# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,100
Open access books available

126,000
International authors and editors

145M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities


WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

## Interested in publishing with us?
## Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

# 11

# Robust Microarray Image Processing

Eugene Novikov, Emmanuel Barillot
*Service Bioinformatique, Institut Curie*
*26 Rue d'Ulm, 75248 Paris Cedex 05,*
*France*

## 1. Introduction

High-density microarrays are a rapidly developing technology in molecular biology allowing one to measure simultaneously the activity of thousands of biomolecules in the cell under different experimental conditions. Two-color comparative microarray experiment is a key point of transcriptome (Yang et al., 2002; Herzel et al., 2001; Hegde et al., 2000), CGH (comparative genome hybridization, Pinkel et al., 1998, Ishkanian et al., 2004) and, more recently, protein (Eckel-Passow et al., 2005) microarray technologies.

In a conventional two-color microarray experiment (Fig. 1) two compared samples are labeled using different fluorescent dyes (typically the red-fluorescent dye, Cy5, and the green-fluorescent dye, Cy3), mixed and then co-hybridized to the DNA clones spotted regularly on the microarray. The array is scanned with a high spatial resolution at the corresponding fluorescent wavelengths, and at each scanned pixel the fluorescence intensities are recorded in two color channels (Cy5 and Cy3). The experiment aims to estimate the ratio of the measured intensities for each spot, reflecting differential gene (cDNA technology) or protein expression or a change in DNA copy number (CGH technology) between the test and control samples for the corresponding gene. These ratios are the primary source of information for the subsequent analysis of the microarray data, such as normalization, clustering, classification, differential expression analysis, etc. The main components of the microarray image analysis pipeline for spots include localization, quantification and quality control.

Spot localization involves: (i) identifying the position of each spot on the array to associate it with the spotted clone; and (ii) establishing the borders between the neighboring spots to allow further independent data processing (extracting quantitative information) for each spot. Although spot localization can in principle be done manually, automating this process is essential, as fast and reliable localization increases overall analysis performance and allows high-throughput applications. Many localization algorithms (Buhler et al., 2000; Yang et al., 2002; Jain et al., 2002; Angulo & Serra, 2003; Brändle et al., 2003; Rueda & Vidyadharan, 2006, Ceccarelli & Antoniol, 2006) have been proposed. Some of them require either prior knowledge of some image-specific parameters or direct user participation to find grids. The others are "fully automatic", meaning that different images can be processed without making adjustments for each particular image. However, even for these algorithms, there are always limitations in the automation process because of unpredictable deviations from the assumed array design, high contamination levels or large numbers of missing spots

Open Access Database www.i-techonline.com

Source: Vision Systems: Segmentation and Pattern Recognition, ISBN 987-3-902613-05-9,
Edited by: Goro Obinata and Ashish Dutta, pp.546, I-Tech, Vienna, Austria, June 2007

that cannot be tolerated by the algorithms. In fact, each of the "fully automatic" algorithms has certain limits, and new attempts will never be stopped to push these limits further.

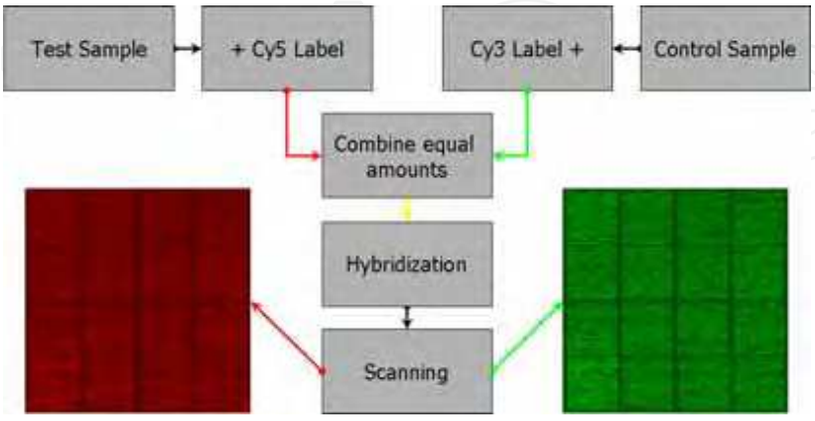


Fig. 1. Two-color comparative microarray experiment.

The aim of the spot quantification is to estimate the ratio. There are two approaches to do that. One is a direct arithmetic ratio of the background-corrected fluorescence intensity estimates in the two color channels (Yang et al., 2002; Bozinov & Rahnenführer, 2002; Angulo & Serra, 2003; Glasbey & Ghazal, 2003; Lehmussola et al., 2006; Axon Instruments, Inc. 2005), and the other is the slope of the linear regression plot of the Cy5 versus Cy3 fluorescence intensities (Jain et al., 2002; Axon Instruments, Inc. 2005). The first approach requires the identification of both the foreground — the measured spot — and the background — typically the level of non-specific hybridization. Large diversity of the algorithms for spot segmentation and background estimation (Lehmussola et al., 2006) highlights the complexity of this problem. The second approach, based on linear regression methods, does not require precise isolation of the spots and identification of the background areas. This method would be rather straightforward, if there were no aberrant or outlier pixels that can strongly affect the slope of the linear regression.

Each ratio estimate should be accompanied by some measure of quality demonstrating the level confidence in the obtained ratios. To determine spot quality we need to have a clear definition of a good spot, or a list of all possible distortions that may spoil the spot. The diversity of instrumental platforms and instrumental and biological factors that may influence the result makes formalization difficult and unlikely to be universal. Several attempts have been made to approach the problem (Buhler et al., 2000; Brown et al., 2001; Wang et al., 2001; Chen et al., 2002; Hautaniemi et al., 2003; Bylesjö et al., 2005). Generally a number of parameters characterizing the spot, such as signal-to-noise ratio, size, circularity, etc., are introduced. These parameters have to be combined into an overall quality value to be used as a confidence level in the follow-up analysis. As individual quality scores generally do not contribute equivalently to the composite quality score, we need to evaluate the weights that control the input of each individual score. For that, training procedures, in which the user classified a set of representative spots into a number of groups ranging from good to bad spots, were proposed (Buhler et al., 2000; Hautaniemi et al., 2003; Bylesjö et al., 2005). This requires an expert to evaluate at least a couple of hundred spots to achieve a good approximation, which is a difficult and time-consuming task.

In this Chapter, we will present a set of advanced algorithms for microarray spot localization, quantification and quality control. We will deal with the rectangular array design. This is the most widespread of the designs used and is also exclusively used within our Institute. In this design, the spots are aligned horizontally and vertically and can be arranged in blocks containing different numbers of spot rows and spot columns. The developed algorithms aim at making analysis more resistant to array contamination and at eliminating user participation at all stages of image processing. The algorithms can be applied to analyze images in one-, two or multi-color microarray experiments. Specific tools have been also developed for ratio evaluation in the two-color comparative experiments.

We present a "fully automatic" spot localization algorithm (Novikov & Barillot, 2006a), which is able to process images of different designs without specific user contribution. We also aimed to make it robust with respect to contamination and missing spots on the array. The developed algorithm is non-supervised and deterministic, ensuring reproducible results. It is assumed that the number of block rows and columns and the number of spot rows and columns within each block are available for analysis as input values.

We have developed a statistical procedure that systematically searches and removes aberrant or outlier pixels (Novikov & Barillot, 2005b). This gives a higher level of confidence in the linear regression ratio estimates. However, as linear regression can give biased estimates when there is a high level of statistical noise (a low correlation between the Cy3 and Cy5 color channels), we still keep estimates from the spot segmentation algorithm. However, after removing aberrant pixels the segmentation algorithm also gives more robust estimates, and there is a greater agreement in the ratio values obtained for both methods.

We have developed a two-level segmentation approach: one intensity level is used to identify spots and the other one separates background areas. Pixels with intensities between these two levels are ignored (buffer zone). We apply the *k*-means adaptive pixel-clustering algorithm (Bozinov & Rahnenführer, 2002) to identify the spot and the background intensity levels. Pixels that are used in the adaptive clustering for the spot and background level estimation are selected from constrained intensity regions. Spot pixels are subject to further geometrical constraints.

We have developed an original set of spot quality characteristics and a model that maps this set into an overall quality value. An automatic training procedure evaluates the contribution of each marginal quality characteristic into the overall quality (Novikov & Barillot, 2005a). This procedure is based on information from replicated spots, located on the same array or over a set of replicated arrays, and assumes that unspoiled replicated spots must have very close intensity ratios, whereas poor spots yield greater diversity in the ratio estimates. Conceptually this approach can be considered as a combination of the "empirical" (based on replicates) and "predictive" (based on quality characteristics) quality assessment methods (Ritchie et al., 2006). The obtained weights can then be used to establish a critical limit for each quality characteristic, such that if a spot's characteristic exceeds its critical limit, the spot is declared a "bad" spot.

The applicability of the developed algorithms has been tested and confirmed using simulated artificial images and experimental images of different array designs used within our Institute and CGH images obtained from the UCSF Cancer Center. These algorithms are included in the software package MAIA (http://bioinfo.curie.fr/projects/maia/), which offers a complete solution for microarray image analysis.

## 2. Spot Localization

As for other automatic spot localization algorithms (Jain et al., 2002; Angulo & Serra, 2003), we take projections of the intensities in the pixel columns on the X (horizontal) axis and in the pixel rows on the Y (vertical) axis. However, instead of taking the overall intensity directly, we correct it by the amount of regularity in the corresponding row or column, so that bright but very irregular regions are systematically penalized. The developed algorithm transforms fluctuations of the intensity in each pixel row or column of the image into a special parameter that takes into account the regularity of these fluctuations.

