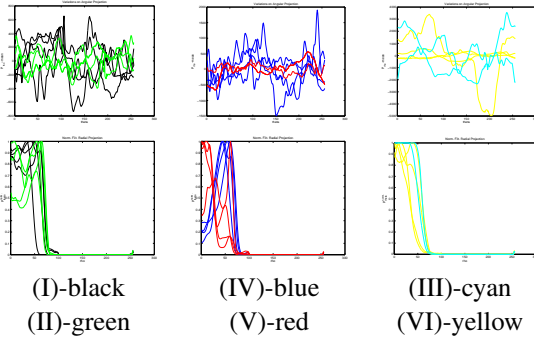


Figure 3. Angular projections $P_\rho(\theta)(f_i^\circ)$ (top row) and radial projections $P_\theta(\rho)(f_i^\circ)$ (down row) for a selection of representative spots of each typology.

MODEL-BASED SPOT CLASSIFICATION

Based on the spot model, we have developed a classification framework for the different spot typologies. The algorithms for feature extraction and classification must be simple and fast: each spot should be individually processed and typical microarrays have thousands of spots. The parameters and the typology will be used to improve and to make more robust the result of segmentation/quantification.

Spots in polar coordinates: According to the models proposed, the polar representation seems to be appropriate to characterise the different spot distributions. Let $f_i^\circ(\rho, \theta)$ be the image polar representation of spot i . Figure 2 gives an example of spot for each typology. We have compared with

the log-polar representation and verified that is more interesting to work on polar images for texture analysis.

Angular and radial projections: The horizontal and vertical projections of image $f_i^\circ(\rho, \theta)$ are then used to describe the spot structures: angular projection $P_\rho(\theta)(f_i^\circ) = \sum_{\rho=0}^R f_i^\circ(\rho, \theta)$ and radial projection $P_\theta(\rho)(f_i^\circ) = \sum_{\theta=0}^{\rho-1} f_i^\circ(\rho, \theta)$. Figure 3 provides the projections $P_\rho(\theta)$ and $P_\theta(\rho)$ for a selection of spots from each typology.

From the analysis of $P_\rho(\theta)$ using Fourier descriptors or morphological parameters (Angulo, 2005), we state that its variation combines the contributions of the background and the spot, including the texture and the shape irregularities. Consequently $P_\rho(\theta)$ is very poor to discriminate the spot categories. As we show below, $P_\theta(\rho)$ is more useful for spot classification.

Morphological filtering of $P_\theta(\rho)$: We start by extracting the background contribution using the top-hat transformation followed by a normalisation, i.e. $P_\theta^*(\rho) = P_\theta(\rho) - \gamma_n(P_\theta(\rho))$ and $\bar{P}_\theta(\rho) = P_\theta^*(\rho) / \max_{P_\theta^*(\rho)}$. The value $\sigma^\dagger = \sum_{\rho=0}^R \gamma_n(P_\theta(\rho)) / \sum_{\rho=0}^R P_\theta(\rho)$ gives an estimate of the regional background. Finally, a pre-filtering step is necessary in order to remove the insignificant extrema, i.e. $\bar{P}_\theta^h(\rho) = \varphi^{rec}(\bar{P}_\theta(\rho) + h; \gamma^{rec}(\bar{P}_\theta(\rho) - h; \bar{P}_\theta(\rho)))$ where typically $h = 2\%$ to 5% . We can now compute several parameters from the processed curves $\bar{P}_\theta^h(\rho)$ such as: an approximation of spot radius, the value for $\rho = 0$, std. dev., the percentage of points equal to 1, etc. which allow detection the main typologies.

