

# Exploring Spatial Information for Improved Microarray Image Segmentation and Quality Assessment

Minyao Sun, Yan Yang

School of Mathematical and Statistical Sciences, Arizona State University,  
Tempe, AZ 85287, USA

## Abstract

Microarray technology has been widely used in biomedical research to study the function of tens of thousands of genes in a single experiment. Prior to downstream analysis (e.g., identification of differentially regulated genes) and the biological interpretation of microarray data, appropriate methods for extracting the hybridization signals from microarray images are essential. However, artifacts caused by dust, fibers, scratches, etc. that are not uncommon on an array can seriously contaminate the signals, whereas tools for automatic quality assessment are lacking. We exploit the spatial information to guide model-based segmentation and spot intensity estimation. Also, we utilize the spatial as well as spot shape information to develop quality assessment measures for detecting potential artifacts. A segmentation method is developed that iterates between mixture model-based clustering and a four-neighbor connected component (FNCC) labeling algorithm to reduce the distorting effect of small disconnected artifacts on cluster memberships. Spot-level quality assessment measures are formulated to automatically detect artifacts of various shapes and sizes. The affected spots are then flagged for downstream analysis. The proposed methods are illustrated with the biomedical data from a human Valley Fever diagnosis study. Our segmentation procedure produces spatially connected clusters free of small bright artifacts. The quality control tools developed can be used to identify dust and fibers that typically expand several spots and are potentially useful for detecting arrays contaminated by scratches. In addition, poor gridding and non-normal spot types (e.g., blank spots, black holes, donut-shaped spots, and irregularly-shaped spots with weak signal) are automatically identified, providing additional information on local and global quality control.

**Key Words:** Artifact detection, connected-component labeling, EM algorithm, mixture model, saturation

## 1. Introduction

As a result of the proposal of Human Genetic Project (HGP) and the development of molecular biology in the 1990s, microarray has been widely introduced into biological and biomedical research. The high throughput technology and downstream analysis has become a crucial tool in the “-omics” era.

A microarray experiment typically consists of array printing, hybridization (binding) with experiment samples, array scanning, image processing, downstream analysis (e.g., identifying differentially regulated genes or classifying a sample as normal or with

cancer), and biological interpretations. Our work focuses on image processing. Gridding, the first step in image analysis, finds the exact location of each spot on an array. Segmentation is the core step that identifies which pixels within a spot (or a target mask) form the foreground signal and which pixels form the background noise. The next step is usually termed as spot intensity estimation which quantifies the expression level of each spot, followed by quality control for assessing whether the estimated spot intensity is reliable for downstream analysis.

Most literature work on microarray image processing has been in the area of image segmentation. The fixed and adaptive circle methods are used in many commercial products, such as GenePix®. The assumption is that the shape of a spot is approximately circular, which is usually unrealistic for all the spots on any array. The seeded region growing method implemented in Yang *et al.* (2001) produces spatially connected components in an array image and does not impose a particular spot shape. However, the selection of the seed may affect the pixel memberships significantly. The histogram approach of Chen *et al.* (1997) identifies foreground and background pixels based only on intensity values and no spatial information is considered. Similar to the histogram method, mixture model-based segmentation relies on pixel intensities to cluster pixels, but in a parametric way (Li *et al.*, 2005; Yang *et al.*, 2011). Thus, it can accurately identify irregularly shaped spots, such as the donut-shaped spots. In addition, Li *et al.* (2005) refined model-based clustering by adding the four-neighbor-connected-component labeling to find a spatially connected component within a target mask as the signal. However, applying connected-component labeling oftentimes changes the raw data within a target mask and may affect the clustering results. In this article, we propose an iterated procedure between mixture model-based segmentation and check of spatially connected components to stabilize the segmentation results.

The output of image processing serves as the input data for downstream analysis. Therefore, the quality of spot intensity estimates directly affects the subsequent analysis and biological interpretations of an experiment. However, artifacts caused by dust, fibers, scratches, etc. that are not uncommon on an array can seriously contaminate the signals. To evaluate the quality of spot summary measures, Wang *et al.* (2001) and Sauer *et al.* (2005) develop quality scores at the spot and array levels. Reimers and Weistein (2005) investigate regional biases and provide tools for visualizing and quantifying the biases. These methods largely focus on unusual spot intensities and sizes without considering irregular spot shapes. In this work we develop spot-level quality assessment measures to automatically detect artifacts of various shapes and sizes. The affected spots are flagged for downstream analysis.

The rest of this article is organized as follows. Section 2 introduces an iterated procedure between pixel clustering and connected-component labeling for improved segmentation. Methods for quality assessment and artifact detection are presented in Section 3. We illustrate the proposed methods with the biomedical data from a human Valley Fever diagnosis study in Section 4. Concluding remarks are provided in Section 5.

## 2. An Iterated EM-FNCC Algorithm for Improved Segmentation

Mixture model-based clustering of pixels has been a popular approach to microarray image segmentation (Chen *et al.*, 1997; Li *et al.*, 2005; Baek *et al.*, 2007). A mixture of parametric distributions, such as the Gaussian family, is usually assumed for the pixel

intensity values within a target mask. The iterative EM algorithm is then employed for maximum likelihood estimation of the model parameters which are subsequently used for measuring spot expression levels. For exploiting the spatial information contained in an image, Li *et al.* (2005) consider connected component labeling to identify spatially connected components as the signals. This added feature enhances the flexibility of the mixture model approach for pixel clustering and can potentially reduce the distorting effect of small artifacts. However, including the pixels that are originally labelled as part of the foreground but disconnected from the rest of the signal, if any, for segmentation may affect the memberships of other pixels.

