BIOINFORMATICS

Vol. 19 no. 5 2003, pages 553–562
DOI: 10.1093/bioinformatics/btg057

Automatic analysis of DNA microarray images using mathematical morphology

Jesús Angulo and Jean Serra*

Centre de Morphologie Mathématique, Ecole des Mines de Paris, 77305 Fontainebleau, France

Received on January 14, 2002; revised on November 12, 2002; accepted on November 14, 2002

ABSTRACT

Motivation: DNA microarrays are an experimental technology which consists in arrays of thousands of discrete DNA sequences that are printed on glass microscope slides. Image analysis is an important aspect of microarray experiments. The aim of this step is to reduce an image of spots into a table with a measure of the intensity for each spot. Efficient, accurate and automatic analysis of DNA spot images is essential in order to use this technology in laboratory routines.

Results: We present an automatic non-supervised set of algorithms for a fast and accurate spot data extraction from DNA microarrays using morphological operators which are robust to both intensity variation and artefacts. The approach can be summarised as follows. Initially, a gridding algorithm yields the automatic segmentation of the microarray image into spot quadrants which are later individually analysed. Then the analysis of the spot quadrant images is achieved in five steps. First, a pre-quantification, the spot size distribution law is calculated. Second, the background noise extraction is performed using a morphological filtering by area. Third, an orthogonal grid provides the first approach to the spot locus. Fourth, the spot segmentation or spot boundaries definition is carried out using the watershed transformation. And fifth, the outline of detected spots allows the signal quantification or spot intensities extraction; in this respect, a noise model has been investigated. The performance of the algorithm has been compared with two packages: ScanAlyze and Genepix, showing its robustness and precision.

Availability: A prototype system integrated in PDI32 (an image processing software for Windows) may be obtained from the authors on request.

Contact: angulo,serra@cmm.ensmp.fr

Supplementary Information: http://cmm.ensmp.fr/∼angulo/

INTRODUCTION

DNA microarrays are an experimental technology for exploring the genome. DNA microarrays provide a simple tool to identify and quantify levels of gene expression for all genes in an organism. The method consists in arrays of thousands of discrete DNA sequences that are printed on glass microscope slides using a robotic 'arrayer'. To compare the relative abundance of each of these gene sequences in two DNA or RNA samples, the two samples are first labelled using different fluorescent dyes (usually cyanine dyes, the red-fluorescent dye Cy5 and the green-fluorescent dye Cy3 with emissions in the 630–660 nm and 510–550 nm respectively). They are then mixed and hybridized with the arrayed DNA spots. After hybridization, fluorescence measurements are made for each dye separately; these measurements are used to determine the ratio, and in turn the relative abundance, of the sequence of each specific gene in the two RNA or DNA samples. This is the technique developed by P. O. Brown, at Stanford (Brown and Botstein, 1999; Eisen and Brown, 1999). There are other microarray systems and methods which differ in several details but share the principle of hybridization and produce the same result, an image of spots. See the review paper (Bowtell, 1999). Fluorescent images can be acquired using different devices including a scanner (laser scanning confocal microscope) and a CCD camera. The digitisation process averages both spatial and temporal intensity and produces for each pixel a signal intensity that represents the total fluorescence in the region corresponding to the pixel. Image analysis is an important aspect of microarray experiment. The extracted intensities can have a potentially large impact on subsequent steps of data mining (Bassett *et al.*, 1999).

In this paper, we present our automatic and non-supervised approach to process microarray images which relies on mathematical morphology operators. Simple techniques of spot segmentation are based on global or local thresholding techniques, as the one implemented in the software package TIGR-Spotfinder (Hegde *et al.*, 2000). Other packages, for instance Dapple (Buhler *et al.*, 2000), are based on several hypotheses about the spots, like the circular shape or a minimal intensity. The spot segmentation approach based on Mann-Whitney hypotesis testing was a pioneer method proposed in (Chen *et al.*,

*To whom correspondence should be addressed.

1997). Recently a variation of this technique combined with a binary hit-or-miss transform has been presented (Vesanen *et al.*, 2002). In the very interesting paper (Yang *et al.*, 2002), the package Spot has been presented and a number of existing image analysis methods used on microarray data were reviewed, examining the statistical properties of different segmentation and background noise extraction methods. Mathematical morphology has been already used for gridding and spot segmentation (Hirata *et al.*, 2001).

SYSTEMS AND METHODS

In this section, a presentation of the microarray image features and a revision of some powerful tools from mathematical morphology are included.

DNA Microarray Images

Spot detection is less simple than it seems since the boundary between spot and background is not 'sharp'. On the one hand, the contrast between the spot region and the background is very different from one spot to another; and moreover, the *volume* (integral of intensity in the hybridized spot region) is also very different. The *area* (number of hybridized pixels corresponding to the spot region) seems to be the magnitude property that defines the spot region. Another problem is due to the fact that the hybridization process is not homogeneous: the spot regions are 'broken'. Furthermore, an important contribution of background noise intensity is observed.

Orthogonality and Image Projections

The spots are placed according to a double orthogonal alignment: the spot groups or spot quadrants and the spots. In this situation, starting from the projections of the image along its rows and columns, a first approach to the location of spot groups firstly and the spots later can be implemented. The horizontal (row) and vertical (column) projections of an image $f(x, y)$ (dimensions X and Y, pixel $\boldsymbol{x} = (x, y)$) are defined as $HP(y) = \sum_{x=0}^{X-1} f(x, y)$ and $VP(x) = \sum_{y=0}^{Y-1} f(x, y)$.

Mathematical Morphology

The present algorithm for the analysis of spot images uses operators from mathematical morphology. A tutorial in the technique can be found in (Serra, 1982, 1988, 1999).

