

A mathematical morphology contribution to the analysis of DNA microarray images

Jesús Angulo and Jean Serra

Centre de Morphologie Mathématique, Ecole des Mines de Paris,
35, rue Saint-Honoré, 77305 Fontainebleau (FRANCE)
{angulo,serra}@cmm.ensmp.fr

Abstract DNA microarrays is an experimental technology which consists of arrays of thousands of discrete DNA sequences that are printed on microscope slides. Image analysis is an important aspect in DNA microarray experiments because the extracted intensities can have a potentially large impact on subsequent data mining steps. Mathematical morphology is a powerful non-linear image analysis technique. In this paper, we present an automatic and fast algorithm for improving the accuracy of spot data extraction from DNA microarrays using mathematical morphology operators.

1 Introduction

DNA microarrays is an experimental technology for identifying and quantifying levels of gene expressions which consists of arrays of thousands of discrete DNA sequences that are printed on glass microscope slides. In order to compare the relative abundance of each of these gene sequences in two DNA samples, the two samples are labelled using different fluorescent dyes; i.e., Cy5 and Cy3 [3].

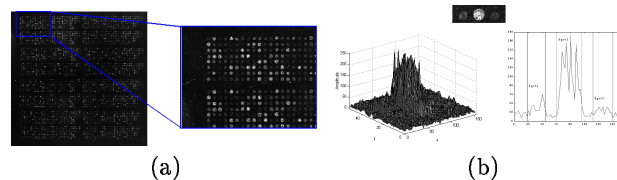


Figure1. (a) Typical image array (size 2200×3000 pixels) and zoom, containing thousands of spots. (b) Variation of intensity in three spots (3D and horizontal cut).

Typical arrays vary from a few hundred to thousands or more spots (each pixel of the image is equivalent to around $10\mu m$), figure 1(a). Image analysis is an important aspect in DNA microarray experiments since the extracted intensities can have a potentially large impact on subsequent data mining steps. The basic goal is to reduce an image of spots of varying intensities into

a table with a measure of the intensity for each spot. The main drawback is the fact that the spots are built using a gridding robot equipped with a series of pins that transfers the small amounts of the DNA. Consequently, the mechanical displacement of the robot may generate some geometric distortion and the impossibility of the global alignment of a template. Other techniques try an automatic spot segmentation using an adaptive intensity thresholding algorithm. However, much important mistakes are caused by the difficult choice of the optimal threshold (see figure 1(b)): the boundary between spot-and-background is not sharp; the contrast (height) between the spot region-and-background and the volume (integral of intensity) are very different from one spot to another; and the non homogeneity of the hybridisation process involves that the spot regions are broken.

A variety of approaches and software tools have been developed for use in processing array images [4] [5] [8] [9].

In this paper, we present our automatic and non-supervised approach for detecting the spot regions and quantifying their associated intensities which relies on mathematical morphology operators.

2 Mathematical morphology

First introduced as a shape-based tool for binary images, mathematical morphology has become a very powerful non-linear image analysis technique with operators capable of handling sophisticated image processing tasks in binary, grey-scale, color and multiresolution imaging modalities. Mathematical morphology is the application of lattice theory to spatial structures. A tutorial in the technique can be found in [7]. This technique is proven to be a very powerful tool in microscopic image analysis.

3 Overview to the algorithm

The morphological approach for processing the microarray images is divided into six sub-algorithms. The full details of these steps are given in [1]. The input is the pair of scanned array images and the output is the intensity associated to every spot for each image. The main steps can be summarised as follows (see figure 2):

Array orthogonal grid Initially a gridding algorithm must yield the automatic segmentation of the further microarray image in subarrays, defining each spot group, which will be individually analysed. Let f_{Cy3} and f_{Cy5} be the Cy3 and Cy5 fluorescent scanned 16-bits images. Our algorithm for image processing requires a single image and it is convenient computationally for the image to be 8-bits. The proposed image f is obtained by means of a *linear weighed by the median values* ν , after a square-root transformation; i.e. $f = (\nu_{Cy3}\sqrt{f_{Cy3}} + \nu_{Cy5}\sqrt{f_{Cy5}})/(\nu_{Cy3} + \nu_{Cy5})$. For

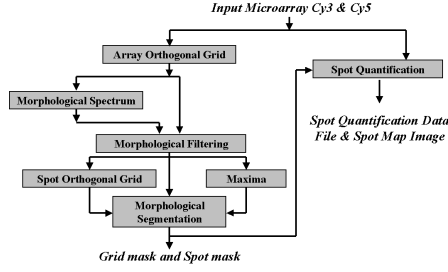


Figure 2. Flow chart of the morphological algorithms for processing the microarray images.

