Journal of Biomedical Optics 7(3), 507–523 (July 2002)

Simulation of cDNA microarrays via a parameterized random signal model

Yoganand Balagurunathan
Edward R. Dougherty
Texas A&M University
Department of Electrical Engineering
College Station, Texas 77843-3128

Yidong Chen
Michael L. Bittner
J. M. Trent
National Institutes of Health
National Human Genome Research Institute

Abstract. cDNA microarrays provide simultaneous expression measurements for thousands of genes that are the result of processing images to recover the average signal intensity from a spot composed of pixels covering the area upon which the cDNA detector has been put down. The accuracy of the signal measurement depends on using an appropriate algorithm to process the images. This includes determining spot locations and processing the data in such a way as to take into account spot geometry, background noise, and various kinds of noise that degrade the signal. This paper presents a stochastic model for microarray images. There are over 20 model parameters, each governed by a probability distribution, that control the signal intensity, spot geometry, spot drift, background effects, and the many kinds of noise that affect microarray images owing to the manner in which they are formed. The model can be used to analyze the performance of image algorithms designed to measure the true signal intensity because the ground truth (signal intensity) for each spot is known. The levels of foreground noise, background noise, and spot distortion can be set, and algorithms can be evaluated under varying conditions.
 [DOI: 10.1117/1.1486246]

Keywords: cDNA microarray; image simulation.

Paper JBO-01046 received July 6, 2001; revised manuscript received Jan. 4, 2002; accepted for publication Jan. 14, 2002.

1 Introduction

Since the inception of cDNA microarray technology[1] as a high throughput method to gain information about gene functions and characteristics of biological samples, many applications of the technology have been reported.[2–10] With the improvement of the technology, including fabrication, fluorescent labeling, hybridization, and detection, many computer software packages for extracting signals arising from tagged mRNA hybridized to arrayed cDNA locations have been designed and applied in various experiments.[11–13] As reported in Ref. 11, a target detection procedure has been implemented that utilizes manually specified target arrays, extracts the background via the image histogram, predicts target shape and then evaluates the intensities from each cDNA location and its corresponding ratio quantity.

While most software packages are satisfactory for routine image analysis and the extraction of information regarding phenomena with highly expressed genes, the desire to discover subtle effects via microarray experiments will ultimately drive experiments towards the limit of the technology,[13] with less starting mRNA and/or more weakly expressed genes. Weak signals and their interaction with background fluorescent noise are most problematic. Problems include the nonlinear trend in expression scatter plots, fishtailing at lower signal range, low measurement quality of expression levels due to uneven local background, and small cDNA-deposition areas. These artifacts, or sources of uncertainty, creep into higher-level statistical data analyses, such as clustering and classification, raising concerns about their validity. Numerous remedies have been proposed, such as carefully designed experiments in which duplications are used to minimize the uncertainty.[14,15] However, given the scarcity of certain biological samples, large duplications of experiments are often impractical. To improve detection and quantification of weak targets, it is important to understand the entire process of microarray formation, from fabrication to the scanning microscope. Use of the knowledge that the average intensity of the background fluorescence is normally distributed to help design a background detection algorithm is one example of incorporating prior knowledge into detection methods.[16]

A complex electrical-optical-chemical process is involved in cDNA-microarray technology, from fabrication of the cDNA slide, to preparing the RNA, to hybridization, to the capture of images created from excitation of the attached fluors. This complex process possesses multiple random factors. Images arising from it must be processed digitally to obtain the gene expression intensities and/or ratios that quantify relative expression levels.[11] The efficacy of the analysis to be carried out on the ratios, be it clustering,[3,17–19] classification,[5,10] prediction,[20,21] or some other, depends on the ability of the imaging algorithm to extract sufficiently accurate and consistent intensity levels from the spots. As is common in imaging applications, it is difficult (or perhaps impossible) to utilize physical ground truth as a standard by which to evaluate algorithm performance. Hence, it is common to proceed by modeling the imaging process to simulate the vari-

Address all correspondence to Edward R. Dougherty. Tel: 979-862-8154; Fax: 979-845-6259; E-mail: e-dougherty@tamu.edu

1083-3668/2002/$15.00 © 2002 SPIE

ous aspects of the real image process.[22–24] Image processing algorithms can be applied to the simulated process to evaluate their performance. One might also concurrently adjust the model parameters to see how changing various random components of the formation process impacts upon the final images, and therefore the ability to extract meaningful information. For instance, an algorithm might have biases at low signal intensities or high noise intensities that are not present at higher signal intensities or lower noise intensities. Here it should be recognized that ''ground truth'' refers to the true signal intensity, not the actual quantity of mRNA in the sample corresponding to the DNA in the spot.

Modeling anything but a very simple physical process is a very challenging task. A physical process is typically influenced, directly or indirectly, by forces whose interrelation is unknown. The resulting model will be a random process. Each realization of the model depends on random variables chosen according to various model distributions. A good quantifiable model must approximate the physical process and have realistic variability to describe the randomness of the system. In the present work, microarray image formation is modeled by a series of random processes influenced by almost two dozen parameters. We will describe the modeling process in terms of the various random variables that determine spot size, shape, and intensity, as well as variables that affect the background, including noise. Each random variable is associated with a distribution. In some cases, one may select the parameters of the distribution (such as mean and variance for a normal distribution) to reflect the image qualities of interest, such as brightness, spot size, noise intensity, etc. In other cases, the distribution of a random variable is dependent on the outcome of some other variable, and it is possible that the parameters governing the distribution of a random variable may themselves be random variables.

Although we postulate various distributions to govern the variables in the model, one may wish to use other distributions to characterize the signal and noise distributions. Moreover, the experimenter is free to choose the parameters of the distributions. Microarray technology is evolving rapidly, and there are already many variations of the technology in use. Hence, model flexibility is mandatory. For instance, for a microarray system that does not produce doughnut holes in the spots, the variables associated with the hole can be nullified. In the case of a stable system in use without change for a sufficiently long period to produce a large number of images, one can apply statistical estimation to determine some model parameters, such as those for spot radius. Clearly, these estimates will only be of value to the specific system from which they have been derived. Hence, they remain outside the simulation package per se.

The simulation algorithm produces spots at a preset grid of locations that resemble the actual microarray. Each block corresponds to a specific pin of the robot hand, and the interblock variation is modeled in the simulation by allowing various model parameters to be randomized by block. At the start of each new block, the parameters of the spots are reset. The intention of the printing process is that spots possess regular circular shapes. Due to mechanical fatigue, the adhesion process for the DNA solution concentration, and biochemical interactions, various perturbations are possible in array preparation, printing, and scanning. Various features of the model simulate these random perturbations.

2 Simulation of cDNA Microarrays

The simulation of the cDNA microarray images is designed for two-color fluorescent systems with a scanning confocal microscope. A block diagram of the overall simulation process is given in Figure 1, which includes four main modules: fluorescent background simulation, simulation of cDNA target spot generation, postprocessing simulation and tagged image file format (TIFF) image output. Each simulation module contains many sequential steps (such as spot formation) or alternative steps (such as different background fluorescence). We will discuss each step according to the order in Figure 1 in the following subsections.

2.1 Background Simulation

The fluorescent background level is an important part of expression-level estimation, since we routinely use the additive model to subtract the local background from the signal intensity measurement. It is understood that when the signal is sufficiently low, the interaction between the fluorescent background and signal affect the estimation process in most image analysis programs, resulting in lower measurement quality in the expression ratio. Many factors contribute to the observed fluorescent background: autofluorescence from the glass surface or the surface of the detection instrument, nonspecific binding of fluorescent residues after hybridization, local contamination from posthybridization slide handling, etc. A perfect system would yield a flat background possessing a normal distribution, while a microscope without an autofocus mechanism may produce a slanted background level if the slides are loaded unevenly. Some other extreme hybridization condition may cause higher nonspecific hybridization to the edge of the hybridization chamber, which effectively creates a parabolic surface of background noise. We leave the local contamination to the processing module in Sec. 2.3.

The background derived from surface fluorescence upon laser excitation is usually governed by the Poisson process, which can be approximated by a normal distribution when the arrival rate, or the accumulation of photons, is large enough.[16] This property can be readily assessed by the histogram of any background region of the microarray images. Therefore, background noise is simulated by a normal distribution whose parameters are randomly chosen to describe the process: $I_b \sim N(\mu_b, \sigma_b^2)$. If multiple arrays are desired, the inter-array difference is modeled by a uniform distribution: $\mu_b \sim U(a,b)$. σ_b is given as a multiple of μ_b: $\sigma_b = k_b \mu_b$. Typically, k_b is about 10% of the mean background level.

Rather than be constant across the entire microarray, the mean of the background noise may vary owing to various scanning effects. It can take different shapes: parabolic, positive slope, or negative slope. In this case a function $g(x,y)$ is first generated (parabolic, positive slope, or negative slope) to form a background surface and normal noise is added to it pixel wise. Thus, the background intensity is of the form $I_b \sim N(\mu_b, \sigma_b^2)$ with $\mu_b = \gamma g(x,y)$, where $\gamma \sim U(a,b)$ is the targeted background noise level. Background deviation is set independently for each channel: $\sigma_{b_1} = k_{b_1} \mu_b$ and σ_{b_2}

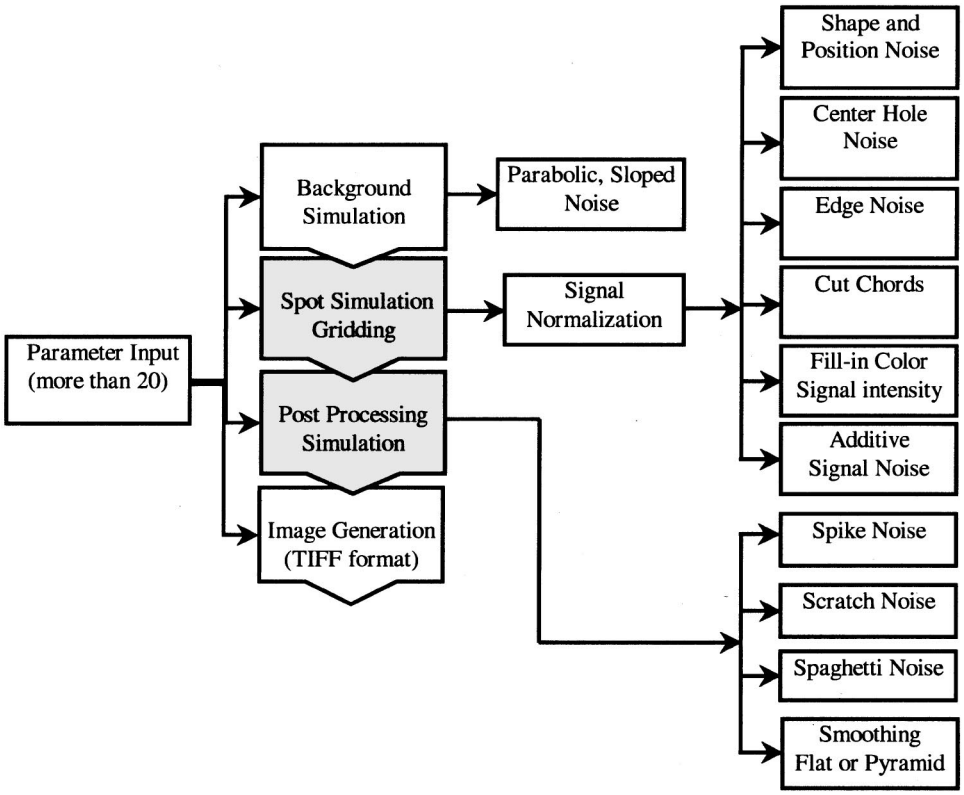


Fig. 1 Figure shows the steps involved in generating the microarray.

$=k_{b_2}\mu_b$. Figure 2 shows various noise backgrounds with $k_{b_1}=k_{b_2}=0.1$. All images are shown in large size on a web page.[27]

In many practical examples, the nonspecific hybridization at the target location may be different from its peripheral region. Although one may have trouble pin-pointing this particular observation under normal conditions owing to signal interference, it is sometimes unmistakable when locations assumed to be weakly expressed, or not expressed at all, carry some nonzero readouts, or the intensity in the center is stronger than the doughnut ring if the printed target is doughnut shaped. We simulate this artifact under a gradient noise condition by allowing the background for the center holes to be at higher levels than the signal intensities. Hence, there is an option to use global background or local background information to set the noise parameter for the center hole. Figure 3 shows the effects of using local and global background parameters. This effect may not appear everywhere in a simulated image; however, it is often sufficient to require appropriate algorithm design in the image analysis program to lessen the penalty. The effects of weak targets will be further studied in later sections.

2.2 Spot Simulation

cDNA deposition routinely follows a rigid grid defined by the robotic print pattern. The simulation algorithm produces spots

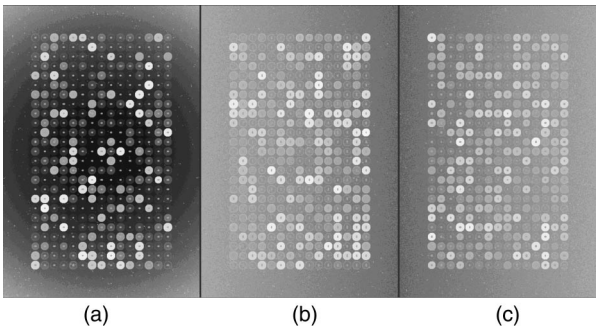


Fig. 2 Figure shows various background noises. The mean SNR is set at 1.0 for the slides. The slides have following settings: (a) parabolic back ground noise, (b) positive slope background, and (c) negative slope background all with global noise parameter. The background deviation factor is set at $k_{b_1}=k_{b_2}=10\%$.

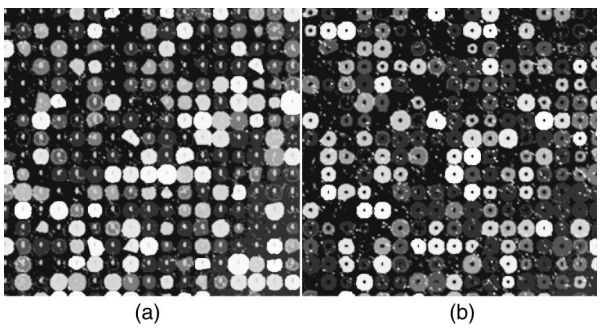


Fig. 3 Example shows different noise settings for spots inner hole. Where (a) uses global background parameter to fill the center hole, (b) uses local background for filling the center hole. The background noise is set to sloped type with SNR of 1.5.

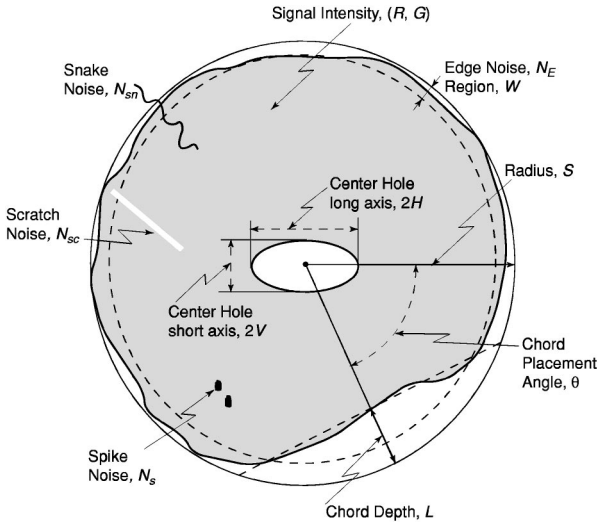


Fig. 4 cDNA microarray spot model.

at preset grid locations that resemble the actual microarray. In principle, print tips are manufactured uniformly; however, their microscopic morphologies, and thus their deposition-binding behaviors, are noticeably different. Each block corresponds to a specific print tip of the robot hand. To take tip variability into account, within each block the spot variation is governed by block parameters, which themselves are random variables. At the start of each new block, the spot parameters are reset according to these random variables.

The key simulation of this study is devoted to the cDNA targets, which nominally possess a circular shape. Owing to many factors, the actual shape may be highly noncircular. The model takes various random perturbations into account: (1) radius variation, (2) spot drifting locally, (3) center core variation, (4) chord removal, (5) edge noise, (6) edge enhancement, (7) signal intensity, and (8) signal response transform. Figure 4 shows a schematic drawing for the cDNA target simulation. The variables in the figure are explained in the following eight subsections.

2.2.1 *Variation of Radius*

Prior to distortion and noise, the cDNA deposition spot is considered to be circular with random radius S. The mean of the radius is set according to the array density and its variance relates to the consistency of spot size. S is modeled by a normal distribution having mean μ_s and variance σ_s^2, $S \sim N(\mu_s, \sigma_s)$, with the standard deviation being a predetermined proportion, k_s, of the mean, or $S \sim N(\mu_s, k_s\mu_s)$. The radius mean is set for every block, and randomized over a small range within the array. The block randomness of μ_s is modeled by a uniform distribution, $\mu_s \sim U(s_a, s_b)$. Figure 5 shows parts of blocks with spot radii depending on the number of spots in a block. For Figures 5(a)–5(c), the block portions are for block sizes (10,15), (25,45), and (25,45), respectively, where (col, row) denotes the number of spots in columns and rows within the block, respectively. Occasionally, a spot overlaps with it neighbors [Figure 5(c)] when k_s is set to a larger proportion. This situation simulates the condition where too much cDNA solution is deposited and/or the drying process may be slow in comparison to the liquid spreading process.

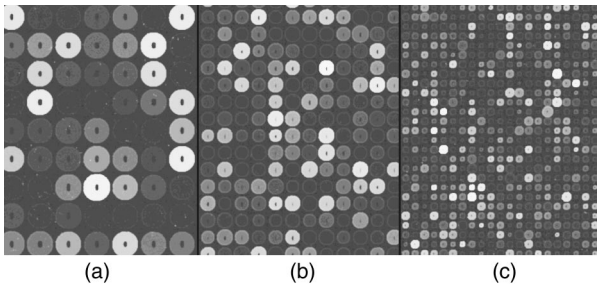


Fig. 5 Figure shows the variability in spot size and spread from its size. The spot radius distribution is automatically set depending on the number of spots in a block (width, height). In the earlier example has (a) (10,15), $\mu_s \sim U[23.3\ 24.3]$, (b) (20,25), $\mu_s \sim U[12.6\ 13.6]$ and (c) (25,45), $\mu_s \sim U[5.45\ 6.45]$, with standard deviation $k_s = 1\%$, 7%, 20% of radius, respectively.

Depending on the robot arm and printing ability of the pins, the interspot distance, G_{sp}, may vary. Owing to the physical mechanics of the robot arm, the block size (pixel units) is fixed in most cases. The interspot distance can be set to accommodate spot size and random variation in spot radii. The effects are illustrated in Figure 6, where the number of rows and columns are fixed.

2.2.2 *Spot Drift*

During the fabrication stage, the deposition of cDNA targets may not follow the predefined grid owing to print-tip rotation, vibration, or other mechanical causes. Other drifts are attributed to the slide's coating properties and the drying rates of the cDNA. This displacement is modeled by possible random translations in the horizontal and vertical directions. Each spot has an equal probability, P_D, of drifting. If a spot is selected for drift, then the amounts of drift in both directions are random multiples of the current spot radius. The horizontal and vertical multiples, δ_x and δ_y, called the ''drift levels,'' are uniformly distributed: δ_x, δ_y, $\sim U(d_a, d_b)$. The horizontal and vertical drifts are $D_x = \delta_x S$ and $D_y = \delta_y S$, respectively. Interspot distance can be set according to the drift to minimize the impact of overlapping spots.

Some microarray scanners capture two fluorescent signals in two passes of scanning. Due to the mechanical homing error, the two fluorescent channels may not align exactly. In

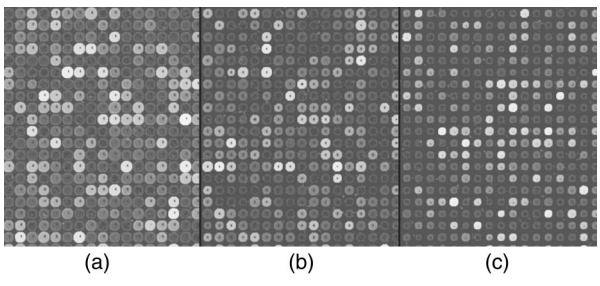


Fig. 6 Figure shows interspot grid spacing, (a) $G_{sp} = 3$ pixels, $\mu_s \sim U[9.5\ 10.5]$, (b) $G_{sp} = 6$ pixels, $\mu_s \sim U[8\ 9]$, (c) $G_{sp} = 10$ pixels, $\mu_s \sim U[6.5\ 7.5]$. The example has (35,20) rows, columns respectively with $k_s = 0.05$.

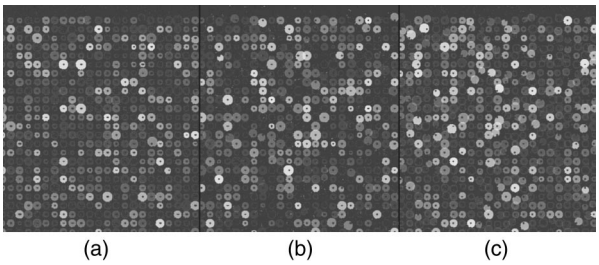


Fig. 7 Figure shows the effect of radius drift (P_d, d_a, d_b). (a) (0.05,5,100), (b) (0.25,15,100), (c) (0.5,50,100). As the activation probability with drift range is set higher, the spots drift away from its center.

these settings, some small offset between the two channels can be observed. This offset may occur at subpixel resolution. To simulate this offset, the model offers a random offset between the centers of the two channels. It is achieved by randomly offsetting the spot center of the second channel by one pixel in either of the horizontal and vertical directions. These offsets are applied following application of the spot drifts. Figure 7 illustrates the spot drift.

It is essential for the image analysis algorithm to determine the exact location of the target spot so that an accurate measurement can be carried out without the interference of the dusty noise around the targets. Some algorithms rely on the assumption that the printing grid is rigid with the cDNA target in the center; others assume an imperfect printing process such that a deformable grid is necessary. The former method is faster and noise insensitive, but may be inaccurate if the slides are fabricated with many displacements; the latter is robust in target position detection, but can be rather slow and noise sensitive. In either case, the simulation outcome will provide a set of evaluation images to assess the tolerance of both algorithmic designs. The slightly misaligned channels also pose a challenge to signal intensity extraction.

2.2.3 *Doughnut Hole*

Owing to the impact of the print tip on the glass surface, or possibly due to the effect of surface tension during the drying process, a significantly lesser amount of cDNA can be deposited in, or attached to, the center of the targets. Consequently, the center of the target emits less fluorescent photons, thereby giving a target the doughnut shape. It is critical for signal intensity extraction whether or not the center hole is assumed, particularly when the signal is weak and there is a large center hole. The simulation allows one hole in the center with varying size, along with a possible off-center displacement. It is not necessary to simulate more than one hole, since the mathematical properties for signal and noise estimation are preserved with this simple condition.

An elliptical shape models the inner core with random horizontal and vertical axes, H and V. The axes are modeled by a normal distribution whose parameters are randomized for each block within a given array: $H \sim N(\mu_H, \sigma_H)$ and $V \sim N(\mu_V, \sigma_V)$. Interarray variability in these radius distributions is modeled by uniformly distributed means: $\mu_H \sim U(a_H, b_H)$, $\sigma_H = \alpha_1 \mu_H$ and $\mu_V \sim U(a_V, b_V)$, $\sigma_V = \alpha_2 \mu_V$,

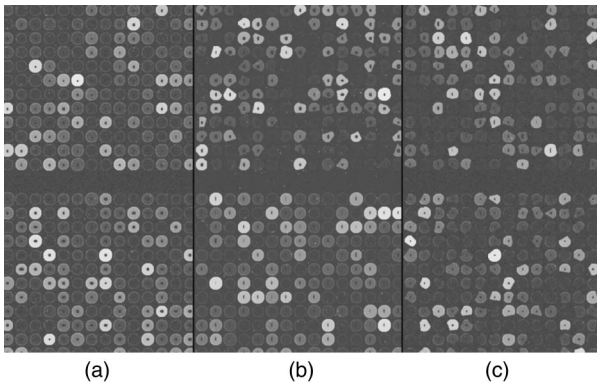


Fig. 8 Figure shows different chord rate settings for each of the slide. The probability weights for (0,1,2,3,4) chord rates were set at following levels. (a) (0.7,0.3,0.0,0,0), (b) (0.2,0.4,0.25,0.15,0), (c) (0.0,0.1,0.4,0.3,0.2), respectively. Chord rate is reset at the beginning of a block.

where the controlling ratios vary over a range, $\alpha_1, \alpha_2 \sim U(P_a, P_b)$. The choice of the parameters governs the hole shapes. The center position of a hole is allowed to drift over a range. The shape is unaffected by the drift because the mechanical print tip to surface contact is unaffected. The amount of drift in the horizontal and vertical directions is modeled similarly to spot drift. Drift levels are set at every block, $(\delta c_{xR}, \delta c_{yR})$ and $(\delta c_{xG}, \delta c_{yG})$, for both channels. The amount of drift is first selected from a uniform range, $\delta c \sim U[i,j]$. Channel and interchannel drifts are modeled by a uniform variate and set for each block: $\delta c_{xG} = \delta c U[-1,1]$, $\delta c_{yG} = \delta c U[-1,1]$, $\delta c_{xR} = \delta c_{xG} + U[-1,1]$, and $\delta c_{yR} = \delta c_{yG} + U[-1,1]$.

2.2.4 *Chord Removal*

Since parts of a spot can be washed off due to various physical effects during the hybridization and washing stages, pieces of a spot may be missing. We would like to simulate this condition for the same reasons that the center hole is simulated. This irregularity is modeled by randomly cutting chords from the circular spots. The number of chords to be removed, N_c, for a spot is selected from a discrete distribution, {0, 1, 2, 3, 4}, where the elements of the distribution occur with probabilities p_0, p_1, p_2, p_3, and p_4, respectively. For images with very few pieces cut off, the zero-chord probability p_0 is very high, and the three- and four-chord probabilities are close to 0 (possibly equal to 0). To model interarray variability, the probabilities can be treated randomly.

Once the number of chords for a spot is determined, the distance, L, of each chord center to the edge is selected from a beta distribution: $L \sim B(\alpha_L, \beta_L)$. Interblock variability is modeled by allowing α_L and β_L to be randomly selected from uniform distributions: $\alpha_L \sim U(a_\alpha, b_\alpha)$, and $\beta_L \sim U(a_\beta, b_\beta)$. Owing to the large family of shapes generated by beta distributions, this provides a wide range of distributions for L. Finally, the chord locations are chosen uniformly randomly according to an angle $\theta \sim U(0, 2\pi)$. Figure 8 illustrates the effect of selecting increased chord rates: (a) $p_0 = 0.70$, $p_1 = 0.30$; (b) $p_0 = 0.20$, $p_1 = 0.40$, $p_2 = 0.25$, $p_3 = 0.15$; (c) $p_0 = 0$, $p_1 = 0.10$, $p_2 = 0.40$, $p_3 = 0.30$, $p_4 = 0.20$.

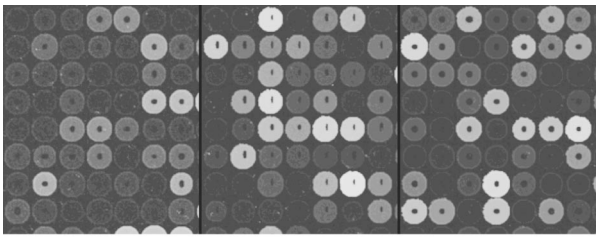


Fig. 9 Figure shows the edge noise on the spots. Noise controlling parameter (δ) can be set from [0,1.0]. The example shows an increased edge noise effect, where (a) $\delta=0.25$, (b) $\delta=0.1$, (c) $\delta=0.03$, where δ is the proportion of maximum intensity.

2.2.5 *Edge Noise*

Owing to the manner in which liquid dries, the spots usually do not have smooth edges. To provide a realistic visual effect, as well as to pose a challenge if edge detection algorithms are under consideration, we simulate this irregular edge effect via parameterized noise using a binary edge-noise algorithm employed in digital document processing.[25] After determining the target shape by cutting the center hole, removing possible chords, and possibly creating drift, and prior to simulating the signal intensity, the spot is still in its binary format, and thus the binary edge-noise algorithm can be applied directly. Edge noise is applied to both the outer perimeter of the spot and the inner perimeter containing the hole.

The algorithm begins by first generating a white noise (mask) image having range [0, max intensity]. A 3×3 averaging filter is applied to the white-noise image to arrive at a noise image N that possesses a degree of correlation resembling the noise characteristics of various physical processes, including printing processes. The edge of a binary image can be considered to consist of two parts, inner and outer borders. In our case, the spot radius is known and so are these borders. The inner border is formed by morphologically eroding the image by a 3×3 structuring element and then subtracting the erosion from the original image. The outer border is formed by morphologically dilating the image by a 3×3 structuring element and then subtracting the original image from the dilation. To apply noise to the inner border, a threshold, mid $+\delta$, just above midpoint is applied to N, this binary image is ANDed with the inner border of the original binary spot S, and the result is XORed with S. Noise is applied to the outer border by thresholding N just below the midpoint $(\text{mid}-\delta)$, complementing, and then ANDing with the outer border of S. This noisy outer border is then ORed with the image possessing inner border noise to yield the edge-degraded binary spot S'. The process is mathematically described by

$$S'=[(N_{\text{mid}+\delta}\cap S_{\text{in}})\Delta S]\cup[(N_{\text{mid}-\delta})^c\cap S_{\text{out}}], \qquad (1)$$

where δ controls the threshold and hence the edge noise, and Δ denotes the symmetric difference. δ is used as controlling parameter. S' is a binary mask giving the spatial domain of the spot. Figure 9 shows edge noise for various δ thresholds.

2.2.6 *Signal Intensity*

Simulation of signal intensity is divided into three steps. First, it is assumed that the fluor-tagged mRNAs cohybridized to a single slide are from the same cell type, and therefore the signals from the two fluorescent channels are supposed to be identical, with some variation. Second, some percentage of genes may be selected as significantly over- or underexpressed. Third, foreground noise is added to the entire array to simulate the normal scanning integration process.

It is well known that the distribution of gene expression levels within a cell closely follows an exponential distribution.[26] Given a microarray containing N genes, the intensity levels I_k, for $k=1,\ldots,N$, assumed to be related to the expression levels of N genes, are simulated by an exponential distribution. This intensity level I_k is considered to be the ground-truth signal that is not directly measurable from the microarray, since from either biological or bio-chemical processes, from mRNA extraction up to the hybridization process, some variation will be introduced into measurement of final fluorescent signal strength. For each microarray, a particular exponential distribution with mean β is first chosen (for a detection system with gray-level up to 65 535, β is usually selected around 3000). Then at each spot location, which we assume to represent one unique gene, one ground-truth signal level I_k is generated from the exponential distribution. For two observable measurements (R_k,G_k) from two fluorescent channels, two numbers are generated from a normal distribution with mean of I_k and standard deviation of αI_k, where α is a predetermined coefficient of variation, which is usually about 5%–30% depending on the assumed biological relation between the two channels.

To include outlier expression levels that reflect certain realistic conditions,[3–10,14] one may select 5%–10% of the spots to be either over- or underexpressed. This condition is achieved by selecting the genes from the entire microarray based on a probability, p_{outlier} (e.g., $p_{\text{outlier}}=0.05$ for 5% outliers), and then selecting the targeted expression ratio for the kth gene

$$t_k=10^{\pm bk}, \qquad (2)$$

where b_k satisfies a beta distribution, $b_k\sim B(1.7,4.8)$, and where the $+/-$ sign is selected with equal probability. Upon obtaining a targeted expression ratio, the algorithm converts the expression intensities from the two fluorescence channels by

$$R_k'=R_k\sqrt{t_k},\qquad G_k'=\frac{G_k}{\sqrt{t_k}}, \qquad (3)$$

where R_k' and G_k' denote the signal values after the conversion.

Upon obtaining the signal intensities for each spot, (R_k',G_k'), each pixel within the spot binary mask derived from steps 2.2.1 to 2.2.5 is filled with the signal intensity. Normally distributed foreground noise is then added pixelwise. This yields, at each pixel, the intensities $SR=R_k+I_{f1}$ and $SG=G_k+I_{f2}$, where $I_{f1}\sim N(\mu_{R_k},\sigma^2_{R_k})$, $I_{f2}\sim N(\mu_{G_k},\sigma^2_{G_k})$ and $\mu_{R_k}\sim R_k'U[f_{a_1},f_{b_1}]$, $\sigma_{R_k}\sim\mu_{R_k}U[f_{c_1},f_{d_1}]$, $\mu_{G_k}\sim G_k'U[f_{a_2},f_{b_2}]$, and σ_{G_k}

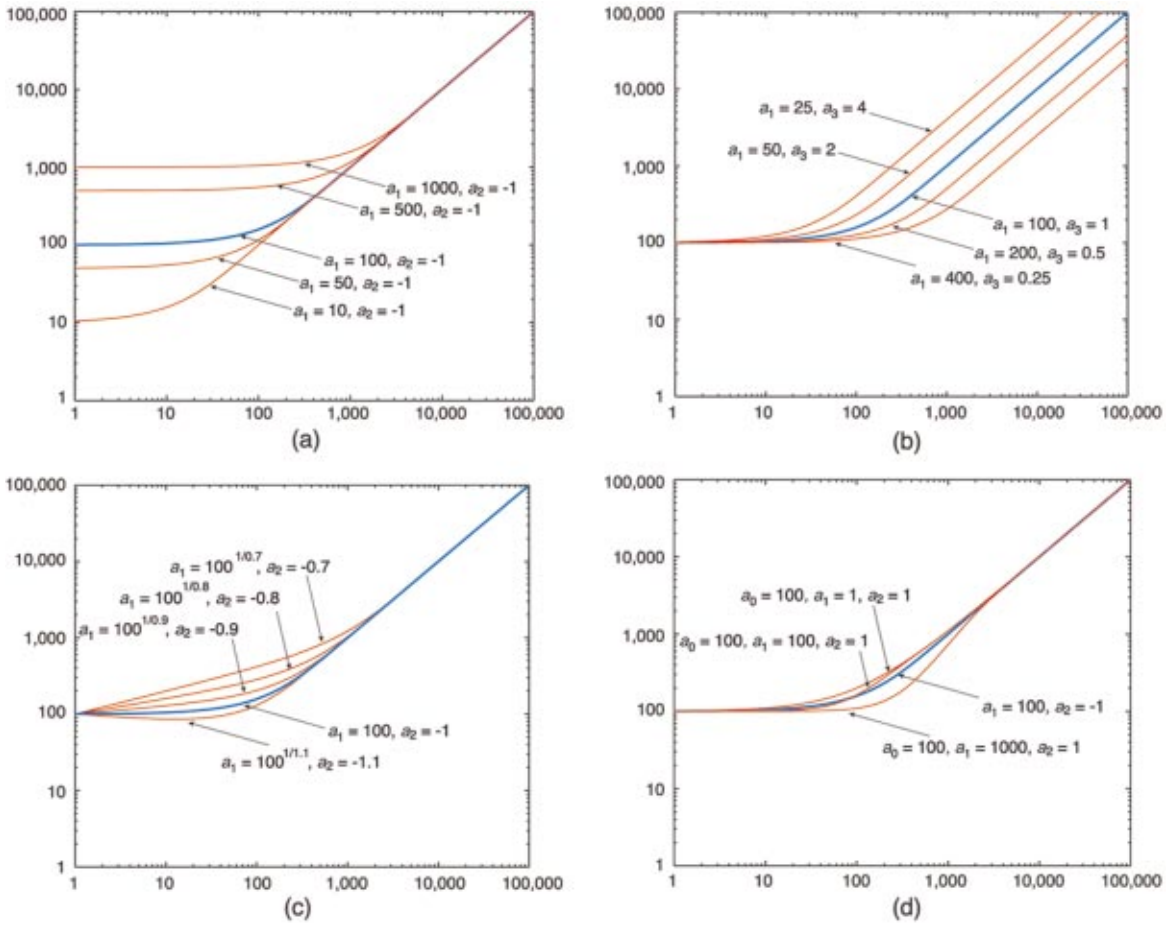


Fig. 10 Fluorescent detection response characteristic functions. In all figures, middle (blue) curve is the reference function with parameters of $(a_0, a_1, a_2, a_3) = (0,100,-1,1)$. Also, in all figures, the x axis is the input signal intensity, and y axis is the observed signal intensity, and both are in $\log_{10}$ scale. (a) Delayed response at various levels, with fixed $a_0 = 0$ and $a_3 = 1$. (b) Different amplification levels, with fixed $a_0 = 0$ and $a_2 = -1$. (c) Different response curvature, with fixed $a_0 = 0$ and $a_3 = 1$. (d) Some other parameter settings, with fixed $a_3 = 1$.

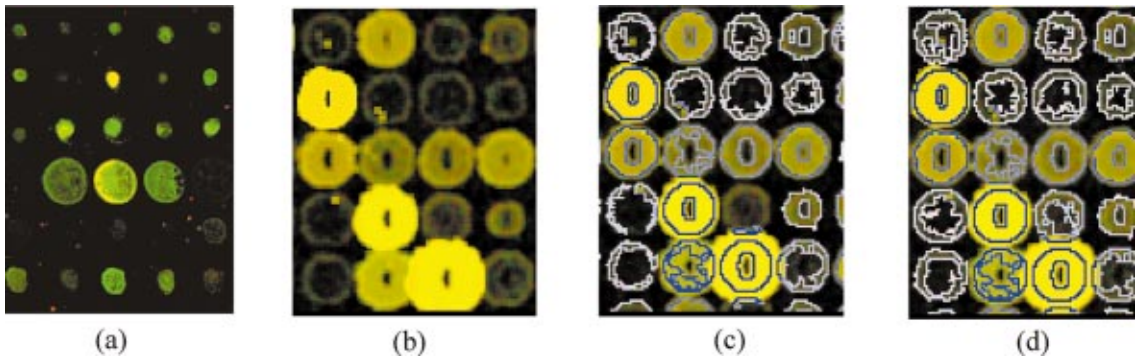


Fig. 19 (a) Part of actual hybridized image with spots larger than average; (b) simulated microarray with larger spots and spots overlapping with their neighbors; (c) original background intensity extraction program produces undetected spot (target in the middle without outer boundary); (d) improved background extraction program more accurately measures the local background intensity and effectively allows detection of weak targets.

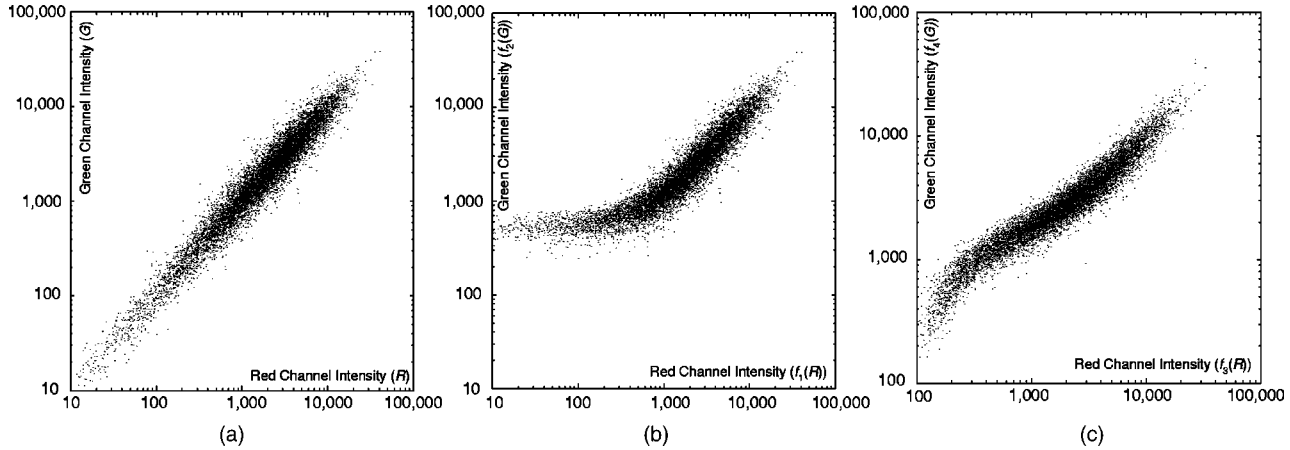


Fig. 11 Possible scatter plot due to various response conversions for different fluorescent channels. 10 000 data points (gene expression levels) were generated by the exponential distribution with mean of 3000. After passing, through two fluorescent channels [with some response characteristic functions as shown in parts (a)–(c)], data variations were added by passing each data point through a normal distribution with the standard deviation to be 15% of mean expression signal. (a) Without any alteration [or equivalently, set parameters for the response function to be $(a_0,a_1,a_2,a_3)=(0,1,-1,1)$], and assume the signal intensities from red channel and green channel are equivalent (a simulated self–self experiment). (b) Banana shape. Intensity in green channel pass a response function with parameters $(a_0,a_1,a_2,a_3)=(0,500,-1,1)$, where red channel takes the parameters (0,10,–1,1). (c) Sinusoid-shape. The red channel's response function with parameters $(0,100^{1/0.7},-0.7,1)$, and the green channel with $(0,100^{1/0.9},-0.9,1)$.

$\sim\mu_{G_k}U[f_{c_2},f_{d_2}]$. In the remainder of the paper, α's are used to denote the uniform variables $\alpha_{m_1}\sim U[f_{a_1},f_{b_1}]$, $\alpha_{m_2}\sim U[f_{a_2},f_{b_2}]$, $\alpha_{s_1}\sim U[f_{c_1},f_{d_1}]$, and $\alpha_{s_2}\sim U[f_{c_2},f_{d_2}]$.

2.2.7 *Channel Conditioning*

Owing to various reasons, such as imprecise quantities of starting mRNA for the two channels, different labeling efficiencies, or uneven laser powers at the scanning stage, in actual microarray experiments there may not be equal intensities even if two channels use exactly the same labeled mRNA. Moreover, one may not be able to assume that the fluorescent intensity is linearly related to the expression level. In fact, it is very difficult to determine the exact form of the response function from expression level to intensity due to the complex combination of bio-chemistry to photon electronics. We choose a family of functions that covers most of the understandable conditions, shown in Figure 10, such as delayed response, saturation (which is an embedded feature in the digital system since no gray level can pass 16-bit binary digits in a typical microarray system), and unbalanced channel intensity. This simulation is intended to facilitate understanding as to what is the best way for expression ratio normalization, whether linear based methods will be sufficient or nonlinear based methods will be necessary. The function family is characterized by four parameters, (a_0,a_1,a_2,a_3), and the function form is given by

$$f(x)=a_3[a_0+x(1-e^{-x/a_1})^{a_2}];\quad a_3>1. \tag{4}$$

Having chosen a function from the family, the expression levels, R' and G', from each fluorescent detection channel are then transformed by the detection system response characteristic function defined by $f_R(x)$ or $f_G(x)$ to obtain the realistic fluorescent intensity observed. The observed fluorescent intensities are

$$R''_k=f_R(R'_k),\qquad G''_k=f_G(G'_k), \tag{5}$$

where f_R or f_G may take different parameters for each fluor-tagging system. The simulation performs the following steps for signal placement to emulate the real process affecting the signal spots.

1. Generate ground truth expression signal I_k $(k=1,\ldots,N)$ for every gene by exponential distribution (see Sec. 2.2.6).
2. Let $R_k\sim N(I_k,\propto I_k)$ and $G_k\sim(I_k,\propto I_K)$. If a self–self experiment needs to be simulated, skip steps 3 and 4.
3. If we simulate an experiment with two different samples, some outlier genes are selected and then their intensities are altered. We obtain (R',G') from (R,G) for all genes [see Sec. 2.2.6, and Eqs. (2) and (3)].
4. If we simulate a fluorescent system with imperfect response characteristics, the intensities are further converted by $R''=f_R(R')$ and $G''=f_G(G')$ (see Sec. 2.2.7).
5. The actual simulated fluorescent intensities for both channels are obtained by applying additional variation via a normal distribution function $SR=R''+N(\mu_R,\sigma_R^2)$, where $\mu_R=\alpha_{m1}R''$, $\sigma_R=\alpha_{s1}\mu_R$, and similarly for signal G (see Sec. 2.2.6).

The scatter plots in Figure 11 show the effects of the channel normalization. By choosing different parameter sets, one can simulate many of the situations observed in real microarray images.

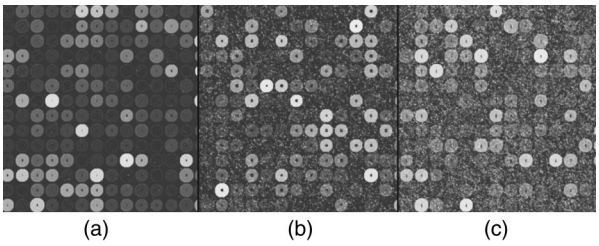


Fig. 12 Figure shows increased spike noise levels L_{spi}. (a) Level of 0.1%, (b) level of 5%, (c) level of 10%, exponential rate range is maintained.

2.2.8 *Edge Enhancement*

Under some fabrication conditions, such as incorrect humidity control, where the cDNA solution tends to accumulate towards the outer edge during the drying process, the spot edge may appear brighter than the rest of the spot. This phenomenon is modeled by randomly enhancing the edge. The number, N_e, of pixels from the edge to be enhanced is fixed. The enhancement, W_{ed}, is added to the original intensity. W_{ed} satisfies a normal distribution, $W_{ed} \sim N(\mu_e,1)$. Randomness between blocks is modeled by making μ_e uniformly distributed, $\mu_e \sim U(l_a, l_b)$.

2.3 Postprocessing Simulation

Most postprocessing steps simulate handling and scanning artifacts: scratch noise resulting from improper handling of microarray slides, spike noise arising from the impurity of mRNA extraction steps or perhaps insufficient washing conditions, snake noise due to the accumulation of dust if the slides have sat in open space too long, and last, but not least, smoothing resulting from many scanners' averaging effects or integration processes. For the most part, these steps model the interaction between signal and noise in the spatial domain, which causes pixel-wise nonlinear degradation. It is expected that the microarray image analysis software shall be able to handle most of the noise conditions outlined here in order to measure the signal precisely.

2.3.1 *Spike Noise*

In a practical biology laboratory, it is not necessary to maintain a dust-free environment. Hence, fine microscopic dust particles are nearly impossible to avoid. On laser excitation, these particles fluoresce to give high intensity spikes. Moreover, in some cases, bad mixtures of cDNA solutions result in precipitation, and these particles fluoresce with a very high intensity. These effects are simulated by adding spike noise at a preset rate. Such intensity spikes are added randomly across the entire slide area, the number of such noise pixels being preset in terms of the total number of pixels in the array. The amount of spike noise in an array is set with reference to the percentage, L_{spi}, of the total number of pixels in the array. Typical low to high noise levels are to be set by selecting 0.1%–10%. Once a pixel is selected for spike noise, the adjacent pixels have a higher probability of being affected. Thus, a random number, W_{spi}, of pixels are chosen in an arbitrary direction to be influenced by this noise. The intensity, N_S, of the spike noise is governed by an exponential distribution with mean μ_{spi}. In Figure 12, the exponential mean is fixed but the spike level is increased through the parts of the figure.

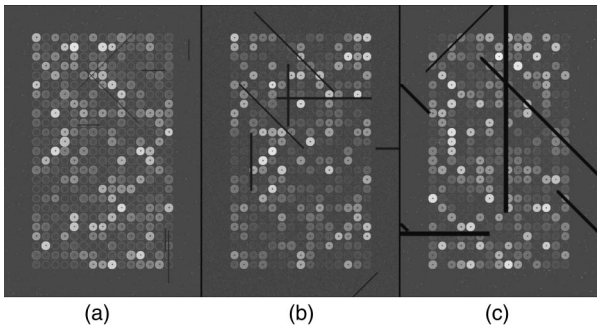


Fig. 13 Figure shows scratch noise with its parameter settings. Number of scratches is maintained to 7 in the earlier examples. Following are the parameter (a) $L_{sc} \sim U[2\ 7]$, $\kappa_{sc}=1.5$, $W_{sc}=3$ pixels, (b) $L_{sc} \sim U[5\ 15]$, $\kappa_{sc}=2.5$, $W_{sc}=7$ pixels, (c) $L_{sc} \sim U[8\ 45]$, $\kappa_{sc}=4.0$, $W_{sc}=15$ pixels. The noise factor $k_{sc}=0.1$.

2.3.2 *Scratch Noise*

Physical handling of the array slides can result in surface scratches. These typically result in low intensity levels. Scratch-noise intensity is parameterized as a ratio, κ_{sc}, giving the background-to-scratch-noise intensity level. Other parameters are the number of strips, strip thickness W_{sc}, and a random strip length, L_{sc}, given as a multiple of the spot size. The latter is modeled as a uniform distribution: $L_{sc} \sim U[L_{sc1}, L_{sc2}]$. Strips are placed at random positions on the array, and are inclined according to a (discrete) uniformly random angle, $\theta_{sc} \in \{0°,45°,90°,135°,180°\}$. Figure 13 shows the noise for incremental parameter settings: (a) $L_{sc} \sim U[2,7]$, $\kappa_{sc}=2.0$, $W_{sc}=$ four pixels; (b) $L_{sc} \sim U[5,10]$, $\kappa_{sc}=3.0$, $W_{sc}=$ seven pixels; (c) $L_{sc} \sim U[7,15]$, $\kappa_{sc}=4.0$, $W_{sc}=$ ten pixels. The number of strips is fixed at 7.

2.3.3 *Snake Noise*

Fine fabric dust particles on the slides can create snake-tailed strips on laser excitation. These strips are normally higher intensity than the signal level. To simulate this noise, an equiprobable multidirectional snake noise has been generated con-

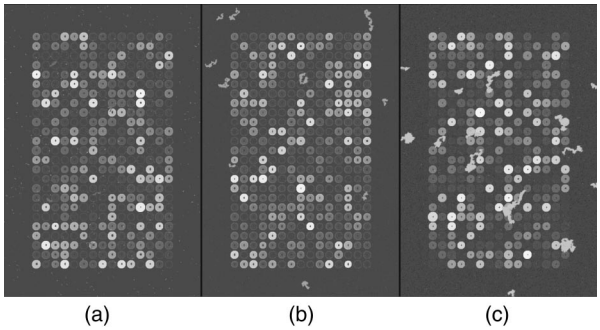


Fig. 14 Example shows different parameter setting for snake noise. In this example (a) $N_{seg}=5$, $L_{sp} \sim U[5\ 10]$, $\kappa_{sn}=0.5$, $W_{sp}=2$ pixels, (b) $N_{seg}=10$, $L_{sp} \sim U[5\ 30]$, $\kappa_{sn}=0.33$, $W_{sp}=3$ pixels, (c) $N_{seg}=15$, $L_{sp} \sim U[5\ 80]$, $\kappa_{sn}=0.25$, $W_{sp}=5$ pixels, respectively. Direction of the tail was randomly chosen with equal probability for each.

1	1	1
1	1	1
1	1	1

(a)

0.5	1	0.5
1	2	1
0.5	1	0.5

(b)

Fig. 15 Example shows the 3×3 convolution kernel for (a) flat function and (b) pyramidal function.

sisting of some number, N_{seg}, of segments. Analogously to scratch noise, the intensity is parameterized as a ratio, κ_{sn}, giving the average-signal-to-snake-noise intensity level, the number of snakes, snake thickness W_{sn}, and a random length, L_{sn}, given as a multiple of the spot size. The latter is modeled as a uniform distribution: $L_{sn}\sim U[L_{sn1},L_{sn2}]$. Figure 14 shows the noise for incremental parameter settings: (a) $N_{seg}=5$, $L_{sn}\sim U[5,10]$, $\kappa_{sn}=0.50$, $W_{sn}=$two pixels; (b) $N_{seg}=10$, $L_{sn}\sim U[5,30]$, $\kappa_{sn}=0.33$, $W_{sn}=$three pixels; (c) $N_{seg}=15$, $L_{sn}\sim U[15,80]$, $\kappa_{sn}=0.25$, $W_{sn}=$five pixels.

2.3.4 *Smoothing Function*

Addition of various noise types makes the microarray highly peaked with high pixel differences. This stark irregularity can be mitigated by smoothing the image with either a flat or pyramidal convolution kernel. The kernels are shown in Figure 15. The effect of smoothing is illustrated in Figure 16, where the three-dimensional (3D) profile of an originally noised image is shown, along with versions smoothed by flat and pyramidal kernels. Either smoothing kernel can be chosen.

2.4 Image Generation and Parameter *I/O*

Parameters governing the effects described in the preceding sections form the input (through a file) to the synthetic array software. These include parameters for array dimensions, shape parameters, and noise processes. All relevant information, such as spot size, position, various drifts (center hole, spot), noise processes, (foreground, spike, snake, scratch, etc.), and chord rate, are recorded for every spot printed on the synthetic array. Block controlling parameters and the array information are also recorded. The recorded information con-

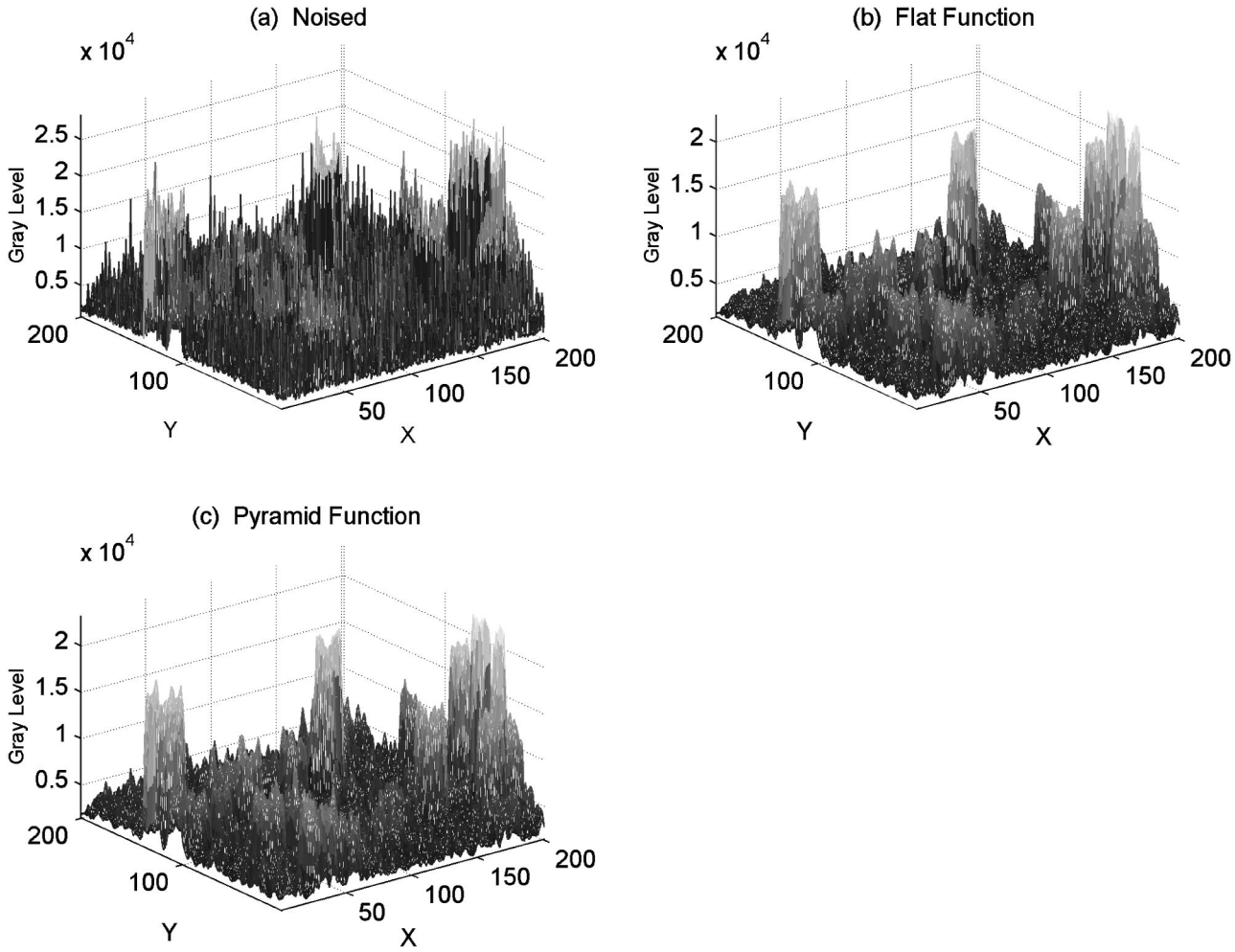


Fig. 16 Example shows the 3D profile before and after smoothing. Where (a) noised, (b) flat function, (c) pyramid function.

Table 1 Parameter settings for the cDNA microarray simulation.

Level	Simulation	Parameter descriptions	Distribution
SPOT	Spot size	S: Spot radius with (μ_s, σ_s^2)	$S \sim N(\mu_s, \sigma_s^2)$
	Spot drift	δ_x, δ_y: Drifting level	$\delta_x, \delta_y \sim U(d_a, d_b)$
		d_a, d_b: percentage of spot radius	
		P_D: Drift activation probability	$D_x = \delta_x SU[-1,1]$
		D_x, D_y: Relative drifting	$D_y = \delta_y SU[-1,1]$
		(X_1', Y_1'): Drifted center coordinates	$\begin{cases} X_1' = X + D_x \\ Y_1' = Y + D_y \end{cases} \begin{cases} X_2' = X_1' + U[-1,1] \\ Y_2' = Y_2' + U[-1,1] \end{cases}$
		(X_2', Y_2'): Second channel, where (X,Y) is predefined spot center coordinates	
	Inner hole size	H,V: Horizontal and vertical axis of the inner elliptical hole	$H \sim N(\mu_H, \sigma_H)$ $V \sim N(\mu_V, \sigma_V)$
	Inner hole drift	X_C, Y_C: Ideal spot center	$X_R = X_C + \delta c_{xR}$
		X_R, Y_R: First channel coordinates	$Y_R = Y_C + \delta c_{yR}$
		X_G, Y_G: Second channel coordinates where	$X_G = X_C + \delta c_{xG}$
		$\delta c_{xG}, \delta c_{yG}, \delta c_{xR}, \delta c_{yR}$: drift level set at the block level	$Y_G = Y_C + \delta c_{yG}$
	Chord removal	P_{N_c}: Chord removal probability (p_k: probability of k chords to be removed from a target spot)	$P_{N_c} = \{p_0, p_1, p_2, p_3, p_4\}$, where $p_0 + p_1 + p_2 + p_3 + p_4 = 1$ $N_c \sim \{0,1,2,3,4\}$
		L: Chord length	$L \sim B(\alpha_L, \beta_L)$
		θ: Chord position	$\theta \sim U(0, 2\pi)$
	Spot intensity	β: Mean intensity for the assumed cell system	$I_k \sim \text{Exp}(\beta)$
		R_k, G_k: kth spot (fixed) signal intensities for both channels	$R_k \sim N(I_k, \sigma_I)$ $G_k \sim N(I_k, \sigma_I)$
		α: Coefficient of variation of signal intensity in the system	$\sigma_I = \alpha I_k$
	Outlier's intensity	p_{outlier}: Outlier activation probability	
		b_k: Outlier control level	$b_k \sim \text{Beta}(1.7, 4.8)$
		t_k: Targeted outlier expression ratio, with equal-probability for +/− sign	$t_k = 10^{\pm b_k}$
		R_k', G_k': kth outlier signal intensities for both channels	$R_k' = R_k \sqrt{t_k}$ $G_k' = G_k / \sqrt{t_k}$
	Channel conditioning	R_k'', G_k'': Prenormalized signal intensity of the spots on red, green channels	$R_k'' = f_1(R_k')$ $G_k'' = f_2(G_k')$
		a_0, a_1, a_2, and a_3, parameters for response characteristic function.	$f(x) = [a_0 + x(1 - e^{-x/a_1})^{a_2}]a_3$; where $a_3 > 1$

Table 1 *(Continued.)*

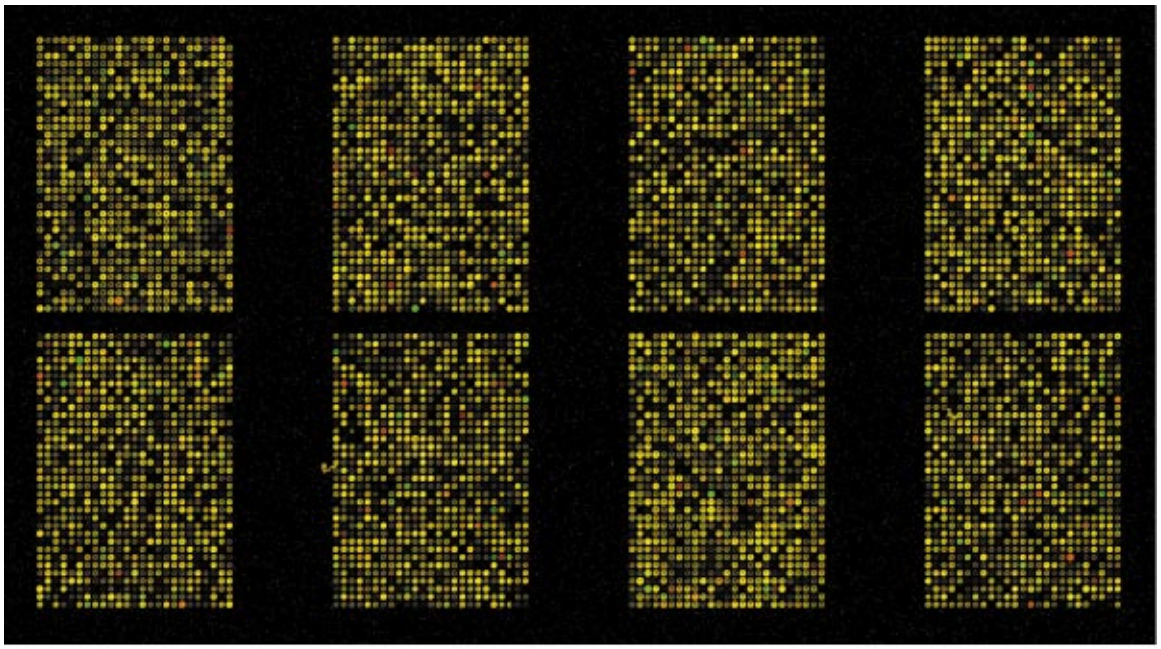
	Spot signal variation—foreground noise	SR_k, SG_k: Pixel-wise (x,y) signal intensity	$SR_k(x,y) \sim R''_k + N(\mu_{R''_k}, \sigma^2_R)$ $SG_k(x,y) \sim G''_k + N(\mu_{G''_k}, \sigma^2_G)$
		α_s: Within spot signal coefficient of variation	$\begin{cases} \mu_{R''_k} = R''_k \alpha_{m_1}; \alpha_{m_1} \sim U[f_{a_1}, f_{b_1}] \\ \mu_{G''_k} = G''_k \alpha_{m_2}; \alpha_{m_2} \sim U[f_{a_2}, f_{b_2}] \end{cases}$ $\begin{cases} \sigma_R = \alpha_{s_1} \mu_{R''_k}; \alpha_{s1} \sim U[f_{c_1}, f_{d_1}] \\ \sigma_G = \alpha_{s_2} \mu_{G''_k}; \alpha_{s_2} \sim U[f_{c_2}, f_{d_2}] \end{cases}$
	Edge enhancement	W_{ed}: Level of enhancement, parameter (μ_e) set for the block	$W_{\text{ed}} \sim N(\mu_e, 1)$
		N_e: Number of pixels enhanced	
	Edge noise	Apply edge noise at the set level (δ_{ed})	
BLOCK	Radius parameters	μ_s, k_s: mean and radius deviation factor	$\mu_s \sim U(s_a, s_b)$ $\sigma_s \sim k_s \mu_s$
		s_a, s_b: bounds of radius, set by block size and inter spot gap	
	Chord parameters	N_c: Chord rate picked with equal probability	$N_c \in U\{0,1,2,3,4\}$ having weights $\{p_0, p_1, p_2, p_3, p_4\}$
		α_L, β_L: Chord distributional parameters	$\alpha_L \sim U(a_\alpha, b_\alpha)$, $\beta_L \sim U(a_\beta, b_\beta)$,
	Inner hole parameters	$\mu_H, \mu_V, \sigma_H, \sigma_V$: Parameters for inner elliptical hole	$\mu_H \sim U(L_a, L_b)\mu_s$, $\mu_V \sim U(L_a, L_b)\mu_s$
		μ_s: Mean spot radius in the block	$\sigma_H = \alpha_1 \mu_s$, $\sigma_V = \alpha_2 \mu_s$
			$\alpha_1 \sim U(P_a, P_b)$, $\alpha_2 \sim U(P_c, P_d)$
	Drift parameters	$\delta c_{xG}, \delta c_{yG}, \delta c_{xR}, \delta c_{yR}$: drift level	$\delta c \sim U[i, j]$
		i, j: Percentage of the spot radius	$\delta c_{xG} = \delta c U[-1,1]$, $\delta c_{yG} = \delta c U[-1,1]$
			$\delta c_{xR} = \delta c_{xG} + U[-1,1]$, $\delta c_{yR} = \delta c_{yG} + U[-1,1]$
	Enhancement	l_a, l_b: Range of intensity ratio. Set mean level of enhancement for a block	$\mu_e \sim U(l_a, l_b)$
ARRAY	Physical dimensions	B_w, B_h: Block size—width, height (distance between first spot centers of any two block)	Typical Setting for a 8 blocks, 2 row array (in pixels):
		M_l, M_r, M_t, M_b: Margin settings (left, right, top, bottom)	$B_h, B_w = 900$ $M_l, M_r, M_t, M_b = 100$
		$N_{\text{pin}}, N_{\text{row}}$: Number of pins in an array, printed equally across N_{row} number of rows	
		NS_w, NS_h: Number of spots along the width (NS_w) and height (NS_h) of the block	

Table 1 *(Continued.)*

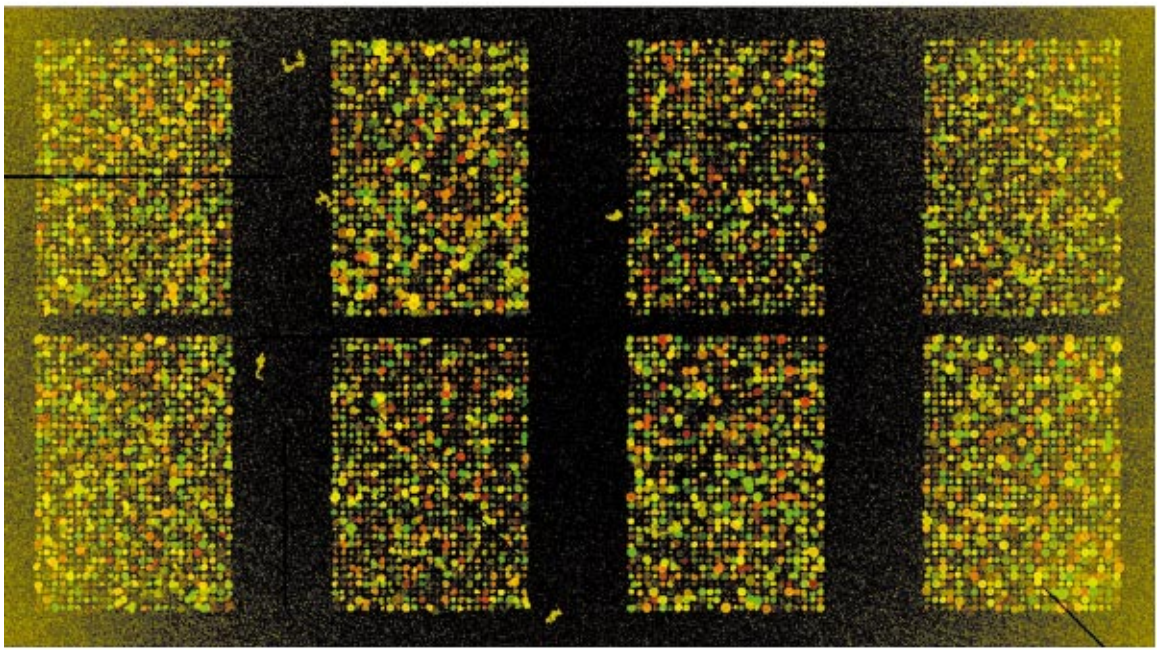
Signal to noise ratio	SNR: Signal to noise level is set for an array	
Interspot distance	G_{sp}: Interspot distance, set for an array	
Background	I_{b_ch1}, I_{b_ch2}: Background intensity, with parameters set for an array	$I_{b\_ch1} \sim N(\mu_b, \sigma_{b_1}^2)$ $I_{b\_ch2} \sim N(\mu_b, \sigma_{b_2}^2)$
	γ: Background level Parameter settings:	$\gamma \sim U[a,b]$
	—Flat fluorescent background	$\mu_b = \gamma$,
	—Functional background $g(x,y)$: choice of parabolic, positive or negative slant surface function	$\mu_b = \gamma g(x,y)$, with, $\sigma_{b_1} = (k_{b_1}\mu_b)$, $\sigma_{b_2} = (k_{b_2}\mu_b)$
Spike noise	L_{spi}: Level of spike noise (set in terms of percentage of total pixels)	
	N_s: Intensity of the spike noise	$N_s \sim \text{Exp}(\mu_{spi})$,
	μ_{spi}: Noise rate	$\mu_{spi} \sim U[e,f]$
	W_{spi}: Width of the noise cluster	$W_{spi} \sim U[g,h]$
Edge noise	δ_{ed}: Set the controlling parameter	δ_{ed} set as a percentage of maximum intensity value
Snake noise	N_{seg}: Number of snake tails in an image	$N_{seg}, \kappa_{sn}, L_{sn}, W_{sn}$
	I_{sn}: Intensity of the noise tail	$I_{sn} \sim N(\mu_{sn}, \sigma_{sn})$,
	κ_{sn}: Average signal-to-snake-noise intensity level	$\mu_{sn} = (I_k/\kappa_{sn})$, $\sigma_{sn} = k_{sn}\mu_{sn}$
	L_{sn}: Length of the segment expressed as multiples of average spot size	$L_{sn} \sim U[L_{sn1}, L_{sn2}]$
	W_{sn}: Width of the snake noise tail	
Scratch noise	N_{sc}: Number of scratch tails in an image	$N_{sc}, \kappa_{sc}, W_{sc}, \theta$
	I_{sc}: Intensity of the scratch noise	$I_{sc} \sim N(\mu_{sc}, \sigma_{sc})$
	κ_{sc}: Average background-to-scratch-noise intensity level	$\mu_{sc} = (\mu_b/\kappa_{sc})$, $\sigma_{sc} = k_{sc}\mu_{sc}$
	L_{sc}: Length of the segment in units of average size of the spots	$L_{sc} \sim U[L_{sc1}, L_{sc2}]$
	W_{sc}: Width of the scratch noise	$\theta \in U\{0°, 45°, 90°, 135°, 180°\}$
	θ: Scratch noise inclination	

tains the true signal for the synthetic microarray. This can be used subsequently to analyze various signal processing tools.

TIFF format is widely used due to platform independence and flexibility of data representation. The synthetic images are generated in TIFF with sample (pixel) resolution of two bytes for every color (*R*,*G*). Both monochrome and color images (*R*, *G* as two block and interlaced *R*, *G*, with dummy *B*) are generated. Standard freeware routines (http://www.libtiff.org) are used to generate these formats. The image file is written in blocks, where the size of the block (commonly called ''strip'') is set equal to the image width. The image data is written in the native order (big-endian, little-endian) of the host CPU on



(a)



(b)

Fig. 17 This example shows full size arrays simulation with different parameter settings: (a) good quality has SNR of 2.0, with normal background, spike noise $L_{spi}=0.3\%$, (b) noisy array with SNR of 1.1 with parabolic background noise, spike noise $L_{spi}=15\%$.

which the library is compiled. Image data quality is maintained by disabling compression and other special options available in these routines and formats.

2.5 Summary of Model Parameters

The cDNA microarray printing process can be categorized and grouped into independent events. Each event is probabilistically described by assigning a distribution, as previously described. Due to the physical nature of the process, there exist variations between events. This variation is described by randomization of the controlling parameters (second level randomization). The parameter randomization can be broadly grouped as (i) randomization at spot level, (ii) randomization at block level, and (iii) randomization at array level. The parameters are grouped and mathematically described in Table 1.

Each noise type is categorized into one of the three groups and individually parameterized. Some are related to another noise parameter; others are independent. Each noise parameter is assigned a statistical distribution fitting its nature. For instance, consider spot radius. Spot radius obeys a normal distribution (μ_s, σ_s^2), where the mean spot radius (μ_s) is randomly picked over a small range (s_a, s_b) at the block level. This spot size range is set for an array depending on a user setting: the number of spots in a block (NS_w, NS_h) at the array level. If a noise type needs to be suppressed, then the corresponding parameters can be set small to nullify its effect. For example, inner spot hole follows a normal distribution along its vertical (μ_H, σ_H) and horizontal (μ_V, σ_V) axes. Its parameters are randomly picked from a preset range (L_a, L_b) and related to the mean spot radius (μ_s) at the block level $[\mu_H \sim U(L_a, L_b)\mu_s, \mu_V \sim U(L_a, L_b)\mu_s]$. For small or negligible doughnut holes, this preset range can be set small, or even null for perfect spots. The table is perused from spot level to the array level, tagging through the corresponding parameters, as indicated in the earlier examples.

3 Examples of Simulated Microarrays and Image Analysis

All of the described process and noise effects are controlled by appropriate parameter selection. Depending on the parameter setting, the arrays can be roughly classified as ideal, average, or noisy. Given a good printing run (no mechanical deposition problems), a relative matured hybridization protocol, and good RNA samples, along with a scanner of minimal optical warping, focusing, and integration problems, we expect a high-quality (ideal) microarray image. The corresponding simulated ideal image will have a flat mean background with typical autofluorescence variation (<10% of mean background level, but no less than square root of the mean background level), minimum spike/scratch/snake noise, little edge enhancement and no channel conditioning problems. For average image quality, one would expect larger background variation and possibly a slanted mean level. There will also be more spike/scratch/snake noise interfering with signal spots. In a noisy setting, besides higher noise levels for various possible interference, one would also expect uneven background level (e.g., parabolic function), heavy spot deformity (chord cuts, edge enhancement, and large inner holes), and different channel conditioning [such as the banana shape in the intensity scatter plot shown in Figure 11(b)].

Figure 17 shows two microarrays generated with $NS_w = 35$ rows and $NS_h = 25$ columns, at $B_h = B_w = 900$ pixels per block. Array boundaries are set at $(M_t, M_l, M_r, M_b) = (100,100,100,100)$. By choosing parameters, two different array qualities have been generated. Part (a) illustrates an ideal microarray image with normal background and parameters $\beta = 3000$, SNR = 2.0, $\alpha = 0.05$, $G_{\text{sp}} = 6$, $P_D = 0.05$, $(d_a, d_b) = (2,15)$, $(k_{b_1}, k_{b2}) = (10,10)$, $P_{\text{outlier}} = 0.05$, $L_{\text{spi}} = 0.3\%$, $\delta_{\text{ed}} = 0.3$:

$$(f_{a_1}, f_{b_1}, f_{c_1}, f_{d_1}) = (2,8,2,6),$$

$$(f_{a_2}, f_{b_2}, f_{c_2}, f_{d_2}) = (2,8,2,8),$$

$$(a_0, a_1, a_2, a_3) = (0,1,-1,1),$$

$$(b_0, b_1, b_2, b_3) = (0,1,-1,1),$$

$$(l_a, l_b, N_e) = (1,3,3),$$

$$(p_0, p_1, p_2, p_3, p_4) = (0.97,0.03,0,0,0),$$

$$(K_{\text{SN}}, L_{\text{SN1}}, L_{\text{SN1}}, W_{\text{SN}}, N_{\text{SN}}) = (0.25,10,50,1,2),$$

$$(K_{\text{SC}}, L_{\text{SC1}}, L_{\text{SC2}}, W_{\text{SC}}, N_{\text{SC}}) = (3,5,35,3,1).$$

Part (b) illustrates a noisy microarray image with parabolic background and parameters: $\beta = 3000$, SNR = 1.1, $\alpha = 0.25$, $G_{\text{sp}} = 4$, $P_D = 0.4$, $(d_a, d_b) = (15,100)$, $(k_{b_1}, k_{b2}) = (25,25)$, $P_{\text{outlier}} = 0.7$, $L_{\text{spi}} = 15\%$, $\delta_{\text{ed}} = 0.03$:

$$(f_{a_1}, f_{b_1}, f_{c_1}, f_{d_1}) = (6,12,8,20),$$

$$(f_{a_2}, f_{b_2}, f_{c_2}, f_{d_2}) = (6,12,8,20),$$

$$(a_0, a_1, a_2, a_3) = (0,500,-1,1),$$

$$(b_0, b_1, b_2, b_3) = (0,10,-1,1),$$

$$(l_a, l_b, N_e) = (10,40,3),$$

$$(p_0, p_1, p_2, p_3, p_4) = (0.05,0.3,0.25,0.25,0.15),$$

$$(K_{\text{SN}}, L_{\text{SN1}}, L_{\text{SN1}}, W_{\text{SN}}, N_{\text{SN}}) = (0.25,60,110,2,10),$$

$$(K_{\text{SC}}, L_{\text{SC1}}, L_{\text{SC2}}, W_{\text{SC}}, N_{\text{SC}}) = (0.25,60,110,2,10).$$

To illustrate how the simulation can be used to analyze microarray image software, we apply the ArraySuite[11] software to extract the image intensities and ratios from the image and then compare these to the corresponding intensities and ratios used for simulation. We use the ideal case to illustrate the utility of the simulation. In Figure 18(a), intensities from one fluorescent channel have been extracted (y axis) and plotted against the simulation signal intensities. The extracted signal generally corresponds well to the simulated signal, with

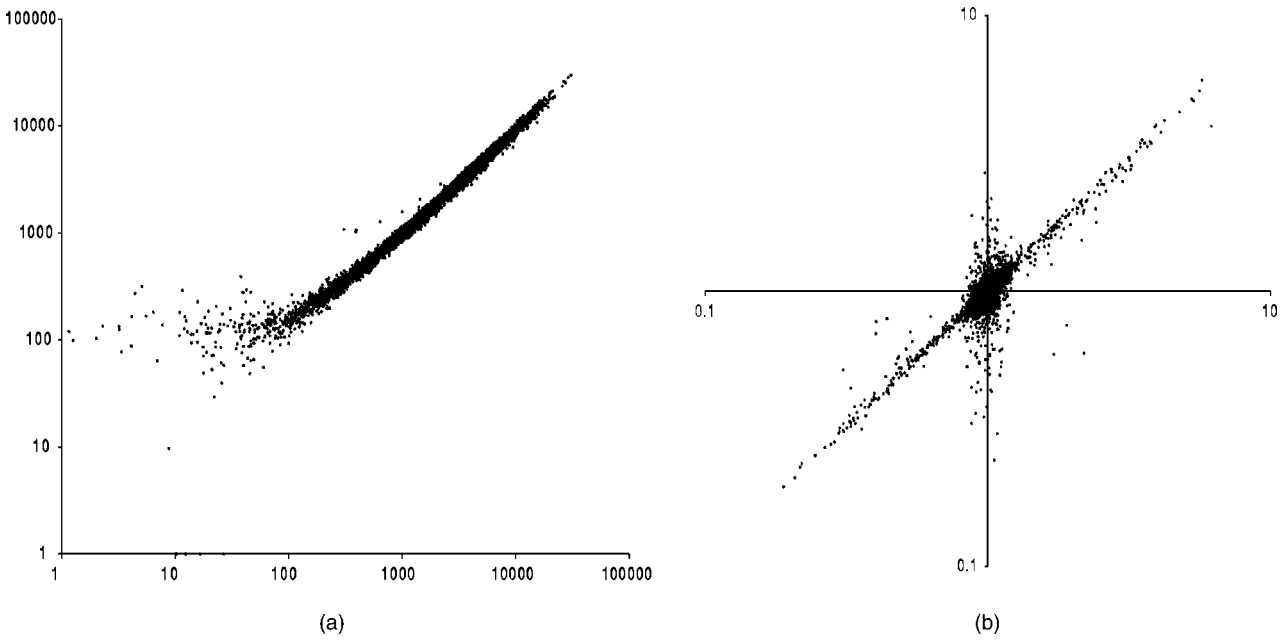


Fig. 18 Comparison between simulated signal (ideal setting) vs extracted signal from microarray image analysis program. (a) Signal extracted from one fluorescent channel (*y* axis) comparing to the signal used for simulation in the same channel (*x* axis). (b) Ratios from microarray image analysis program (*y* axis) comparing to the ratios generated by the simulation (*x* axis).

some variation. After excluding intensities less than 300, the mean and standard deviation of the difference between the two $\log_{10}$-transformed intensities are 0.016 (or $10^{0.016} = 1.038$) and 0.038 (or $10^{0.038} = 1.09$), respectively. The ratio comparison is given in Figure 18(b). When signal intensity is weak (less than 300), various noise components in the simulation process affect the accuracy of the signal extraction program. Since the problem is unavoidable, a measurement quality metric is necessary to provide confidence in downstream data analysis. In this case, we see that if the signal intensity is less than 300, then the noise interaction is significant.

4 Conclusion

Modeling and simulation of microarray image formation is a key to benchmarking various signal processing tools being developed to estimate cDNA signal spots. Using a model to describe the signal ground truth not only helps in evaluating these tools, but also facilitates the understanding of various process interactions. To illustrate how the image-simulation program presented in this paper can be used in the development of image-analysis software, we describe an actual case.

The simulation program has been used extensively in the design of the microarray image-analysis program used at the National Human Genome Research Institute. This has been done by testing the accuracy of the analysis program on simulated images exhibiting troublesome noise conditions and then tuning the program to achieve better results. One such application concerns large and overlapping spots, as illustrated in Figure 19(a), which shows part of an actual hybridized image in which some spots are substantially larger than intended owing to randomness in the cDNA deposition procedure. This defect causes various problems, one being poor background estimation. We illustrate this problem by simulating an image with large spot size variation and drifting conditions [Figure 19(b)]. If the image analysis program extracts the local background by averaging the region around the bounding box (which was used as a starting condition in an earlier version of the NHGRI program), an elevated background average may be obtained since the bounding box may overlap neighboring targets that are large in size and strong in expression level. An additional problem is that some weak targets may not be detected [Figure 19(c)]. Based on these considerations, the program has been modified to calculate the four average intensities from the four corners and the four average intensities from the four sides of the bounding box, and then take the minimum among all of these as the initial estimation of the local background. A histogram-based method is then invoked around the initial estimated background to further improve the estimation. The output from Figure 19(b) according to the modified program is shown in Figure 19(d): the weak target is detected and there is improved local background estimation for all spots.

References

1. M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, ''Quantitative monitoring of gene expression patterns with a complementary DNA microarray,'' *Science* **270**, 467–470 (1995).
2. J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent, ''Use of a cDNA microarray to analyse gene expression patterns in human cancer,'' *Nat. Genet.* **14**(4), 457–60 (1996).
3. P. T. Spellman et al., ''Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization,'' *Mol. Biol. (Moscow)* **9**(12), 3273–3297 (1998).
4. J. Khan, R. Simon et al., ''Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays,'' *Cancer Res.* **58**(22), 5009–5013 (1998).
5. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, ''Molecular classification of cancer:

class discovery and class prediction by gene expression monitoring,'' *Science* **286**(5439), 531–537 (1999).
6. V. R. Iyer et al., ''The transcriptional program in the response of human fibroblasts to serum,'' *Science* **283**(5398), 83–87 (1999).
7. M. Bittner, P. Meltzer et al., ''Molecular classification of cutaneous malignant melanoma by gene expression profiling,'' *Nature (London)* **406**(6795), 536–540 (2000).
8. A. A. Alizadeh et al., ''Distinct types of diffuse large *B*-cell lymphoma identified by gene expression profiling,'' *Nature (London)* **403**(6769), 503–511 (2000).
9. I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, Y. Yakhini, A. Ben-Dor, E. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Pittaluga, S. Gruvberger, N. Loman, O. Johannsson, H. Olsson, B. Wilfond, G. Sauter, O. Kallioniemi, A. Borg, and J. Trent, ''Gene-expression profiles in hereditary breast cancer,'' *N. Engl. J. Med.* **344**(8), 539–548 (2001).
10. J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, ''Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,'' *Nat. Med. (N.Y.)* **7**(6), 673–679 (2001).
11. Y. Chen, E. R. Dougherty, and M. Bittner, ''Ratio-based decisions and the quantitative analysis of cDNA microarray images,'' *J. Biomed. Opt.* **2**(4), 364–374 (1997).
12. P. Kalocsai and S. Shams, ''Use of bioinformatics in arrays,'' *Methods Mol. Biol.* **170**, 223–236 (2001).
13. See www.imgresearch.com, genome-www.stanford.edu/microarray, www.axon.com, www.imagingresearch.com, and www.nutecsciences.com.
14. D. J. Duggan, M. L. Bittner, Y. Chen, P. S. Meltzer, and J. M. Trent, ''Expression profiling using cDNA microarrays,'' *Nat. Genet.* **21**(1 Suppl), 10–14 (1999).
15. M. K. Kerr and G. A. Churchill, ''Statistical design and the analysis of gene expression microarray data,'' *Genet. Res.* **77**(2), 123–128 (2001).
16. F. W. D. Rost, *Fluorescence Microscopy*, Cambridge University Press, Cambridge (1995).
17. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, ''Cluster analysis and display of genome-wide expression patterns,'' *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863–14868 (1998).
18. A. Ben-Dor, R. Shamir, and Z. Yakhini, ''Clustering gene expression patterns,'' *J. Comput. Biol.* **6**(3/4), 281–297 (1999).
19. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, ''Interpreting pattern of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation,'' *Proc. Natl. Acad. Sci. U.S.A.* **96**(6), 2907–2912 (1999).
20. S. Kim, E. R. Dougherty, M. L. Bittner, Y. Chen, K. Sivakumar, P. Meltzer, and J. M. Trent, ''A general framework for the analysis of multivariate gene interaction via expression arrays,'' *J. Biomed. Opt.* **5**(4), 411–424 (2000).
21. S. Kim, E. R. Dougherty, Y. Chen, K. Sivakumar, P. Meltzer, J. M. Trent, and M. Bittner, ''Multivariate measurement of gene-expression relationships,'' *Genomics* **67**, 201–209 (2000).
22. D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, Wiley, Chichester (1995).
23. *Advances in Theory and Applications of Random Sets*, D. Jeulin, Ed., World Scientific, New York (1997).
24. E. R. Dougherty, *Random Processes for Image and Signal Processing*, SPIE, Bellingham, WA (1999).
25. R. P. Loce and E. R. Dougherty, *Enhancement and Restoration of Digital Documents*, SPIE, Bellingham, WA (1997).
26. J. O. Bishop, J. G. Morton et al., ''Three abundance classes in Hela cell messenger RNA,'' *Nature (London)* **250**(463), 199–240 (1974).
27. http://arrayanalysis.nih.gov/resources/pub_download/jbo3_supplement.htm