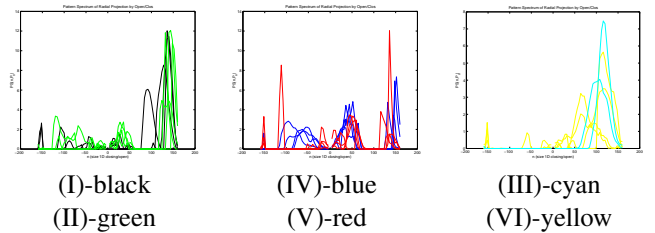


Figure 4. Pattern spectra of radial projection, $PS(n_\rho, P_\theta(\rho))$, for a selection of representative spots of each typology.

Furthermore, the variation of $P_\theta(\rho)$ can be analysed by means of 1D granulometries or pattern spectra. A granulometry is a family of openings of increasing size $\{\gamma_n\}_{n \geq 0}$ and the pattern spectrum of f is the following mapping $PS_\gamma(f, n) = \frac{m(\gamma_n(f)) - m(\gamma_{n+1}(f))}{m(f)}$, $n \geq 0$ and where $m(g)$ is the

integral of g . A dual definition $PS_\varphi(f, -n)$ is associated to a family of closings and then both curves are represented together $\{-n, 0, n\} \rightarrow PS(f, n) = \{PS_\varphi(f, -n), 0, PS_\gamma(f, n)\}$. Note that the computation of these 1D openings/closings is very fast. In Figure 4 are shown the corresponding pattern spectra for the selection of spots. The parameters computed from $PS(n_\rho, P_\theta(\rho))$ (moments, partial sums, significant points, etc.) combined with those obtained directly from $P_\theta(\rho)$ allow a spot classification into the different categories considered and without needing the spot segmentation, more details in (Angulo, 2005).

GENERALIZED DISTANCE GLOBAL MINIMAL PATH ALGORITHMS

Limitations of watershed transformation for detecting lines: According to the analysis by (Vincent, 1998), extracting a continuous track (=“crest-line”) going from the top to the bottom of the image by means of a constrained watershed using as markers the right and left sides of the image presents several limitations: (1) it fails when SNR is low (= sensitivity of watershed line to noise); (2) the watershed between two markers A and B depends on the position of the saddle points (for all the paths joining A to B with minimal elevation, the highest pixels along those paths are the saddle points) between the markers, and their location is one of the main factors determining the location of the line; (3) the criteria used to build the watershed are based on grey levels, and the length of watershed lines is irrelevant. Length constraints can be introduced in the segmentation by using global minimal paths algorithms. This approach is also useful to detect “disconnected” crest-line between two markers.

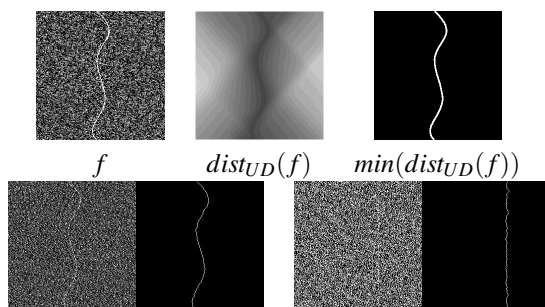


Figure 5. Top, generalised distance function and global minimal paths. Bottom, two examples of GMP detection.

Generalised distance function, GDF: The algorithm is based on a modification of the classic two-pass sequential distance function algorithm

of Rosenfeld and Pfaltz (1968) so that: (1) edge cost is taken into account; (2) raster and anti-raster scans are iterated until stability. Let us to denote by $N^+(p)$ (resp., $N^-(p)$) the neighbors of pixel p scanned before p (resp., after p) in a raster scan, for a 8-connected grid (neighborhood graph). In this graph, to each edge between two neighboring pixels p and q of an image f one associates the cost value $C_f(p, q) = f(p) + f(q)$ (or any other increasing function, such as $\max(f(p), f(q))$ or $\min(f(p), f(q))$). More specifically, the algorithm of GDF to set X in image f proceeds as follows,

- Initialise result image d : $d(p) = 0$ if $p \in X$ and $d(p) = +\infty$ otherwise;
- Iterate until stability:
 - Scan image in raster order \rightarrow For each pixel p , do: $d(p) \leftarrow \min\{d(p), \min\{d(q) + C_f(p, q), q \in N^+(p)\}\}$
 - Scan image in anti-raster order \rightarrow For each pixel p , do: $d(p) \leftarrow \min\{d(p), \min\{d(q) + C_f(p, q), q \in N^-(p)\}\}$

The algorithm typically converges in two or three iterations (relatively efficient).