In the framework of digital grids, a *gray tone image* can be represented by a function $f : D_f \rightarrow T$, where D_f is a subset of Z^2 and $T = \{t_{\min}, \ldots, t_{\max}\}$ is an ordered set of gray-levels, i.e. a subset of Z. $f(\boldsymbol{x})$ is the gray value of the image at point $\boldsymbol{x} = (x, y)$. From a morphological point of view, the set of all the points $\{\boldsymbol{x}, f(\boldsymbol{x})\}$ belonging to $Z^2 \times Z$ can be seen as a topographic surface S. The lighter the gray value of f at point $\boldsymbol{x}$, the higher the altitude of the corresponding point $\{\boldsymbol{x}, f(\boldsymbol{x})\}$ on the surface. Let B be a subset of Z^2 and λ a scaling factor. λB is called *structuring element* B of size λ.

The basic morphological operators are

- Dilation: $\delta_B(f(\boldsymbol{x})) = \sup_{\boldsymbol{y} \in B}\{f(\boldsymbol{x} - \boldsymbol{y})\}$;
- Erosion: $\varepsilon_B(f(\boldsymbol{x})) = \inf_{-\boldsymbol{y} \in B}\{f(\boldsymbol{x} - \boldsymbol{y})\}$.

The two elementary operations of *erosion* and *dilation* can be composed together to yield a new set of operators having desirable feature extractor properties which are given by

- Opening: $\gamma_B(f) = \delta_B[\varepsilon_B(f)]$;
- Closing: $\varphi_B(f) = \varepsilon_B[\delta_B(f)]$.

The morphological *openings* $\gamma_{\lambda B}$ (*closings* $\varphi_{\lambda B}$) filter out light (dark) structures from the images according to a predefined size λ and shape criterion B.

A morphological tool that complements the opening and closing operators for feature extraction (extract the marked particles) is the morphological reconstruction, implemented using the *geodesic dilation* operator based on restricting the iterative dilation of a function marker f by B to a function mask g, $\delta_g^n(f) = \delta_g^1 \delta_g^{n-1}(f)$, where $\delta_g^1(f) = \delta_B(f) \wedge g$. The *reconstruction* by dilation is defined by $\gamma^{rec}(g, f) = \delta_g^i(f)$, such that $\delta_g^i(f) = \delta_g^{i+1}(f)$ (idempotence). The *filters by reconstruction* involve the notion of connectivity and preserve the 'edges' of the structures.

Families of openings or closings of increasing size are at the basis of the granulometric analysis. The *granulometries* or *size distributions* allow a good knowledge of the objects or structures present in the images.

A *plateau* in a gray tone image is a connected component of uniform altitude. A *maximum* is a plateau without higher neighbours. Among the various features that can be extracted from an image, the *maxima* and the *minima* are of primary importance in the morphological segmentation and in the modern morphological filtering techniques because they mark the structures present in the image (Grimaud, 1992).

The *dynamics* is a measure of contrast which maps each maximum with a contrast value. The dynamics is closely linked to the H-reconstruction or *contrast opening* of f of size H, defined as $\gamma_H^d(f) = \gamma^{rec}(f, f - H)$ (Grimaud, 1992). The gray tone *area opening* of an image f of size λ, denoted $\gamma_\lambda^a(f)$ can be seen as an opening by reconstruction with a structuring element which locally adapts its shape to the image structures (Vincent, 1992).

The *extinction value*, $\mathcal{E}$, is a granulometric operator which maps the extrema of gray tone images with a measurement of their persistence (for which step the structure is filtered out) when families of openings or

closings of increasing size are applied (Vachier and Meyer, 1995). In this case, the openings are by contrast, area and volume (combination of both) and the corresponding extinction values are denoted $\mathcal{E}^d$, $\mathcal{E}^a$ and $\mathcal{E}^v$ respectively.

The watershed line is one of the most powerful tools for segmenting images (Beucher and Meyer, 1992; Beucher, 1999). The watershed line associates a catchment basin to each minimum of the function. Using the watershed on a gray tone image without any preparation leads to a strong over-segmentation. The best solution to this problem consists in initially determining markers for each region of interest, including the background of the image. The image corresponding to the markers is denoted g. It is then possible to construct a watershed line associated with these markers $Wshed(f, g)$. Another important choice is the function used in the flooding process. Usually, the selected function f is some form of image gradient.

Visualisation, segmentation and quantification

The end product of a comparative hybridization experiment is a pair of scanned array images: two 16-bits TIFF files. Usually, for *display* purposes the two 16-bits images are compressed into two 8-bits images using a square root transformation. By this transformation, a false-colored 24-bits RGB image is composed: blue values are set to zero, red values are used for the Cy5 image, and green values for the Cy3 image. Our algorithm for microarray *image pre-processing* require a single image and it is convenient computationally for the image to be 8-bits. There are several alternatives for handling the pair of 16-bits images. We propose to combine both images in order to obtain a single 8-bits image using a linear combination weighted by the median values (Yang *et al.*, 2002). In the *quantification*, the segmentation mask is used together with the original 16-bits images for extracting the background and the spot intensities.

ALGORITHM

Initially, a gridding algorithm yields the automatic segmentation of the microarray image in sub-arrays, defining each spot quadrant, which are individually analysed in five steps.

Array Orthogonal Grid

A first operation is the reduction of the microarray size by image decimation with averaging of size K ($K = 4$ has shown to be suitable for this kind of images). The decimated image $\widetilde{f}$ is only used as a means to segment the spot quadrants, then once the coordinates of each quadrant or block have been defined the initial sub-array is extracted and individually analysed. We propose to enhance the spot quadrants of $\widetilde{f}$ by means of a simple morphological operation (before that, filtering out the image with a median filter, square of size 3×3, which reduces the noise): the supremum of a vertical closing φ_n^{π} and a horizontal closing φ_n^0 of $\widetilde{f}$ with a structuring element of size n; i.e. $\widetilde{f}^{\bullet} = \varphi_n^{\pi}(\widetilde{f}) \vee \varphi_n^0(\widetilde{f})$. As a result, the orthogonal neighbouring spots inside a block are merged, Figure 1b.