We propose an iterative procedure between model-based segmentation and connected component labeling to provide potentially more stable final segmentation results. In each iteration step, a censored Gaussian mixture model is first applied to cluster valid pixel intensity data within a target mark; saturated pixels, when present, are treated as censored observations (Yang *et al.*, 2011). Based on the clustering results, FNCC is used to identify spatially disconnected foreground pixels, which are then regarded as potential artifacts and removed from future iterations. In addition, bright small artifacts are also deleted with a check on its cluster size. The modified dataset is used for the next iteration. The iteration between segmentation and FNCC stops when either no more pixels are to be removed or the number of iterations reaches a specified maximum (e.g., 10).

## 2.1 Mixture Model-Based Image Segmentation

In mixture model-based clustering, we assume that the pixel intensities within a target mask are independent and identically distributed from a  $K$ -component Gaussian mixture model;  $K \leq 3$ .  $K = 1$  suggests a blank target mask with no identifiable signal.  $K = 2$  represents an ideal spot associated with signal and noise, respectively. The cluster with an intermediate mean in a 3-component model usually corresponds to the fuzzy edge of a bright spot or the inner hole of a donut-shaped spot. Let  $Y_i$  be the intensity value for pixel  $i$ ,  $i = 1, \dots, n$ , and  $y_i$  be the observed value of  $Y_i$ , then the mixture density function can be written as

$$f(y_i; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \phi(y_i; \mu_k, \sigma_k),$$

where  $\pi_k, k = 1, \dots, K$ , are the mixing weights ( $0 \leq \pi_k \leq 1, \sum_{k=1}^K \pi_k = 1$ ),  $\phi(y_i; \mu_k, \sigma_k)$  represents the density function of a normal distribution with mean  $\mu_k$  and standard deviation  $\sigma_k$ , and  $\boldsymbol{\theta} = (\pi_1, \dots, \pi_{K-1}, \mu_1, \dots, \mu_K, \sigma_1, \dots, \sigma_K)^T$  contains all parameters that need to be estimated. The log-likelihood function is then given by

$$l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^n \left\{ \log \sum_{k=1}^K \pi_k \phi(y_i; \mu_k, \sigma_k) \right\},$$

where  $\mathbf{y} = (y_1, \dots, y_n)^T$ .

In practice, signal saturation may occur, especially for microarray slides with large spot-to-spot variations. If the intensity of a pixel exceeds the maximal detectable value of the scanner (e.g.,  $2^{16} - 1 = 65,535$  for 16-bit images), this maximum will be recorded. To partially recover the lost information and correct for saturation-induced bias, Yang *et al.* (2011) include a right-censored normal distribution as the  $K$ th component in the mixture

model with density function

$$\phi_S(y_i; \mu_3, \sigma_3) = I(y_i < S)\phi(y_i; \mu_3, \sigma_3) + I(y_i = S)\left\{1 - \Phi\left(\frac{S - \mu_3}{\sigma_3}\right)\right\},$$

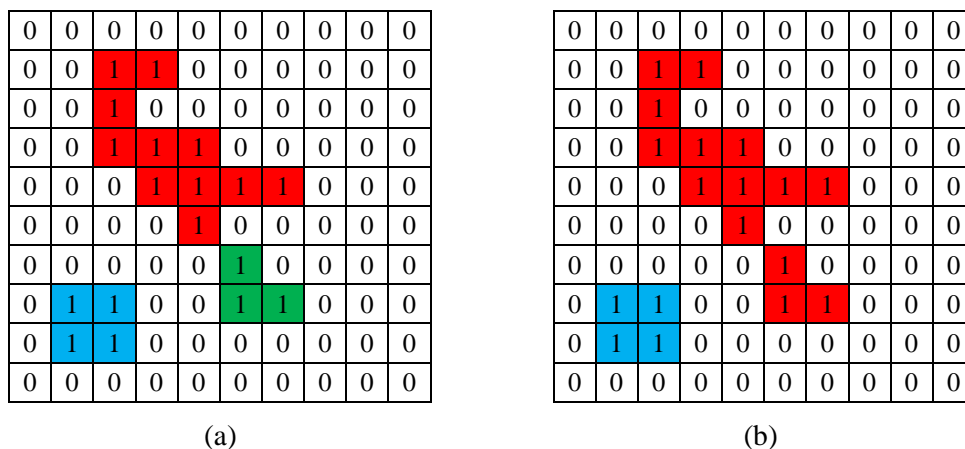
where  $S = 65,535$  is the saturation point,  $\Phi(\cdot)$  is the standard normal cumulative distribution function, and  $I(A)$  is the indicator function which equals 1 if event  $A$  occurs and 0 otherwise. The parameters  $(\pi_k, \mu_k$  and  $\sigma_k)$  can be estimated by maximizing the likelihood function using the EM algorithm (Yang *et al.*, 2011). Once the parameter estimates are obtained, the optimal number of clusters  $K$  can be determined based on likelihood-based information criteria, such as the Bayesian information criterion (BIC).

### 2.2 Connected-Component Labeling

A two-pass algorithm is used for implementing connected component labeling. The first pass goes through all the pixels in an image and records their connectivity. In FNCC the connectivity check is done for the top and the left neighbors. The eight-neighbor connected component (ENCC) procedure additionally checks the top left and top right pixels. Temporary labels are assigned to different connected subsets of pixels during this step. If the target pixel shares the same property, in our case, falling in the foreground cluster, with its neighbors, then it obtains the smaller/smallest label of its neighbors. The connectivity between the subsets with different labels is also recorded.