Global minimal paths, GMP: Each path P in the 8-connect graph has associated a cost $C_f(P)$, equal to the sum of the cost of its successive edges. We can now define the distance $d_f(p, q)$ between two pixels p and q in the image f as: $d_f(p, q) = \min\{C_f(P), P \text{ path between } p \text{ and } q\}$.

For the simple problem of finding a path of minimal cost (or global minimal path, GMP) going from the top row U to the bottom row D of the image, we use the following result: a pixel p belongs to such minimal path if and only if $d_f(p, U) + d_f(p, D) = d_f(U, D)$. This is the approach introduced by (Vincent, 1998). To extract such Up/Down GMP in image f , we can therefore proceed as follows:

- Compute GDF to set U in image f : for each pixel p , compute $d_f(p, U)$;
- Compute GDF to set D in image f : $d_f(p, D)$;
- Sum these two distance functions, $d_f(U, D)(p) = d_f(p, U) + d_f(p, D)$;
- Find u_{min} , the minimal value of $d_f(U, D)$ and threshold the result in order to keep only the pixels whose values in $d_f(U, D)$ is equal to u_{min}

Since the extracted minimal paths are preferentially located on dark pixels (i.e., have low cost), the original image with the bright track must be inverted before computing the two generalised

distance functions. From an algorithmic point of view, the problem is reduced to computing two grey-weighted generalised distance transforms. Figure 5 shows some examples, illustrating the robustness against to the noise.

At any location along a track, according to the neighborhood graph used, it is assumed that the absolute value of the angle between the track and the vertical direction is less than or equal to 45° . It guarantees a certain smoothness to the extracted tracks. This segmentation can be interpreted in terms of an optimality criteria framework (Vincent, 1998): (1) the pixel values along the track (to maximise), (2) the length of the track (to minimise), (3) the raggedness of the track (to minimise).

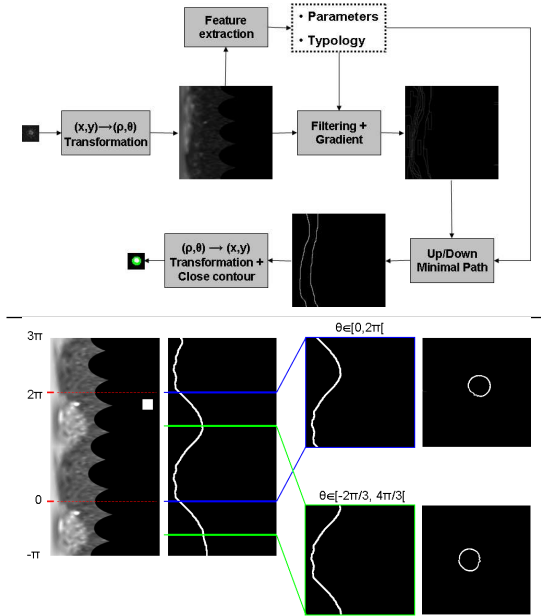


Figure 6. Top, flowchart of algorithm for model-based spot segmentation in polar coordinates. Down, selection a period of circular minimal path to define close contour.

MODEL-BASED SPOT SEGMENT. BY MINIMAL PATHS

Starting from the gradient of a filtered version of the spot in polar coordinates, the aim now is to segment its contour using the GMP technique. To achieve a robust algorithm several issues must be considered in detail (see figure 6).