### 2.1 Spot regularity profiles

**Regularity components**. For each pixel row or column we choose an intensity threshold, $T$, and isolate continuous regions of pixels with intensities, $I_l$, higher than $T$ (bright regions): $I_l > T$, and lower than $T$ (dark regions): $I_l \le T$, $l$=1,…,$m$, where $m$ is the number of pixels per row or column. Each bright region can be characterized by its center position $\mu_n(T)$, length $\lambda_n(T)$ and mean intensity $F_n(T)$. For each dark region we estimate its mean intensity, $B_n(T)$.
We then define four components based on these estimates that contribute to the regularity parameter. The most important component is the overall intensity of the bright regions:

$$S(T) = \frac{1}{N(T)} \sum_{n=1}^{N(T)} F_n(T) - \frac{1}{N_B(T)} \sum_{n=1}^{N_B(T)} B_n(T) \tag{1}$$

where $N(T)$ and $N_B(T)$ are the numbers of bright and dark regions at the threshold level, $T$. The three following parameters deal with the regularity of the bright regions. The first parameter penalizes deviations from the expected spot size, $D$, of the bright regions:

$$W_1(T, D) = \frac{1}{N(T)} \sum_{n=1}^{N(T)} \left( \frac{\lambda_n(T)}{D} - 1 \right)^2 \tag{2}$$

The second parameter ensures that inter-spot distance is not too small. That is, the centers of two bright regions ($\mu_n(T)$ and $\mu_{n+1}(T)$) should not be closer than the expected spot size, $D$:

$$W_2(T, D) = \frac{1}{N(T)} \sum_{n=1}^{N(T)-1} \left( 1 - \frac{\mu_{n+1}(T) - \mu_n(T)}{D} \right)_+^2 \tag{3}$$

where $(x)_+ = x$, if $x > 0$ and $(x)_+ = 0$, if $x \le 0$. The third parameter controls the number of bright regions:

$$W_3(T, H) = \left( N(T)/N(H) - 1 \right)_+ \tag{4}$$

where $H$ is the inter-spot distance and $N(H)$ is the expected number of spots in the corresponding pixel row or column. $N(H)$ can be estimated by dividing the number of row or column pixels by $H$. As we do not expect the number of bright regions to be more than

*N(H)*, this has to be penalized. On the other hand, we cannot impose a lower bound for *N(T)*, as some spots may be missing, but the structure is preserved.

**Overall regularity parameter.** The intensity component (1) and the three regularity components (2), (3) and (4) are combined into an overall regularity parameter:

$$R(T,D,H) = S(T)\exp\left\{-\gamma_1 W_1(T,D) - \gamma_2 W_2(T,D) - \gamma_3 W_3(T,H)\right\} \quad (5)$$

where $\gamma_1$, $\gamma_2$ and $\gamma_3$ are weights determining the contribution of each regularity component. Since all these components are relative quantities, we expect that none will be over-weighted, and hence the weights can be equalized: $\gamma=\gamma_1=\gamma_2=\gamma_3$, where $\gamma$ is provided by the user. In our analysis we always take $\gamma = 2$, and we have had no problems with the localization for different experimental designs. However, the robustness of the analysis would be increased if $\gamma$ (or even $\gamma_1$, $\gamma_2$ and $\gamma_3$) were chosen more specifically.

The threshold level, *T*, can be best determined using a special optimization procedure which searches for *T* from the interval $[I_{min};I_{max}]$ maximizing *R(T,D,H)*:

$$R(D,H) = \max_{T\in[I_{\min};I_{\max}]} R(T,D,H) \quad (6)$$

where $I_{min} = \min(I_l)$ and $I_{max} = \max(I_l)$, $l=1,\ldots,m$. Eq. (6) represents the final expression for the regularity parameter. We then calculate a set of regularity parameters for each pixel row *i* or column *j*, leading to a regularity profile in the Y ($R_i(D,H)$) and X ($R_j(D,H)$) directions.

**Spot size *D* and inter-spot distance *H*.** Although possibly available from the experimental design, spot size, *D*, and inter-spot distance, *H*, are not required as prior values. We assume only that *D* and *H* are related as $D = H(1-\alpha)$, where $\alpha$ is the ratio of the inter-spot gap to the inter-spot distance and should be provided by the user. A very precise value of $\alpha$ is not essential. We always take $\alpha = 0.25$, and it appeared to be very stable with respect to different array designs. As *D* is directly available from *H*, we can omit *D* from the notation of the regularity parameter, so that *R(H)* will be used instead of *R(D,H)*.

We can obtain $H_0$, an initial approximation for *H*, by dividing the total number of pixels in the X or Y direction of the array by the total number of spots in the corresponding direction. This is only a rough estimate, but it is sufficient for building the regularity profiles, $R_k(H)$, where $k = i$ for the Y direction and $k = j$ for the X direction (Eqs. (5) and (6)).

We could have, using the profiles obtained, estimated *D* by dividing the number of pixel rows or columns with high regularity by the total number of spots in the Y or X directions, respectively. However, the spots are almost never perfectly aligned and they can get mixed up and become unrecognizable on the one-dimensional axis irrespective of the cutoff level chosen for the regularity profile. This leads to overestimation of the lengths of the regions with high regularity and consequently to an overestimate of *D*.

If all spots within each block overlapped completely in the projections, we could estimate *H* as the ratio of the number of pixel rows or columns with a regularity higher than the selected level to the total number of spots in Y or X directions, respectively. However, as the spots within a block may, even after projecting pixel rows and columns, be separated by dark gaps, the length of the bright regions, needed to evaluate *H*, may be underestimated. To ensure realistic *H* we overlap the spots by superimposing the given profile with itself shifted to the left or right by a certain number of pixels. Complete overlapping of the

neighborhood spots can be achieved by setting the number of pixels used in the profile shifting to the correct value for the inter-spot distance, *H*. We assume that the neighborhood spots are completely overlapped when the number of dips (regions with a regularity lower than the selected level) in the overlapped regularity profile should not be larger a limit defined as the number of blocks plus one. A small number of dips can indicate that neighboring blocks are also indistinguishable.

We search for the highest level of regularity profile that gives the largest number of dips but not larger than the defined limit. The corresponding *H* is then considered as the final estimate. If number of dips is larger than the defined limit for any level of regularity (and correspondingly for any *H*), then the regularity level giving the lowest number of dips is selected, despite being greater than the defined limit. This situation occurs for relatively bright contamination in the positions where there are no spots according to the array design.

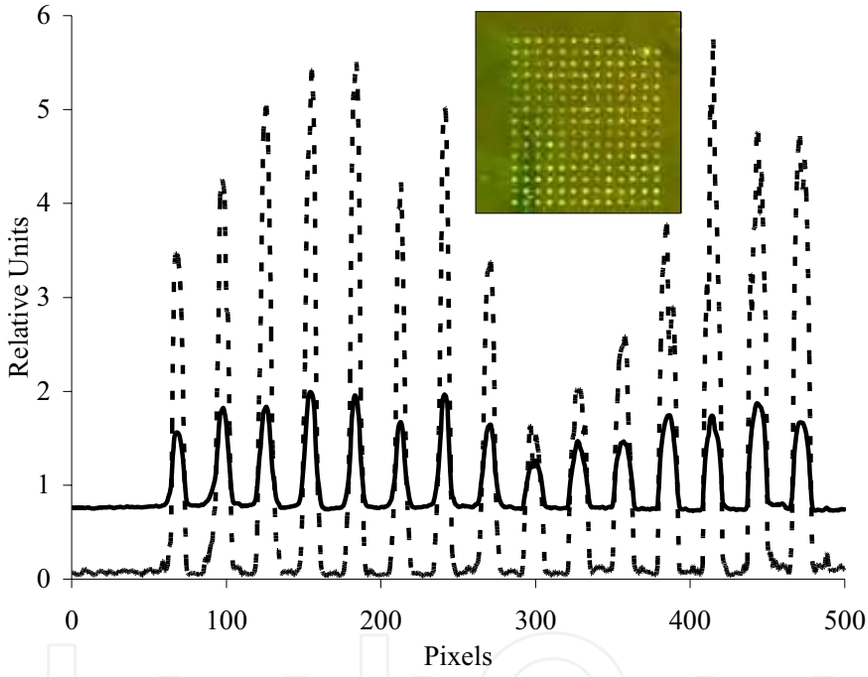


Fig. 2. Intensity (solid line) and regularity (dashed line) profiles for microarray image segment (inset) obtained by projecting on Y axis.

The advantage of using regularity profiles instead of simple intensity profiles is demonstrated in Fig. 2. The regularity profiles (dashed lines) ensure a larger dynamic range (signal to background) than the intensity profiles (solid lines). This leads to better identification of the background regions where it would be expect to find a separation between different spot rows or columns.

Note that each of the approaches that use intensity projections (e.g. Jain et al., 2002; Angulo & Serra, 2003; Brändle et al., 2003) could be reinforced if, instead of simple projections, measures based on the regularity parameter were used.

### 2.2 Generation of the localization grid

**Block separation.** First, we use the regularity profiles to look for the borders between the blocks. To increase robustness, the whole array is divided into segments (Fig. 3). If we need to identify the borders between the blocks in the X direction, we take segments in the Y direction with the height of the segment, in pixels, being equal to the height of the image in pixels divided by the number of blocks in the Y direction (NBY). We identify the block borders in the Y direction by taking segments in the X direction with the width of the segment, in pixels, equal to the width of the image in pixels divided by the number of blocks in the X direction (NBX).

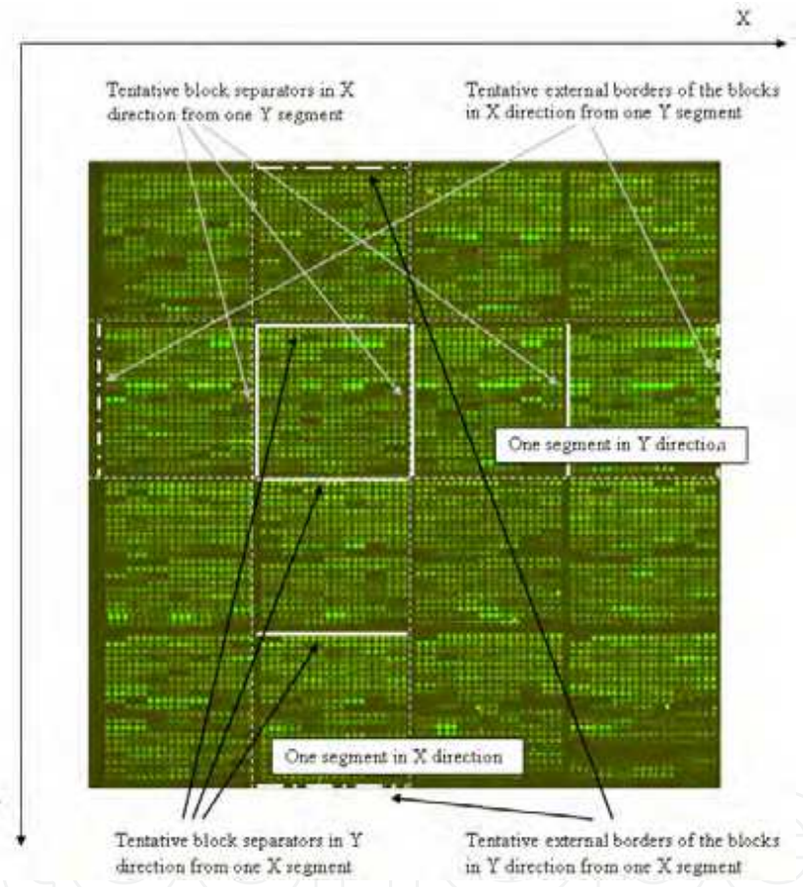


Fig. 3. An example of the separation of the microarray image into segments. There are four sets of tentative block separators and four sets of tentative external borders in the X direction, as four segments (according to the number of blocks) are isolated in the Y direction. Similarly, four tentative block separators and four sets of tentative external borders can be built in the Y direction.