The second pass reassigns all the labelled pixels with the lowest equivalent label. Figure 1(a) provides a simple example for FNCC. All the squares labelled “1” represent the foreground. Three separate groups of pixels, in red, blue and green, respectively, are identified according to the definition of neighborhood in FNCC. Since the connection through the vertex is included, the red and green subgroups are considered to be linked in ENCC (see Figure 1(b)). As a result of applying the two-pass algorithm to the image, all connected subsets of pixels from the foreground cluster will be detected.

A lower bound on the number of pixels in the foreground cluster can be imposed to detect bright small artifacts (e.g., 10% of the pixels within a target mask). Subsets fallen below this threshold are treated as potential artifacts.



**Figure 1:** Four- and eight-neighbor connected component labeling

### 3. Quality Assessment and Artifact Detection

Quality assessment is the last but a fairly important step in image processing. It attaches quality measures to each spot and flags out unreliable spot expression data to be used for downstream analysis.

#### 3.1 Identification of Non-Normal Spot Types

Spots that deviate from expected shapes, sizes, locations or intensity values emerge at a considerable frequency in microarray images with today’s technology. We define some common non-normal types of spots in this section and develop measures to identify them. The tools also facilitate artifact detection to be presented next.

**Donut-shaped spot.** One special type of spots that are common in practice is the donut-shaped spot. The spot has a dim inner hole and a relatively bright outer ring. Although its shape is non-standard, the ring may contain useful biological information. Note that the signal of a donut-shaped spot is bounded by two “circles” (instead of just one): a bigger or regular “circle” that separates the ring from the surrounding background, and a smaller “circle” that tells the ring apart from the inner hole. We identify potential donut-shaped spots by checking whether or not there are two consecutive boundaries or circles for the spot. A spot is labelled donut shaped if spatially disconnected boundary pixels are present according to the ENCC method. Such spots will be treated differently when calculating the circularity measure, one element of the proposed quality measures.

**Blank spot.** A spot is called blank or empty if the associated target mask only contains background pixels. Blank spots are identified when the mixture component  $K$  is selected to be one by BIC after the iterated EM-FNCC procedure stabilizes.

**Black Hole.** A spot is referred to as a black hole if pixels near the center of a target mask have lower intensities than the surrounding pixels. In the segmentation stage the cluster having the highest mean intensity and passing the cluster size check is labelled as the foreground. Thus, black holes need to be identified and require special handling.

Another type of non-normality is related to bad gridding. Although ideally all foreground pixels for a spot should be completely within a target mask created in the gridding stage, deviations between the actual and pre-specified spot locations oftentimes exist due to the inaccuracy of the slide printing process. Such deviations may cause part of the foreground of an intact spot to fall into an adjacent target mask. To identify the issue of bad gridding, we calculate the percentage of foreground pixels that are on the boundary of a target mask and compare it against a threshold. If the percentage for a spot is higher than the threshold, then the spot is flagged due to bad gridding. A sizable number of spots being flagged under this category would suggest a re-do of the gridding step. Black holes can also be identified with this measure but a different criterion. The following describes the decision rules for bad gridding and black holes:

$$n_{FG}/(n_{FG} + n_{BG}) \leq p, \quad (1)$$

$$n_{FG}/(n_{FG} + n_{BG}) \geq 1 - p, \quad (2)$$

$$p < n_{FG}/(n_{FG} + n_{BG}) < 1 - p, \quad (3)$$

where  $n_{FG}$  is the number of foreground pixels on the target mask boundary,  $n_{BG}$  is the number of background pixels on the boundary, and  $p$  is the threshold. For spots satisfying

criterion (1), the gridding is acceptable. Spots satisfying criterion (2) are identified as black holes. When criterion (3) is met, bad gridding is signaled. Note that, besides the spots with true gridding issues, dust and fibers may also give rise to high percentages when they run across several spots. If the spots are well aligned and their locations are well identified, the criteria above will be an efficient approach to the detection of those artifacts, that is, dust and fibers. The threshold used in our data example was 0.15. Asymmetric thresholds can be used as well. In the absence of gridding problems, a lower threshold will increase the sensitivity of artifact detection.

### 3.2 Artifact Detection

Artifacts, such as dust and fibers, are usually very bright on a microarray image. During the segmentation stage, small bright artifacts are detected using a cluster size check and removed from further analysis. In the case of large bright artifacts spanning over several target masks, the spots will not pass the gridding check introduced earlier and will thus be flagged. Next we discuss how to detect dust, fibers or scratches that touch the signals.

When artifacts overlap with the foreground pixels, an irregular spot shape is oftentimes observed. The ideal spot shape in our applications is circular. Thus, a convenient method to identify irregular spot shapes is the circularity measure  $C$ :

$$C = 4\pi A/P^2.$$

Here  $A$  is the area of the signal or the total number of foreground pixels, and  $P$  is the perimeter of the signal or the number of boundary pixels that separate the foreground and the background. For a donut-shaped spot, the circularity measure  $C$  tends to be low due to a reduced area and an inflated perimeter, even with an approximate circular shape. We therefore modify the circularity measure for donut-shaped spots as follows:

$$C_{DS} = 4\pi(A_r + A_i)/P^2.$$

In the formula above,  $A_r$  is the area of the ring,  $A_i$  is the area of the inner hole, and  $P$  is defined as earlier but care needs to be taken to make sure that the boundary separating the ring from the inner hole is not counted.

A perfectly circular spot has a circularity measure equal to 1. In application, however, it is difficult to tell good spots apart from spots potentially contaminated by artifacts solely based on a circularity cutoff. A high cutoff tends to make a significant portion of spots unusable even in the absence of dust and fibers, whereas a low threshold fails to detect contaminated spots effectively. An exploratory analysis investigating both the intensity and the circularity for all the spots from a few blocks contaminated by artifacts indicates that contaminated spots are generally much brighter than the rest and have relatively low circularity. Thus, we combine a circularity cutoff with an intensity threshold for detecting artifacts. More details and examples illustrating the proposed methods are provided in the next section.

