Published in final edited form as:

IEEE Trans Nanobioscience. 2010 March ; 9(1): 51–58. doi:10.1109/TNB.2010.2040627.

Microarray analysis at single molecule resolution

Leila Mureşan,

Department of Knowledge-based Mathematical Systems, Johannes Kepler University, Linz, Austria

Jarosław Jacak,

Institute of Biophysics, Johannes Kepler University

Erich Peter Klement,

Department of Knowledge-based Mathematical Systems, Johannes Kepler University, Linz, Austria

Jan Hesse, and

Center for Biomedical Nanotechnology, Upper Austrian Research, Linz, Austria

Gerhard J. Schütz

Institute of Biophysics, Johannes Kepler University

Abstract

Bioanalytical chip-based assays have been enormously improved in sensitivity in the recent years; detection of trace amounts of substances down to the level of individual fluorescent molecules has become state of the art technology. The impact of such detection methods, however, has yet not fully been exploited, mainly due to a lack in appropriate mathematical tools for robust data analysis. One particular example relates to the analysis of microarray data. While classical microarray analysis works at resolutions of two to 20 micrometers and quantifies the abundance of target molecules by determining average pixel intensities, a novel high resolution approach [1] directly visualizes individual bound molecules as diffraction limited peaks. The now possible quantification via counting is less susceptible to labeling artifacts and background noise.

We have developed an approach for the analysis of high-resolution microarray images. It consists first of a single molecule detection step, based on undecimated wavelet transforms, and second, of a spot identification step via spatial statistics approach (corresponding to the segmentation step in the classical microarray analysis). The detection method was tested on simulated images with a concentration range of 0.001 to 0.5 molecules per square micron and signal-to-noise ratio (SNR) between 0.9 and 31.6. For SNR above 15 the false negatives relative error was below 15%. Separation of foreground/background proved reliable, in case foreground density exceeds background by a factor of 2. The method has also been applied to real data from high-resolution microarray measurements.

Keywords

Single molecule imaging; microarrays

COPYRIGHT ©2010 IEEE.

This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of UKPMC's products or services. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to pubspermissions@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

I. Introduction

Microarray technology is used in medical diagnostics and basic research for analyzing the global transcriptional state of biological samples. The massively parallel detection approach allows the determination of several thousand expression levels in a single experiment. Technologies for sample preparation like Fluorescence Activated Cell Sorting and Laser Capture Microdissection allow to isolate small subpopulations of cells and enable researchers to investigate heterogeneities within their samples. For the global expression analysis of such small samples standard low-resolution methods require time consuming and possibly distorting [2] pre-amplification steps. Recent developments in readout- [3] and platform/array-technology [1], [4], dramatically extended the range of directly accessible concentrations by increasing detection efficiency and the resolution to the optical diffraction limit.

In general, microarray technology is based on specific binding of fluorescent-tagged target molecules on different locations of the array and the subsequent determination of target molecule abundance by measuring fluorescence on the respective area. Classical methods use the pixel intensity values inside the pre-determined spot regions of the microarray scans, which is an indirect measure for the presence of hybridized molecules. The analysis is divided in the following tasks: identification of spot locations in the pattern, segmentation of sub-images containing single spots into signal and background pixels, construction of summaries of pixel intensity via appropriate statistics. Some further steps, typically background subtraction and normalization, are intended to remove all non-biological variation of the data. Several overviews of the classical microarray image analysis are available (e.g. [5]-[7]).

Conventional analysis is based on models of the microarray signal formation, like the ones proposed by [8]-[12]. These models include several aspects of the acquired data such as image intensity, spot shapes, noise. However, the identification of true signal and the control of the unspecific intensity variation is not a trivial task. Some of the main problems are related to the difficulty of signal detection, especially in the case of low signal-to-noise ratio (SNR) typical for small amounts of target molecules, correct background estimation and background subtraction, handling artifacts and accounting for the variability of the number of fluorophores per molecule (for example by dye swap). In this work we show that our approach brings considerable improvement in solving the aforementioned problems.