If the blocks are well separated, we can proceed in the following way. For each segment we identify positions separating the blocks by looking for the maximal intervals between the peaks in the regularity profiles. Thus we obtain NBY (in X direction) or NBX (in Y direction)

possible sets of block separations. The best set is the one that has the most regular structure. We calculate the median width of the blocks in NBY sets and the median height of the blocks in NBX sets, and the set, in either the horizontal or vertical separation that gives the smallest deviation from the corresponding median is selected as a final one.

However, this approach is not applicable for arrays where the distance between two neighboring blocks is similar to the distance between the neighboring spots. In this case we take advantage of the fact that the blocks are regularly distributed over the array, and we place the borders equidistant between the external borders of the blocks. These regions have to be long enough to be considered as initial spots in the blocks. We require that the first high-level region must be longer than $\beta D$, where $\beta$ is provided by the user and characterizes the filtering properties on the edges of the array. A default value of $\beta = 0.2$ had been found to be the most relevant for the microarray images of different designs and noise levels that we have tested. The external borders of the blocks are calculated for all segments described above (Fig. 3), and the median estimates are taken. We use two localization iterations to increase the precision of block separation. The first approximation of the grid is used to adjust the borders of the blocks at the second iteration.

**Spot localization.** After blocks are separated, we have to identify the borders between the spots within each block. Although it may appear straightforward to use regularity (or intensity) profiles to draw lines at the positions of minimal regularity of the corresponding profiles to separate the neighboring spots this often results in errors, because the positions of the minima can be due to random regularity fluctuations. Therefore, we have developed a robust procedure searching for the spot separations. It uses the same optimization procedure as for the overall regularity parameter, but instead of the intensity, $I_l$, we use regularity profiles in the X ($R_j(H)$) or the Y ($R_i(H)$) directions. An example of the row regularity profile (Y direction) for a one block (shown in inset of Fig. 2) is given in Fig. 2 in dashed line. Applying a set of criteria represented by Eqs. (1), (2), (3) and (4) for each block we can build up a vertical regularity parameter $R_Y(R_i^*,H)$ (Eq. (5)) using a row regularity profile, $R_i(H)$, and a horizontal regularity parameter $R_X(R_j^*,H)$ (Eq. (5)) using a column regularity profile $R_j(H)$. The parameters $R_Y(R_i^*,H)$ and $R_X(R_j^*,H)$ are dependent on the threshold levels $R_i^*$ and $R_j^*$, and should ensure the highest regularity of the regularity profiles $R_i(H)$ and $R_j(H)$ (see Eq. (6)). However, in difference to Eq. (6), $R_i^*$ in $R_Y(R_i^*,H)$ is determined from the interval between $\min(R_i(H))$ and $\max(R_i(H))$, where $i$ is the row number; and $R_j^*$ in $R_X(R_j^*,H)$ belongs to the interval between $\min(R_j(H))$ and $\max(R_j(H))$, where $j$ is the column number.

Note that the optimized values of $R_Y(R_i^*,H)$ and $R_X(R_j^*,H)$ are of no use in this context. The middle positions of the intervals in the regularity profiles lower than the optimal threshold level are taken as the positions separating spot rows or columns.

## 3. Spot Quantification

After spot localization step, we assume that the spots are identified and well localized in squares (called spot cells), so that each spot cell can be processed independently of the others. We calculate the ratio of the spot using either a linear regression or a segmentation (spot contouring or spot isolation) approach.

### 3.1 Ratio estimation based on linear regression

The linear regression approach represents the ratio as the slope of the linear regression fit of the pixel intensities in two channels (Fig. 4). We use orthogonal regression (Kendall & Stuart, 1979, Dissanaike & Wang, 2003) since measured fluorescence intensities are statistically distorted in both color channels. Spot segmentation is unnecessary with this method, as background pixels are concentrated at the origin of the linear regression plot and do not influence the slope of the regression line (Fig. 4). However, outlier or aberrant pixels within the spot cells, even in small numbers, can strongly influence the regression line, thus biasing the ratio. With the aim to fully exploit the advantages of the linear regression approach we tried to reinforce this procedure by systematically filtering out aberrant pixels.

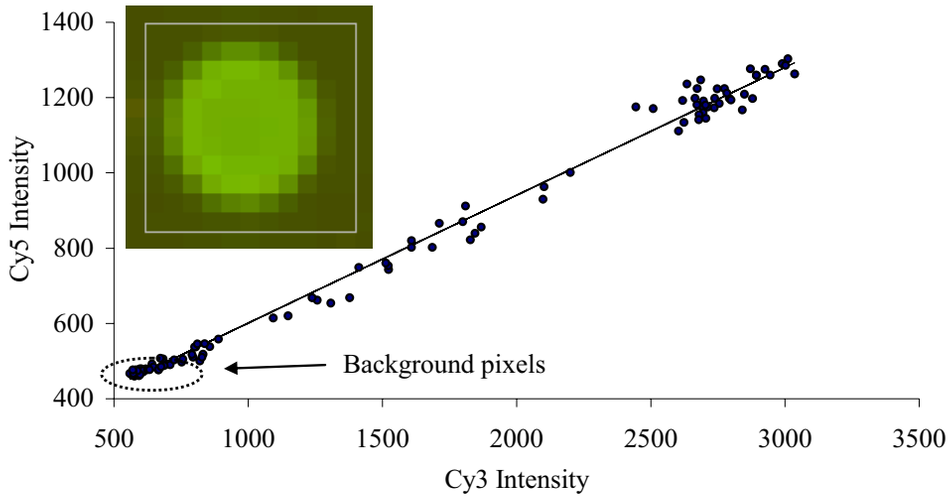


Fig. 4. Estimation of the ratio using linear regression fit for a good spot with a correlation coefficient of 0.99 (ratio = 0.339). The background pixels are grouped near the origin of the linear regression plot.

Different approaches exist to detect statistical outliers in experimental data (Rousseeuw & Leroy, 2003; Atkinson & Riani, 2000). Well-advanced high-breakdown algorithms (Rousseeuw & Leroy, 2003) or forward search algorithms (Atkinson & Riani, 2000) are based on repetitive resampling of experimental data and iterative linear regression approximation. This makes these algorithms computationally infeasible for microarray image analysis, where thousands of spots, each one containing 100-500 data points (pixels), should be processed in seconds. Therefore, we have to look for more approximate algorithms, which, however, can ensure higher efficiency. For microarray images, we expect that the majority of the spots should not have outliers, and the number of outliers for possibly contaminated spots should not be too high. Therefore it would be advantageous to have an algorithm that could quickly identify outlier presence, without being involved in time-consuming iterations. With this aim we have adopted the backward search algorithm with single-case diagnostics (Rousseeuw & Leroy, 2003). The advantage of this algorithm is that if the procedure can not identify an outlier at the first iteration, it proceeds to the next spot, thus saving processing time. Although single-case diagnostics are known to be less efficient

(Rousseeuw & Leroy, 2003) for the data with tight groups of outliers, in our work we rarely had problems: in microarray image, even if several aberrant pixels form a spatial cluster (Fig. 5), they are often very different at the intensity scale (at least in one of two color channels). As outlier intensities are widely distributed, the removal of even one of them changes the quality of the linear regression noticeably, facilitating the one-pixel (or single-case) backward search procedure for spot quantification.

The backward search procedure, in our implementation, examines suspicious pixels by evaluating the quality of the linear regression fit with and without the suspicious pixel. We quantify the fit quality by the residual variance, $s^2$. The smaller $s^2$ is, the closer the linear regression line is to the experimental data. The ratio of the $s^2$ values is calculated for the fit with the tested pixel and for the fit without. If this ratio is larger than a critical value of the *F*-distribution at a user-defined confidence level, the pixel will be marked as aberrant. We select pixels with the highest intensity in either of two channels first and then select pixels having the largest deviation from the fitted regression line. To take into account the fact that the distortions caused by pixels from the top of the intensity scale and by pixels lying off of the linear regression line, may be different, we apply different confidence levels for the *F*-statistics for these pixels. In our analysis we use 0.01 as a confidence level for the pixels from the top of the intensity scale and 0.1 for the pixels lying off of the linear regression line.

For the high-intensity pixels we also perform another test to determine how far their intensities are from the averaged intensity of the other pixels within the spot cell. This detects pixels, far away from the other pixels, that do not distort the linear regression line. Although these pixels may not change the ratio, they could be considered as aberrant pixels, as we expect to see an almost continuous distribution of pixels intensity (Fig. 4). The procedure performs iteratively until no more aberrant pixels are detected.

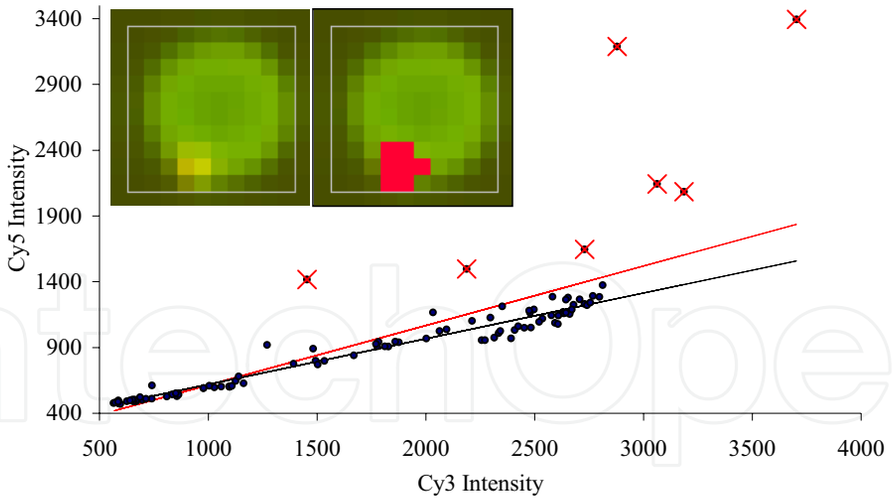


Fig. 5. Estimation of the ratio using linear regression fit for a spot with aberrant pixels (red crosses). The estimated ratio with the aberrant pixels is 0.45 (a), when the aberrant pixels are removed it decreases to 0.37 (b). The estimated ratios for the other two spots from the same triplicate are 0.342 and 0.332.