## 4. Results and Discussion

In this section, the results from analyzing biological data are presented. The iterated EM-FNCC procedure is compared with the original EM method. The potential improvement on increasing the accuracy of segmentation results is demonstrated by data from a human

Valley Fever diagnosis study. We also show that the proposed quality measures are able to identify several non-normal spot types and to detect potential artifacts at the spot and sub-array levels.

The original images were first processed in Matlab following Yang *et al.* (2011) to obtain pixel intensity values and coordinates within each target mask.

#### **4.1 A Human Valley Fever Diagnosis Study**

The original microarray data used in this project came from the Center for Innovations in Medicine in the Biodesign Institute at Arizona State University. The set of slides was peptide microarray from a human Valley Fever diagnosis study. 20mer peptides with distinct sequences were printed onto the microarray by NanoPrint™ microarrayer printer. Proteins from the human blood sample were captured by specific peptides on the array during hybridization and labelled by antibodies with fluorescent dye. The images were created by Agilent's DNA Microarray Scanner at 100% laser power and 70% PMT voltage. Under the specific settings of the scanner, some pixels in the images saturated at the intensity value 65,535. The spot information including the center coordinates were also provided by the research center.

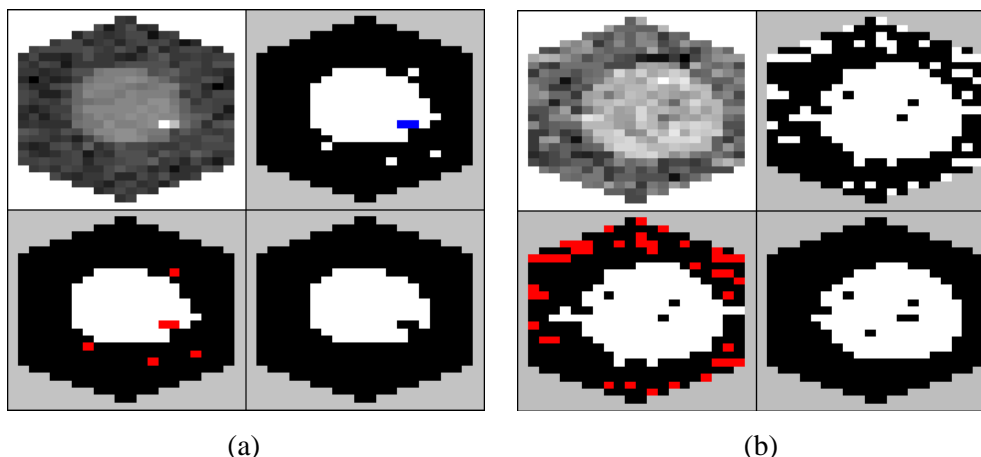
The peptide microarray slides used in the human Valley Fever diagnosis study contain 48 blocks. Each block consists of 454 spots and 30 blank positions lined up in a 22 by 22 array. Included in the 454 real spots, the last 6 spots are landing lights which are used for locating the block and normalization. The gridding step (more details will be provided later) divides a block into small pieces, called target masks, for processing the spots individually. Each target mask contains 400-500 pixels (target masks on the edge of a block may contain more pixels). The segmentation and intensity extraction procedure are carried out at the spot level.

#### **4.2 Improved Image Segmentation by Iterated EM-FNCC Method**

The original EM segmentation method is able to detect small bright artifacts by a cluster size check. If the total number of pixels assigned with the highest cluster falls below the size cutoff, the cluster will be reset to “-1”. Therefore, the former second highest cluster will become the highest. No iterated segmentation is applied to the result after cluster size check.

The new iterated EM-FNCC method keeps the core segmentation part in EM method including the cluster size check. Moreover, small sets of bright pixels interspersed in the background can be filtered out by a connected-pixel area check. Figure 2 shows us two examples of how the new method improves the segmentation results. In both case, the top left part is the microarray image reconstructed based on the signal extracted by Matlab. The segmentation result from EM algorithm is on the top right. The white pixels which belong to the highest cluster form the foreground. The background part in black consists of pixels fallen in lower cluster. If there exists a small group of pixels with highest cluster which do not pass the cluster size check, these pixels will be considered as artifacts and labelled with the color blue. The two images at the bottom are results from the iterated EM-FNCC approach. The left one is obtained by implementing one round of segmentation. The color code used here are similar with the one for the old method

except that the artifacts are filled with red. The last image is the final result after the iteration is done. The iteration will stop whenever there is no more change between rounds of segmentation or a preset maximum number of iteration is reached.



**Figure 2:** Raw and segmented images: (a) array 29P01143\_P1304, block 20, spot 68; (b) array 29P01140\_P1153, block 15, spot 344

**Table 1:** Mean foreground intensities by different segmentation methods.  $I_1$ ,  $I_2$ , and  $I_3$  are mean intensities of foreground pixels obtained from the EM algorithm, one round of the EM-FNCC algorithm, and the iterated EM-FNCC method, respectively.  $\Delta_{2,1} = (I_2 - I_1) / I_1 \times 100\%$  and  $\Delta_{3,2} = (I_3 - I_2) / I_2 \times 100\%$  represent relative changes.