The size n of the closing has to be larger than the diameter of the spot d_{Spot} (similar to the distance between neighbouring spots) and smaller that the distance between spot groups $d_{SpotGroups}$ in order to avoid the fusion of blocks. A good criterion is to take $n = 2d_{Spot}$, since we did not find in our microarray database any image with $2d_{Spot} \geq d_{SpotGroups}$. The value of d_{Spot} depends on the microarray but we introduce below a morphological spectral technique for computing the spot size; i.e. the spot area $a_{Spot} = \lambda_s$. Therefore, $d_{Spot} = 2\sqrt{\frac{a_{Spot}}{\pi}}$, and we have to take into account that we are working on an image decimated $\div 4$. In the current example, $d_{Spot} \simeq 5$ and $d_{SpotGroups} = 30$.

On the enhanced image $\widetilde{f}^{\bullet}$, the orthogonal projections are computed. Let $P^{\widetilde{f}^{\bullet}}(i)$ be the horizontal or vertical projection. The one-dimensional morphological signal processing in order to obtain the grid is performed in three steps:

(1) Intra-block filtering: A first opening of size n_{ib} removes the intra-block variations. Obviously, n_{ib} has to be smaller than the width of the block, $P_{ib}^{\widetilde{f}^{\bullet}} = \gamma_{n_{ib}}(P^{\widetilde{f}^{\bullet}}(i))$.

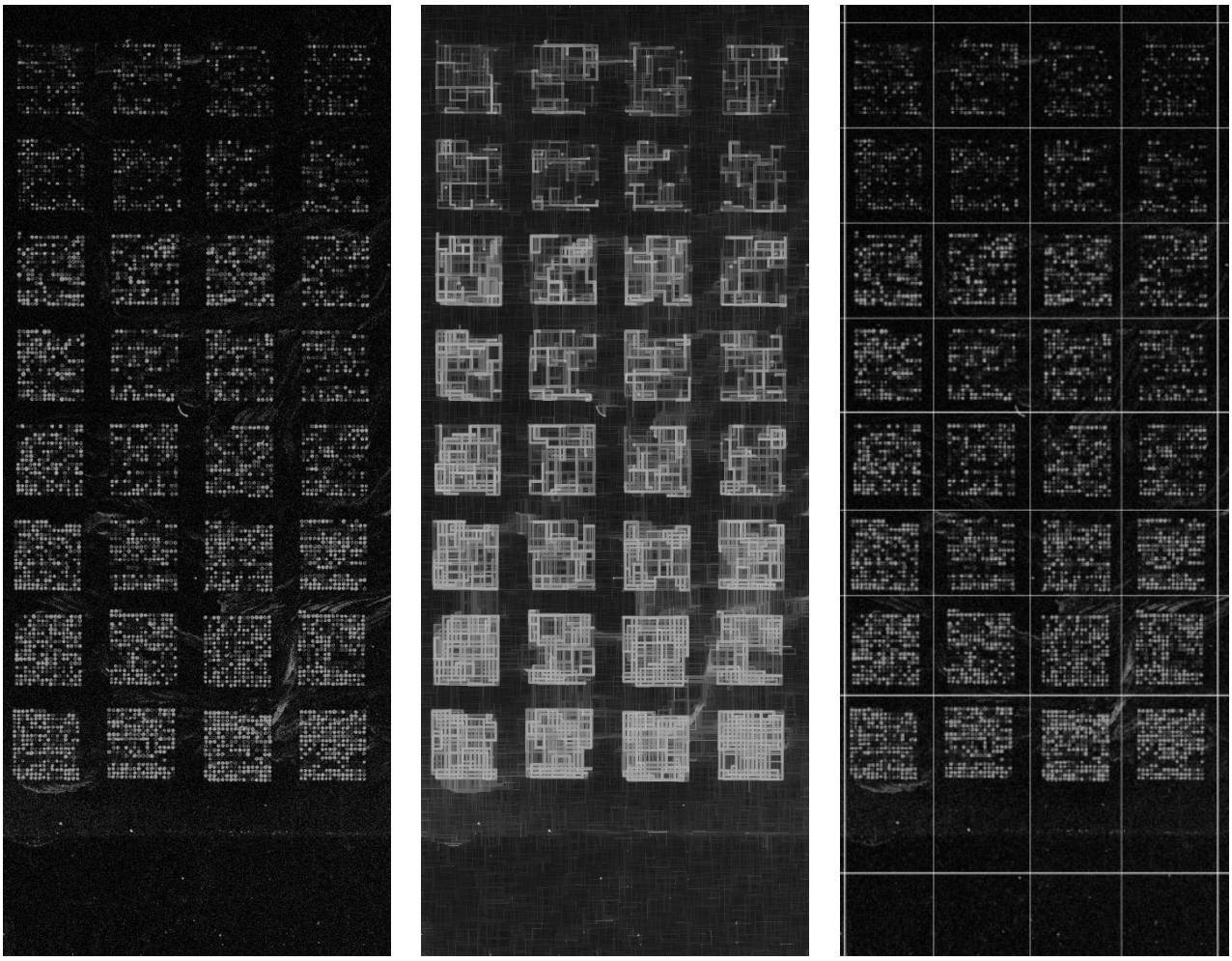
(2) Block extraction: The residue of this signal and another opening of size n_b (larger than the width of the block) extracts the blocks, $P_b^{\widetilde{f}^{\bullet}} = P_{ib}^{\widetilde{f}^{\bullet}} - \gamma_{n_b}(P_{ib}^{\widetilde{f}^{\bullet}})$.

(3) Thresholding: The thresholding process on this signal is straightforward. As criterion, the optimal threshold value u_P is defined as 20% of the average of $P_b^{\widetilde{f}^{\bullet}}$, $u_P = 0.2\frac{1}{N}\sum_{k=1}^{N} P_b^{\widetilde{f}^{\bullet}}(k)$.

Using the binary signal it is straightforward to define the orthogonal region for each spot group or spot block (the boundary goes through the middle of each interval of zeroes). In Figure 1c the result for the previous example is included and on http://cmm.ensmp.fr/∼angulo/research/dnamicro.htm two other microarray examples can be viewed. We fix n_{ib} as the size of five spots, $n_{ib} = 5d_{Spot}$; in the case of n_{ib}, the choice is not so important, for instance $n_b \gg n_{ib} \rightarrow n_b = 10n_{ib}$. In the examples, the values are $n_{ib} = 25$ and $n_b = 250$.