An example of the outlier detection is presented in Fig. 5. It is important to note that the regression approach is capable of detecting contamination pixels that are geometrically inseparable from the spot. Therefore, the developed procedure can be considered not only as a procedure for correcting ratio recovery, but also as a procedure to repair the spot and to improve the quality of experimental material. It requires, however, that the contamination clearly deviates from the straight regression line, which is defined by the majority of "good" pixels from the spot. The filtering procedure can detect up to ~30% of aberrant pixels with respect to the number of spot pixels. For the spots with larger number of aberrant pixels, a safer way would be to flag out these spots rather than to try to identify all aberrant pixels. Besides much higher computational complexity (and hence processing times), high-breakdown filtering algorithms may have difficulties to distinguish between contaminating pixel clusters and useful spots, when these become comparable in size, and contamination is highly correlated in two color channels.

One potential problem of linear regression approach is when one image (Cy3) is shifted relative to the other (Cy5). As this shift increases, the correlation between the two channels decreases rapidly, and linear regression fit becomes poorly defined. To solve this problem we have developed a special procedure for the automatic identification and removal of shift between two images. The procedure moves one image with respect to the other one to obtain the largest correlation coefficient for a number of representative spots. These spots are selected according to two criteria: they should be bright enough, but not beyond the dynamic range of the registered intensities; and they should not contain pixels a lot brighter than most of the pixels in the corresponding spot cell.

### 3.2 Ratio estimation using spot segmentation

The spot segmentation approach identifies spots and background areas. The ratio is then defined as

$$r = (F_{Cy5}-B_{Cy5})/(F_{Cy3}-B_{Cy3}) \quad (7)$$

where $F_{Cy5}$($F_{Cy3}$) is either the mean or median estimate of the spot intensity in the Cy5(Cy3) channel, and $B_{Cy5}$($B_{Cy3}$) is either the mean or median estimate of the background intensity in the Cy5(Cy3) channel.

We have developed a multi-level segmentation approach where a segmentation algorithm is first applied to isolate spots and then to identify background pixels. The algorithm is applied to the combined image: $F_i = F_i^{Cy5}A_{Cy5}+F_i^{Cy3}A_{Cy3}$, where $F_i$ is the combined intensity of the $i$-th pixel, $F_i^{Cy5}$($F_i^{Cy3}$) is the intensity of the $i$-th pixel in the Cy5(Cy3) color channel, and $A_{Cy5}$ and $A_{Cy3}$ are the normalization constants: $A_k = \min(M_{Cy5},M_{Cy3})/M_k$, $k$ = {Cy5,Cy3}, where $M_{Cy5}$($M_{Cy3}$) is the mean intensity of the pixels located along the borders of the given spot cell in the Cy5(Cy3) color channel.

The spot is isolated by establishing the signal level, $L_s$, such that all pixels with intensities higher than $L_s$ will be classified as potentially belonging to the spot. We used the $k$-means adaptive pixel-clustering algorithm (Bozinov and Rahnenführer, 2002) to do this. However, we had problems when this algorithm was applied to segment spots with relatively smooth edges. Some pixels may be clearly brighter than the background, but not bright enough to be included into the spot. To regularize the solution, we establish an intensity limit, $U$, such that only pixels with the intensities higher than $U$ participate in the spot segmentation.

We use Chebyshev's inequality (Fisher & van Belle, 2003) to define $U$ as $M+W/(1.35p^{½})$, where $p$ is a user-defined confidence level for the intensity distribution of background

pixels, $M$ is the median and $W$ is the inter-quartile distance of pixel intensities located along the borders of the given spot cell (these pixels are expected to be purely background pixels). Then pixels with the intensities higher than $U$ are classified according to the $k$-means adaptive pixel clustering algorithm to estimate $L_s$.

After selecting the bright pixels some geometrical constraints need to be imposed. We define a spot circle, centered on the center of mass of all the bright pixels from the given spot cell, with the radius $(0.5Z/\pi)^{1/2}$, where $Z$ is the number of pixels with intensities higher than $L_s$. If it turns out that the number of bright pixels within the circle is relatively small ($<0.5Z$), we increase the radius by one until the number of pixels covered by the circle becomes equal or higher than $0.5Z$. For spots with a circular shape it should happen at the first trial. More attempts are needed for spots with more peculiar shapes (e.g. donut-like). The bright pixels within this circle are considered as belonging to the spot. All other bright pixels in the same spot cell are considered as potential space outliers. Further steps resemble the seeded region growing (Yang et al., 2002). The space outliers are converted into spot pixels only if one of their neighbors is already a spot pixel. It performs iteratively building up a cluster of bright pixels, which are geometrically inseparable from the originally defined spot pixels. These pixels constitute a spot and the remaining bright pixels are considered as space outliers that should be ignored during further analysis.

Spot pixels with excessively high or low intensity with respect to the majority of spot pixels can also be discarded. The admissible range is defined as "median of spots pixels" ± "inter-quartile distance of spot pixels"/$(1.35p^{½})$, where $p$ is a user-defined confidence level for spot pixels. This filtering is appropriate for flat spots with large amount of pixels.

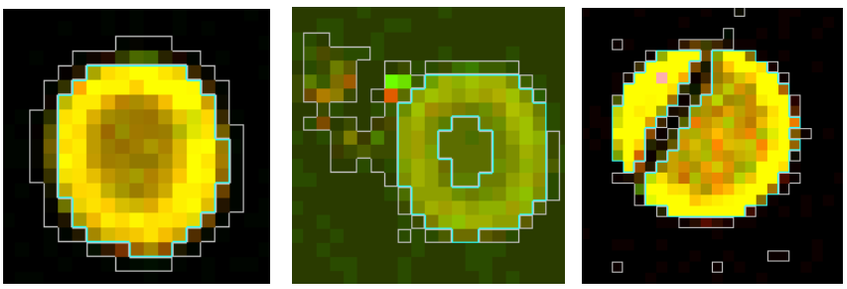

Fig. 6. Segmentation examples: pixels within turquoise contours represent spots and pixels outside gray contours represent background areas.

Finally, we identify the areas used to calculate the background levels, $B_{Cy5}$ and $B_{Cy3}$. The different approaches for calculating the background vary considerably (Yang et al., 2002; Bozinov & Rahnenführer, 2002; Bengtsson & Bengtsson, 2006; Axon Instruments, Inc. 2005). We search for the background level, $L_b$, such that all pixels with intensities lower than $L_b$ are classified as background, and pixels with intensities from the interval $[L_b;L_s]$ comprise the buffer zone ignored in further quantification. $L_b$, in our implementation, is estimated from the $k$-means adaptive clustering applied to pixels with intensities from the interval $[M;U]$. This procedure identifies background areas within a spot cell. Similar to (Axon Instruments, Inc. 2005) the background estimates, $B_{Cy5}$ and $B_{Cy3}$, are taken from all background areas within approximately two spot-cell-size regions centered at the current spot.

Several examples of segmentation for spots of different shapes and geometries are shown in Fig. 6. As one can see, the developed algorithm is able to produce predictable contours for broad range of different spots.

### 3.3 A combined approach to unique ratio estimation

The performance of the linear regression approach depends on the level of statistical noise in the detected images and hence on the level of correlation between two (Cy3 and Cy5) color channels. For images with a high correlation coefficient (~0.90), the linear regression approach is often better than the segmentation approach, and filtering is more effective, as any contamination is better recognized by the linear regression fit. For noisier images, the regression approach is less efficient in filtering and may also produce biased estimates. For such images, the segmentation algorithm generally demonstrates better performance.

A general strategy to estimate the ratios can be composed of two steps. First, linear regression filtering is applied to each spot. This removes aberrant pixels for highly correlated signals, and leaves the data largely unaltered for noisy images. Then segmentation approach is used for the final ratio estimation according to Eq. (7), where $F_{\{Cy5,Cy3\}}$ and $B_{\{Cy5,Cy3\}}$ are the mean estimates for the spot and background intensities, respectively. Mean estimates are more precise (Fisher & van Belle, 2003), but can be affected by outliers. However, as the outliers have been already removed by the linear regression filtering, we can use the mean values. Although estimation using the segmentation estimator may be not as good as the linear regression estimator for highly correlated spots, the difference is generally so unimportant that we can sacrifice some quality for generality. We call this two-step algorithm the regression filtered segmentation estimator (RFSE).

In general, the idea to perform preliminary filtering of microarray images is not new. There have been a number of publications reporting application of the median filter (Glasbey & Ghazal, 2003), top-hat filter (Yang et al., 2002; Glasbey & Ghazal, 2003) or a set of morphological operators (Angulo & Serra, 2003). However, all these techniques, while reducing noise in images, also change intensity levels of the majority of pixels on the array, regardless of whether these pixels are outliers or not. For example, existing filtering procedures may dissolve micro-cluster of aberrant pixels (like the one shown in Fig. 5), so that it will not be seen any more. However, exceptionally high intensities from the outlier cluster will implicitly influence the intensity of both, the neighboring "good" pixels and the new "good" pixels that will substitute the outliers. This may result in biased intensity and ratio estimates. Contrary to that, our approach specifically eliminates outlier pixels, otherwise not distorting data. It also allows for visual examination of the contaminating pixels to evaluate sources of possible problems in microarray experiment.

## 4. Spot Quality

Each ratio estimate should be accompanied by some value of quality reflecting the level confidence in the obtained ratios. This value is derived from a set of quality characteristics generated by spot quantification procedures (linear regression and spot segmentation).

### 4.1 Spot characterization by quality parameters

The generated quality characteristics ($x$) may be defined on any domain, but we scale them ($q(x)$) to fit the range between 0 (bad spot) and 1 (good spot). This facilitates further quality analysis. For scaled quality characteristics we use another term: quality parameters.

**Coefficient of determination** ($CD$) of linear regression signifies the degree of linear relationship between the intensities in the Cy3 and Cy5 channels. High values of $CD$ (approaching 1) are expected for good spots. Low values suggest either relatively bright but non-correlated contamination, or strong statistical noise normally characterizing low-level (or missing) spots. $q(CD) = CD$.

**Durbin-Watson statistic** ($DWS$) evaluates the presence of the first-order autocorrelation in the residuals of the linear regression fit. It ranges from 0 to 4, 0 being a positive correlation and 4 being a negative correlation. A $DWS$ value close to two indicates that the residuals are uncorrelated and the model is appropriate. Large deviations from two, resulting from systematic patterns in the residuals plot suggest that the spot cannot be modeled in terms of a simple linear regression. $q(DWS) = 1-|DWS-2|/2$.

**Spot contamination** is the number ($SC$) of the aberrant pixels (within the spot contours) flagged out by the filtering procedure. $q(SC) = 1-SC/Z$, where $Z$ is the number of pixels within the spot contour.

**Diameter** of the spot: $D = 2(Z/\pi)^{1/2}$. As the true value for the spot diameter may be difficult to establish, we use a typical value taken as the median diameter over all spots on the array. Spots with exceptionally small or large diameters should be penalized. $q(D) = exp\{D_T-D\}$, if $D>D_T$ and $q(D) = exp\{D-D_T\}$, if $D<D_T$ where $D_T$ is the typical diameter.