Spot	$I_1$	$I_2$	$I_3$	$\Delta_{2,1}$	$\Delta_{3,2}$
Array 29P01143_P1304 block 20 spot 68	4903.695	4997.099	5024.067	2.4547%	0.5397%
Array 29P01140_P1153 block 15 spot 344	6956.661	7346.601	7529.813	8.2389%	2.4938%

In Figure 2(a), we can see that the four separate pixels identified as foreground with old method is correctly classified as artifacts after one round of new segmentation. More foreground pixels are deleted at the end of the iteration. Also, since the new method preserves the cluster size check, whatever is considered questionable will be still potential artifacts in the modified segmentation. In this case, the differences among the average intensities of the foreground pixels identified by original and improved segmentation methods are not significant. Figure 2(b) reveals another situation in which the spatial information must be involved. Due to the high signal of the real background part within the target mask, a large amount of pixels are misclassified as foreground by EM method. The data extracted from these pixels need to be excluded from the following biological analysis. The modified segmentation tool successfully removed the misclassified pixels. By comparing the two images at the bottom, we can see that the new method is also able to get rid of those “tentacle” around the edge of the real spot. It is not rigorous to make rules for distinguishing real foreground from all pixels in the mask; however, the method we provided here is able to select those pixels which look like foreground at this point. Table 1 lists the percent change of the average intensity among different segmentation methods. For these two spots, it is obvious that the first round of

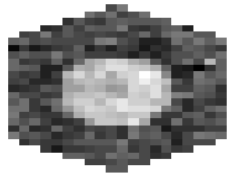
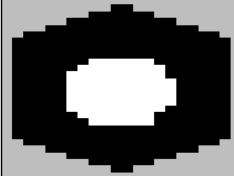
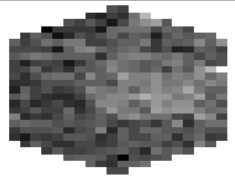

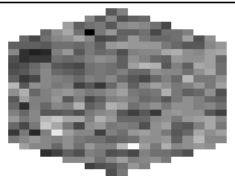
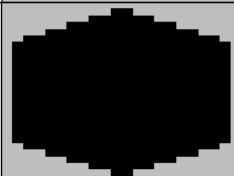
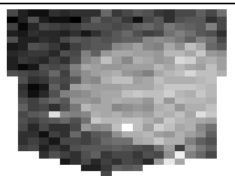

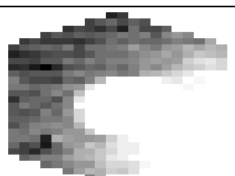

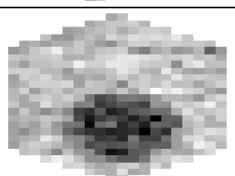
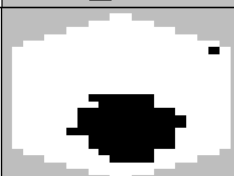
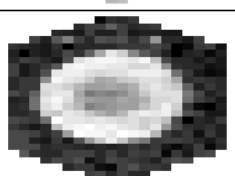



segmentation is essential compared with the following iteration. However, in special case like “blank”, without the whole iteration process, the spot may not be labelled correctly. The improved segmentation results also demonstrate the ability of the proposed method for identifying potential artifacts on pixel level.

### 4.3 Quality Assessment and Artifact Detection

#### 4.3.1 Identification of spot types

**Table 2:** Special types of spots

Mask	Segmentation	Spot Label	Note	
(i)			normal	
(ii)			normal	dark spot with low circularity
(iii)			blank	
(iv)			bad gridding	spot touching mask boundary
(v)			bad gridding	contaminated spot
(vi)			black hole	
(vii)			donut	

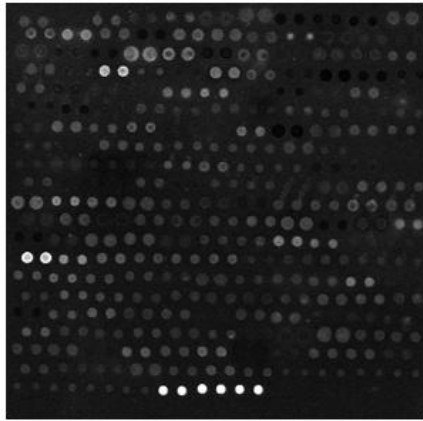
Segmentation results in Table 2 shows some typical types of spots we've processed in the research and the corresponding results generated by the iterated EM-FNCC. All the spots are selected from block 15 on microarray slide 29P01140\_Cocci\_P1153 and block 3 on slide 29P01147\_Cocci\_P8. The grayscale image on the left for each spot is regenerated based on the intensity of the pixels. The iterated EM-FNCC images shown on the right side indicate which pixels in the mask are recognized as foreground. The same color code is applied here as we discussed for the result in Figure 2.

Mask (i) contains a “perfect” spot the shape of which is very close to a circle. There is no foreground pixel touching the boundary of the mask. The contrast between the foreground and background is obvious. This type of spot is relatively easy to process since most of the popular segmentation methods are able to identify the foreground accurately. Mask (ii) represents another common case in which the true foreground pixels are relatively dark. Part of the foreground with low intensity can be misidentified as background so that the segmented foreground pixels may form an irregular shape. A spot contaminated by scratch may show a similar pattern. The segmentation will yield a blank spot for mask (iii). Our program recognized the 120 blank spots out of 1936 total spots in four blocks with a type I error of 0.83%. The only misidentification is due to a contaminated mask. Scratches on the slides may also lead to a “blank”. Mask (iv) and (v) are examples of bad gridding in which the percentage of foreground pixels on the boundary of the target mask are between 15% and 85%. The former is a mask with real gridding issue. That is, the input center location for gridding terribly mismatches the center of the spot. The second case indicates that the gridding check may have the potential to identify contaminated masks. The white part on the lower right corner in mask (v) is caused by a bright artifact. A “black hole” spot is shown in mask (vi) which is a circle containing pixels with unusual low intensity. The reason why “black hole” spots are formed is probably related to the printing process. The last type of spot presented (vii) is the donut shaped. Research has shown that the formation of the donut shape is possibly related to physical phenomena during the printing and hybridization process (Pappaert et al., 2006; Popov, 2005; Dugas et al., 2005). Also, the inner hole of a “donut” spot is classified as the medium cluster in segmentation. Therefore, our method simply treats the middle-cluster pixels as background.