Interpolation: The spots are small image structures, typically their diameter is approximately equal to 15 pixels and their bounding boxes of size

25×25 pixels. In polar coordinates, the radial variation is consequently limited to around 7 pixels. These small magnitudes limit the possibilities to obtain a regularised or multi-level segmentation of the spots by using Up/Down GMP's. Let $f_i(\mathbf{x})$ be the original spot sub-image, we propose to interpolate it by using a bi-linear schema to increase the size of structures to be segmented, $f_i^{\uparrow k}(\mathbf{x})$ (a factor $k = 4$ constitutes an interesting value). The cartesian-to-polar conversion is then computed from this image, followed by the Up/Down GMP and the inverse conversion. Obviously, the closed contour must be decimated by the same factor k in order to obtain the original spot size.

Circular minimal path to close contour: In order to obtain a closed contour for the spot region, we must impose a circular minimal path, i.e., in polar coordinates and with the Up/Down GMP, the initial radial value ρ_{up} (for $\theta = 2\pi$) and the final one ρ_{down} ($\theta = 0$) are equal. Several algorithms have been proposed in the literature to calculate circular minimal paths, relatively sophisticated and solved by dynamic programming (multiple search algorithm, branch and bound algorithm, etc.) (Sun and Pallottino, 2003). We propose to apply a simpler algorithm to allow using GMP approach to define closed spot contours.

The original polar images $[0, 2\pi[$ can be cycled, extending the image along its angular direction by adding the top part of the image on the bottom and the bottom part on the top, and consequently repeating another period of the image. When the Up/Down GMP is applied to this cycled image, the continuity provided by the added cycle yields almost always a circular path. In fact, even if $\rho_{up} \neq \rho_{down}$, but $|\rho_{up} - \rho_{down}| \leq \Delta_\rho$ (Δ_ρ being a small value, typically 2 to 5 pixels), the contour can be ‘‘closed’’ applying previously a dilation of size Δ_ρ before computing the transformation to Cartesian coordinates. Moreover, the cycled image allows to select different periods of the minimal path to find a circular minimal path or at least the minimal path with lowest Δ_ρ . In practice, the translation along the angular axis θ in polar coordinates involves a rotation in Cartesian coordinates, i.e., if the selected period of $\theta \equiv [0 + \alpha, 2\pi + \alpha[$ the image of the closed contour should be rotated α radians. To avoid the vagueness due to the rotation, we usually consider five simple cases ($\alpha = 0, \pi/2, \pi, -\pi/2$ and $-\pi$) and we choose the α which has lower Δ_ρ , see figure 6.

Filtering and gradient in polar coordinates: As we have shown, the polar image $f_i^c(\rho, \theta)$ is cycled to ensure the periodicity of the angular coordinate. The polar image filtering (i.e., type and sizes of filters) is a critical step in order to achieve a robust segmentation method.

An anisotropic effect in polar coordinates is obtained by applying two separable directional filters (unidimensional filtering) in the angular and radial coordinates. Usually, for the polar image of spots, the vertical (according to the angular coordinate) filtering has a size n_θ which is notably higher than the size n_ρ of horizontal filtering (radial direction). We have compared three different families of filters: Gaussian diffusion, morphological operators (opening/closing + levelling) and sliding average. In fact, the average filter is the simplest and fastest approach which simplifies the structure such a way that the GMP corresponds to the main spot contour. It seems that the sizes $n_\rho = 16$, $n_\theta = 48$ ($\simeq \pi/3$) yield a satisfactory trade-off for this spot whose diameter is approx. equal to 7 pixels ($7 \times 4 = 28$ pixels in the interpolated version). If the adequate vertical size of filtering can be consider as independent of the spot diameter, the choice for the horizontal one well-adapted to one spot is obviously associated to an estimate of its radius, obtained from the radial projection (see previous section). Concerning the gradient, the external gradient is always applied, $g^+(f(\mathbf{x})) = \delta_1(f(\mathbf{x})) - f(\mathbf{x})$.