segmenting the spot groups, the image f is reduced in size by *image decimation with averaging* (size $K = 4$); i.e., $\tilde{f} = f \div 4$. Then, the spot groups are enhanced by means of the supremum of a vertical closing φ_n^π and a horizontal closing φ_n^0 ; i.e., $\tilde{f}^\bullet = \varphi_n^\pi(\tilde{f}) \vee \varphi_n^0(\tilde{f})$. The size n of the closing and other size parameters depend on the microarray and below a morphological spectral technique for computing the spot size is given. On the enhanced image, the horizontal and vertical projections are computed. The aim of the subsequent unidimensional morphological filtering is to simplify the projection signal by removing the contribution of spots and background noise. Let $P^{\tilde{f}^\bullet}(i)$ be the horizontal or vertical projection, the signal processing in order to obtain the grid is performed in three steps: *intra-block filtering* using an opening of size n_{ib} ; i.e., $P_{ib}^{\tilde{f}^\bullet} = \gamma_{n_{ib}}(P^{\tilde{f}^\bullet}(i))$; *block extraction* by means of a top-hat of size n_b , which extracts the blocks; i.e., $P_b^{\tilde{f}^\bullet} = P_{ib}^{\tilde{f}^\bullet} - \gamma_{n_b}(P_{ib}^{\tilde{f}^\bullet})$; and *thresholding*, the optimal threshold value u_P is defined as 20% of the average of $P_b^{\tilde{f}^\bullet}$. After thresholding, a binary unidimensional signal (blocks and background) is used for defining the origin $(x_0(j), y_0(j))$ and the dimensions $[xsize(j), ysize(j)]$ for each block j which are extracted from the image f . Then the analysis of the spot group images $\{f^j\}$ is achieved in five steps.

Spot-size distribution law In [2], the morphological extinction spectra (histograms of extrema) which are characterised by three measures (contrast, area or volume) and their ability for the analysis of genome images have been introduced. Based on the area extinction spectrum in logarithmic scale, a new tool is defined: *the spot-size distribution law*, $SS[\lambda, f^j]$, where λ is the spot size (area) and $SS[\lambda, f^j] = n_\lambda$ denotes the normalised number of occurrences at the extinction value λ (a probability density function). The $SS[\lambda, f^j]$ is usually composed of several modes (the interesting mode is the first mode and the other ones can be considered as harmonics of the fundamental frequency). This fundamental mode, or mean spot size, provides the threshold area value for the subsequent filtering.

Morphological filtering by area The contribution of background can be important and introduces mistakes in the building of the grid and in the detection of frontiers of spots. The background noise extraction is obtained using a *morphological filtering by area* γ_n^a . Using this filtering technique by area on the spot group image f^j ; i.e., $f_s^j = \gamma_{\lambda_s}^a(f^j)$, the structures with area greater than the chosen threshold λ_s are preserved in f_s^j . 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 $SS[\lambda]$. Another important advantage of morphological filtering by area opening is the implicit selection of maxima, after this filtering by reconstruction there is one and only one maximum associated to each spot.

Spot orthogonal grid The spots inside a spot group are placed according to an orthogonal alignment and again, using the horizontal and vertical projections the spot grid is obtained. The algorithm for the spot orthogonal gridding is as follows. Let $P(i)$ be the horizontal or vertical projection: (1) calculate the mean value of the elements in $P(i)$; i.e., $\bar{P} = \frac{1}{N} \sum_{i=1}^N P(i)$; (2) subtract the mean from the projection; i.e., $P_\eta(i) = P(i) - \bar{P}$; (3) morphological reconstruction of $P(i)$ using $P_\eta(i)$ as marker; i.e., $P^{rec}(i) = \gamma^{rec}(P(i); P_\eta(i))$; (4) take the residue of the initial projection $P(i)$ and the reconstruction $P^{rec}(i)$; i.e., $P_{TH}(i) = P(i) - P^{rec}(i)$; (5) estimate the optimal threshold value u_P , defined as the $\alpha\%$ of the average of the residue $P_{TH}(i)$; i.e., $u_P = \frac{\alpha}{100} \frac{1}{N} \sum_{i=1}^N P_{TH}(i)$; (6) find the binary reference signal $P_u(i)$ by thresholding process at u_P on the residue signal $P_{TH}(i)$ and using the i middle of each interval equal to 1 in $P_u(i)$, draw the straight lines corresponding to the orthogonal grid. After different test in the database, the choice of $\alpha = 50\%$ appears to be a suitable optimal threshold value.

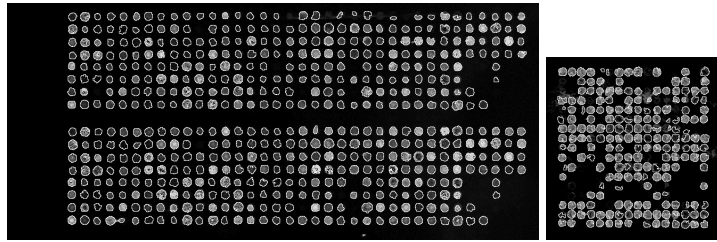


Figure3. Two examples of spot segmentation.

Spot morphological segmentation by watershed transformation The spot segmentation: spot boundary definition, is performed using the watershed transformation: (1) the *function to flood* is the external gradient g^+ , defined as the difference between the dilated image and the original

image; i.e., $g^+(f_{ss}^j) = \delta_B(f_{ss}^j) - f_{ss}^j$ (before that, the filtered image f_s^j is again simplified by a morphological leveling; i.e., $f_{ss}^j = \Lambda(f_s^j)$); (2) the *outer markers* are the filled borders of the grid (background markers) together with the orthogonal grid (spot region markers) and for the *inner markers* a specific algorithm has been developed on the basis of an individual image analysis of the spot bounding boxes defined by the spot grid and the maxima after filtering by area; the global markers to impose are done by $mk = g_r \vee mki$; (3) construction of the *watershed line*, $sm = Wshed(g^+(f_{ss}^j, mk))$, where sm are the line boundaries of each spot. In figure 3, two examples of spot segmentation are depicted.

Spot quantification and noise extraction The segmentation layer sm obtained from the previous algorithms is used on the initial 16-bits images f_{Cy3} and f_{Cy5} for the spot quantification and noise extraction. 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 short, there are two alternatives to study. First, consider that the background noise is uniform on the array; and second, consider that the background noise is not uniform, and therefore a local background estimate is necessary. Obviously, the *global background noise* approach is simpler: quantify only a mean value of noise, whereas that for the *local background noise* model is necessary to quantify the noise in a region for each spot. Based on the Matheron's Geostatistics Theory [6], if we suppose that the image $f(x, y)$ is a regionalised variable, in order to estimate the mean of this variable on the area S as $\mu = 1/S \sum \sum f(x, y)$, the estimated variance has to follow the law $\sigma^2(0/S) = 1/S^2 C$ (variance must vary inversely with the square area). In order to verify this hypothesis, some tests which involve to sample the background using different surface areas of sample and the calculation of the variance have been made (see data example in figure 4). As conclusion, the local variations of the background noise at the scale of the spot size do not allow a global estimate of the average noise. The orthogonal spot grid yields an alternative segmentation: the spot bounding boxes BS_i which are considered the influence regions of spots S_i . These regions can be used for quantifying the local noise associated to each spot. In practice, an enveloping zone of safeguard (to avoid the bias of background) is obtained by the residue of a dilation of the spot region and the noise \tilde{n}_i is estimated in the region $BS_i - \delta_n(S_i)$ which has an area of \tilde{A}_i pixels (the typical size of dilation is $n = 3$). The global expression for the signal intensity is given by $s_i = \hat{s}_i - n_i = \hat{s}_i - \bar{N}_i A_i = \hat{s}_i - \frac{\tilde{n}_i}{\tilde{A}_i} A_i$.

4 Conclusions

We have presented an approach for automatic and robust extraction of the microarray spot data based on mathematical morphology. The described

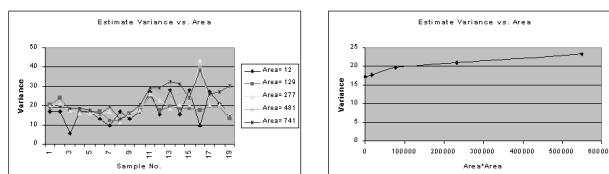


Figure 4. Sampling of background using different surface areas of sample and estimate of variance: left, estimated values of variance and right, square of area against variance (mean values).

methodology allows to solve the main problems of these images: spot shape problems, high global background, locally weak signal, etc. As said above, our approach is entirely automatic and the spot segmentation relies on an adaptive technique in size and position. The experimental evaluated performance of spot segmentation and quantification shows that the use of these algorithms is generally equal or better than the use of conventional manual techniques [1].

References

1. Angulo, J. and Serra, J. (2002a) Automatic analysis of DNA microarray images using mathematical morphology. *Submitted to Bioinformatics*, January 14th 2002 (revised version on August 26th), 50 pp.
2. Angulo, J. and Serra, J. (2002b) Morphological spectrum applied to genome arrays quantification. *Submitted to Journal of Visual Communication and Image Representation*, January 14th 2002, 20 pp.
3. Brown, P. O. and Botstein, D. (1999) Exploring the New World of the genome with DNA microarrays. *Nature Genetics*, Vol. 21 (supplement), pp. 33–37.
4. 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*, Vol. 2, pp. 364–374.
5. Hirata JR, R., Barrera, J., Hashimoto, R. F., Dantas, D. O. (2001) Microarray Gridding by Mathematical Morphology. *Proc. SIBGRAPI, International Symposium on Computer Graphics, Image Processing and Vision 2001*, Florianopolis, IEEE Computer Society, pp. 112–119.
6. Matheron, G. (1975) *Random Sets and Integral Geometry*. Wiley, New York.
7. Serra, J. (1982,1988) *Image Analysis and Mathematical Morphology*. Vol I and II. *Academic Press*, London.
8. 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*, 12 pp.
9. Yang, Y. H., Buckley, M. J., Dudoit, S. and Speed, T. P. (2002) Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, Vol. 11, pp. 108–136.