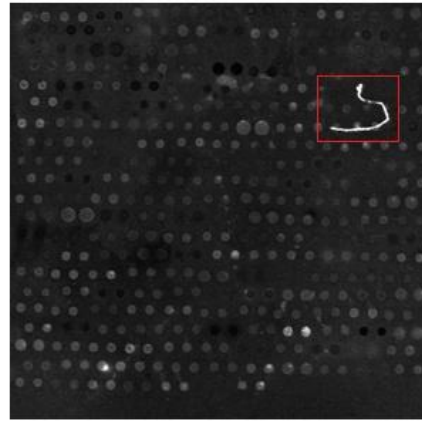
#### 4.3.2 Artifact detection

Four blocks from different arrays are selected for purpose of demonstration. The original microarray images of these blocks are shown in Figure 3. Block (a) does not encounter significant contamination issues and could be desirable for a reference block. The other blocks contain three typical artifacts which frequently appear in a microarray experiment. Several spots in block (b) are contaminated by a piece of fiber while a bright dust appears in block (c). Small unexpected porous particles are good binding surface for biological molecules. A large amount of the antibody with fluorescent dye can attach to these particles which form artifacts in the scanning phase. Even if the bright artifacts and the spots do not overlap, the segmentation result may still be distorted. These two types of contamination usually happen in the target mask level and spots around may be affected. Block (d) undergoes severe scratch problem and more than 10% of the spots are ineligible for downstream analysis. This type of artifact is common in the block and slide level. All the blocks other than the reference one are particularly good cases for testing the feasibility and efficiency of the artifact detection tools. We applied the previously discussed segmentation method and other measurements to these blocks. The circularity

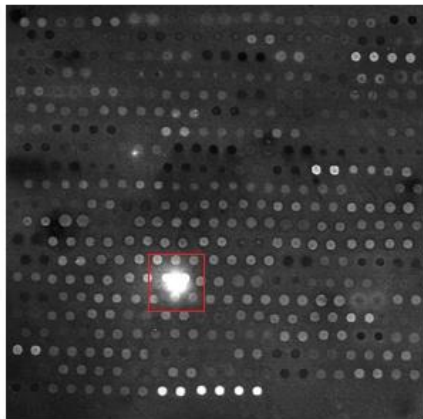
of all spots identified as “blank”, “black hole”, and “bad gridding” are assigned with the value “0”. For other spots (both “donut” and non-donut shape), the value is measured by the method provided in section 3.2. The average intensity of each spot is calculated based on the signal of the foreground pixels detected by segmentation except that the intensity of the “blank” spots is reset to “1”.



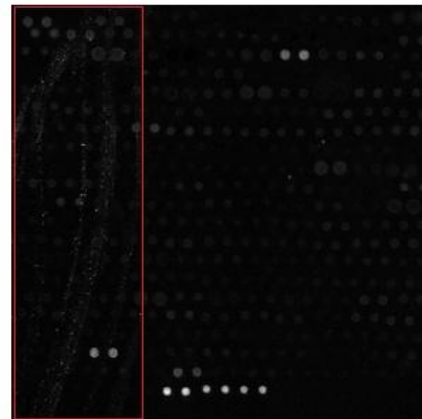
Block (a), control



Block (b), artifact (fiber)



Block (c), artifact (dust)

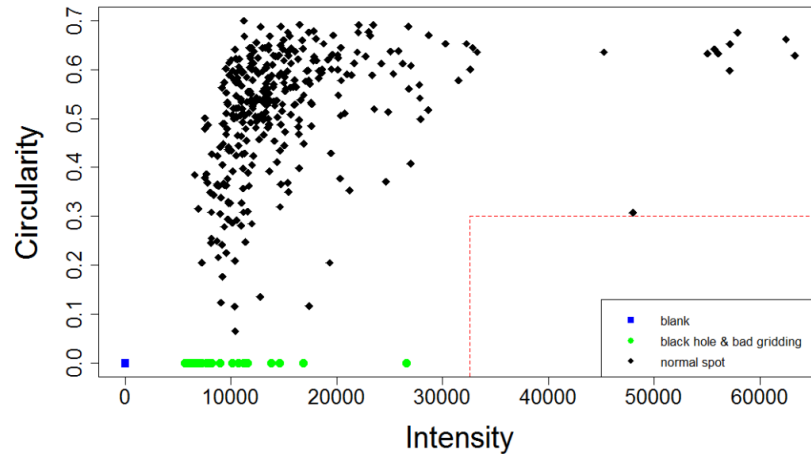


Block (d), artifact (scratch)

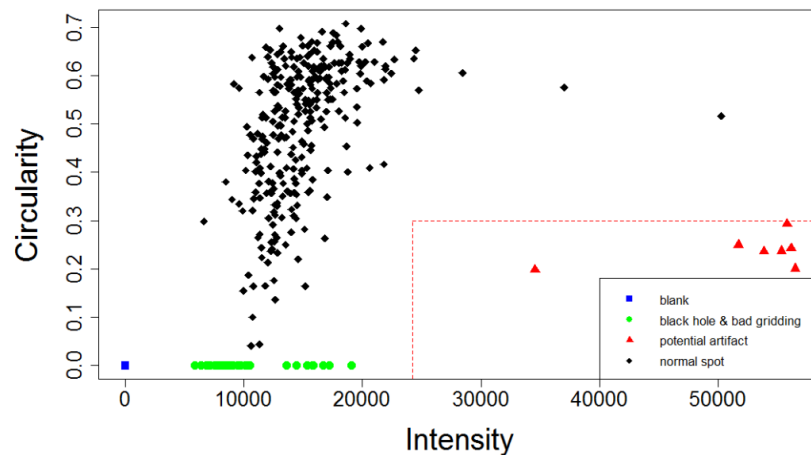
**Figure 3:** Original microarray image of four selected blocks