Spot typologies for segmentation: The *homogeneous spots* (regular or saturated) are easily segmented using the present approach. The *inhomogeneous spots* (cracked or fragmented) need an estimate of spot diameter and of texture degree to adapt the size of horizontal/vertical anisotropic filtering. In the case of *empty spot* (or absent spot), we propose to calculate also a GMP to segment the background and try to compute a parameter of intensity. These classes of spots only need one contour. The segmentation of *doughnut-like spots* (i.e., presenting a hole) and *egg-like spots* (i.e., with a peak of intensity) needs the computation of a multiple contour, i.e., multiple minimal path.

Several alternatives can be applied for the spot segmentation in two or more regions. From a mathematical morphology viewpoint, this involves filtering the spot, removing the hole/peak, and therefore enhancing its main contour. In order to do that, we use the “close-holes” operator. This operator fills all holes in an image f that do not touch the image boundary f_∂ (used as a marker) and therefore provides a parameter free approach to detect holes in an image: $\psi^{ch}(f) = [\delta_{f_\partial}^{rec}(f_\partial)]^c$, where $\delta_g^{rec}(m)$ is the geodesic reconstruction by dilation of the reference image g with the marker image m . For a binary image, the definition of grains and holes is clear; for the case of grey level images, a “hole” is defined as a set of connected points surround by connected components of value strictly greater than the values of the hole. This operator is a morphological closing and therefore

removes the dark structures (valleys of intensity). A dual version of this operator allows the definition of a dual close-holes operator to remove the peaks of intensity.

We can also work on the residues of these morphological operators. That is, to be able the segment, on the one hand, the spot without hole or grain and on the other hand, the hole and the grain. The final algorithm proposed is based just on working on three different images on which we compute eventually, and according to the typology, up to three Up/Down GMP.

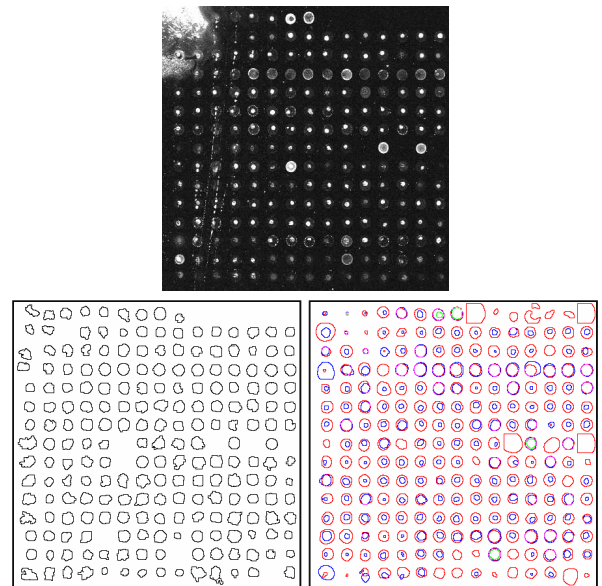


Figure 7. *Top, original microarray (intensity $\times 10$). Down left, segmentation using classical morphological approach (by Angulo and Serra (2003)). Down right, segmented spots according to the present algorithm (main contour in red, peak contour in blue and hole contour in green).*

In figure 7 is given an example of segmentation of a (very “bad” quality) microarray image, including regular spots, doughnut-like spots and a majority of egg-like spots. The main spot contours (in red) are well segmented in comparison with the classical morphological segmentation. The second contour or hole contour (in green) of doughnut-like spots is always well segmented. In the case of egg-like spots, the second contour which correspond to the peak (in blue), is in general satisfactory but, due to the noise, a few contours are wrong.

CONCLUSIONS, PERSPECTIVES

The morphological spot modelling allows us to calculate quality control parameters to detect the

accidents of preparation; define distances between spots and spot kernels for image-based machine learning and classifying algorithms; propose new ways of visualisation and analysis of spot images, etc.