**Geometrical symmetry** parameter measures deviation of the contoured spot from the ideal circle. We divide both the real spot and the ideal circle into eight segments (pie slices defined as $[k\pi/4;(k+1)\pi/4]$, $k = 0,\ldots,7$) and we count the number of pixels belonging to the spot ($Z_{si}$, $i = 1,\ldots,8$) and to the circle ($Z_{ci}$, $i = 1,\ldots,8$) for each segment. The sum of the absolute relative differences $GS = \sum|Z_{si}-Z_{ci}|/Z_{ci}$ is then taken as an indicator of quality. For ideal circular spots $GS$ should approach 0, whereas highly deformed (un-circular) spots can be recognized by high $GS$ values. $q(GS) = exp(-GS)$.

**Intensity symmetry** of the spot is defined as $IS = \sum|F_i-F|/F$, where $F_i$, $i = 1,\ldots,8$ are the mean intensities for the same 8 segments and $F$ is the mean intensity within the spot. Although a spot may have perfect circular shape, it may contain very bright (or dark) and highly concentrated groups of pixels originating from pieces of dust or other contamination. $q(IS) = exp(-IS)$.

**Coefficient of variation of two ratio estimates:** $CVR = 2^{1/2}|RR-RS|/(RR+RS)$. Despite the different methods of ratio estimation (one by the linear regression approach ($RR$), and the other by the segmentation algorithm ($RS$)), the variation between the two obtained ratios should be as small as possible. Large variations between the two estimates may indicate a problematic spot. $q(CVR) = exp(-CVR)$.

**Uniformity of the background** along the grid lines separating neighborhood spots is defined as $UB = \sum|B_i-B|/B$, where $B_i$, $i = 1,\ldots,8$ are the mean intensities in 8 segments of the grid line around the spot, and $B$ is the mean intensity for the whole grid line around the spot. Large $UB$ values may discover presence of relatively bright contamination around the spot, large variability in the background or merged neighboring spots. $q(UB) = exp(-UB)$.

**Absolute level of background** ($AB$) calculated from the local area around the spot ($AB = \max(B_{Cy5},B_{Cy3})$) is compared to the median background level over all spots on the array.

Spots with exceptionally high *AB* values may indicate the presence of the contamination areas, which are larger than the size of the spot. $q(AB) = exp(1-AB/AB_T)$, if $AB>AB_T$ and $q(AB) = exp(AB/AB_T-1)$, if $AB<AB_T$. where $AB_T$ is the typical background level.

**Signal** (*S*) is defined as $S = \min(F_{Cy5} - B_{Cy5}, F_{Cy3} - B_{Cy3})$. $q(S) = 1$, if $S>S_T$ and $q(S) = exp(S/S_T-1)$, if $S<S_T$, where $S_T$ is the median signal over all spots on the array.

The developed quality parameters, although not optimal, have led to reasonable results for most of the experimental and simulated situations we tested. Of course, there may be a possibility to formalize some of these parameters more precisely and/or to develop new parameters accounting for other types of distortions.

### 4.2 Spot quality analysis

We consider two aims of spot quality analysis. The first is to combine the marginal quality parameters into an overall quality value. This value can be used either to flag out directly spots with a quality lower than a user-defined threshold, or, in the follow-up image analysis procedures (normalization, classification, clustering, etc.) as a parameter characterizing the level of confidence in the obtained Cy5/Cy3 ratios. The second aim is to identify a critical range for each quality characteristic. If a certain quality characteristic of the spot falls in this range, the corresponding spot is classified as a "bad" spot.

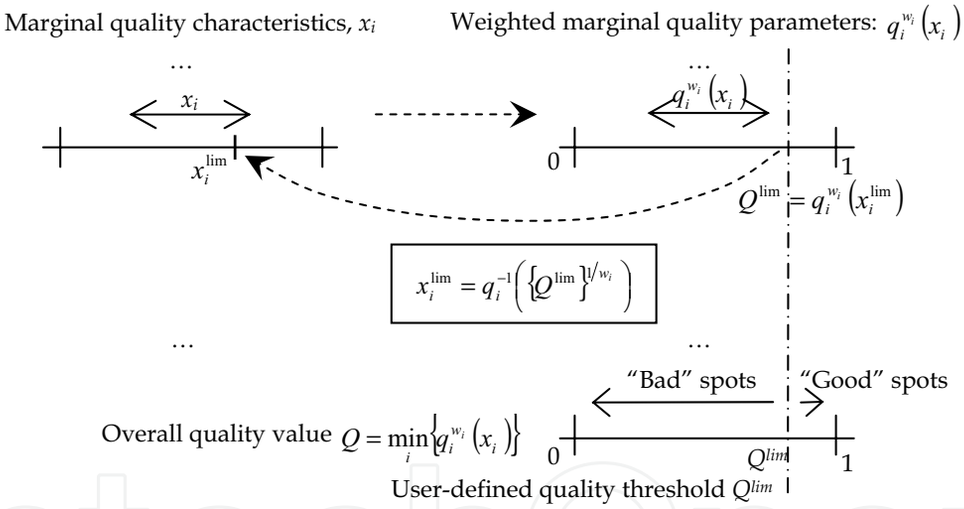


Fig. 7. The correspondence between the quality characteristics, quality parameters and overall quality value.

**Overall quality.** We used the following definition for the overall quality value:

$$Q = \min_i\{q_i^{w_i}\}, \tag{8}$$

where $q_i = q_i(x_i) \in [0;1]$ are the marginal quality parameters for $x = \{CD, DWS, SC, D, GS, IS, CVR, UB, AB, S\}$ and $w_i$ are the weights that control the input of the corresponding quality

components into the overall quality value. A link between the weight $w_i$ and the critical value $x_i^{lim}$ can be established for each quality characteristic:

$$w_i = \ln\{Q^{\lim}\} / \ln\{q_i(x_i^{\lim})\} \text{ or } x_i^{\lim} = q_i^{-1}\left(\{Q^{\lim}\}^{1/w_i}\right) \quad (9)$$

where $Q^{lim} \in [0;1]$ is the user-defined overall quality threshold, and $q_i(x_i^{lim})$ is the quality parameter calculated for $x_i^{lim}$. The critical value $x_i^{lim}$ sets up the limit such that if a certain characteristic $i$ exceeds this limit, the corresponding quality parameter $q_i(x_i^{lim})$ will become lower than $Q^{lim}$. The correspondence between $x_i$, $x_i^{lim}$, $q_i(x_i)$, $q_i(x_i^{lim})$, $w_i$, $Q$ and $Q^{lim}$ is demonstrated in Fig. 7.

**Quality weights $w_i$.** The experimental quality parameters, $q_i$, are directly available from the quantification procedure, whereas the weights $w_i$ (or the critical values $x_i^{lim}$) are unknown and are not easily guessed or derived from theory. Therefore, the problem of spot quality analysis becomes a problem of weights ($w_i$) estimation. This can only be solved if additional information is available. Here we consider three possibilities:

1. The additional information may come, for example, from the user expertise. The user has to classify the spots manually (Buhler et al., 2000; Hautaniemi et al., 2003; Bylesjö et al., 2005) and assign a quality value to each spot from a representative subset. These values are then used for training the model (8) leading to a combination of the weights ($w_i$) such that the overall quality values reproduce the user classification reasonably well.
2. We can manually apply different combinations of the weights $w_i$ and visually appreciate, which spots have been flagged out. The trials must be continued until most of the user classified "bad" spots are eliminated by the chosen combinations of the weights.
3. The weights can be estimated automatically using information available from replicated spots on the same array or over a set of replicated arrays. Unspoiled replicate spots should have very similar ratio values. Large differences between the observed ratios in the replicate spots would signal that some spots from this replicate were irregular. We formalize this approach by first defining the quality value for the replicate:

$$Q_k = \min_j\left\{\min_i\left\{q_{kji}^{w_i}\right\}\right\} \quad (10)$$

where $q_{kji}$ is the $i$-th quality parameter of the $j$-th replicated spot in the $k$-th replicate. Then we require that the ratio variation coefficient in the $k$-th replicate, $V_k$, is proportional to the logarithm of $Q_k$:

$$V_k \sim -\ln\left[\min_j\left\{\min_i\left\{q_{kji}^{w_i}\right\}\right\}\right] \quad (11)$$

The log transform is the most "natural" way to convert [0;1] scale of $Q_k$ into [0;∞) scale of $V_k$. Finally, exponential transform of Eq. (11) yields

$$\exp\left(-V_k / V\right) = \min_j\left\{\min_i\left\{q_{kji}^{w_i}\right\}\right\} \quad (12)$$

where $V$ is the user-defined characteristic ratio variation coefficient. The weights $w_i$ can be estimated from the best fit of the experimental quality values $Q_k$ to the exponentially transformed ratio variation coefficient $V_k$ (Novikov & Barillot, 2005a). If certain quality factors do not influence the shape of the experimental quality curve $Q_k$ (Eq. (10)), the corresponding weights will be set close to 0. If a certain effect shows up in only a small number of spots, it may be neglected by the optimization procedure, and the corresponding weight will be erroneously small. In this case, manual correction of the weights would be necessary.

In our quality analysis algorithm, user participation is limited to the definition of the characteristic ratio variation coefficient, $V$. This is somewhat simpler than deciding on the quality of several hundred spots, which is used to teach the algorithm in the manual approach. However, as with other solutions, this algorithm requires representative images to train the model. It is impossible to evaluate confidently the weight of the contribution of the diameter quality parameter, for example, if all spots in the array have the same diameter. Therefore, a careful selection of training images containing a realistic diversity of all possible distortions and artifacts is needed.

In (Novikov & Barillot, 2005a) we have also demonstrated possibilities to perform quality analysis based on replicated spots from different arrays and a possibility to apply quality weights obtained from the analysis of one training image, which should contain replicated spots, to other arrays, which may not contain replicates. The latter example attempts to reproduce an important possibility of designing microarray experiments. A small number of training arrays with replicated spots and representative diversity of possible artifacts can be measured and analyzed. The obtained results can then be used to evaluate the quality of other arrays of similar design, which may not contain replicated spots.

**Follow-up image analysis.** As it was mentioned earlier, the overall quality value, $Q$ (Eq. (8)), can be used as a parameter characterizing the level of confidence in the obtained Cy5/Cy3 ratios. If, for example, $n$ ratios should be averaged, the weighted mean would ensure a more robust estimate for the average:

$$r = \sum_{l=1}^{n} Q_l r_l \Big/ \sum_{l=1}^{n} Q_l \tag{13}$$

where $r_l$ is the Cy5/Cy3 ratio and $Q_l$, is the corresponding overall quality value ($l$ = 1,…,$n$). The weighted coefficient of variation is defined as

$$V = \frac{1}{r}\sqrt{\sum_{l=1}^{n} Q_l \left(r_l - r\right)^2 \Big/ \sum_{l=1}^{n} Q_l} \tag{14}$$

Note that the ratio variation coefficient $V_k$ can be determined from Eq. (14), if we set $Q_l$ = 1, $l$ = 1,…,$n$, with $n$ being the number of spots in a replicate.