A scatter plot containing the information of the average intensity and circularity of all 484 spots is made for each block (Figures 4, 5, 6 and 7). In all four plots, a set of spots are overlapped and appear near the bottom left corner (blue). These spots are the “blank”. All the other spots with circularity “0” are “black hole” and “bad gridding” (green). With the purpose of detecting artifact spot, a circularity cutoff of 0.3 (the red horizontal line) and an intensity cutoff of 97.5 percentile (the blue vertical line) are applied. The region below the circularity cutoff and on the right side of the blue line (red) may contain potential artifacts. The relatively dark spots with a circularity value fallen in the interval (0, 0.3) (pink) could be questionable in some cases, particularly, when huge scratches go through the block (block (d) in Figure 3). The upper part in the plot (black) contains the spots which pass our threshold and are eligible for downstream biological illustration.

After comparing the scatter plots of block (a) (Figure 4) and block (b) (Figure 5), we can easily tell that the second block contains 7 suspicious spots while the other does not by the two parallel checking criteria. Further investigation verifies the “identity” of the suspicious spots - the spots covered by the piece of fiber in the image.

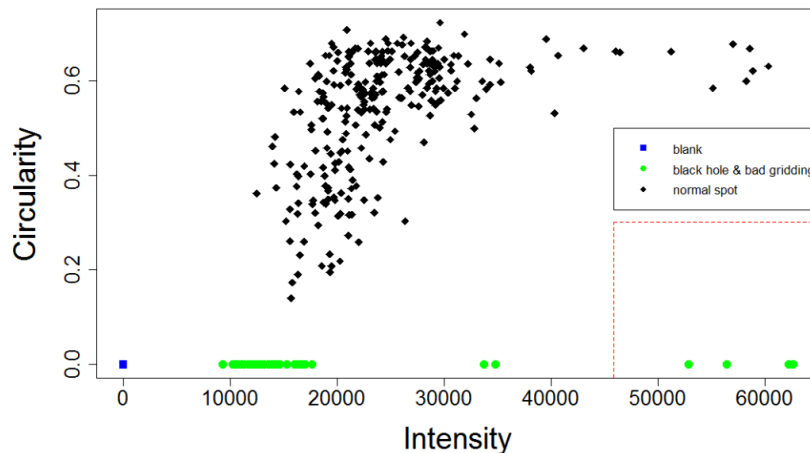


**Figure 4:** Circularity against mean foreground intensity for block (a). All pixel values were set to 1 for one-cluster spots (blank spots and spots with very weak signals).



**Figure 5:** Circularity against mean foreground intensity for block (b). All pixel values were set to 1 for one-cluster spots (blank spots and spots with very weak signals).

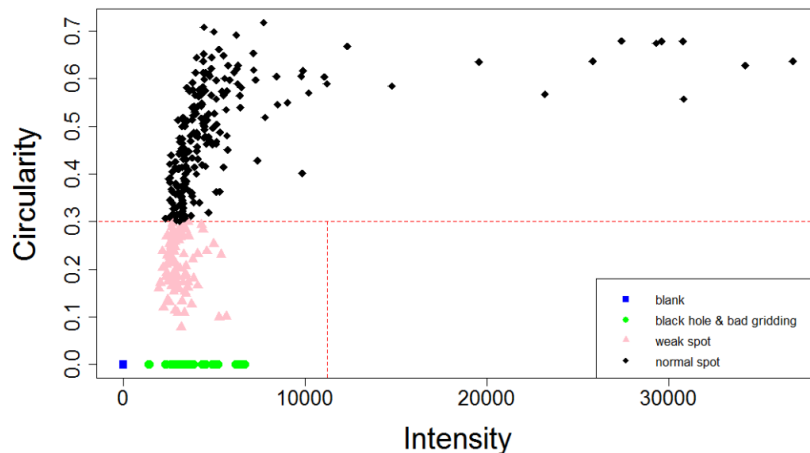
Block (c) brings us another situation for artifact detection. Four spots labelled with “bad gridding” fall in red (Figure 6). By checking the original microarray image, we found that these four spots are close to a bright artifact and are severely contaminated. Two more spots are affected by the artifact. They can be found in green in the plot because of the relatively low average intensity. Despite the fact that different criteria for artifact detection are used here, all six suspicious spots are selected successfully by our program.



**Figure 6:** Circularity against mean foreground intensity for block (c). All pixel values were set to 1 for one-cluster spots (blank spots and spots with very weak signals).

No spot from block (d) appears in red in Figure 7. However, a major difference between the scatter plots of block (a) and (d) is that much more spots crowd in green and pink for the latter case. The original location of these spots in green and pink confirms that they intersperse in the left part of the block which is damaged by scratches. It is almost impossible to distinguish the contaminated spots from others. Therefore, we could use the single threshold of circularity and all spots in pink should be excluded from the downstream analysis if a large area of the block is affected by artifact issue.

Although we successfully detected several types of artifacts with the 97.5 percentile threshold, it may not be a suitable cutoff in other experiments. In our microarray, each block contains 484 spots (including “blank”), so the 97.5 percentile corresponds to approximately the top 12 brightest spots. If large amount of spots are contaminated by bright articles, the threshold need to be reduced.