Spot-Size Distribution Law

The *morphological extinction spectrum* $\mathcal{ES}[\lambda]$ is a granulometry dealing with families of openings by reconstruction, i.e. mapping any extreme with its extinction value. In



(a) (b) (c)

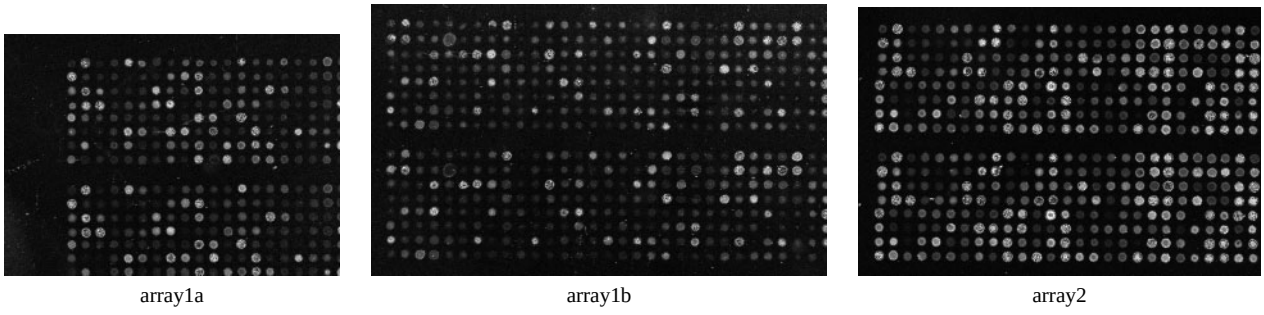
Fig. 1. Procedure for array orthogonal grid definition: (a) Decimation and low-pass filtering ÷4, the initial image has a size of 1842 × 4512 pixels and the present decimated image, 460 × 1128 pixels. (b) Spot groups morphologically enhanced by means of supremum of horizontal and vertical closings. (c) Array grid obtained.

Angulo and Serra (2002a), we have introduced these histograms of extrema which are characterised by three measures: contrast, area or volume, and their ability for the analysis of genome images. Based on the area extinction spectrum in logarithmic scale, we presented a new tool: the *spot-size distribution law*, $SS[\lambda]$, where λ is the spot size (area) and $SS[\lambda] = n_\lambda$ denotes the normalised number of occurrences at the extinction value λ. Consequently, $SS[\lambda]$ is a probability density function. Figure 2a depicts three images from two microarrays. In Figure 2b, the area extinction spectra of the image examples are shown.

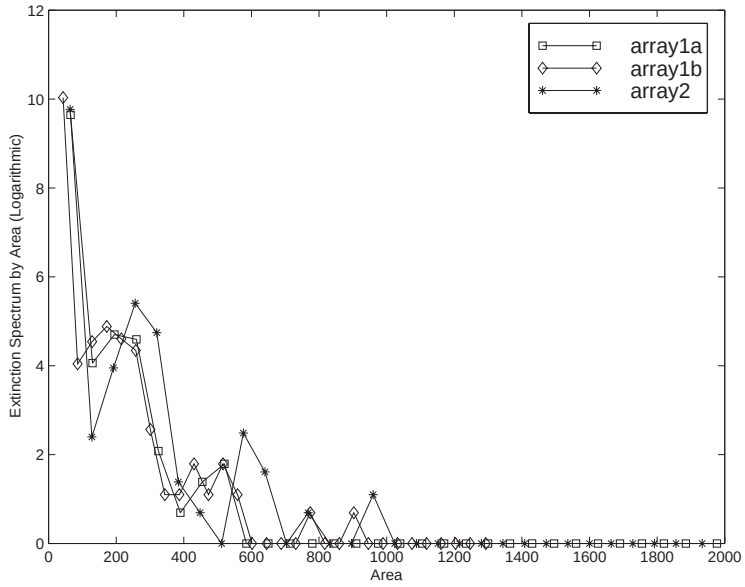
The $SS[\lambda]$ is usually composed of several modes (multimodal histogram). In the case of images extracted from the same array the spectra are almost superimposed. Owing to the definition of area extinction value as a filter by reconstruction, some extrema have associated an extinction value corresponding to the area of several neighbouring spots, getting this impression of multimodal. However, the interesting mode is the first mode and the other ones can be considered as *harmonics* of the fundamental morphological frequency. This *fundamental mode*, or mean spot size, provides the threshold area value for the subsequent filtering.

Morphological Filtering by Area Extinction Value

Besides distorting the quantified intensity, the background noise could introduce mistakes in the building of the grid and in the detection of frontiers of spots. Therefore the aim of the filtering step is to remove the noise trying at once not to distort the spot boundaries. The extinction values



(a)



(b)

Fig. 2. Meaning of Spot-Size Distribution: (a) Examples of microarray images, *array1a* and *array1b* from the same array and *array2* from other one. (b) Area extinction spectra of the examples in logarithmic scale.

associated with the area opening, called *area extinction values*, $\mathcal{E}^a$, allow to study the size distribution of the structures on the image without taking shape information into consideration (only the area or size) and this is the above presented spot-size distribution law, $SS[\lambda]$. The significant extrema are selected by thresholding their extinction values, then the image is filtered by geodesic reconstruction using the extrema as markers (Vachier and Meyer, 1995). Using this filtering technique by area on the spot quadrant image f, $f' = \gamma^a_{\lambda_s}(f)$, the structures with area greater than the chosen threshold λ_s are preserved in f'.

Therefore the threshold value λ_s optimal for each image is an important choice which depends on the size of spots. And this is the rationale for computing $SS[\lambda]$; λ_s is automatically obtained from the *spot-size distribution law*, as an estimate of the mean spot size. Different experimental tests have shown that after the cut-off of the higher harmonic, take λ_s equal to the 50% of the mean size yields good results. It is possible that a statistical measure of dispersion would provide a more sound theoretical basis for the spot size threshold.