## 5. Testing image processing algorithms

### 5.1 Image Simulation

In (Novikov & Barillot, 2005b) we have described a software component for Monte-Carlo simulation of microarray images. The simulator accounts for statistical noise and different types of distortions, such as non-specific hybridization and dust. As the values of the ratios are exactly known in the simulation experiments, it allows us to test and compare

objectively different ratio estimation algorithms. The general model for the two-color (Cy3, Cy5) microarray image is given by:

$$F_{Cy3}(i,j) = \sum_{k=1}^{N_S} g\left(i, j, c_k^{sx}, c_k^{sy}, \rho^s, I^s\right) + \sum_{k=1}^{N_D} g\left(i, j, c_k^{dx}, c_k^{dy}, \rho^d, I^d\right) \tag{15}$$

$$F_{Cy5}(i,j) = r\sum_{k=1}^{N_S} g\left(i, j, c_k^{sx}, c_k^{sy}, \rho^s, I^s\right) + \sum_{k=1}^{N_D} g\left(i, j, c_k^{dx}, c_k^{dy}, \rho^d, I^d\right) \tag{16}$$

where $N_S$ is the number of spots and $N_D$ is the number of dust clusters, $c_k^{sx}$ and $c_k^{sy}$ are the coordinates of the center of a spot, $c_k^{dx}$ and $c_k^{dy}$ are the coordinates of the center of a dust cluster, $\rho^s$ and $\rho^d$ are the approximate radiuses of the spot and dust cluster, respectively, $I^s$ and $I^d$ are the fluorescence intensity in the center of the spot in the Cy3 color channel and in the center of the dust cluster, respectively, and $r$ is the ratio of the test and control samples. Dust is represented by the random distribution over the array of clusters of pixels of varying brightness. We consider that these pixel clusters have an identical shape to the spots and therefore the same analytical representation is used for an ideal spot shape and dust cluster:

$$g(i,j,c^x,c^y,\rho,I) = I\exp\left(-\left\{\left(i-c^x\right)^4 + \left(j-c^y\right)^4 + \left(i-c^x\right)^2\left(j-c^y\right)^2\right\}\Big/2\rho^4\right) \tag{17}$$

The parameters characterizing the spots ($c_k^{sx}$, $c_k^{sy}$, $\rho^s$, $I^s$ and $R$) are user-defined. For example, the coordinates $c_k^{sx}$ and $c_k^{sy}$, the radius $\rho^s$ and the ranges for $x$ and $y$ for each spot are defined from a user-defined array design. The user should also specify the number of dust clusters $N_D$ on the array. The other parameters characterizing the dust are random variables, and the probability laws for their generation is a matter of choice. We use uniform distributions for $\rho^d$ (in the interval 0 to $\rho_m$) and $I^d$ (in the interval 0 to $I_m$), where $\rho_m$ and $I_m$ are a user-defined maximal dust cluster radius and maximal dust intensity, respectively. We also assume that $c_k^{dx}$ and $c_k^{dy}$ are uniformly distributed over the array. Statistical laws of the dust characteristics can generally be different in the two (Cy3, Cy5) channels.

In the developed simulation model we also account for the nonspecific hybridization and statistical noise:

$$\widetilde{F}_k(i,j) = F_k(i,j) + B_k + \eta_{Bk}B_kG_B + \sigma(i,j)G_S \tag{18}$$

where $k$ represents either Cy3 or Cy5, $B_i$ and $\eta_{Bk}$ are the user-defined average and noise-to-signal ratio of nonspecific fluorescence intensity in the color channel $k$, $\sigma(i,j)$ is the standard deviation of the pixel statistical noise, and $G_B$ and $G_S$ are independent Gaussian random variables with zero mean and unit standard deviation. The exact representation for $\sigma(i,j)$ is defined by the experimental set-up. There are currently three possibilities: $\sigma(i,j)$ can be (i) constant, (ii) proportional to the signal, or (iii) proportional to the square root of signal. The type and quantitative characteristics of the statistical noise are defined by the user.

### 5.2 Evaluation of the noise resistance using artificial images

All artificial images were generated using the same array design: 4x12 blocks and 21x21 spots within each block with the inter-spot distance of 15 pixels and the inter-block gap of 20 pixels. For all spots in the generated arrays the spot radius, $\rho^s$, was about 4 pixels, the intensity, $I^s$, in the Cy3 color channel was 5000 and the ratio, $r$, of the Cy5 and Cy3 channels was 3. Non-specific hybridization was generated using $B_k$ = 1000 and $\eta_{Bk}$ = 0.5. The standard

deviation of the statistical noise, $\sigma(i,j)$, at each pixel was proportional to the signal at the corresponding pixel with the noise-to-signal ratio of 0.1. We also added randomly distributed dust clusters with the maximal intensity, $I_m$ = 65535, and maximal radius, $\rho_m$ = 2 pixels. Generated images differ in the number of dust clusters, $N_D$.

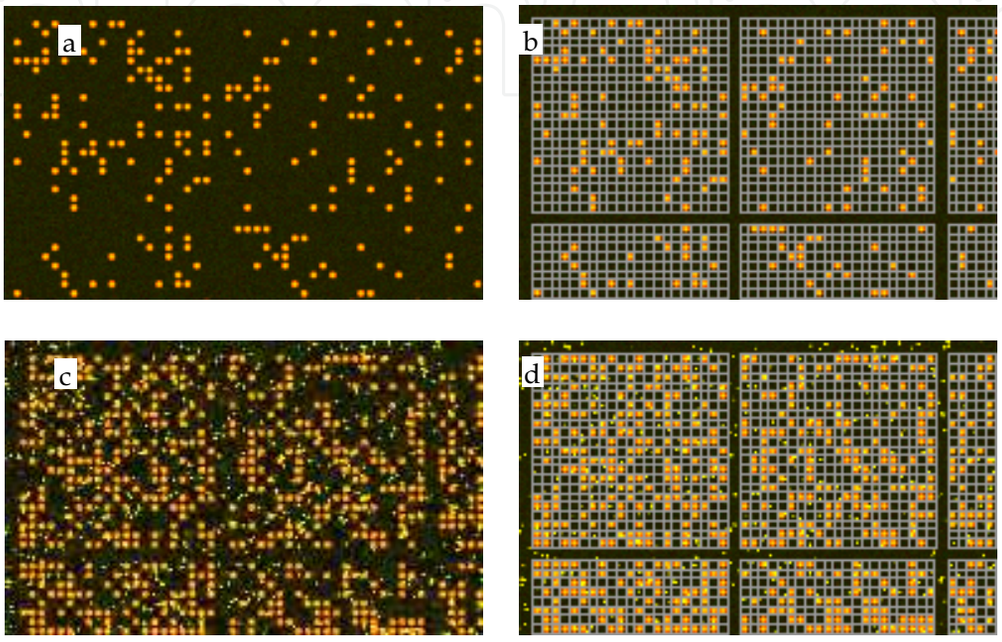


Fig. 8. Fragments of artificial microarray images with 4x12 blocks and 21x21 spots per block: a) the fraction of the bright spots is equal to 15%; no contamination; b) the same image with the generated grid; c) randomly distributed contamination spots are added; the percentage of the bright correct spots is 40% and the number of the contamination spots is equal to the number of the correct spots ($NS = ND$); d) the same image with the generated grid.

**Localization.** We studied the influence of the amount of bright (visible) spots and the level of contamination on the spot localization. Two exemplary artificial images are presented in Fig. 8. One (Fig. 8a) containing only 15% of bright spots randomly distributed over the image, and the other one (Fig. 8c) with randomly distributed contamination spots. For the contaminated array, and the number of dust clusters was equal to the number of true spots ($N_S = N_D$).

Grid placement depends on the distribution of the spots over the array. Therefore, we generated 100 images, each with a random spot distribution, and counted the amount of grids that needed user intervention. For the images without contamination, only 10 of 100 images gave misplaced grids. This happened when first or last spot rows or columns are empty, so that the algorithm shifted the grid by one row or column. For contaminated images grid misplacement occurred in 7 of 100 images. This took place when false spots were recognized as the real spots by the algorithm. Examples of the correctly generated grids in both cases are given in Figs. 8b and 8d.

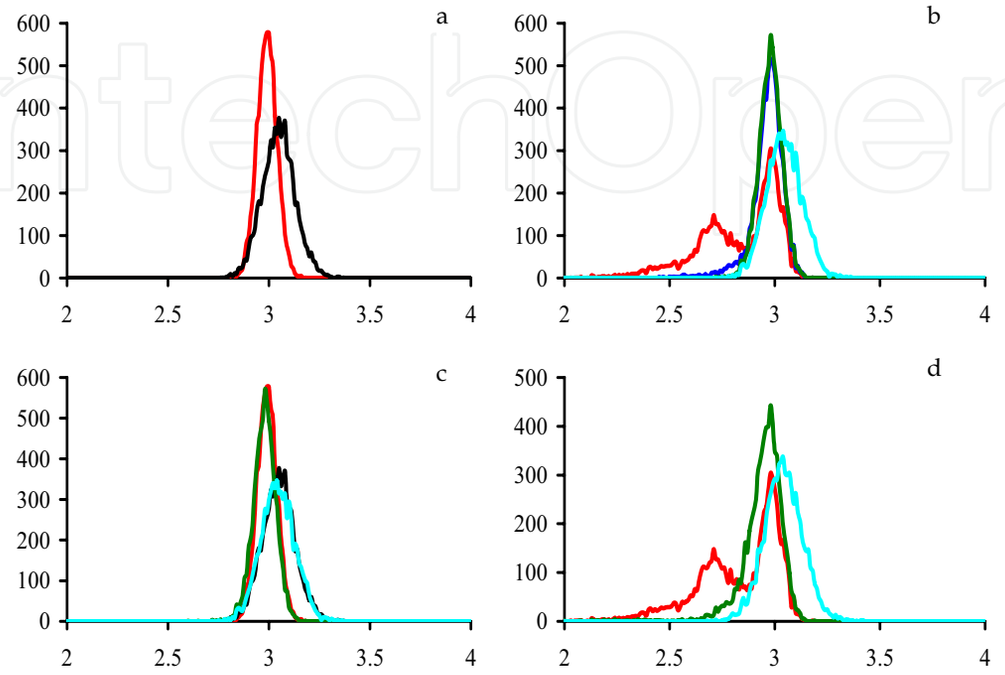


Fig. 9. Histograms of the ratio estimates: a) ratio of means (red) and ratio of medians (black) for the dust-free image; b) ratio of means (red), un-weighted RFSE (blue), weighted RFSE (green) and weighted ratio of medians (turquoise) for the contaminated image; c) weighted RFSE (green) and weighted ratio of medians (turquoise) for the contaminated image, ratio of means (red) and ratio of medians (black) for the dust-free image; d) ratio of means (red), weighted ratio of means (green) and weighted ratio of medians (turquoise) for the contaminated image.