**Figure 7:** Circularity against mean foreground intensity for block (d). All pixel values were set to 1 for one-cluster spots (blank spots and spots with very weak signals).

## 5. Conclusions

The iterated EM-FNCC method discussed in this report takes advantage of the censored Gaussian mixture model which allows us to apply missing data analysis to the saturated pixels. The newly added FNCC module enables the program to remove small disconnected groups of pixels from foreground part. By iterating the segmentation and area check, our method is able to further exclude small artifacts and identify “blank” target masks. Several simple measurements provide the program with ability to recognize certain types of unusual spots so that data which is not eligible for further biological analysis will be filtered out. We also developed the parallel checking criteria for artifact detection. Spots contaminated by large artifacts, such as fiber or dust, were successfully identified according to the criteria. When a significant amount of spots are affected by huge scratch, an only circularity threshold is used in order to removing more questionable spots. However, we suggest using the replicate data if large scale contamination happens.

For the microarray slides we used for the project, the peptides were spotted on the microarray in an orange-crate packing pattern. We finished the gridding in Matlab based on the location of the center for each spot. The locations of the centers were obtained from the auto alignment function in the standard microarray image analysis software GenePix<sup>®</sup> Pro. The function is not reliable especially when the spots are not precisely printed. A biased location of the spot center may cause the gridding issue which will be detected by our segmentation procedure. Therefore, “bad gridding” is not a type of artifact and the problem will not arise when an accurate gridding result is available. If the issue is solved in the gridding step and every spot falls in its own target mask, then the so-called “gridding check” will only be used for identifying bright artifacts in the space among spots (see block (c) discussed in section 4.3.2).

The artifact caused “bad gridding” is not the only case of mislabelling. A real “black hole” spot is found to be labelled as “blank” or “bad gridding” occasionally. Fortunately, the mislabelling mentioned above does not affect the result of our image analysis since none of the three types of spots, “blank”, “black hole”, and “bad gridding”, will be processed to the next step. However, misidentifying dark spots as “blank” will cause information loss. It is possible to increase the Photomultiplier Tube (PMT) setting in microarray scanner so that the contrast between the real foreground and background will be amplified. Nevertheless, more missing data will be shown because of the saturation issue under high PMT gain. Thus, the researchers may balance the benefit of enlarged signal against the loss of information in this case.

It is difficult to determine whether a pixel should be treated as foreground or not simply by the data of intensity, even from perspective of biology. Also, more special types of spots may be shown in the research. Our segmentation method and the following QA only offer a computational and statistical approach to identify the foreground part in the microarray image. More interpretation from biologist is required to apply the result to downstream analysis.

The two thresholds (circularity and average intensity) for artifact detection given in the early example can only be used as a reference. In general, all spots located far away from the majority and near the right bottom corner should be suspicious. Locating these spots in the original microarray image could be helpful for determining whether they are

artifacts or not. It is possible to modify the criterion for identifying an artifact by setting a cutoff combining the information of circularity and intensity.

### Acknowledgements

This research was supported in part by NSF grant DRL-0909630.

### References

- Baek, J., Son, Y. S., McLachlan, G. J. 2007. Segmentation and intensity estimation of microarray images. *Bioinformatics*, 23(4): 458-465.
- Blekas, K., Galatsanos, N. P., Likas, A., Lagaris, I. E. 2005. Mixture Model Analysis of DNA Microarray Images. *IEEE Transactions on Medical Imaging*, 24(7): 901-909.
- Chen, Y., Dougherty, E. R., Bittner, M. L. 1997. Ratio-Based Decisions and The Quantitative Analysis of cDNA Microarray Images. *Journal of Biomedical Optics*, 2(4): 364-374.
- Dugas, V., Broutin, J., Souteyrand, E. 2005. Droplet Evaporation Study Applied to DNA Chip Manufacturing. *Langmuir*, 21: 9130-9136.
- He, L., Chao, Y., Suzuki, K., A Run-Based Two-Scan Labeling Algorithm. 2008. *IEEE Transactions on Image Processing*, 17(5): 749-756.
- Katzer, M., Kummert, F., Sagerer, G. 2003. Methods for Automatic Microarray Image Segmentation. *IEEE Transactions on Nanobioscience*, 2(4): 202-214.
- Li, Q., Fraley, C., Bumgarner, R. E., Yeung, K. Y., Raftery, A. E. 2005. Donuts, scratches and blanks: robust model-based segmentation of microarray images using a gamma-t mixture model. *Bioinformatics*, 21(12): 458-465.
- Pappaert, K., Ottevaere, H., Thienpont, H., Van Hummelen, P., Desmet, G. 2006. Diffusion limitation: a possible source for the occurrence of doughnut patterns on DNA microarrays. *BioTechniques*, 41: 609-616.
- Popov, Y. O. 2005. Evaporative Deposition Patterns: Spatial Dimensions of the Deposit. *Physical Review E*, 71: 036313-1-036313-17.
- Reimers, M., Weinstein, J. N. 2005. Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases. *BMC Bioinformatics*, 6: 166-173.
- Sauer, U., Preininger, C., Hany-Schmatzberger, R. 2005. Quick and simple: quality control of microarray data. *Bioinformatics*, 21(8): 1572-1578.
- Wang, X., Ghosh, S., Guo, S. 2001. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, 29(15): e75-1-8.
- 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*, 11(1): 108-136.
- Yang, Y., Stafford, P., Kim, Y. J., Segmentation and intensity estimation for microarray images with saturated pixels. *BMC Bioinformatics*, 2011. 12: 462-471.