Figure 1 gives an insight into the difference between our single molecule and the classical microarray approach. It shows the images of four simulated spots at diffraction limited resolution (200nm, upper row) and the same images downsampled to the scale of the classical microarray techniques (4 μ m lower row). Due to the low SNR, background noise and the fluctuations of the single molecule signal intensities in the downsampled images the foreground/background contrast is low, making the segmentation impossible at low concentrations (leftmost images). However, the analysis is still possible at the original resolution. The single bright peaks, representing the molecules, can be well detected and their concentration inside the spot of interest can be estimated. Note that the rectangular sub-image, obtained after the gridding step, usually contains beside the spot location (foreground) a background region where the peaks correspond to unspecifically bound molecules or dirt, that should not influence the concentration estimation of the hybridized molecules.

The technology of high resolution microarrays offers direct access to the number of hybridized molecules. In the classical case this information is hidden in the low-resolution

pixel intensity value, that integrates background information, possible artifacts as well as the intensities of the hybridization signal inside the region corresponding to the respective pixel.

In this work we provide a method for the analysis of high resolution microarray images. The detection of single molecules is based on sparsity-adaptive wavelet thresholding, applied after a variance stabilization step. The estimation of the abundance of single molecules is performed on the detection results, and separates specific hybridization from clutter (unspecifically bound signal, impurities, etc.). We show that the approach outperforms classical counterparts for low concentration samples. The results are not hampered by the effect of the unknown background and, normalization and dye swap post-processing steps are not necessary.

The methods were validated by analyzing, on one hand, simulated data with known ground truth and, on the other hand, high-resolution microarray images.

II. Methods

We describe here an original framework to measure hybridization on high-resolution microarray data. The approximate positions of the microarray spots are found via classical gridding methods (see for example [7], [13]) on a low resolution (software-binned) image of the microarray pattern. The gridding procedure returns a rectangular sub-image corresponding to the approximate area containing the spot of interest. Such a sub-image is shown in fig. 2(a). We propose an approach that finds the position and shape of the spot as well as an estimate of the hybridization signal and of the background.

Our approach relies on two independent steps. First we present a wavelet based method to detect single molecules in each sub-image. Wavelet transform offers an attractive solution for the detection of small bright features, for e.g. in astronomical images [14] or in the case of microscopy, for detection of sub-cellular structures [15]. The detection is based on the property of the wavelet transform to concentrate the information in a few wavelet coefficients, and subsequently thresholding the pixels corresponding to signal from background.

Second, we separate the detected molecules inside the spot of interest (the hybridization signal) from the unspecifically bound ones. We use two concentration estimation approaches based on spatial statistics. The first algorithm matches the empirical moments with the moments of a mixture of two Poisson distributions representing counts of molecules outside and inside the spot. The second algorithm separates spot-bound single molecules from dirt based on nearest neighbour distances of all the detected peak locations, via an Expectation-Maximization (EM) approach. Since the surface was made anti-adsorptive for target molecules, we can assume that the concentration of peaks outside the spot is lower than the concentration of the hybridized molecules inside the spot ([1]).

Each step of the high resolution image analysis is illustrated on a real spot in fig. 2.

A. Detection of single molecules

1) Isotropic undecimated wavelet transform (IUWT)—The wavelet transform is based on dilations and translations of a “father” and “mother” wavelet: $\phi_{jk}(t) = 2^{-j/2} \phi(2^{-j}t - k)$ and $\psi_{jk}(t) = 2^{-j/2} \psi(2^{-j}t - k)$ for $j, k \in \mathbb{Z}$. The family $\{\phi_{jk}, k = 0, 1, \dots, 2^j - 1; \psi_{jk}, j = j, k = 0, 1, \dots, 2^j - 1\}$ forms an orthonormal basis of $L^2([0, 1])$. Any function $f(t) \in L^2([0, 1])$ can be arbitrarily well approximated by a wavelet series:

$$f(t) = \sum_k a_{jk} \phi_{jk}(t) + \sum_{j=-\infty}^J \sum_k d_{jk} \psi_{jk}(t), \text{ where}$$

$$a_{jk} = \langle f, \phi_{jk} \rangle, k=0, 1, \dots, 2^j - 1 \quad (1)$$

$$d_{jk} = \langle f, \psi_{jk} \rangle, \quad j \geq J, k=0, 1, \dots, 2^j - 1 \quad (2)$$

represent the approximations and detail coefficients, respectively.