The results of model-based spot segmentation are satisfactory (sensible and robust) and improve the previous approaches, allowing an automatic adaptation to all the situations. An additional control step will be included in the approach in order to evaluate the pertinence of the minimal paths extracted (using the value of the gradient along the path, the contrast between the regions separated by the path, etc.) which will be useful for the data post-processing and quantification.

Classically, the “intensity” of the spot is given by computing the integral or the mean/median (and the variance) of the grey-level image points inside the spot region. In this new approach the spot according to its typology can be segmented into several regions, and consequently the “intensity” of the spot will be characterised by a vector of several parameters (i.e., median and variance for each region).

This kind of quantified data (spot shape/texture features and typology, multi-region spot segmentation, multiple parameter of hybridization by spot, etc.) opens new possibilities to refine the existing microarray platforms and especially to adapt the high-level data analysis algorithms.

Acknowledgements. The author gratefully thanks Fernand Meyer for his valuable suggestions. This work is part of the French Project *GEMBIO-Bioinformatique* 2003-2006 (Mathematical methods for the analysis of biochip data: towards medical and therapeutic diagnosis and prognostic) supported by the Conseil General des Mines.

REFERENCES

- Angulo J, Serra J (2003). Automatic analysis of DNA microarray images using mathematical morphology. *Bioinformatics* 19:553–562.
- Angulo J (2005). Automated spot classification in cDNA images using mathematical morphology. *Internal Note N-19/05/MM* CMM-Ecole des Mines de Paris, 28p.
- Appleton B, Talbot H (2005). Globally Optimal Geodesic Active Contours. *Journal of Mathematical Imaging and Vision* 23: 67–86.
- Balagurunathan Y, Dougherty ER (2002). Simulation of cDNA microarrays via a parameterized random signal model. *Journal of Biomedical Optics* 7: 507–523.
- Brown PO, Botstein D (1999) Exploring the NewWorld of the genome with DNA microarrays. *Nature Genet.* 21 (Suppl.):33–37.
- Chen Y, Dougherty ER, Bittner ML (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* 2: 364–374.
- Ekstrom CT, Bak S, Kristensen C, Rudemo M (2004). Spot shape modelling and data transformation for microarrays. *Bioinformatics* 20: 2270–2278.
- Glasbey C, Khondoker M (2005). Correction for pixel censoring in cDNA microarray. In *Proc. of the 20th International Workshop on Statistical Modelling*, University of Western Sydney Press: 17–31.
- Li Q, Fraley C, Bumgarner RE, Yeung KY, Raftery AE (2005). Donuts, scratches and blanks: robuts model-based segmentation of microarray images. *Bioinformatics* 21: 2875–2882.
- Luengo-Oroz MA, Angulo J, Flandrin G, Klossa J (2005). Mathematical morphology in polar-logarithmic coordinates. In *Proc. of the 2nd Iberian Conference on Pattern Recognition and Images Analysis (IbPRIA'05)*, Estoril, Portugal, Springer LNCS 3523: 199–206.
- Rosenfeld A, Pfaltz J (1968). Distance functions on digital pictures. *Pattern Recognition* 1: 33–61.
- Steinfath M, Wruck W, Seidel H, Lehrach H, Radelof U, O'Brien J (2001). Automated image analysis for array hybridization experiments. *Bioinformatics* 17: 634–641.
- Sun C, Pallottino S (2003). Circular shortest paths by branch and bound. *Pattern Recognition* 36: 2513–2520.
- Vincent L (1998). Minimal Path Algorithms for the Robust Detection of Linear Features in Gray Images. In *Proc. of International Symposium on Mathematical Morphology (ISMM'98)*, Amsterdam, Kluwer: 331–338.
- Wierling CK, Steinfath M, Elge T, Schulze-Kremer S, Aanstad P, Clark M, Lehrach H, Herwig R (2002). Simulation of DNA array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis. *BMC Bioinformatics* 3: 1–17.
- Yang YH, Buckley MJ, Dudoit S, Speed TP (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics* 11: 108–136.