Another important advantage of morphological filtering by area opening is the implicit selection of maxima. Let m be the binary image which contains the maxima after area filtering; i.e. $m = \mathrm{Max}(f')$. These *spot markers* are used as initial markers for the segmentation by watershed transformation.

Spot Orthogonal Grid

The spots inside the spot quadrant are placed according to an orthogonal alignment and again, using the horizontal and vertical projections, a first approach to the locus of spot can be implemented by definition of the spot grid. The

algorithm for the spot orthogonal gridding is as follows. Let $P(i)$ be the horizontal or vertical projection signal:

(1) Calculate the mean value of the elements in $P(i)$, $\overline{P} = \frac{1}{N}\sum_{i=1}^{N} P(i)$.

(2) Subtract the mean from the projection, $P_\eta(i) = P(i) - \overline{P}$.

(3) Morphological reconstruction of $P(i)$ using $P_\eta(i)$ as marker, $P^{rec}(i) = \gamma^{rec}(P(i); P_\eta(i))$.

(4) Take the residue of the initial projection $P(i)$ and the reconstruction $P^{rec}(i)$, $P_{TH}(i) = P(i) - P^{rec}(i)$.

(5) Estimate the optimal threshold value u_P, defined as the α% of the average of the residue $P_{TH}(i)$, $u_P = \frac{\alpha}{100}\frac{1}{N}\sum_{i=1}^{N} P_{TH}(i)$.

(6) Find the binary reference signal $P_u(i)$ by thresholding process at u_P on the residue signal $P_{TH}(i)$.

(7) Using the i middle of each interval of ones in $P_u(i)$, draw the straight lines corresponding to the orthogonal grid.

After different tests in the database, the choice of $\alpha = 50\%$ appears to be a suitable optimal threshold value (Angulo and Serra, 2002b). The orthogonal grid, together with the frontiers of the image (from the border to the first lines of the grid), are used as outer markers for segmenting by watershed transformation; the defined image is denoted g_r.

Morphological Segmentation by Watershed Transformation

According to the classical paradigm of morphological segmentation (Beucher, 1999), the algorithm for segmenting the spot images is as follows:

(1) Define a gradient function to flood: The filtered image f' is again simplified by a morphological levelling (Meyer, 1998). The new image is denoted f''. The external gradient g^+ is defined as the difference between the dilated image and the original image; i.e. $g^+(f'') = \delta_B(f'') - f''$.

(2) Obtain the markers: The outer markers are the filled borders of the grid together with the orthogonal grid, g_r.
For the inner markers, we propose a specific algorithm. On the basis of an individual image analysis of the spot bounding boxes defined by the spot grid, the procedure for each spot i is the following,
   (a) Take the bounding box of spot i (A_i is the area of this region).
   (b) Extract the corresponding image region (V_i is the volume of this region).
   (c) From m, number of maxima inside this region $N_{\text{maxima}}(i)$, if
      - $N_{\text{maxima}}(i) = 0$: The spot i is classified as *absent spot* and no marker is assigned.
      - $N_{\text{maxima}}(i) = 1$: The spot i is classified as *clear spot* and a marker is defined.
      - $N_{\text{maxima}}(i) > 1$: The spot i is classified as *vague spot* and a marker is defined.
   (d) For the cases $N_{\text{maxima}}(i) \geq 1$, the marker is calculated as the morphological centroid (a dilated pixel from the ultimate erosion) of a thresholded binary spot approximation region, where the threshold value is done by $\mu_i = V_i/A_i$. The set of all the inner markers is the binary image mk_i.

   The global markers to impose are done by $mk = g_r \vee mk_i$.

(3) Determine the watershed with markers: Construction of the watershed line for $g^+(f'')$ associated to the markers of the spots and the background mk, $sm = Wshed(g^+(f''), mk)$, where sm are the line boundaries of each spot.

Spot Quantification and Noise Extraction

The segmentation layer sm obtained from the presented algorithm is used on the initial 16-bits images for the spot quantification and noise extraction. The intensities provided by the array image can be quantified by measuring the average or integrated intensities of the spots. The ratio of fluorescent intensities for a spot is interpreted as the ratio of concentrations for its corresponding DNA in the two cell populations. The motivation behind background adjustment is the fact that a spot measured intensity includes a contribution not specifically due to the hybridization, but to other unwanted phenomena; for a deep study on some of these phenomena see (Doel, 2002). The *measured signal intensity* $\hat{s}_i$ of spot i is defined as $\hat{s}_i = \sum_{(x,y)\in S_i} \hat{f}(x,y)$, where $\hat{f}(x,y)$ is the array image intensity at pixel $\boldsymbol{x} = (x,y)$ and S_i is the image region of spot i from the segmentation layer. The spot measured intensity can be expressed as the sum of a *signal intensity value*, s_i, and a *noise intensity value*, n_i, such that $\hat{s}_i = s_i + n_i = s_i + \bar{N}_i A_i$, where $\bar{N}_i$ is the *average noise* of spot i and A_i is the *area* (number of pixels) of spot i.

In Angulo and Serra (2002b) and based on Matheron's geostatistics theory (Matheron, 1975), we have presented a complete study in order to verify that only a local background estimate is adequate. The orthogonal spot grid yields an alternative segmentation: the spot bounding boxes BS_i which can be considered as the influence regions of spots S_i. These regions can be used for quantifying the local noise associated to each spot. However,

if we consider as background all the pixels that are not within the spot region but are within the orthogonal spot region, small segmentation mistakes could bias the background quantification. Thus, in order to avoid that the residual spot signal forges the noise estimate, a region of safeguard is considered. In practice, this enveloping zone is obtained by the residue of a dilation of the spot region; i.e. $\delta_n(S_i) - S_i$, and the noise $\check{n}_i$ is estimated in the region $BS_i - \delta_n(S_i)$ which has an area of $\check{A}_i$ pixels. The typical size for the dilation is $n = 3$. The global expression for the signal intensity is given by

$$s_i = \hat{s}_i - n_i = \hat{s}_i - \bar{N}_i A_i = \hat{s}_i - \frac{\check{n}_i}{\check{A}_i} A_i .$$

IMPLEMENTATION

The present approach has been implemented as a prototype software package for the analysis of DNA microarray images. The input is the pair of scanned images and the output is the data file with the quantification parameters.