**Quantification and Quality. W**e investigated the influence of the level of contamination on the spot quantification. We used the same array design as before (Fig. 8) with one exception: all true spots were bright and visible. We compared RFSE ratio with the ratio (7) where $F_{\{Cy5,Cy3\}}$ and $B_{\{Cy5,Cy3\}}$ are either the mean (ratio of means) or median (ratio of medians) estimates. We also compared the weighted and un-weighted mean estimates for the average $r$ (Eq. (13)). The un-weighted characteristics were obtained from Eq. (13) by setting all $Q_l$, $l$ = 1,…,$n$ to 1. The weighted characteristics were calculated with the overall quality values $Q_l$ available from the quality analysis algorithm. As all spots from the simulated image can be considered as replicates, having the same theoretical ratio ($r$ = 3), we artificially split up the total number of spots into the groups of three closely placed spots. These groups, regarded as independent triplicates, can be used to calculate the experimental quality values $Q_k$ (Eq. (10)) and to build up the corresponding quality plot, $Q_k$ *versus* $V_k$, according to Eq. (12). The weights $w_i$ are estimated from the best fit in Eq. (12). For each group we calculated the weighted and un-weighted means of ratios using Eq. (13). These averaged ratios were

collected in histograms presented in Fig. 9. We expect the best estimators to provide distributions centered on the true ratio ($r$ = 3) with the least spread around this value.
As expected, the ratio of medians gave a broader distribution for the dust-free image (Fig. 9a). Neither regression filtering nor quality control could improve observed estimates: the histograms of obtained ratios with or without filtering or with or without quality control were indistinguishable in the figure. For the contaminated image (Fig. 9b), ratio of means without filtering or quality control produced an additional peak (red line) reflecting contribution of dust clusters. RFSE estimate eliminates that peak (blue line) and the application of quality weights further improves the estimation (green line). These measures are so efficient that the resulting histogram after regression filtering and quality weighting became almost equivalent to the histogram of the ratios for the dust-free image (Fig. 9c). The ratio of medians is a robust estimate, but less accurate than RFSE. Fig. 9d demonstrates the power of quality control. Linear regression filtering was not applied in this case. The histogram of ratios of means had the same peak of aberrant ratios. Once weights have been applied, the peak disappeared.
Depending on the image, or even on each particular spot, different ratio estimators, such as the ratio of means or ratio of medians, may ensure a better performance; however, in practice it is difficult to predict with confidence the best estimator. RFSE approach gives a unique ratio estimate, which is always comparable to the best of other ratio estimators.

### 5.3 Robust processing of experimental images

**Localization.** We tested spot localization algorithm for arrays with different spot sizes, experimental designs and levels of contamination (numerous examples can be found on our web site http://bioinfo.curie.fr/projects/maia/). In all cases the spot localization procedure was carried out automatically with no user intervention. We only supplied the number of blocks in rows and columns and the number of spots in rows and columns within each block when switching from one image to another one. Comparison of the performance of our spot localization algorithm with others can be found in (Novikov & Barillot, 2006a). Although the developed procedure has proved to be very robust with respect to different types of microarray distortions, there is no guarantee that it will perform well for any array. Therefore, interactive tools are available to repair erroneous grids.
**Quantification and Quality.** We quantified two experimental images (Fig. 10) of different array design and signal-to-noise levels. One image (Fig. 10A) was provided as demonstration example for UCSF Spot 2.0 (downloadable from http://jainlab.ucsf.edu/Downloads.html). It contains 4x4 blocks with 21x21 spots per block, with a spot cell size of about 10 pixels. Cy3 and Cy5 color channels are strongly correlated, with the average correlation coefficient for the spots being about 0.97. Bright contamination spots can be seen irregularly scattered over the array. The magnified image of one such spot is shown in Fig. 5. Each clone was spotted in triplicate. The replicated spots are placed as neighbors in a row. The second image (measured in the Institute Curie, downloadable from http://bioinfo.curie.fr/projects/maia/) contains 12x4 blocks with 15x15 spots per block (Fig. 10B), with a spot cell size of about 30 pixels. The average correlation between the channels in the spots was about 0.85, being somewhat lower than for the first image, although there are no obvious contamination spots. Each clone was prepared in triplicate with the replicated spots put in three vertically distributed sub-arrays.

It is difficult to remain objective while doing comparative study for the experimental images. As the true ratio values are unknown, the only useful measure of quality is the variation in ratio estimates between the replicated spots, which should be reasonably low. Therefore we take the coefficient of variation (Eq. (14)) of the replicates as a quantitative measure of the ratio estimation consistency. However, this measure may not be totally objective: (i) the estimates may be consistent, but systematically biased (the true values of the ratios are unknown); (ii) three replicated spots of very poor quality may give very similar ratio values just by chance (the number of replicates is low). The average over all replicates at the given array coefficient of variation is taken as a global indicator of the Cy5/Cy3 ratio consistency of the array.

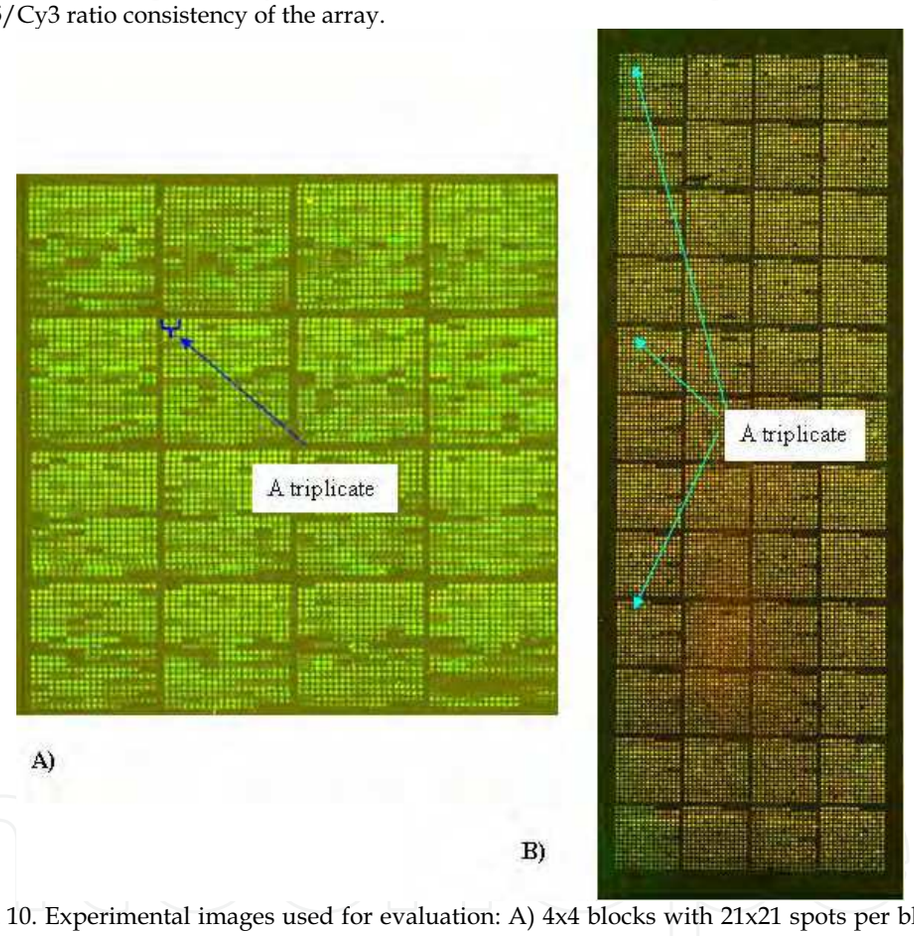


Fig. 10. Experimental images used for evaluation: A) 4x4 blocks with 21x21 spots per block, spot cell size is about 10 pixels; B) 12x4 blocks with 15x15 spots per block, spot cell size is about 30 pixels. The locations of triplicates are indicated.

We compared the averaged coefficient of variation for three ratio estimates (RFSE, ratio of means and ratio of medians) with or without quality control. The weights, $w_i$, of the marginal quality parameters for $Q_k$ were identified using Eq. (12) with $V \approx 0.07$ for image A, and with $V \approx 0.2$ for image B.

The results are summarized in Table 1. RFSE algorithm ensures the smallest coefficient of variation for both images and quality control improves performance for all three ratio estimates. We found a greater improvement for image A than for image B. This was not a surprise, as image B is characterized by a reasonably high signal-to-noise level, and it does not contain any obvious contaminated spots. However, even in this case the quality measures cannot be ignored, as there are still a few low-intensity spots that need to be specially treated (probably rejected). By contrast, image A has obvious randomly distributed pieces of dust, and the developed filtering procedure (RFSE) and quality measures proved to be powerful enough to repair or to disregard the contaminated spots, thus increasing the consistency of the Cy5/Cy3 ratio estimates. The fact that quality control does not show up much better performance is due to rather good general quality of the images, and a few problematic triplicates cannot influence very much the averaged coefficients of variation. For example, in image A, we have less than 9% of triplicates with the ratio variation coefficients larger that the selected $V$ (~0.08), and 7% for image B ($V \approx 0.2$).

| Image | Quality weights | RFSE | Ratio of means | Ratio of medians |
|---|---|---|---|---|
| A | Without | 0.0196 | 0.0324 | 0.0410 |
| | With | 0.0172 | 0.0245 | 0.0381 |
| B | Without | 0.119 | 0.120 | 0.133 |
| | With | 0.108 | 0.109 | 0.122 |

Table 1. The averaged coefficient of variation of the ratio triplicates for two images A and B (see Fig. 10).

Results on comparison of the performance of our quantification approaches with the approaches available from other image analysis packages can be found in (Novikov & Barillot, 2005a; Novikov & Barillot, 2005b).

## 6. Software

The developed algorithms have been implemented in the MAIA (microarray image analysis) software package (Novikov & Barillot, 2006b). Demonstration version of the software can be downloaded from http://bioinfo.curie.fr/projects/maia/. A full version is freely available to non-commercial users upon request from the authors. The package is written in Java (interface) and C++ (algorithms), and. runs on Windows 95/98/Me/NT/2000/XP platforms (may be used under Unix after recompiling C++ code) and needs the Java Runtime Environment. The whole quantification procedure (including filtering, segmentation and ratio estimation) for one 4Mb image pair (Cy3/Cy5, ~7300 spots; each spot cell is ~10 pixels) takes ~3 sec on 3.00GHz Pentium® 4 CPU with 1 GB of RAM; for a 40Mb image pair (~10800 spots; each spot cell is ~30 pixels) takes up to 20 seconds of processing.