The functions ϕ and ψ fulfill the dilation equations (see [16]):

$$\phi(x/2) = \sqrt{2} \sum_k h_k \phi(x-k), \quad \psi(y/2) = \sqrt{2} \sum_k g_k \phi(x-k) \quad (3)$$

with h_k a discrete low-pass and g_k a discrete band-pass filter (followed by downsampling), the approximation and detail coefficients can be computed as:

$$a_{j+1,k} = \sum_i h_{i-2k} a_{ji}, \quad d_{j+1,k} = \sum_i g_{i-2k} a_{ji}. \quad (4)$$

Note that the described wavelet transform is anisotropic 1D and not translation invariant. However, these two properties are essential to a good detection scheme. We propose therefore to consider the isotropic undecimated wavelet transform (IUWT). The “à trous” scheme is thus used [14] and wavelet coefficients are now computed over the entire grid as:

$$a_{jk} = \langle f, \phi_{jk} \rangle, \quad \phi_{jk} = 2^{-j/2} \phi(2^{-j}(t-k)) \quad (5)$$

$$d_{jk} = \langle f, \psi_{jk} \rangle, \quad \psi_{jk} = 2^{-j/2} \psi(2^{-j}(t-k)). \quad (6)$$

The recursive computation of the dyadic wavelet transform becomes:

$$a_{j+1,k} = a_{jk} * \tilde{h}_{jk}, \quad d_{j+1,k} = a_{jk} * \tilde{g}_{jk}. \quad (7)$$

where $\tilde{h}_{jk}$ ($\tilde{g}_{jk}$, respectively) is obtained by inserting $2^j - 1$ zeros between each sample of h_k (g_k).

In order to preserve isotropy the filters h and g and the father and mother function ϕ and ψ have to be nearly isotropic. A popular choice is based on the B_3 spline scaling function, $h_k = [1/16, 1/4, 3/8, 1/4, 1/16]$, and for the 2D case a separable filter $h_{(k,l)} = h_k h_l$ and $g_{(k,l)} = \delta_{k,l} - h_{(k,l)}$, where $\delta_{k,l} = 1$ if $(k, l) = (0, 0)$ and 0 otherwise.

The wavelet detail coefficients are given by: $d_{j+1,(k,l)} = a_{j,(k,l)} - a_{j+1,(k,l)}$ and the reconstruction is the sum of all details and the coarsest approximation:

$$f_{(k,l)} = a_{J,(k,l)} + \sum_{j=0}^J d_{j+1,(k,l)}. \quad (8)$$

When there is no confusion, a single index will be used to denote the 2D index (k, l) . The first index, j , denotes the scale. Further details on IUWT can be found in [17].

2) Thresholding based on False Discovery Rate (FDR)—The wavelet transform provides a sparse representation of signals as the number of significant coefficients is small.

The remaining coefficients of low amplitude can then be considered as noise and eliminated via thresholding. Hard thresholding of wavelet coefficients d_{jk} can be written as

$$d_{jk}^* = \begin{cases} \widehat{d}_{jk}, & |\widehat{d}_{jk}| \geq T \\ 0, & |\widehat{d}_{jk}| < T. \end{cases} \quad (9)$$

Since signals produce significant wavelet coefficients, correlated across wavelet planes, while noise is supposed to be uncorrelated. A pixel i is considered signal if all its corresponding wavelet coefficients d_{ji} , $j = 1, \dots, J$ are exceeding the threshold T , e.g. $T = c \cdot \widehat{\sigma}_j$, where $\widehat{\sigma}_k$ is the robust estimate of the noise variance at scale k and c is a constant. As estimate of the noise variance in [18], [19] is proposed:

$$\widehat{\sigma}_j = \text{median}(|d_{ji} - \text{median}(d_{ji})|) / 0.674 \text{ and } c \text{ is appropriately chosen, e.g. } c = 3.$$

The difficulty of the detection task lies in the fact that it has to be robust for a whole range of single molecule concentrations. The unknown concentration of single molecules in the image influences the sparsity of the signal, and implicitly the value of the parameter T that has to be chosen in order to obtain a correct detection. Therefore the detection method has to be driven by the (unknown) sparsity of the data. Some recent thresholding algorithms are sparsity adaptive as for instance Stein Unbiased Risk Estimator (SURE), FDR [20], [21] and empirical Bayes methods [22], [23].