The system has been built using PDI32, an image processing software developed at *Image and Video Processing Group* (Politechnic University of Valencia, Spain) and at *Center of Mathematical Morphology* (School of Mines of Paris, France). The integration of these techniques into another microarray image/data manipulation software system is an easy task since the algorithms are based on functions of the exportable PDI32 library.

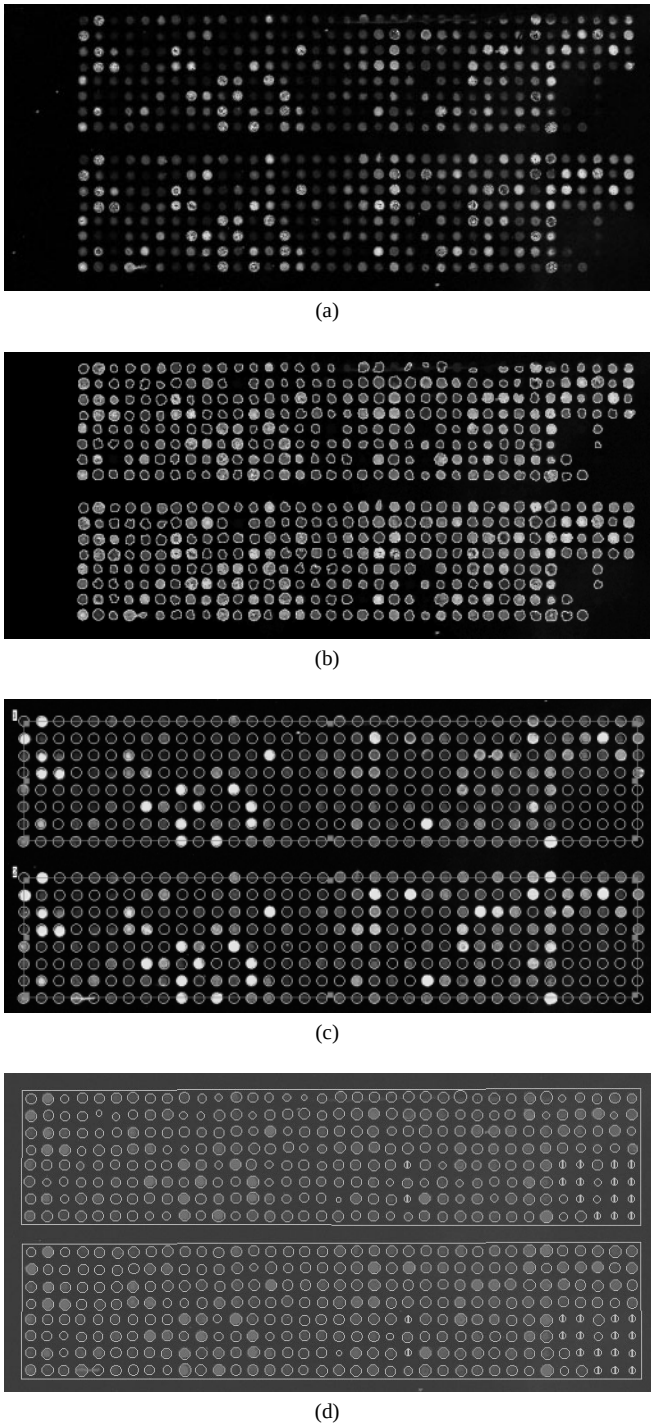
DISCUSSION AND CONCLUSIONS

We continue with the discussion of experimental results by means of a comparative study, including the limitations and perspectives of application.

Results of a comparative study

The algorithms have been developed and tested using a selection of DNA scanned microarrays from three different laboratory systems. In this section, three microarrays (one from each source) are used to compare the merits of the present algorithms with two very widely microarray image analysis methods: ScanAlyze (Eisen, 1999) and GenePix (Axon Instruments, Inc., 2002); for detailed description of these packages see (Yang *et al.*, 2002).

The analysed microarrays correspond to the ones depicted in Figure 1 and figures which can be viewed on http://cmm.ensmp.fr/∼angulo/research/dnamicro.htm, and are named from now on Microarray 3, Microarray 2 and Microarray 1. Apart from its computational efficiency (tables of execution times are available) and the full automation, the set of algorithms introduced in this paper has many other advantages. In Figures 3 and 4 are included the initial spot blocks and the results of segmentation for a specific spot quadrant from Microarrays 2 and 3. The segmentation results of Microarray 1 can be viewed on http://cmm.ensmp.fr/∼angulo/research/dnamicro.htm.



(a) (b) (c) (d)

Fig. 3. Comparison of segmentation algorithms using the spot block No. 1 (double) from Microarray 2: (a) Initial spot block. (b) Spot segmentation using the present approach. (c) Spot segmentation using ScanAlize. (d) Spot segmentation using GenePix.