## 7. Conclusions

In this work we have presented a complete solution for robust, high-throughput, two-color microarray image processing comprising procedures for automatic spot localization, spot quantification and spot quality control.

The spot localization algorithm is fully automatic and robust with respect to deviations from perfect spot alignment and contamination. As an input, it requires only the common array design parameters: number of blocks and number of spots in the x and y directions of the array. Although fully automatic, there is no guarantee that it will perform well for any array. Therefore, we offer some interactive tools to repair grid in case if it is erroneous.

Robust ratio estimation comprises two steps. First, linear regression filtering is used to identify and remove aberrant pixels, and then more traditional segmentation approaches are applied for final estimation. Using the two-step quantification algorithm, we ensure a unique ratio estimate, which is as robust as estimates based on medians and as precise as estimates based on means. Linear regression filtering relies on the fact that the two color channels are expected to be highly correlated. Any contamination, which is uncorrelated in the two channels, can be easily recognized by the algorithm and removed. For noisy (weakly correlated) data, the filter is transparent for the data. Moreover, in this case, linear regression estimates can be biased. Therefore we apply a spot segmentation step to establish the final estimate.

The spot quality algorithm provides a value of spot quality reflecting the level of confidence in the obtained ratio estimate at each spot. The unique spot quality value is derived from a set of ten marginal quality parameters characterizing certain features of the spot. The contribution of each quality parameter in the overall quality is automatically evaluated based on the visual classification of the spots, or using information available from the replicated spots, located on the same array or over a set of replicated arrays. Therefore the developed procedure allows us not only to quantify spot quality, but also to identify different types of spot deficiency occurring in microarray technology. The quality values can be used either directly to flag out some spots with the quality lower than the user-defined threshold, or in the follow-up analysis as a weight controlling the contribution/influence of the obtained ratio estimates.

There are many possibilities to advance the developed algorithms. For example, several spot localization parameters ($\gamma$, $\alpha$ and $\beta$), that are currently fixed in predefined values, can be iteratively adjusted to achieve the highest regularity of the generated grid. To enhance spot quantification, we can envisage more sensitive (than the single-case diagnostics for the linear regression model) algorithms for aberrant pixel detection. These perspectives are facilitated by further standardizing microarray technology, so that images are becoming more regular, and more specific models for spots and arrays can be developed and justified. As it was shown, different features of the spot (intensity, size, circularity, etc.) can be quantitatively characterized. These characteristics, besides ratios, may contain useful information for the follow-up analysis. One possibility to utilize this information is presented in this paper: we used them to derive spot quality values. However, we believe that more sophisticated analytical tools can be applied to use spot information in other applications. Exploration of these possibilities creates an interesting perspective for future developments.

## 8. Acknowledgements

We would like to thank our colleagues from the different laboratories of the Institute Curie: (F. Radvanyi, CNRS/IC 144; O. Delattre, INSERM/IC 830; M. Dutreix, CNRS/IC 2027) and Prof. D. Pinkel (UCSF Comprehensive Cancer Center), who have provided numerous microarray images allowing considerable improvement of the algorithms.

## 9. References

Angulo, J. & Serra, J. (2003). Automatic analysis of DNA microarray images using mathematical morphology. *Bioinformatics*, Vol. 19, 553-562.

Atkinson, A. & Riani, M. (2000). *Robust Diagnostic Regression Analysis*, Springer.

Axon Instruments, Inc. (2005). GenePix Pro 6.0. http://www.axon.com, User's Guide and Tutorial.

Bengtsson, A. & Bengtsson, H. (2006). Microarray image analysis: background estimation using quantile and morphological filters. *BMC Bioinformatics*, Vol. 7, 96.

Bozinov, D. & Rahnenführer, J. (2002). Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering, *Bioinformatics*, Vol. 18, 747-756.

Brändle, N.; Bischof, H. & Lapp, H. (2003). Robust DNA microarray image analysis. *Machine Vision and Applications*, Vol. 15, 11-28.

Brown, C.S.; Goodwin, P.C. & Sorger, P.K. (2001). Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences*, Vol. 98, 8944-8949.

Buhler, J.; Ideker, T. & Haynor, D. (2000). Dapple: improved techniques for finding spots on DNA microarrays. *UW CSE Technical Report UWTP 2000-08-05*.

Bylesjö, M.; Eriksson, D.; Sjödin, A.; Sjöström, M.; Jansson, S.; Antti, H. & Trygg, J. (2005). MASQOT: a method for cDNA microarray spot quality control. *BMC Bioinformatics*, Vol. 6, 250.

Ceccarelli, M. & Antoniol, G. (2006). A deformable grid-matching approach for microarray images. *IEEE Transactions on Image Processing*, Vol. 15, 3178-3188.

Chen, Y.; Kamat, V.; Dougherty, E.R.; Bittner, M.L.; Mel*tzer, P.S. & Trent, J.M. (2002). Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics*, Vol. 18, 1207-1215.

Dissanaike, G. & Wang, S. (2003). A critical examination of orthogonal regression. http://ssrn.com/abstract=407560.

Eckel-Passow, J.E.; Hoering, A.; Therneau, T.M. & Ghobrial I. (2005). Experimental design and analysis of antibody microarrays: applying methods from cDNA arrays. *Cancer Research*, Vol. 65, 2985-2989.

Fisher, L.D. & van Belle, G. (1993). *Biostatistics. A Methodology for the Heath Sciences.* John Willey & Sons.

Glasbey, C.A. & Ghazal, P. (2003). Combinatorial image analysis of DNA microarray features, *Bioinformatics*, Vol. 19, 194-203.

Hautaniemi, S.; Edgren, H.; Vesanen, P.; Wolf, M.; Järvinen, A.K.; Yli-Harja, O.; Astola, J.; Kallioniemi, O. & Monni, O. (2003). A novel strategy for microarray quality control using Bayesian networks. *Bioinformatics*, Vol. 19, 2031-2038.

Hegde, P.; Qi, R.; Abernathy, K.; Gay, C.; Dharap, S.; Gaspard, R.; Hughes, J.E.; Snesrud, E.; Lee, N. & Quackenbush, J. (2000) A concise guide to cDNA microarray analysis. *BioTechniques*, Vol. 29, 548-562.

Herzel, H.; Beule, D.; Kielbasa, S.; Korbel, J.; Sers, C.; Malik, A.; Eickhoff, H.; Lehrach, H. & Schuchhardt, J. (2001) Extracting information from cDNA arrays. *Chaos*, Vol. 11, 98-107.

Ishkanian, A.S.; Malloff, C.A.; Watson, S.K.; DeLeeuw, R.J.; Chi, B.; Coe, B.P.; Snijders, A.; Albertson, D.G.; Pinkel, D.; Marra, M.A.; Ling, V.; MacAulay, C. & Lam, W.L. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genetics*, Vol. 36, 299–303.

Jain, A.N.; Tokuyasu, T.A.; Snijders, A.M.; Segraves, R.; Albertson, D.G. & Pinkel, D. (2002). Fully automated quantification of microarray image data. *Genome Research*, Vol. 12, 325-332.

Kendall, M.G. & Stuart, A. (2003). *The Advanced Theory of Statistics,* Vol. 2, McMillan, 1979.

Lehmussola, A.; Ruusuvuori, P. & Yli-Harja, O. (2006). Evaluating the performance of microarray segmentation algorithms. Bioinformatics, Vol. 22, 2910-2917.

Novikov, E. & Barillot, E. (2005a). An algorithm for automatic evaluation of the spot quality in two-color DNA microarray experiments. *BMC Bioinformatics*, Vol. 6, 293.

Novikov, E. & Barillot, E. (2005b) A robust algorithm for ratio estimation in two-color microarray experiments. *Journal of Bioinformatics and Computational Biology*, Vol. 3, 1411-1428.

Novikov, E. & Barillot, E. (2006a). A noise-resistant algorithm for grid finding in microarray image analysis. *Machine Vision and Applications*, Vol. 17, 337-345.

Novikov, E. & Barillot, E. (2006b). Software package for automatic microarray image analysis (MAIA). *Bioinformatics*, Vol. 23, 639-640.

Pinkel, D.; Segraves, R.; Sudar, D.; Clark, S.; Poole, I.; Kowbel, D.; Collins, C.; Kuo, W.L.; Chen, C.; Zhai, Y.; Dairkee, S.H.; Ljung, B.M.; Gray, J.W. & Albertson, D.G. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, Vol. 20, 207–211.

Ritchie, M.E.; Diyagama, D.; Neilson, J.; van Laar, R.; Dobrovic, A.; Holloway, A. & Smyth, G.K. (2006). Empirical array quality weights in the analysis of microarray data. *BMC Bioinformatics*, Vol 7, 261.

Rousseeuw, P.J. & Leroy, A.M. (2003). *Robust Regression and Outlier Detection*, John Willey & Sons.

Rueda, L. & Vidyadharan, V. (2006). A hill-climbing approach for automatic gridding of cDNA microarray images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 3, 72-83.

Wang, X.; Ghosh, S. & Guo, S.W. (2001). Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, Vol. 29, e75.

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


**Vision Systems: Segmentation and Pattern Recognition**

Edited by Goro Obinata and Ashish Dutta

ISBN 978-3-902613-05-9

Hard cover, 536 pages

**Publisher** I-Tech Education and Publishing

**Published online** 01, June, 2007

**Published in print edition** June, 2007

Research in computer vision has exponentially increased in the last two decades due to the availability of cheap cameras and fast processors. This increase has also been accompanied by a blurring of the boundaries between the different applications of vision, making it truly interdisciplinary. In this book we have attempted to put together state-of-the-art research and developments in segmentation and pattern recognition. The first nine chapters on segmentation deal with advanced algorithms and models, and various applications of segmentation in robot path planning, human face tracking, etc. The later chapters are devoted to pattern recognition and covers diverse topics ranging from biological image analysis, remote sensing, text recognition, advanced filter design for data analysis, etc.

**How to reference**

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Eugene Novikov and Emmanuel Barillot (2007). Robust Microarray Image Processing, Vision Systems: Segmentation and Pattern Recognition, Goro Obinata and Ashish Dutta (Ed.), ISBN: 978-3-902613-05-9, InTech, Available from: http://www.intechopen.com/books/vision_systems_segmentation_and_pattern_recognition/robust_microarray_image_processing

INTECH
open science | open minds

**InTech Europe**

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

**InTech China**

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2007 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen