Estimating Gene Signals From Noisy Microarray Images

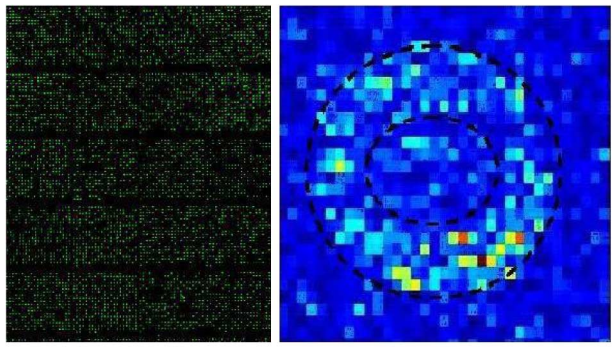
Pinaki Sarder, *Student Member, IEEE*, Arye Nehorai, *Fellow, IEEE*, Paul H. Davis, and Samuel L. Stanley, Jr.

***Abstract*—In oligonucleotide microarray experiments, noise is a challenging problem, as biologists now are studying their organisms not in isolation but in the context of a natural environment. In low photomultiplier tube (PMT) voltage images, weak gene signals and their interactions with the background fluorescence noise are most problematic. In addition, nonspecific sequences bind to array spots intermittently causing inaccurate measurements. Conventional techniques cannot precisely separate the foreground and the background signals. In this paper, we propose analytically based estimation technique. We assume *a priori* spot-shape information using a circular outer periphery with an elliptical center hole. We assume Gaussian statistics for modeling both the foreground and background signals. The mean of the foreground signal quantifies the weak gene signal corresponding to the spot, and the variance gives the measure of the undesired binding that causes fluctuation in the measurement. We propose a foreground-signal and shape-estimation algorithm using the Gibbs sampling method. We compare our developed algorithm with the existing Mann–Whitney (MW)- and expectation maximization (EM)/iterated conditional modes (ICM)-based methods. Our method outperforms the existing methods with considerably smaller mean-square error (MSE) for all signal-to-noise ratios (SNRs) in computer-generated images and gives better qualitative results in low-SNR real-data images. Our method is computationally relatively slow because of its inherent sampling operation and hence only applicable to very noisy-spot images. In a realistic example using our method, we show that the gene-signal fluctuations on the estimated foreground are better observed for the input noisy images with relatively higher undesired bindings.**

***Index Terms*—cDNA microarray, Gibbs sampling, low PMT voltage image, spot segmentation.**

I. Introduction

In microarray experiments, noise is increasingly becoming a problem, as biologists now are studying their organisms not in isolation (e.g., pure RNA from a single species grown in culture), but in the context of a natural environment. Namely, the amoebic RNA signal is more difficult to ascertain in the presence of stochastic, confounding host-RNA noise, such as when RNA is measured from n amoebae surrounded by m liver cells $(n \ll m)$, as well as high background noise produced by the imaging scanner because of the low photomultiplier tube (PMT) voltage setting. In this paper, we develop a Gibbs-sampling method for estimating the foreground signal and the shape information from such noisy microarray spot images.



(a) (b)

Fig. 1. (a) RGB image of a oligonucleotide-based microarray. (b) Intensity image of a single spot where the circular outer periphery and the elliptical center hole are shown using dashed lines.

A. Oligonucleotide Microarray

Oligonucleotide microarray technology is a powerful tool for the analysis of differences in the gene expression levels of a multitude of genes in parallel. Hybridized oligonucleotide microarrays are prepared by automatically printing thousands of distinct oligonucleotides, each representing different genes, as several gridded, predefined spots in an array format on glass microscope slides [1]. Messenger RNAs present in a particular sample of cells are extracted and used to form fluor-tagged cDNA *in vitro* using the *reverse transcription* method. Tagged cDNAs are then hybridized to the array of oligonucleotides, and the gene expression level is quantified at the site of each immobilized cDNA [1]. Fig. 1(a) shows a typical oligonucleotide microarray red–green–blue (RGB) image, where each spot shows the gene-expression signal corresponding to a particular gene. Fig. 1(b) presents the intensity image of a single noisy spot. In general, processing of such images requires following three prior information.

Shape: During the manufacturing process, a robot finger places the oligonucleotide on the slide, resulting in variability in the placement. Because of surface tension, significantly less oligonucleotide may be deposited at the center of the target. Consequently, the center of the hybridized target emits fewer

Manuscript received November 5, 2007; revised January 31, 2008. *Asterisk indicates corresponding author.*

P. Sarder is with the Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO 63130 USA (e-mail: psarde1@ese.wustl.edu).

A. Nehorai is with the Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO 63130, USA (fax: (314) 935-7500, e-mail: nehorai@ese.wustl.edu).

P. H. Davis is with the Department of Biology, University of Pennsylvania, Philadelpha, PA 19104 USA (e-mail: paulhd@sas.upenn.edu).

S. L. Stanley, Jr. is with the Departments of Medicine and Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110 USA (e-mail: stanleys@wusm.wustl.edu).

This work was supported by the National Science Foundation Grant CCR-0330342 and the Imaging Sciences Pathway program of Washington University in St. Louis.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNB.2008.2000745

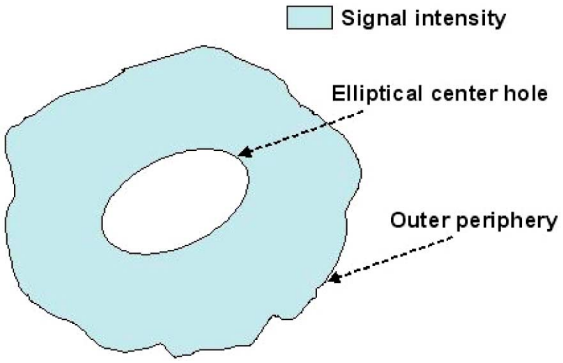


Fig. 2. A schematic view of a oligonucleotide-based microarray spot with an elliptical center hole.

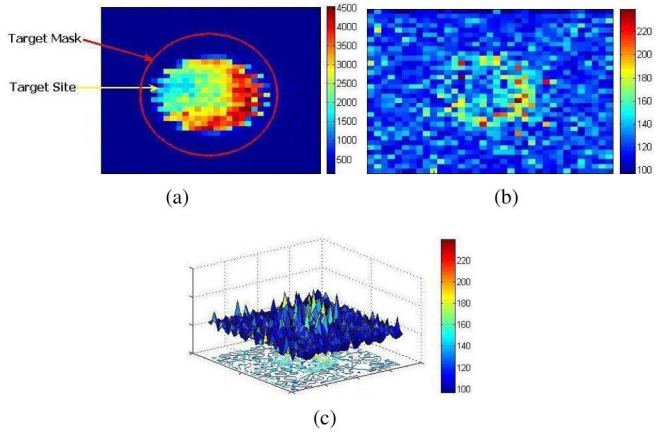


Fig. 3. Gene signals from (a) high and (b) low signal-to-noise ratio spots. (c) The intensity image of (b) with the signal intensities represented by height along the pixels on the focal plane.

fluorescent photons, thereby giving the target the shape of a doughnut. Therefore, it is critical to consider the center hole in signal-intensity estimation methods, especially when the signal is weak and the center hole is large. In practice, the center holes have an elliptical shape (see Fig. 2) [2]. In a few cases, there may even be more than one hole.

Background Noise: The oligonucleotide microarray images are collected by scanning the signal intensities of the corresponding spots using dedicated fluorescence scanners [3]. The major scanner settings for increasing the spot intensities are the laser power and the voltage of the PMT. In almost all scanners, within a limited intensity range from 200 to 50 000 (mean spot intensity), gene expressions are independent of the PMT voltage. This usable intensity range is considerably smaller than the maximum detection range of the PMTs. However, spot and background intensities outside this range will produce errors in the measured expression levels. The brightest spots reach saturation level at high PMT settings, and differences in expression levels cannot be ascertained. In order to avoid saturation, the images are acquired at low-PMT settings. As a consequence, the captured images of the weakest spots become noisy [2].

Foreground Noise: In this paper, we assume that the intensity measurement of each spot is a function of the specific gene available within each sample. The random fluctuation in the foreground occurs because of the undesired binding of the host RNA. It is often difficult to identify the foreground gene-expression region (shape) in low signal-to-noise ratio (SNR) situations, since the signal is weak and there is no marked transition between the foreground and background noise.

B. Literature Review

In order to estimate gene-signal intensities in each spot, local segmentation of the image is used to distinguish foreground pixels (signals) from the background. In conventional software, this segmentation method creates a local *target mask* [see Fig. 3(a)] on the gene-signal region comprising a set of foreground pixels for every spot. Then, quantification is performed to extract raw data intensities from the signal areas and their relative backgrounds. The image-processing challenge is to extract the shape of the spot [denoted as the *target site* in Fig. 3(a)] emitting the gene signals. Most software resources assume during the processing that the *target mask* itself contains the gene signals. Some others use the Mann–Whitney (MW) test to differentiate the *target site* from the *target mask* [1].

The existing literature abounds in methods for automatic segmentation of the microarray images. In [4], the authors propose Markov random field (MRF) and active-contour-based methods. In [5], the authors explore an order-statistics-based technique. A correlation-statistics-based method is proposed in [6]. In a complementary work, the authors use a wavelet-denoising method for microarray image enhancement [7]. In [8], the authors propose a noise-reconstruction-based method. A *k*-means clustering-based microarray image-segmentation method is described in [9]. The main disadvantage of the preceding methods is that they perform well only for high SNR images. In addition, conventional adaptive-thresholding techniques are unsatisfactory in low-SNR microarray spot images since it is difficult to differentiate the foreground and the background for such cases [see Fig. 3(b) and (c)]. Standard morphological methods also fail to capture the shape information because of the weak signal.

In a recent work [10], the researchers present an expectation maximization (EM)/iterated conditional modes (ICM)-based method for processing noisy microarray spot images. In their work, the authors do not assume any spot-shape information for processing images. In this paper, we present an improved and simplified version of their method by introducing *a priori* spot-shape information for the microarray spots using parametric doughnut shapes.

Estimating the gene-signal intensity accurately is essential for its use in biological analysis. For example, in ratio-based expression analysis, often the gene-signal intensity of a *control* may be transcribed poorly (say, with a value of 0) using conventional software at low SNR. However, in the *experiment* let the gene be transcribed with a value of ten. Hence, the gene is inactive in the *control*, but active in the *experiment*, which should be considered significant. In these instances, however, generating a fold ratio is impossible since 10/0, the ratio of the gene signal intensities, is undefined. Therefore, a more analytically based estimation is necessary.

C. Overview of Our Method

In this paper, we consider the following analytical strategy for estimating gene-signal intensities from oligonucleotide microarray spot images:

- a *parametric* doughnut-shape model for the spot shape and location;
- a *parametric* model for the foreground and background signals;
- a Gibbs sampling-based algorithm for estimating the unknown shape and signal parameters from a given spot image.

We test our proposed algorithm numerically and compare the results with the existing MW- and EM/ICM-based methods [1], [10]. Our proposed method significantly outperforms these existing methods at low SNR. Our algorithm performs better because it contains prior spot-shape information, whereas the other methods (MW and EM/ICM) do not have that flexibility. Namely, we observe that the performance of the center-hole estimation is overly sensitive using the EM/ICM algorithm in very low SNR images, whereas our proposed method does not have that limitation. In a realistic example using our proposed method, we show that the gene-signal fluctuations at the estimated foreground are better observed as host redundancy increases in the noisy input images. Our research verifies the fact that statistical signal processing can play a significant role in estimating noisy microarray image data.

One application of our proposed work is in infectious disease research where many amoebic genes produce very low-intensity signals in the measurement. Biologists often discard such noisy spot measurements because no existing methods guarantees the desired segmentation performance [11]. However, our proposed approach performs better than the existing methods. Note that our method is slower than the existing algorithms. Hence, we propose using conventional methods for segmenting high-SNR spot images and our proposed method for segmenting very noisy spot images.

The paper is organized as follows. In Section II, we present our proposed method for modeling microarray spot shapes and signals. Then, we describe the measurement model with noise. In Section III, we present a Gibbs sampler for estimating the shape and signal parameters of a given spot. In Section IV, we review existing MW- and EM/ICM-based methods. In Section V, we present our results using real data on *Entamoeba* oligonucleotide microarrays that were collected at the Washington University School of Medicine Microarray core facility [11]. In Section VI, we present numerical examples for quantitative and qualitative comparison of the parameter estimation using our proposed, MW-, and EM/ICM-based methods for low-SNR images. Finally, we conclude in Section VII.

II. Spot Shape and Signal Modeling

In this section, we first present a gridding method to obtain a rough estimate of the position of each spot in the microarray by finding a rectangular grid. Then, we discuss our proposed parametric model of the spot shape and location. Finally, we present the statistical measurement model comprising the foreground and background signal.

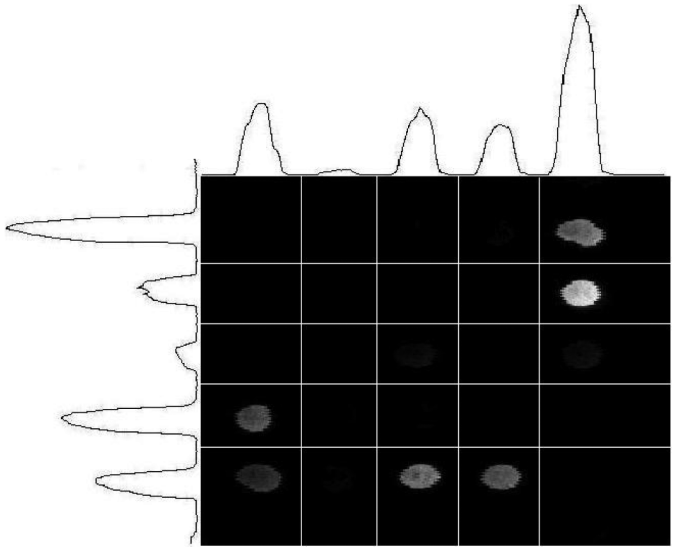


Fig. 4. Illustration of the gridding algorithm [12]. The image is projected onto the x axis and y axis. The off-peaks in the two projections define the lines of the grid.

Gridding: We adopt a similar method to that proposed in [12] for gridding. We manually select the image portion of interest from the microarray. We project this image onto the x and y axes. The projection looks like a series of peaks separated by off-peaks. Finally, the grid is formed by plotting a line in each off-peak. We present an illustration of the gridding algorithm in Fig. 4.

Spot-Shape modeling: We model the spot shape using a parametric circle with an elliptical center hole resembling a doughnut shape. Parametric formulation of the spot introduces prior information in the gene signal estimation algorithm, as we show in the next section. In most cases, microarray spot shapes are circular without any center hole. The remainder are mostly doughnut shaped. Spots with more than one center hole are possible, but very rare in practice. Hence, we confine ourselves to modeling the spots using a single center hole.

We assume that the signal region $R(\boldsymbol{\tau})$ is given by

$$R(\boldsymbol{\tau}) = \{\boldsymbol{r} : (\boldsymbol{r}-\boldsymbol{r}_0)^T(\boldsymbol{r}-\boldsymbol{r}_0) \le r_1^2, \\ (\boldsymbol{r}-\boldsymbol{r}_0)^T\boldsymbol{\Sigma}(d,A,\phi)^{-1}(\boldsymbol{r}-\boldsymbol{r}_0) \ge 1\} \quad (1)$$

where $\boldsymbol{r} = [x,z]^T$ and $\boldsymbol{r}_0 = [x_0,z_0]^T$ denotes a pixel location and the center of the circle and the ellipse in Cartesian coordinates, respectively; ``^{T}'' is a matrix transpose operation; r_1 is the radius of the circular spot; and $\boldsymbol{\Sigma}(d,A,\phi)$ is defined as

$$\boldsymbol{\Sigma}(d,A,\phi) = \boldsymbol{\Phi}(\phi)\begin{bmatrix} d^2 & 0 \\ 0 & A^2/d^2\pi^2 \end{bmatrix}\boldsymbol{\Phi}(\phi)^T, \\ \boldsymbol{\Phi}(\phi) = \begin{bmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{bmatrix} \quad (2)$$

where $d > 0$ is an axis parameter, $A > 0$ the area, and $\phi \in [-\pi/4, \pi/4]$ the orientation parameter (in radians) of the ellipse. Here, d and $A/d\pi$ are the axes

of the elliptical hole. The inverse of $\mathbf{\Sigma}(\cdot)$ is defined as $\mathbf{\Sigma}(d, A, \phi)^{-1} = \mathbf{\Phi}(\phi)[\begin{smallmatrix} 1/d^2 & 0 \\ 0 & d^2\pi^2/A^2 \end{smallmatrix}]\mathbf{\Phi}(\phi)^T$.

We denote the unknown shape-parameter vector as $\boldsymbol{\tau} = [\boldsymbol{r}_0{}^T, d, A, \phi, r_1]^T$, the rectangular grid containing the kth spot $R_k(\boldsymbol{\tau})$ and its neighborhood $R_k^{\mathrm{c}}(\boldsymbol{\tau})$ as $R_k \bigcup R_k^{\mathrm{c}}(\boldsymbol{\tau})$, where $\bigcup$ denotes the union operation. It is worth mentioning that our proposed spot-shape model can be extended to the more general case using multiple overlapped center holes.

Signal Modeling: The gene signal in the kth spot, ignoring the background noise in $R_k \mathrm{U} R_k^{\mathrm{c}}(\boldsymbol{\tau})$, is given by

$$y_k(\boldsymbol{r}; \boldsymbol{\tau}) = \begin{cases} f_k(\boldsymbol{r}) & \text{if } \boldsymbol{r} \in R_k(\boldsymbol{\tau}), \\ 0 & \text{if } \boldsymbol{r} \in R_k^{\mathrm{c}}(\boldsymbol{\tau}) \end{cases} \tag{3}$$

where $y_k(\boldsymbol{r}; \boldsymbol{\tau})$ is the measurement and $f_k(\boldsymbol{r})$ the kth gene's expression. For notational convenience, we will omit the subscript k in the remainder of this paper, since we present a generalized analysis of the gene signal estimation for each spot location. The measurement-noise model is given by

$$y(\boldsymbol{r}; \boldsymbol{\tau}, \boldsymbol{\theta}) = \begin{cases} f(\boldsymbol{r}; \boldsymbol{\theta}) + w(\boldsymbol{r}) & \text{if } \boldsymbol{r} \in R(\boldsymbol{\tau}), \\ w(\boldsymbol{r}) & \text{if } \boldsymbol{r} \in R^{\mathrm{c}}(\boldsymbol{\tau}) \end{cases} \tag{4}$$

where $\boldsymbol{\theta} = [\mu, \sigma]^T$ is the vector of unknown foreground spot signal parameters and $f(\boldsymbol{r}; \boldsymbol{\theta})$ the independent identically distributed (i.i.d.) Gaussian random variable in $R(\boldsymbol{\tau})$ with unknown mean μ and variance σ^2 [13]. The parameter μ denotes the gene expression level and σ^2 signifies the random fluctuation as caused by the undesired binding of the host. The local background noise values $w(\boldsymbol{r})$ in $R_k \mathrm{U} R_k^{\mathrm{c}}(\boldsymbol{\tau})$ are modeled as independent from pixel to pixel and identically distributed additive Gaussian random variables with known mean μ_w and variance σ_w^2. We assume that $f(\boldsymbol{r}; \boldsymbol{\theta})$ and $w(\boldsymbol{r})$ are independent of each other at every pixel location. Hence, the unknown spot shape, location, and signal parameters are $\boldsymbol{\psi} = [\boldsymbol{\tau}^T, \boldsymbol{\theta}^T]^T$.

Data Preprocessing: We estimate the background-noise parameters locally from the noise-only data. Then, we subtract the estimated μ_w from the available data in $R_k \bigcup R_k^{\mathrm{c}}(\boldsymbol{\tau})$. In this way, the local background noise $w(\boldsymbol{r})$ in $R_k \bigcup R_k^{\mathrm{c}}(\boldsymbol{\tau})$ become i.i.d. Gaussian random variables with zero mean and known estimated variance σ_w^2.

Summary: We adopt a shape bounded by a circle with an elliptical center hole and also take into account the Gaussian signal and noise models. Similar frameworks are applicable to other analysis fields as well [14]. We ignore the randomness along the periphery for modeling the oligonucleotide deposition spot. The elliptical shape model for the center hole is well suited to random horizontal and vertical axes. In [13], a more general modeling of the periphery considering a random variation is assumed; however it requires a larger number of parameters and, as a consequence, the solution to the reverse problem becomes more computationally intensive.

III. Estimation

In this section, we discuss a Bayesian approach for estimating the unknown parameters in $\boldsymbol{\psi}$. The Bayesian approach is based on the Gibbs sampling method as discussed in [14] for nondestructive evaluation (NDE) defect signal analysis.

We denote the probability density function (pdf) of a Gaussian random variable a with mean α and variance β^2 as $p(a) = \mathcal{N}(a; \alpha, \beta^2)$ and the conditional pdf of a random variable a given random variable b as $p(a \,|\, b)$. Then, the conditional pdf of any observation $y(\cdot)$ given $\boldsymbol{\psi}$ is

$$p(y(\cdot)\,|\boldsymbol{\psi}) = \begin{cases} \mathcal{N}\left(y(\cdot); \mu, \sigma^2 + \sigma_w^2\right) & \text{if } \boldsymbol{r} \in R(\boldsymbol{\tau}), \\ \mathcal{N}\left(y(\cdot); 0, \sigma_w^2\right) & \text{if } \boldsymbol{r} \in R^{\mathrm{c}}(\boldsymbol{\tau}). \end{cases} \tag{5}$$

We assume the available measurements are $\{y(x, z); 1 \le x \le L, 1 \le z \le M\}$ and the vector form of the lumped measurements is $\boldsymbol{y}$. The likelihood $L(\boldsymbol{y}\,|\boldsymbol{\psi})$ of the measurement $\boldsymbol{y}$ given $\boldsymbol{\psi}$ is

$$\begin{aligned} L(\boldsymbol{y}\,|\boldsymbol{\psi}) &= \prod_{\boldsymbol{r} \in R(\boldsymbol{\tau})} \mathcal{N}\left(y(\cdot); \mu, \sigma^2 + \sigma_w^2\right) \\ &\quad \times \prod_{\boldsymbol{r} \in R^{\mathrm{c}}(\boldsymbol{\tau})} \mathcal{N}\left(y(\cdot); 0, \sigma_w^2\right) \\ &\propto \left(1 + \frac{\sigma^2}{\sigma_w^2}\right)^{-\frac{N(\boldsymbol{\tau})}{2}} \\ &\quad \cdot \exp\left\{-\frac{1}{2} \sum_{\boldsymbol{r} \in R(\boldsymbol{\tau})} \left[\frac{(y(\cdot) - \mu)^2}{\sigma^2 + \sigma_w^2} - \frac{(y(\cdot))^2}{\sigma_w^2}\right]\right\} \end{aligned} \tag{6}$$

where $N(\boldsymbol{\tau}) = \sum_{\boldsymbol{r} \in R(\boldsymbol{\tau})} 1$.

- *Prior specification*: We denote the prior pdf of a random variable a as $\pi_a(a)$. We assume the parameters in $\boldsymbol{\psi}$ are independent *a priori* and we assume uniform distribution priors for all the parameters, e.g., i) $\pi_\mu(\mu) = \mathrm{uniform}(0, \mu_{\mathrm{MAX}})$; ii) $\pi_\sigma(\sigma) = \mathrm{uniform}(0, \sigma_{\mathrm{MAX}})$; iii) $\pi_{x_0}(x_0) = \mathrm{uniform}(x_{0,\mathrm{MIN}}, x_{0,\mathrm{MAX}})$; iv) $\pi_{z_0}(z_0) = \mathrm{uniform}(z_{0,\mathrm{MIN}}, z_{0,\mathrm{MAX}})$; v) $\pi_d(d) = \mathrm{uniform}(0, d_{\mathrm{MAX}})$; vi) $\pi_A(A) = \mathrm{uniform}(A_{\mathrm{MIN}}, A_{\mathrm{MAX}})$; vii) $\pi_\phi(\phi) = \mathrm{uniform}(\phi_{\mathrm{MIN}}, \phi_{\mathrm{MAX}})$; viii) $\pi_{r_1}(r_1) = \mathrm{uniform}(0, r_{1,\mathrm{MAX}})$. Hence, the joint prior distribution of the parameters in $\boldsymbol{\psi}$ is given by

$$\begin{aligned} \pi_{\boldsymbol{\psi}}(\boldsymbol{\psi}) = \pi_\mu(\mu)\pi_\sigma(\sigma)\pi_{x_0}(x_0)\pi_{z_0}(z_0) \\ \times \pi_d(d)\pi_A(A)\pi_\phi(\phi)\pi_{r_1}(r_1). \end{aligned} \tag{7}$$

- *Posterior pdf of $\boldsymbol{\psi}$ given $\boldsymbol{y}$*: Hence, the posterior pdf of $\boldsymbol{\psi}$ given the observations in $\boldsymbol{y}$ is

$$\begin{aligned} p(\boldsymbol{\psi}\,|\boldsymbol{y}) &\propto \pi_{\boldsymbol{\psi}}(\boldsymbol{\psi}) p(\boldsymbol{y}\,|\boldsymbol{\psi}) \\ &\propto \pi_{\boldsymbol{\psi}}(\boldsymbol{\psi}) L(\boldsymbol{y}\,|\boldsymbol{\psi}). \end{aligned} \tag{8}$$

  We draw samples to estimate the unknown parameters in $\boldsymbol{\psi}$ from the posterior pdf in (8).
- *Sampling the parameters in $\boldsymbol{\psi}$*: Sampling from (8) is a large dimensional problem. This motivates us to draw samples

from the joint posterior using a Gibbs sampling method [15]. The sequence (see, for example, [14]) is as follows:

1) We first draw $\sigma^{(t)}$ from $p(\sigma\,|\mu^{(t-1)},\boldsymbol{\tau}^{(t-1)},\boldsymbol{y})$ using *rejection sampling* [15].
   - $p(\sigma\,|\mu,\boldsymbol{\tau},\boldsymbol{y})$
   $$\approx \left(\sigma^2+\sigma_w^2\right)^{-N(\boldsymbol{\tau})/2} \times \exp\left[-\frac{\sum_{\boldsymbol{r}\in R(\boldsymbol{\tau})}(y(\cdot)-\mu)^2}{2(\sigma^2+\sigma_w^2)}\right] \times i_{(0,\sigma_{\mathrm{MAX}})}(\sigma) = q(\sigma\,|\mu,\boldsymbol{\tau},\boldsymbol{y}).$$
   - *Rejection sampling*:
     a) We draw σ from $\pi_\sigma(\sigma)=\mathrm{uniform}(0,\sigma_{\mathrm{MAX}})$;
     b) We draw u from uniform(0,1);
     c) We repeat steps a) and b) until $u \le (q(\sigma\,|\mu,\boldsymbol{\tau},\boldsymbol{y})/m(\mu,\boldsymbol{\tau}))$, where
     $$m(\mu,\boldsymbol{\tau}) = \max_{\sigma} q(\sigma\,|\mu,\boldsymbol{\tau},\boldsymbol{y}) = q\left(\sqrt{\widehat{\sigma_q^2}}\,|\mu,\boldsymbol{\tau},\boldsymbol{y}\right),$$
     $$\widehat{\sigma_q^2} = \min\{\max[0,\widehat{\sigma^2}],\sigma_{\mathrm{MAX}}^2\}$$
     and
     $$\widehat{\sigma^2} = \frac{\sum_{\boldsymbol{r}\in R(\boldsymbol{\tau})}(y(\cdot)-\mu)^2}{N(\boldsymbol{\tau})} - \sigma_w^2.$$
2) We then draw $\mu^{(t)}$ from $p(\mu\,|\sigma^{(t)},\boldsymbol{\tau}^{(t-1)},\boldsymbol{y})$, which is a *truncated Gaussian distribution* [16]. The pdf $p(\mu\,|\sigma,\boldsymbol{\tau},\boldsymbol{y})$ is equivalent to $\mathcal{N}(\mu;\widehat{\mu},(\sigma^2+\sigma_w^2/N((\boldsymbol{\tau})))i_{(0,\mu_{\mathrm{MAX}})}(\mu)$, where $\widehat{\mu}=(\sum_{\boldsymbol{r}\in R(\boldsymbol{\tau})}y(\cdot)/N(\boldsymbol{\tau}))$.
3) Finally, we draw $\boldsymbol{\tau}^{(t)}$ from $p(\boldsymbol{\tau}\,|\sigma^{(t)},\mu^{(t)},\boldsymbol{y})$ using a *shrinkage slice sampling* [17].
   - $p(\boldsymbol{\tau}\,|\boldsymbol{\theta}^{(t)},\boldsymbol{y}) \approx \pi_{\boldsymbol{\tau}}(\boldsymbol{\tau})L(\boldsymbol{y}\,|\boldsymbol{\tau},\boldsymbol{\theta}^{(t)})$.
     a) We define the starting *hyperrectangle* as follows: $x_{0,\mathrm{L}}=x_{0,\mathrm{MIN}}$; $x_{0,\mathrm{U}}=x_{0,\mathrm{MAX}}$; $z_{0,\mathrm{L}}=z_{0,\mathrm{MIN}}$; $z_{0,\mathrm{U}}=z_{0,\mathrm{MAX}}$; $d_{\mathrm{L}}=0$; $d_{\mathrm{U}}=d_{\mathrm{MAX}}$; $A_{\mathrm{L}}=0$; $A_{\mathrm{U}}=A_{\mathrm{MAX}}$; $\phi_{\mathrm{L}}=-\pi/4$; $\phi_{\mathrm{U}}=\pi/4$; $r_{1,\mathrm{L}}=0$; $r_{1,\mathrm{U}}=r_{1,\mathrm{MAX}}$.
     b) We draw an auxiliary random variable $u^{(t)}$ from uniform$(0,L(\boldsymbol{y}\,|\boldsymbol{\tau}^{(t-1)},\boldsymbol{\theta}^{(t)}))$.
     c) We draw $\boldsymbol{\tau}=[x_0,z_0,d,A,\phi,r_1]^T$ from uniform$(x_{0,\mathrm{L}},x_{0,\mathrm{U}})$, uniform$(z_{0,\mathrm{L}},z_{0,\mathrm{U}})$, uniform$(d_{\mathrm{L}},d_{\mathrm{U}})$, uniform$(A_{\mathrm{L}},A_{\mathrm{U}})$, uniform$(\phi_{\mathrm{L}},\phi_{\mathrm{U}})$, and uniform$(r_{1,\mathrm{L}},r_{1,\mathrm{U}})$, respectively.
     d) If $\boldsymbol{\tau}$ is within the starting *hyperrectangle*, i.e., $L(\boldsymbol{y}\,|\boldsymbol{\tau},\boldsymbol{\theta}^{(t)}) \ge u^{(t)}$, we return $\boldsymbol{\tau}^{(t)}=\boldsymbol{\tau}$. Otherwise we *shrink* the original *hyperrectangle* as follows:
        — if $x_0 \le x_0^{(t-1)}$, we set $x_{0,\mathrm{L}}=x_0$; otherwise we set $x_{0,\mathrm{U}}=x_0$.
        — if $z_0 \le z_0^{(t-1)}$, we set $z_{0,\mathrm{L}}=z_0$; otherwise we set $z_{0,\mathrm{U}}=z_0$.
        — if $d \le d^{(t-1)}$, we set $d_{\mathrm{L}}=d$; otherwise we set $d_{\mathrm{U}}=d$.
        — if $A \le A^{(t-1)}$, we set $A_{\mathrm{L}}=A$; otherwise we set $A_{\mathrm{U}}=A$.
        — if $\phi \le \phi^{(t-1)}$, we set $\phi_{\mathrm{L}}=\phi$; otherwise we set $\phi_{\mathrm{U}}=\phi$.
        — if $r_1 \le r_1^{(t-1)}$, we set $r_{1,\mathrm{L}}=r_1$; otherwise we set $r_{1,\mathrm{U}}=r_1$.
        — we repeat from step c.
   - Any floating-point underflows that occur while evaluating the expression $L(\boldsymbol{y}\,|\boldsymbol{\tau},\boldsymbol{\theta}))$ in MATLAB are adjusted numerically.
4) We repeat from *Step* 1 until a sufficient number of samples (T_0) have been drawn.

The samples $\boldsymbol{\psi}^{(0)},\boldsymbol{\psi}^{(1)},\boldsymbol{\psi}^{(2)},\ldots$ produce a guaranteed stationary (invariant) posterior distribution of $p(\boldsymbol{\psi}\,|\boldsymbol{y})$ [18].

- *Sampling the signals* $f(\boldsymbol{r};\boldsymbol{\theta})$: We estimate the signals $f(\boldsymbol{r};\boldsymbol{\theta})$ for each pixel using a *composition sampling* from the posterior pdf $p(f(\cdot)\,|\boldsymbol{y}) = \int p(f(\cdot)\,|\boldsymbol{\psi},\boldsymbol{y})p(\boldsymbol{\psi}\,|\boldsymbol{y})\mathrm{d}\boldsymbol{\psi}$ as mentioned in [14]. The process is as follows:
  1) We draw $\boldsymbol{\psi}^{(t)}$ as mentioned before.
  2) We draw $f(\cdot)^{(t)}$ from $p(f(\cdot)\,|\boldsymbol{\psi}^{(t)},\boldsymbol{y})$ such that
     - for $\boldsymbol{r}\in R(\boldsymbol{\tau}^{(t)})$ we draw $f(\cdot)^{(t)}$ from
     $$\mathcal{N}\left(f(\cdot);\frac{(\sigma^{(t)})^2 y(\cdot)+\sigma_w^2\mu^{(t)}}{(\sigma^{(t)})^2+\sigma_w^2},\left[\frac{1}{(\sigma^{(t)})^2}+\frac{1}{\sigma_w^2}\right]^{-1}\right).$$
     - for $\boldsymbol{r}\in R^{\mathrm{c}}(\boldsymbol{\tau}^{(t)})$ we set $f(\cdot)^{(t)}=0$.

  Samples $f(\cdot)^{(0)},f(\cdot)^{(1)},f(\cdot)^{(2)},\ldots$ yield a Markov chain with a stationary posterior distribution equal to $p(f(\cdot)\,|\boldsymbol{y})$.
- *Estimating* $\boldsymbol{\psi}$ *and* $f(\cdot)$: We define t_0 as the burn-in period. Hence, the minimum mean-square estimates (MMSE) of $\boldsymbol{\psi}$ and $f(\cdot)$ are computed as follows:
$$\widehat{\boldsymbol{\psi}} = \frac{1}{T_0-t_0}\left(\sum_{t=t_0+1}^{T_0}\boldsymbol{\psi}^{(t)}\right), \tag{9}$$
$$\widehat{f(\cdot)} = \begin{cases}\frac{1}{T_0-t_0}\left(\sum_{t=t_0+1}^{T_0}f(\cdot)^{(t)}\right) & \text{if } \boldsymbol{r}\in R(\widehat{\boldsymbol{\tau}}),\\ 0 & \text{otherwise,}\end{cases} \tag{10}$$
where $\widehat{\boldsymbol{\tau}}$ is the MMSE of $\boldsymbol{\tau}$ as defined in (9).

IV. Comparison of MW, EM/ICM, and our Proposed Estimation Methods

In this section, we first present the MW-test-based segmentation method [1] analytically. Then, we present the EM/ICM-based method as proposed by Gottardo *et al.* [10]. Finally, we present a comparative study of MW, EM/ICM, and our proposed estimation methods.

A. Mann–Whitney Segmentation Method

In [1] the authors propose a MW-test-based segmentation method for gene-signal estimation. First, the independent measurements $X_1,X_2,\ldots,X_n$ and $Y_1,Y_2,\ldots,Y_m$ are collected

from two random variables X and Y with sample means μ_X and μ_Y, respectively. The rank-sum statistic W is defined as the sum of ranks of all the X samples in the combined ordered sequence of the X and Y samples. The testing problem is defined as follows:

$$\begin{aligned} H_0&: \mu_X - \mu_Y = 0 \\ H_1&: \mu_X - \mu_Y > 0. \end{aligned} \tag{11}$$

Rejection of H_0 occurs when $W \geq w_{\vartheta,n,m}$, the critical value corresponding to the significance level ϑ [19].

A predefined *target mask* is used to identify a portion of the image of the spot and its background that contains the *target site*. Eight samples are randomly selected from the known background (outside the *target mask*) as $Y_1, Y_2, \ldots, Y_8$, and the lowest eight samples are picked within the *target mask* as $X_1, X_2, \ldots, X_8$. The rank-sum statistic W is calculated and, for a given significance level ϑ, compared with $w_{\vartheta,n,m}$. Under the null hypothesis, we have

$$\frac{W - \frac{n(n+1)}{2} - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \propto \mathcal{N}(0,1) \tag{12}$$

if both m and n are large [19]. If the null hypothesis is not rejected, then one sample is discarded at random from the eight potential target region's samples and the lowest eight remaining samples are selected from the *target mask*. The Mann–Whitney test is repeated until the null hypothesis is rejected. When H_0 is rejected, the *target site* is decided, with significance level ϑ, to be the eight samples causing the rejection, together with all pixels in the *target mask* whose values are greater than or equal to the minimum value of the eight. If the null hypothesis is never rejected, then we conclude that there is no significant signal at the *target site*. Once a *target site* is determined, gene expression is measured by the median of the *target site* minus the median of the background area (outside the *target mask* area).

B. Gottardo Segmentation Method [10]

We summarize briefly the segmentation method as proposed by Gottardo *et al.* (see [10] for more information on this method). For a given spot, the measurement model at every pixel location is proposed as [10]

$$y(\cdot) = \eta + \iota x(\cdot) + \epsilon(\cdot)/\sqrt{\nu(\cdot)} \tag{13}$$

where $(\cdot)$ denotes a pixel location, η is the background effect, ι quantifies the gene signal corresponding to the spot, $x(\cdot)$ is 1 to classify the pixels as belonging to the spot and 0 otherwise, $\epsilon(\cdot)$ follows $\mathcal{N}(\epsilon(\cdot); 0, 1/\lambda^{\epsilon})$, and $\nu(\cdot)$ follows a Gamma distribution, $\mathcal{G}(\kappa/2, \kappa/2)$. The random variables $\epsilon(\cdot)$ and $\nu(\cdot)$ are independent of each other and i.i.d. from pixel to pixel. Hence, $\epsilon(\cdot)/\sqrt{\nu(\cdot)}$ follows a t-random variable with κ degrees of freedom and variance λ^{ϵ}. A modified symmetric first-order Ising model is used to estimate the pixel classification level $x(\cdot)$. The spot pixels are forced to lie within a circle of fixed radius r_g and center c_g. The lumped vector forms of $x(\cdot)$, $\epsilon(\cdot)$, and $\nu(\cdot)$ are $\boldsymbol{x}$, $\boldsymbol{\epsilon}$, and $\boldsymbol{\nu}$. In [10], the authors propose an EM/ICM-based microarray spot-image segmentation algorithm for estimating the unknown parameters $[\eta, \iota, \boldsymbol{x}^T, \lambda^{\epsilon}, c_g, \boldsymbol{\nu}^T]^T$ assuming κ and r_g values are known.

C. Comparison

Our proposed parametric method is clearly an improvement over the existing nonparametric MW-test-based segmentation method which only works well at high SNR. We justify this claim in Section VI where we show that both the MW- and EM/ICM-based segmentation methods do not perform as well as our proposed method in very low-SNR images. Since our proposed method is an improved and simplified version of the EM/ICM-based segmentation method, we confine ourselves to compare with that method in the rest of this subsection. The segmentation method as proposed by Gottardo *et al.* is a pixel-by-pixel process whereas our method is more parametric. The forward model (13) is not analytically tractable for developing a user friendly MCMC-based signal-estimation algorithm. The EM/ICM-based algorithm was developed for multiple-center-hole case. Our proposed method can be extended to such case at the cost of added computational load. Note that cDNA microarray spots with more than one center hole are very rare in practice.

Gottardo *et al.* assume the radius of the spot is fixed and known, whereas we assume that the circular outer-periphery radius is an unknown parameter. As an advantage, if the signal level in a spot is insignificant, the spot-outer-periphery radius parameter r_1 in (1) is expected to be estimated as a value near to zero using our algorithm.

In our analysis we take into account the random fluctuation of the gene signal in the spots by modeling the undesired binding of the host. As a consequence, we estimate the signals in each spot-pixel location using a *composition sampling* method, assuming random fluctuation of the gene signals. On the other hand, the Gottardo *et al.* segmentation method does not account for that in their analysis and models the gene signal in the spot as a deterministic constant.

We observe that estimation of the center holes using the EM/ICM method is overly sensitive to the initialization of the unknown parameters in very low-SNR images. Namely, such sensitivity occurs because the EM/ICM algorithm employs a pixel-by-pixel processing. In contrast, our proposed method can overcome such a problem because of the realistic and parametric spot-shape information that we employ in our analysis. As a consequence, more accurate prior knowledge is employed during the initialization of the estimation using our method. In general, our algorithm is time intensive and hence we propose using conventional methods for segmenting high-SNR spot images and our method for segmenting very noisy spot images.

V. Results Using Real Data

A 70-base-pair oligonucleotide microarray designed to analyze 6242 genes from the *protozoan* human gut parasite *Entamoeba histolytica* was used for image signal analysis [11]. The average computed melting temperatures for all oligos was 80.8C, with a standard deviation of 2.73 (range 70.5–95.5C). The oligonucleotides were manufactured by *Illumina* (San Diego, CA) and were printed in triplicate on 100-cell-associate epoxy slides (Santa Clara, CA) by the Washington University School of Medicine Microarray core facility. RNA was isolated from approximately 5×10^6 log-phase *Entamoeba histolytica*

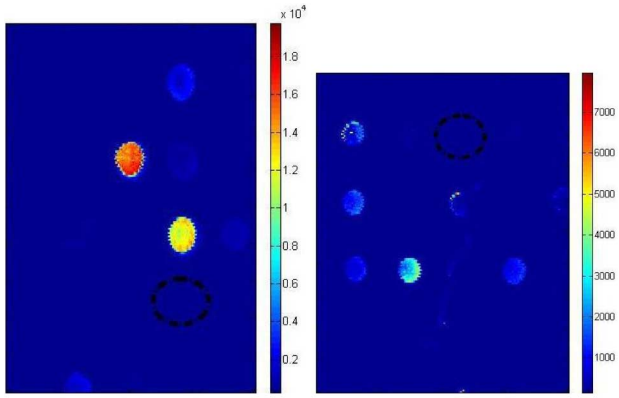


Fig. 5. Two different regions of *Entamoeba* microarray intensity image data exhibiting gene signals in low signal-to-noise ratio. Signals in the dash-dotted regions in (a) and (b) are not visible and the corresponding genes' expressions cannot be discerned.

HM-1:IMSS grown in 15-ml glass flasks using the *Qiagen RNeasy kit* (Valencia, CA) following the manufacturer's protocol, including a *DNase* treatment. Past studies suggested that in amoebae more than 30% of genes are transcribed at detectable levels when grown in culture [20], [21]. RNA quantity and quality were obtained from an absorbance ratio at 260 nm and 280 nm. RNA quality was confirmed for each sample using an *Agilent* 2100 bioanalyzer (Palo Alto, CA) according to the manufacturer's instructions. Cy3- and Cy5-labeled cDNA was created using the *Genisphere 3DNA array350 kit* (Hatfield, Pennsylvania). Slides were scanned using a *ScanArray Express HT* scanner (Perkin Elmer, Boston, MA) to detect Cy3 and Cy5 fluorescence. Laser power was kept constant, and PMT was varied for each experiment to achieve optimal signal intensity with lowest possible background fluorescence. In order to differentiate expression levels among highly expressed genes, the data were collected at low-PMT settings. We applied our proposed estimation algorithms to noisy parts of the microarray image data.

In Fig. 5(a) and (b), we show intensity images of two different parts of the raw data from Cy3 and Cy5 fluorescence, respectively. In most regions, gene signals are hardly visible compared with those of the few highly expressed genes' signals in some spots. We use two randomly chosen spots and their neighboring regions for analysis (see elliptical dash-dotted regions in Fig. 5(a) and (b), respectively).

We denote the randomly chosen spots and their neighboring regions as data-sets A and B, respectively. The images have dimensions of 35×35 pixels in each. Realistic and parametric modeling of the spot-shapes allows us to initialize the prior shape-parameter pdf's accurately. We chose prior pdf's with $\mu_{\rm MAX} = \text{maximum}(\{y_{\rm i} \ \forall \ i \in (1, LM)\}), \sigma_{\rm MAX} = 2\sigma_w, \phi_{\rm MIN} = -\pi/4, \phi_{\rm MAX} = \pi/4$. We chose $x_{0,\rm MIN}$, $x_{0,\rm MAX}$, $z_{0,\rm MIN}$, and $z_{0,\rm MAX}$ around the neighborhood of $\{x_0 = 0, z_0 = 0\}$. We picked $r_{1,\rm MAX} \sim 12$ pixels using a prior knowledge from the high SNR spots. The size parameters of the center hole ellipse, $d_{\rm MAX}$, $A_{\rm MIN}$, and $A_{\rm MAX}$, are chosen to span inside the outer periphery. Note that Markov chain calculation may not converge to a true value for too small an $A_{\rm MIN}$ value.

We used a Intel dual-core CPU (Clocks: 2.4 GHz and 1.58 GHz; RAM: 1.99 GB) for all the computer simulations in this paper. We compare the estimated spot-shapes after running individual Gibbs samplers for 10 000, 1000, 500, and 100 cycles, all starting with different initialization points, while evaluating our proposed MCMC-based MMSE estimation. We discarded 8000, 800, 400, and 80 samples, respectively; therefore the burn-in periods were $t_0 = 8000, 800, 400$, and 80, respectively. We estimated the MMSE of posterior pdf's $p(\boldsymbol{\psi}\,|\boldsymbol{y})$ and $p(f(\cdot)\,|\boldsymbol{y})$ as well as the unknown parameters of $\boldsymbol{\psi}$ using (9) and (10) from the last 2000, 200, 100, and 20 samples of the respective Gibbs samplers. We eliminated the weak-estimated signals to zero values if $\{\widehat{f(x,z)} \leq 0.75 \times \text{MAX}_{x,z}(\widehat{f(x,z)})\}$, where the threshold 0.75 was chosen arbitrarily. We introduced this step in our analysis for making a rough estimate of other center-holes (if they at all exist).

In Figs. 6 and 7 we present the signal estimation results for these data-sets using our method. We computed the sample estimates of the background noise mean μ_w and variance σ_w^2 as (120.59, 122.58) and (119.91, 181.74), respectively. In these figures, we present the noisy images and our estimated images for data-sets A and B, respectively. Here, we ran separate Gibbs samplers of 10 000, 1000, 500, and 100 cycles for each data-set. In Fig. 8, we present convergence plots of the Markov chains for data-set A with 100 draws for parameters a) x_0, b) z_0, c) r_1, d) d, e) A, f) ϕ, g) μ, and h) σ. We computed the SNRs of the data-sets A and B as 2.9 dB and -21.52 dB, respectively, using (14) (see Section VI). Note that the estimated center hole might not be very accurate for the data-set B since this data-set is overly noisy. In Table I we present the estimated gene signal means and computation times for data-sets A and B with 10 000, 1000, 500, and 100 Markov draws.

We conclude that our method: i) clearly segments the foreground spot shapes from the respective backgrounds and ii) also estimates the foreground signals using Gibbs sampler with 1000 runs. The data-set A is less noisy and hence the estimation performance using this data-set does not vary much (see Table I and Fig. 6). However, the data-set B is very noisy and estimation performances with 500 and 100 draws using this data-set do not appear very satisfactory (see Table I and Fig. 7). Despite of this deficiency, we cannot use long time in real-life analysis for a single-noisy spot since the whole microarray might contain thousands of such spots. Hence, we recommend using 500-cycle Gibbs sampler that takes around reasonable 10 min to process images of dimension 35×35 pixels. We justify this claim using a numerical example in Section VI.

VI. Numerical Examples

In this section we present two numerical examples. In Example 1, we compare the estimation accuracy of our proposed method with MW- and EM/ICM-based methods. This analysis is performed for a spot shape with two elliptical nonoverlapping center holes using the estimation method we proposed in Section III. In Example 2, we address a more realistic example where we generate noisy data for parasitic amoebae surrounded

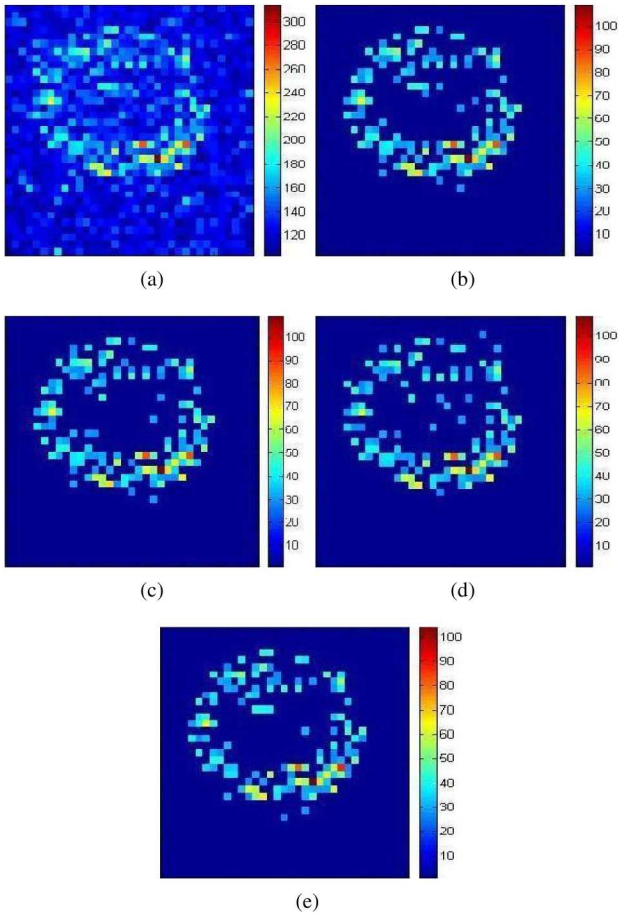


Fig. 6. Estimation results using our proposed algorithm of Markov chain Monte Carlo-based minimum mean-square error algorithm for the data-set A. (a) Noisy data. (b)–(e) estimated shape, signals, and location after 10 000, 1000, 500, and 100 draws, respectively. The estimated images are presented using the methodology described in Section V.

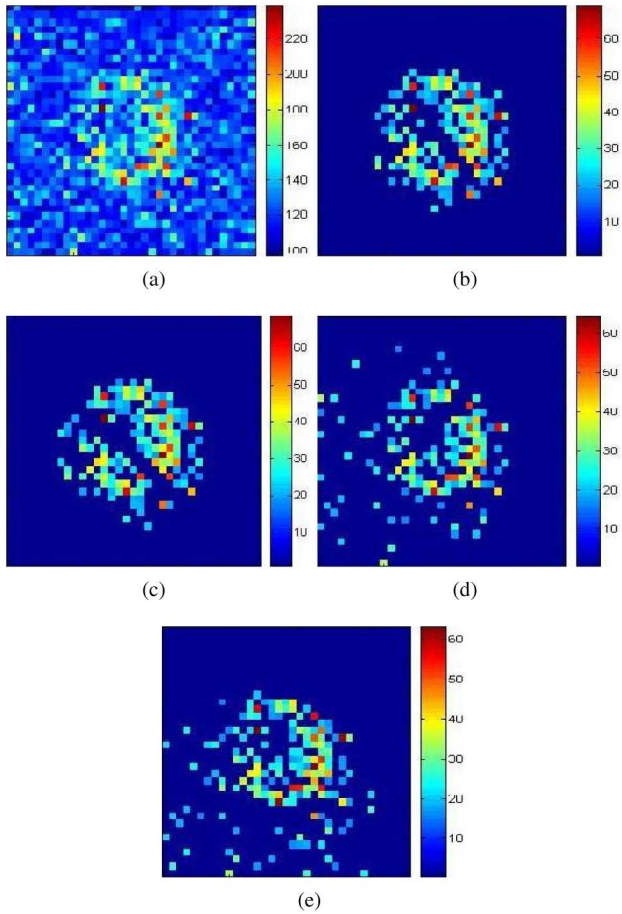


Fig. 7. Estimation results using our proposed algorithm of Markov chain Monte Carlo-based minimum mean-square error algorithm for the data-set B. (a) Noisy data. (b)–(e) estimated shape, signals, and location after 10 000, 1000, 500, 100 draws, respectively. The estimated images are presented using the methodology described in Section V.

by a host of varying amount. Here, we generate the spot shape considering a more realistic model as proposed in [13]. We qualitatively compare the estimated image using this data with the ideal amoeba image data.

Example 1: In this example we aim to show that at low SNR our method outperforms the existing methods. We generated the simulated image of dimensions 25×25 pixels, assuming the spot shape with two elliptical nonoverlapping center holes [see Fig. 9(a)]. We used the foreground signal mean $\mu = 20$, which resembles the gene signal, and variance $\sigma^2 = 3$. In Fig. 9(b), we present the noisy version of this image with noise variance $\sigma_w^2 = 300$. Here we use noise mean $\mu_w = 0$ without loss of generality. In Fig. 9(c), we present the estimated image from this noisy image using the EM/ICM algorithm. Here, the estimated foreground signal mean is $\hat{\mu} = 15.98$. In Fig. 9(d), we present the estimated image using the MW-test-based image segmentation method with $\vartheta = 0.05$. We observe that the separation of the foreground and background is impossible.

In Fig. 9(e), (f), (g), and (h), we present the segmentation results using our proposed method as outlined in Section III with *a priori* spot-shape information, assuming two elliptical center holes, with the flexibility that the center holes can merge with each other. We drew 4000, 1000, 500, and 100 samples, respectively, for evaluating our proposed Gibbs sampler. In these figures, we estimated the foreground signal means $\hat{\mu} = 19.02, 19.25, 19.49$, and 15.38, respectively. Note that here we present the estimated $\widehat{f(x,z)}$ directly unlike eliminating the weak signals as we performed in Section V. Here we used a similar initialization strategy as described for the real-data case. In Table II we present estimated gene signal means and computation times for different simulations that we performed in this example. From this result (see Table II and Fig. 9), we conclude that our proposed method performs very well using the 500-cycle Gibbs sampler that takes around 6.21 min to process images of dimensions 25×25 pixels. Such a result is ascertained given that we initialize our algorithm with good starting points. We already discussed in Section V that accurate initialization is always feasible in our analysis for the case of real data.

In Fig. 10, we present a quantitative comparison of the estimation accuracy of these three methods. We define the SNR as follows:

$$\frac{N(\boldsymbol{\tau})\mu^2}{N(\boldsymbol{\tau})\sigma^2 + N\left(\mu_w^2 + \sigma_w^2\right)}. \tag{14}$$

In our analysis we define mean-square error (MSE) as $E[f(\boldsymbol{r}) - \widehat{f(\boldsymbol{r})}]^2$, where $E(\cdot)$ denotes the statistical mean.

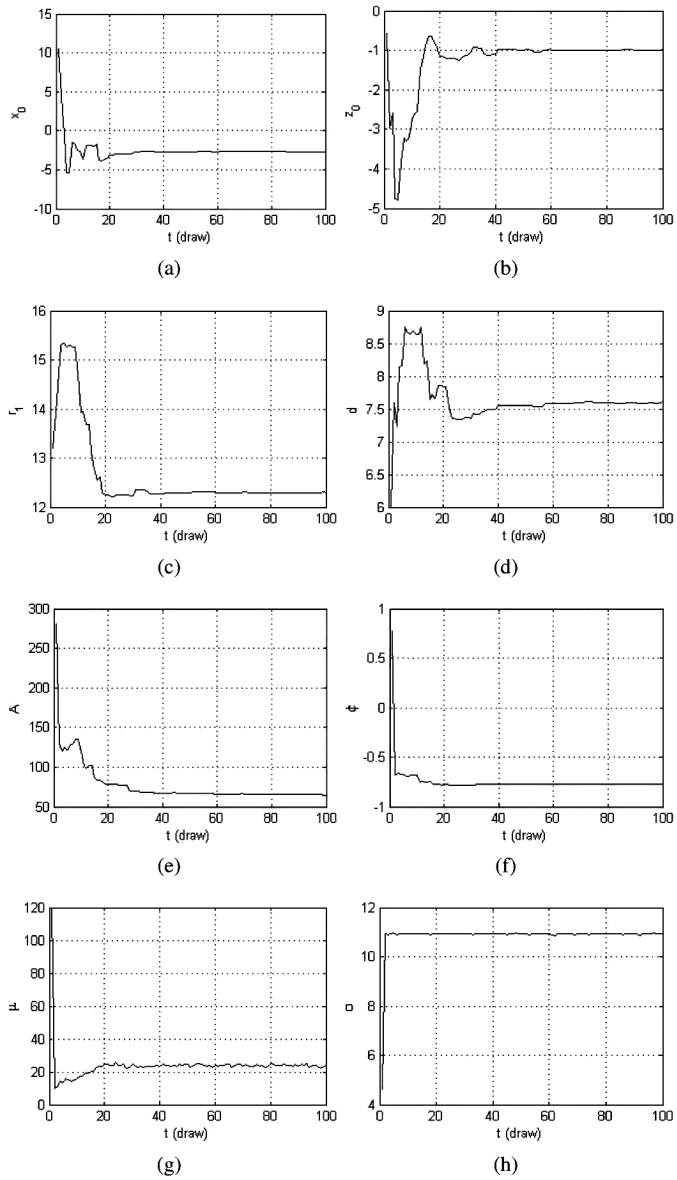


Fig. 8. Convergence plots of the Markov chain for parameters a) x_0, b) z_0, c) r_1, d) d, e) A, f) ϕ, g) μ, and h) σ, respectively, using data-set A.

TABLE I
ESTIMATED GENE SIGNAL MEANS AND COMPUTATION TIMES (IN MINUTES) FOR DATA-SETS A AND B AFTER 10 000, 1000, 500, AND 100 MARKOV DRAWS USING OUR PROPOSED METHOD

Draws	10000	1000	500	100
Data-set A: $\hat{\mu}$	40.24	40.19	39.98	38.83
Computation time (min)	226.94	21.76	10.67	1.99
Data-set B: $\hat{\mu}$	30.63	30.45	27.8	27.12
Computation time (min)	230.01	21.23	10.27	1.68

We perform 20 realizations per SNR. We vary the background noise level to obtain noisy images with different SNR values. Though the MW-test-based method performs worst in the beginning, starting from -20 dB it starts outperforming the EM/ICM-based method. Our proposed method performs the best.

The EM/ICM cannot efficiently estimate the spot shape in large noise. Also this algorithm is very sensitive in estimating the center holes because of employing a pixel-by-pixel processing. In conclusion, though our method is time intensive than

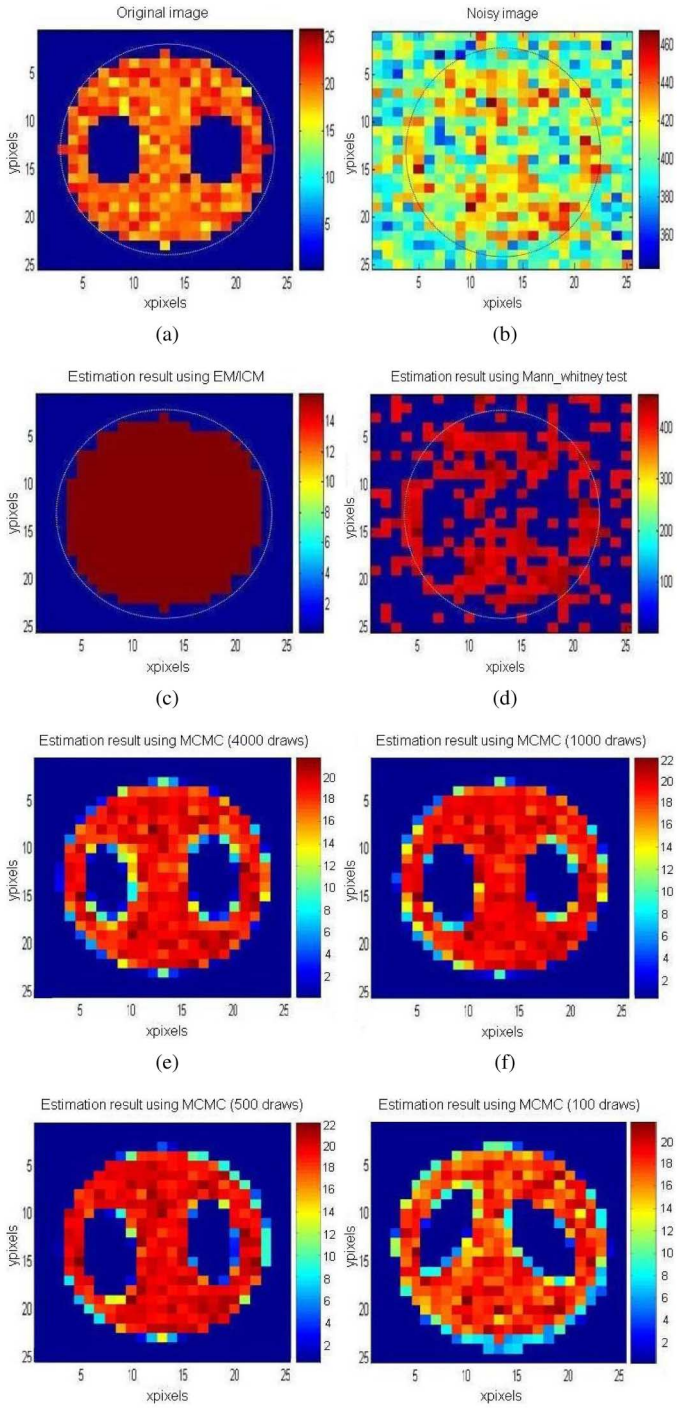


Fig. 9. (a) Simulated image of dimensions 25×25 pixels with the foreground signal mean $\mu = 20$ and variance $\sigma^2 = 3$. (b) The noisy version of this image with noise variance $\sigma_w^2 = 300$ and mean $\mu_w = 0$. (c) The estimated image from the noisy image using EM/ICM algorithm. The estimated foreground signal mean is $\hat{\mu} = 15.98$. (d) The estimated image using MW-test based image segmentation method using $\vartheta = 0.05$. (e)–(h) The segmented images using our proposed method after running individual Gibbs samplers for 4000, 1000, 500, and 100 cycles, respectively. The estimated foreground signal means are $\hat{\mu} = 19.0219.25, 19.49$, and 15.38, respectively.

compared to existing methods but outperforms them with significant margins. In our future work we aim at developing a fast version of our proposed algorithm.

Example 2: In this example we qualitatively show how our proposed method is useful in a more realistic environment. For

TABLE II
ESTIMATED GENE SIGNAL MEANS AND COMPUTATION TIMES IN EXAMPLE 1

Methods	MCMC	MCMC	MCMC
Draws	4000	1000	500
$\hat{\mu}$	19.02	19.25	19.49
Computation time	49.7min	12.51min	6.21min
Methods	MCMC	EM/ICM	MW
Draws	100	–	–
$\hat{\mu}$	15.38	15.98	–
Computation time	1.09min	16.23sec	0.05sec

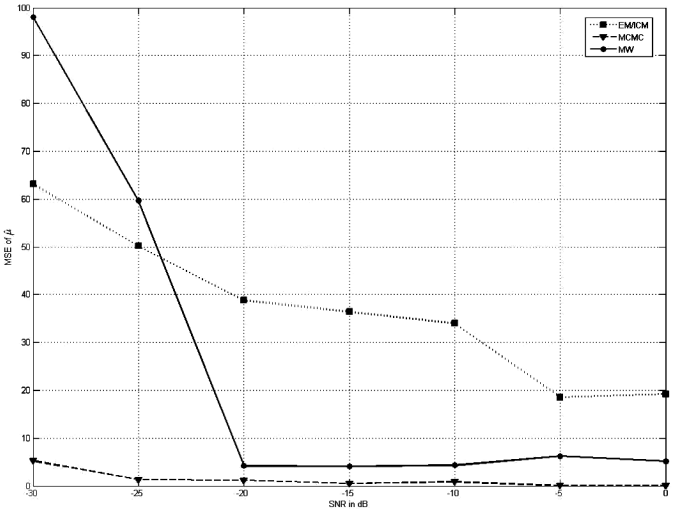


Fig. 10. A quantitative comparison of the mean-square-error of the estimated $\hat{\mu}$ using our proposed MCMC-based, MW-test-based, and EM/ICM-based methods. We use $\vartheta = 0.05$ for evaluating the MW-test-based segmentation method.

this analysis, we consider the case of clinically measured human gut parasite *Entamoeba histolytica* data. In such data, host RNA obscures the ground-truth. As a result, the measured *Entamoeba* RNA image varies measurably from the truth. Our motivation in this example is to show that the application of statistical signal processing can decrease that variance from the truth.

In this example we generate data using the spot shape model as we proposed in Section II. We further distort the true spot shape to make it more realistic. In order to do that, we i) eliminate one chord using a randomly chosen chord length and position and ii) introduce an edge noise effect in the spot by randomly keeping or removing the spot pixels along the spot edge [13].

We generate data by assuming that the truth attached to the gene-signal quantification level is $5a$ in the spot where a is a known constant (see Fig. 11, first row). This spot can be assumed as an outcome of a purified *Entamoeba* RNA image. Human RNA is sticky and binds weakly/intermittently to the spot, causing fluctuations/false readings in the foreground signal. In general, the host RNA quantity is large in the measured clinical sample, reducing the amount of labeled *Entaomeba* RNA hybridizing to the spot. As a result, the measurement image becomes noisy. We generate noisy image data at SNRs of 5 dB, 0 dB, and -5 dB, respectively. Such images are generated assuming the following mixtures: i) a amount of *Entamoeba* and 4 a amount of host; ii) 0.5 a amount of *Entamoeba* and 4.5 a amount of host; and iii) 0.25 a amount of *Entamoeba* and 4.75 a amount of host. We vary the foreground signal variance in the images as $\sigma^2 = c/100$ where c is the host amount in the clinical mixtures. We estimate the unknown parameters for these images using our proposed method. Then, we compare the uncorrected and signal-processed samples with the original pure *Entamoeba* sample.

In Fig. 11 we present the analysis result. The ground-truth is shown in the first row. In the second row we present the estimation results at 5, 0, and -5 dB, respectively, for the mixture i) image data. The results for mixtures ii) and iii) are presented in the third and fourth rows, respectively. Our proposed method estimates the spot shapes efficiently for all the generated noisy images of all the mixtures. In addition, we notice that the signal fluctuations at the estimated foreground are better observed as the host redundancy increases in the input noisy images (see Fig. 11, last row). On the other hand, when the host redundancy is less, the estimated foreground signal fluctuation is not well observed at low SNR (see Fig. 11, second and third rows). We estimate the means of the foreground signals satisfactorily in these nine cases. We conclude that statistical signal processing can play a significant role in estimating spot shapes and signals in noisy microarray image data as we present in this example.

VII. CONCLUSION

We have presented a novel mechanism for microarray image analysis that has several potential advantages for biological investigators. The drastic reduction in stochastic noise will increase the accuracy of all measured ratios compared to the methods currently used for signal quantification. Most significantly, oligonucleotide and similar microarray images analyzed with our algorithm can experience log increases in gene-expression dynamic range by expanding the lower limit. This will be accomplished by decreasing noise from spots that would otherwise be excluded from microarray analysis due to SNRs that are too low for reliable quantification. The drastic reduction in noise and accurately defined area of signal will additionally result in a more accurate quantification, and therefore a more accurate resultant ratio, from spots where at least one channel has low SNR. Other researchers, using less rigorous algorithms, have found that the quality of measured ratios from low expression spots is unreliable. By differentiating low SNR spots from no-signal spots, microarray and other similar images could be more reliably employed in sensitive biodetection assays [22]. In addition, by combining more accurate signals from differentially stringent hybridization conditions, off-target hybridization thermodynamic estimates could then more accurately suggest the degree of sequence misidentification. Our algorithms for microarray analysis should make these applications feasible.

In our future work we will apply our proposed method to the real microarray image data of a mixture of *Entamoeba* RNA and host human RNA to determine the effects of interference. We have already analyzed a soft version of this experiment in Example 2 in Section VI. This RNA mixture will vary significantly from amoebic RNA isolated without any host cells. One would expect some true transcriptional difference to exist based on the organism's adaptation to its environment; however, we do not anticipate that the true biological transcriptional profile

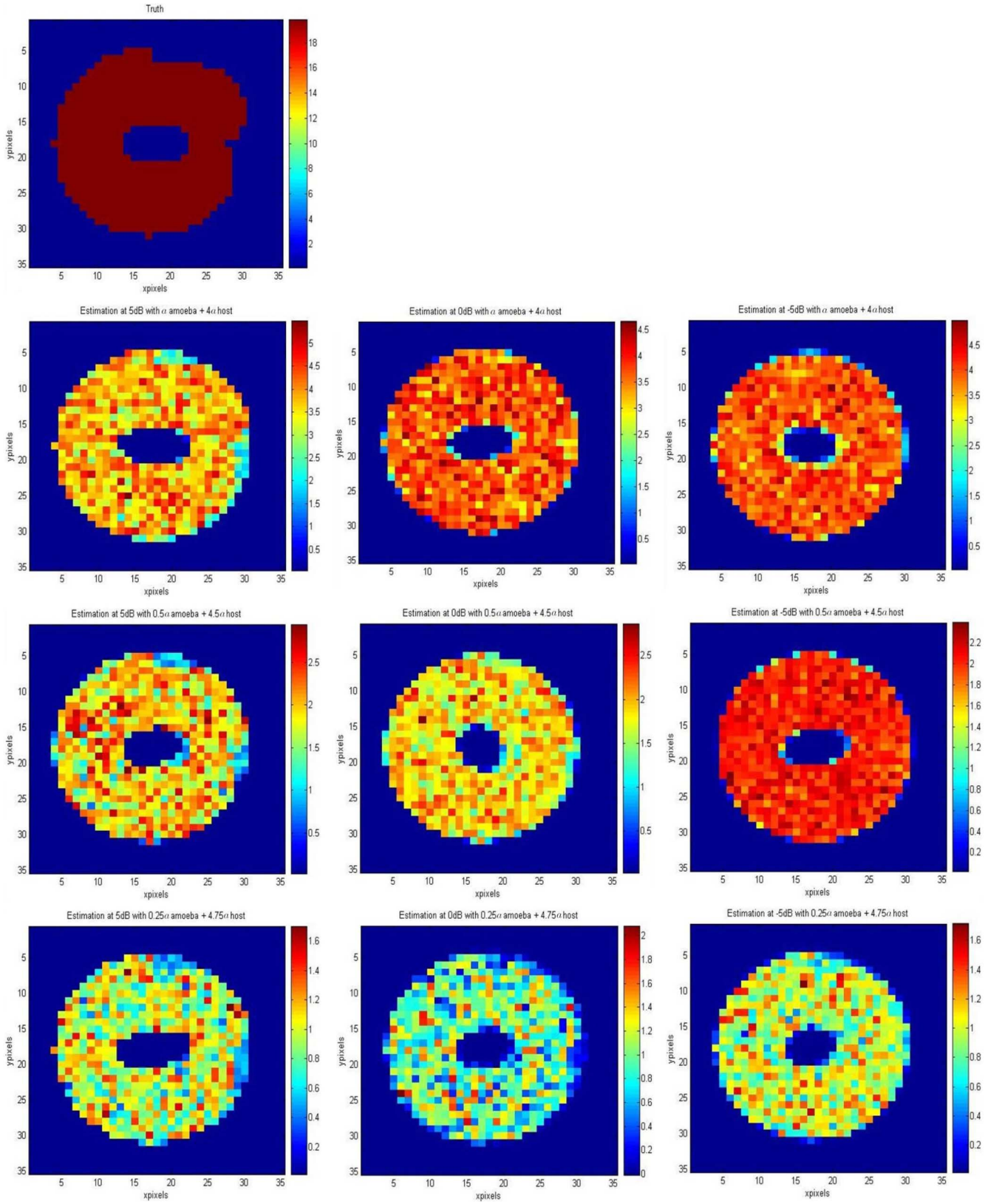


Fig. 11. Row 1: Simulated spot-image of purified *Entamoeba* RNA with the truth attached as the gene signal quantification level is 5 a in the spot where a is a known constant; Row 2: Segmented images using our proposed method for the clinical mixture composed of a amount of *Entamoeba* and 4 a amount of host at SNRs 5 dB, 0 dB, and -5 dB; Row 3 and Row 4: Similar analysis result as shown in Row 2 for clinical mixtures composed of 0.5 a amount of *Entamoeba* and 4.5 a amount of host (Row 3) and 0.25 a amount of *Entamoeba* and 4.75 a amount of host (Row 4), respectively.

would be as distinct as the dual-source RNA profile of the host and the amoeba.

Our algorithm is relatively slow but is more accurate than existing methods. In order to analyze the total-genome-microarray

images of any organism, we propose using our method for processing low-SNR spot images and conventional methods for processing high-SNR spot images. In our future computational development, we aim at increasing the computational speed of our method.

REFERENCES

[1] Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *J. Biomed. Opt.*, vol. 2, no. 4, pp. 364–374, 1997.

[2] H. Lyng, A. Badiee, D. H. Svendsrud, E. Hovig, O. Myklebost, and T. Stokke, "Profound influence of microarray scanner characteristics on gene expression ratios: Analysis and procedure for correction," *BMC Genomics*, vol. 3, no. 1, pp. 5–10, 2004.

[3] [Online]. Available: http://las.perkinelmer.com/Content/RelatedMaterials/Brochures/BRO_ScanArrayExpress.pdfURL:

[4] M. Katzer, F. Kummert, and G. Sagerer, "Methods for automatic microarray image segmentation," *IEEE Trans. NanoBiosci.*, vol. 2, no. 4, pp. 202–214, Dec. 2003.

[5] R. Lukac, K. N. Plataniotis, B. Smolka, and A. N. Venetsanopoulos, "A multichannel order-statistics technique for cDNA microarray image processing," *IEEE Trans. NanoBiosci.*, vol. 3, no. 4, pp. 202–214, Dec. 2004.

[6] R. Nagarajan and M. Upreti, "Correlation statistics for cDNA microarray image analysis," *IEEE Trans. NanoBiosci.*, vol. 3, no. 3, pp. 232–238, Sep. 2006.

[7] X. H. Wang, R. S. H. Istepanian, and Y. H. Song, "Microarray image enhancement by denoising using stationary wavelet transform," *IEEE Trans. NanoBiosci.*, vol. 2, no. 4, pp. 184–189, Dec. 2003.

[8] P. O'Neil, G. D. Magoulas, and X. Liu, "Improved processing of microarray data using image reconstruction techniques," *IEEE Trans. NanoBiosci.*, vol. 2, no. 4, pp. 176–183, Dec. 2003.

[9] R. Nagarajan and C. A. Peterson, "Identifying spots in microarray images," *IEEE Trans. NanoBiosci.*, vol. 1, no. 2, pp. 78–84, Jun. 2002.

[10] R. Gottardo, J. Besag, M. Stephens, and A. Murua, "Probabilistic segmentation and intensity estimation for microarray images," *Biostatistics*, vol. 7, no. 1, pp. 85–99, 2006.

[11] P. H. Davis, J. Schulze, and S. L. Stanley, Jr., "Transcriptomic comparison of two entamoeba histolytica strains with defined virulence phenotypes identifies new virulence factor candidates and key differences in the expression patterns of cysteine proteases, lectin light chains, and calmodulin," *J. Int. Parasitol.*, submitted to.

[12] Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed, "Comparison of methods for image analysis on cDNA microarray data," *J. Comput. Graph. Stat.*, vol. 11, pp. 108–136, 2002.

[13] Y. Balagurunathan, E. R. Dougherty, E. Chen, M. L. Bittner, and J. M. Trent, "Simulation of cDNA microarrays via a parameterized random signal model," *J. Biomed. Opt.*, vol. 7, no. 3, pp. 507–523, 2002.

[14] A. Dogandžić and B. Zhang, "Bayesian NDE defect signal analysis," *IEEE Trans. Signal Process.*, vol. 55, no. 1, pp. 372–378, Jan. 2007.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. New York: Chapman & Hall, 2004.

[16] J. Geweke, "Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints," in *Comput. Sci. Stat.: Proc. 23rd Symp. Interface*, Seattle, WA, Apr. 1991, pp. 571–578.

[17] R. M. Neal, "Slice sampling," *Ann. Statist.*, vol. 31, pp. 705–741, Jun. 2003.

[18] J. B. Elsner, X. Niu, and T. H. Jagger, "Detecting shifts in hurricane rates using a Markov chain monte carlo approach," *J. Climate*, vol. 17, no. 13, pp. 2652–2666, 2004.

[19] R. V. Hogg and A. T. Craig, *Introduction to Mathematical Statistics*, 4th ed. New York: Macmillan, 1998.

[20] R. C. MacFarlane, P. H. Shah, and U. Singh, "Transcriptional profiling of entamoeba histolytica trophozoites," *Int. J. Parasitol.*, vol. 35, no. 53, pp. 533–542, 2005.

[21] C. Weber, G. Guigon, C. Bouchier, L. Frangeul, S. Moreira, O. Sismeiro, C. Gouyette, D. Mirelman, J. Y. Coppee, and N. Guillen, "Stress by heat shock induces massive down regulation of genes and allows differential allelic expression of the Gal/GalNAc lectin in Entamoeba histolytica," *Eukaryot. Cell.*, vol. 5, no. 5, pp. 871–875, 2006.

[22] A. Urisman, K. F. Fischer, C. Y. Chiu, A. L. Kistler, S. Beck, D. Wang, and J. L. DeRisi, "A computational strategy for species identification based on observed DNA microarray hybridization patterns," *Genome Biol.*, vol. 6, no. 9, p. R78, 2005.

Pinaki Sarder (S'03) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 2003. He is currently a Ph.D. student and research assistant in the Department of Electrical and System Engineering at Washington University in St. Louis (WUSTL).

His research interests are mainly in statistical signal processing and biomedical imaging.

Mr. Sarder is a recipient of the Imaging Sciences Pathway program fellowship for graduate students at WUSTL starting from January 2007 to December 2007.

Arye Nehorai (S'80–M'83–SM'90–F'94) received the B.Sc. and M.Sc. degrees in electrical engineering from the Technion, Israel, and the Ph.D. degree in electrical engineering from Stanford University, Stanford CA.

From 1985 to 1995 he was a faculty member with the Department of Electrical Engineering at Yale University. In 1995 he joined as Full Professor the Department of Electrical Engineering and Computer Science at the University of Illinois, Chicago (UIC). From 2000 to 2001 he was Chair of the department's Electrical and Computer Engineering (ECE) Division, which then became a new department. In 2001 he was named University Scholar of the University of Illinois. In 2006 he became Chairman of the Department of Electrical and Systems Engineering at Washington University in St. Louis, St. Louis, MO (WUSTL). He is the inaugural holder of the Eugene and Martha Lohman Professorship and the Director of the Center for Sensor Signal and Information Processing (CSSIP) at WUSTL since 2006.

Dr. Nehorai was Editor-in-Chief of the IEEE TRANSACTIONS ON SIGNAL PROCESSING during 2000–2002. In 2003–2005 he was Vice President (Publications) of the IEEE Signal Processing Society, Chair of the Publications Board, member of the Board of Governors, and member of the Executive Committee of this Society. From 2003 to 2006 he was the founding editor of the special columns on Leadership Reflections in the *IEEE Signal Processing Magazine*. He was corecipient of the IEEE SPS 1989 Senior Award for Best Paper with P. Stoica, coauthor of the 2003 Young Author Best Paper Award, and corecipient of the 2004 Magazine Paper Award with A. Dogandzic. He was elected Distinguished Lecturer of the IEEE SPS for 2004–2005 and received the 2006 IEEE SPS Technical Achievement Award. He is the Principal Investigator of the new multidisciplinary university research initiative (MURI) project entitled Adaptive Waveform Diversity for Full Spectral Dominance. He has been a Fellow of the Royal Statistical Society since 1996.

Paul H. Davis received the B.S. degree in microbiology from Brigham Young University, Provo, UT, in 2001 and the Ph.D. degree in molecular microbiology from Washington University in St. Louis, St. Louis, MO, in 2006.

He is currently a Postdoctoral Fellow in the Department of Biology at the University of Pennsylvania, Philadelphia, PA. His major research interests include understanding human parasitic pathogenesis in the context of transcriptional and genotypic variation.

Samuel L. Stanley, Jr. received the B.A. degree in biology from the University of Chicago, Chicago, and the M.D. degree from the Harvard Medical School, Boston, MA.

He is currently the Vice Chancellor for Research with Washington University in St. Louis, MO, the Director of the Midwest Regional Center of Excellence in Biodefense and Emerging Infectious Diseases, Professor of Medicine and Molecular Microbiology, Washington University School of Medicine. His areas of interest include understanding the molecular basis for pathogenesis, genetic controls of virulence pathways, and markers for genetic susceptibility to disease.