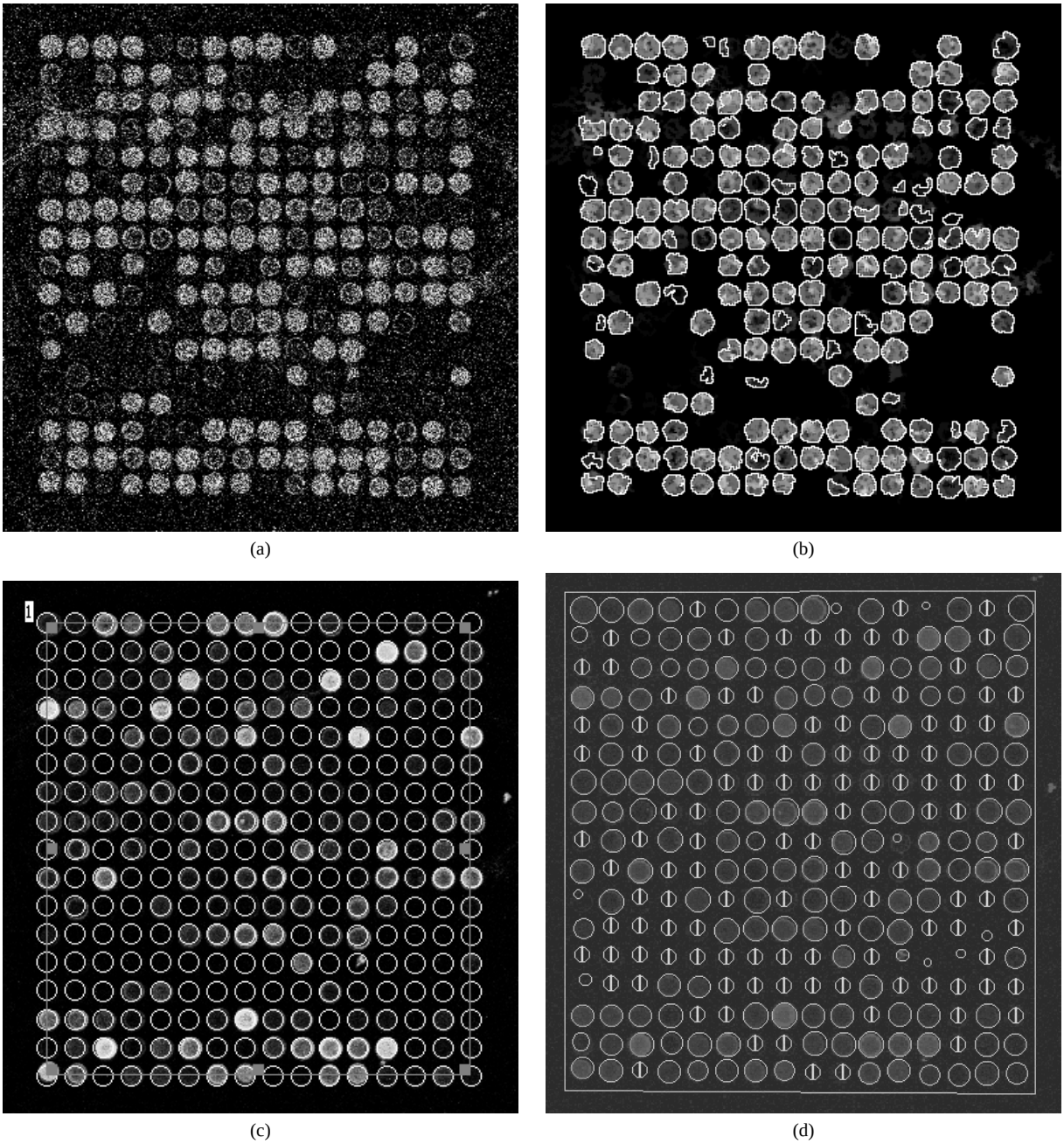
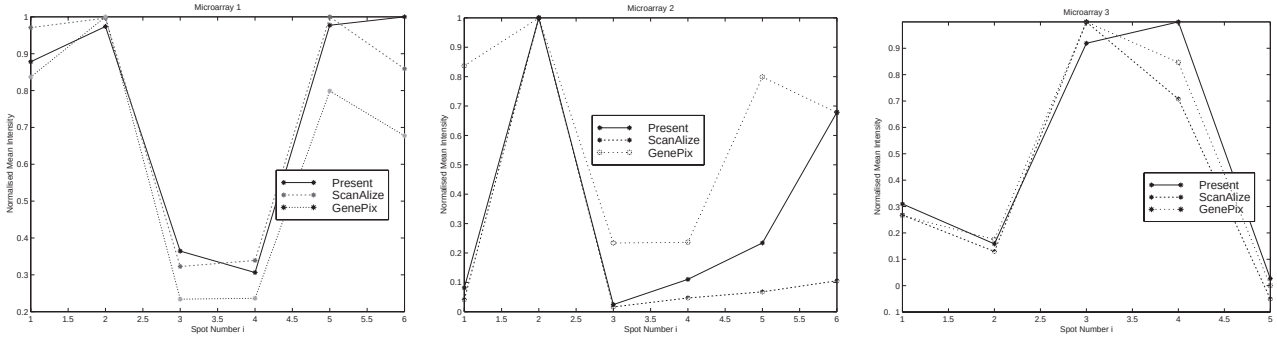


Fig. 4. Comparison of segmentation algorithms using the spot block No. 32 from Microarray 3: (a) Initial spot block. (b) Spot segmentation using the present approach. (c) Spot segmentation using ScanAlize. (d) Spot segmentation using GenePix.

First, we can observe that the use of a fixed size circle, as it is the case for ScanAlyze, is in general a very poor approach; moreover placing manually the spot grid on the image is not easy (the option of refinement is very fortuitous) and there are always mistakes. In good quality images, for instance Microarray 2, the errors can be negligible but with severe problems of noise or in presence of misalignment, Microarray 3, the bias can be important. Note that in ScanAlyze algorithm, all the spots are considered valid and the issue of absent spot should be considered later in the statistical data analysis. We think that it is better to detect and mark the absent spots (not hybridization or very weak level) directly from the image information using an objective criterion, such that GenePix (the special marking (|) is used in GenePix for the absent spots) or our own approach. As you can see



(a) (b) (c)

Fig. 5. Comparison of the spot quantification for the three examples (the parameter represented is the normalised mean intensity after background correction in a dye): (a) Six random spots from Microarray 1. (b) Six random spots from Microarray 2. (c) Five random spots from Microarray 3.

in the examples, the results provided by GenePix and our algorithm are similar enough, although the present technique introduces generally more absent spot. This is related to λ_S the size of morphological filtering by area. However, as you can verify on the results of Microarray 1, in most cases the proposed solution by GenePix for the difficult spots is quite dubious: a very small spot. This is in connection with the limitation of GenePix to use always a circle as spot shape.

The comparison of Figure 3 is very interesting since it concerns a duplicate block-quadrant; and we can observe the algorithm robustness. The results obtained by means of GenePix are very good: the same absent spots are detected in the two blocks (almost, the lower block has one more absent spot). Using our approach, the differences are focusing on the first row since an important linear artefact involves a strong decrease of intensity on the first block and consequently, some absents spots. In general our technique is also very robust, as we can state with the results of Figure 4a very low quality array.

Finally, we would like also to show some results of spot quantification for the three examples. Figure 5 depicts the measured values for some spots from each image. As we can see, the obtained values using our approach are an intermediate measurement between Genepix and ScanAlyze (in certain situations, the background could be under-estimated or over-estimated). We think that our technique is more stable because we estimate the background on the morphological filtered image, similar to the technique proposed in the Spot package (Yang *et al.*, 2002) which estimate the background using a morphological opening.

Conclusions and perspectives

The experimental evaluated performance of spot segmentation shows that the use of present algorithms is generally equal or better than the use of conventional manual techniques. Nevertheless, it is difficult to give a single figure of merit of this kind of techniques. Our successful implementation suggests that these techniques could be valuable as part of a highly automatic, high-throughput system for global microarray analysis, including data analysis process. The algorithm is general enough to be used without any modification on other biochip and miniaturisation technologies: tissue microarrays, Lab-on-Chip, etc.

We are currently investigating supplementary ways of improvement aimed at the measurement of quality assurance. On the other hand, we are also working towards morphological methods that allow the normalisation and quantification of differential gene expression directly on the image.

ACKNOWLEDGEMENTS

The authors would like to thank Ana Dopazo of CNIO, Madrid, Spain, and Nicolas Ugolin of CEA, France, for providing the images used in this study.

REFERENCES

Angulo,J. and Serra,J. (2002a) Morphological spectrum applied to genome arrays quantification. *Journal of Visual Communication and Image Representation*, Submitted.

Angulo,J. and Serra,J. (2002b) A mathematical morphology contribution to the analysis of DNA microarray images. *5th ESMTB Conference Mathematical Modelling & Computing in Biology and Medecine*, Milano, Italy, July 2002, Proceedings to be published.

Axon Instruments, Inc (2002) GenePix Pro 4.0. http://www.axon.com/, Documentation.

Bassett,D.E., Eisen,M.B. and Boguski,M.S. (1999) Gene expression informatics—it's all in your mine. *Nat. Genet.*, **21** (Suppl.), 51–55.

Beucher,S. and Meyer,F. (1992) The morphological approach to segmentation: the watershed transformation. In Dougherty,E. (ed.), *Mathematical Morphology in Image Processing*. Marcel-Dekker.

Beucher,S. (1999) Image segmentation and mathematical morphology. http://cmm.ensmp.fr/∼beucher/wtshed.html, Centre de Morphologie Mathématique, Ecole des Mines de Paris.

Bowtell,D.D.L. (1999) Options available—from start to finish—for obtaining expression data by microarray. *Nat. Genet.*, **21** (Suppl.), 25–32.

Brown,P.O. and Botstein,D. (1999) Exploring the New World of the genome with DNA microarrays. *Nat. Genet.*, **21** (Suppl.), 33–37.

Buhler,J., Ideker,T. and Haynor,D. (2000) Dapple: Improved Techniques for Finding Spots on DNA Microarrays. *UW CSE Technical Report UWTR 2000-08-05*. pp. 12.

Chen,Y., Dougherty,E.R. and Bittner,M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, **2**, 364–374.

van den Doel,L.R. (2002) *Quantitative Microscopic Techniques for Monitoring Dynamic Processes in Microarrays*, Ph.D. thesis, Delft University of Technology.

Eisen,M.B. and Brown,P.O. (1999) DNA Arrays for Analysis of Gene Expression. *Meth. Enzymol.*, **303**, 179–205.

Eisen,M.B. (1999) ScanAlyze http://rana.lbl.gov/EisenSoftware.htm, Software and Documentation.

Grimaud,M. (1992) A new measure of contrast: dynamics. *Proceedings of SPIE, Image Algebra and Morphological Image Processing III San Diego*, **1769**, 292–305.

Hegde,P., Qi,R., Abernathy,K., Gay,C., Dharap,S., Gaspard,R., Earle-Hughes,J., Snesrud,E., Lee,N. and Quackenbush,J. (2000) A Concise Guide to cDNA Microarray Analysis. *Biotechniques*, **29**, 548–562.

Hirata JR,R., Barrera,J., Hashimoto,R.F. and Dantas,D.O. (2001) Microarray Gridding by Mathematical Morphology. *Proceedings of SIBGRAPI, International Symposium on Computer Graphics, Image Processing and Vision 2001*. IEEE Computer Society, Florianopolis, pp. 112–119.

Maragos,P. (1989) Pattern Spectrum and Multiscale Shape Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 701–716.

Matheron,G. (1975) *Random Sets and Integral Geometry*. Wiley, New York.

Meyer,F. (1998) *The levelings*, Mathematical Morphology and its Applications to Image and Signal Processing, Proceedings of ISMM'98, Kluwer, Dordrecht, pp. 199–206.

Serra,J. (1982, 1988) *Image Analysis and Mathematical Morphology*, Vol I and Vol II: Theoretical Advances, Academic Press, London.

Serra,J. (1999) Course on morphological operators. http://cmm.ensmp.fr/∼serra/cours/index.htm, Centre de Morphologie Mathématique, Ecole des Mines de Paris.

Vachier,C. and Meyer,F. (1995) Extinction value: a new measurement of persistence. *Proceedings of 1995 IEEE Workshop on Nonlinear Signal and Image Processing*, **I**, 254–257.

Vesanen,P., Tiainen,M. and Yli-Harja,Olli (2002) On Calibration-Free Methods in Segmentation of cDNA Microarray Images. *IS&T/SPIE Symposium Image Processing: Algorithms and Systems*, Proceedings of SPIE Vol. 4667, pp. 12.

Vincent,L. (1992) *Morphological Area Openings and Closings for Grayscale Images*, Shape in Picture, NATO Workshop, Driebergen, Holland.

Yang,Y.H., Buckley,M.J., Dudoit,S. and Speed,T.P. (2002) Comparison of methods for image analysis on cDNA microarray data. *J. Comput. Graphical Statistics*, **11**, 108–136.