The wavelet coefficient thresholding approach can be reformulated from a multiple hypothesis testing point of view. To each wavelet coefficient a ('no-signal') hypothesis H_{jk} : $d_{jk} = 0$ is attached. The hypotheses are rejected if d_{jk} corresponds to signal (and the corresponding wavelet coefficient kept in the expansion). Ideally, no coefficients corresponding to noise should be rejected.

The False Discovery Rate is defined as the expectation of the proportion of erroneously kept coefficients among all the coefficients kept in the representation. Applying the Benjamini-Hochberg method as described in [21] one maximizes the number of kept coefficients controlling meanwhile the FDR to a predefined level q . The algorithm consists of the following steps:

1. For each $\widehat{d}_{jk}$ calculate the two-sided p -value:

$$p_{jk} = 2 \left(1 - \Phi \left(|\widehat{d}_{jk}| / \sigma \right) \right)$$

2. Order ascendingly the computed p_{jk} -s, $p_{(1)} \quad p_{(2)} \quad \dots \quad p_{(m)}$
3. Find the largest i such that $p_{(i)} \leq (i/m)q$ and denote it i_0 . Compute $\lambda_{i_0}(j) = \sigma \Phi^{-1}(1 - p_{(i_0)}/2)$.
4. Threshold all coefficients at level $\lambda_{i_0}(j)$.

Again the significant pixels are those that have non-zero coefficients in all the J detail levels (except the finest, which usually is contaminated by noise). The binary image obtained from the J detail levels, $B = \prod_{j=1}^J \{d_{jk} > \lambda_{i_0}(j)\}$ is an indicator image for the support of the detected single molecules.

A denoised image is additionally obtained after applying the reconstruction step (equation 8) with thresholded detail coefficients. The wavelet detection algorithm has a certain "resolution", two molecules that are spatially close together will be detected as one. In order to correct for the estimation of the number of molecules, the binary image obtained after the

detection step is combined with the denoised image such that all the local maxima of the denoised image inside the support of the binary mask are considered distinct single molecules (see fig. 2, c).

3) Variance stabilization—Wavelet methods are typically designed for additive Gaussian noise: $X_i = \mu_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma)$. However low intensities (small photon counts) collected by the sensor are not well modeled by Gaussian noise. A combination of Poisson (shot-noise) and Gaussian noise is more appropriate to describe photon count variations and read-out noise. The main difference is the heteroskedasticity of the new model (the variance of the noise depends on the signal).

In order to take into account the characteristics of the noise, variance stabilizing transforms are applied prior to wavelet detection to the input image which transform the heteroskedastic noise into Gaussian noise of variance approximately equal to one. In case of a Poisson noise model (suitable to describe the photon count model) the well known Anscombe transform

can be used: $t(X_i) = 2\sqrt{X_i + 3/8}$ (it underestimates the intensity for values under 30).

Modeling both the photon count noise as well as the read-out noise, one obtains the mixed Poisson-Gaussian image model [14], $X_i = a \cdot N_i + \epsilon_i$, where $a > 0$ represents the gain of the detector, $N_i \sim P_{oi}(\mu_i)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma)$ that can be stabilized via the generalized Anscombe

transform (GAT): $t_G(x) = \frac{2}{\alpha} \sqrt{\alpha x + \frac{3}{8}\alpha^2 + \sigma^2} - \alpha\mu$. The parameters α, μ, σ are determined from the image itself via robust fitting as described in [24].

B. Foreground/background separation and estimation of single molecule concentration

Not all the peaks detected in the subimage belong to the spot of interest (see fig 2, c). The background might be heavily contaminated by unspecifically bound signal, impurities, etc. (clutter), which when unaccounted for could seriously distort the hybridization results. Therefore, peaks detected in the subimages have to be assigned either to foreground or background. In the concentration estimation step, we thus model both the spot and the background concentration.

In order to distinguish between peaks within the spot, representing true hybridization signal and those representing clutter a spatial mixture model is used. A similar approach used for segmentation of classical microarrays, but based on Gaussian mixture models for pixel intensity values, was described in [25], where the mixtures had two (signal/background) or three (signal/background/artifacts) components.

The peak locations obtained after the wavelet transform correspond either to peaks situated in the spot or to peaks in the background. Assuming that in case of strong hybridization there are much more peaks inside the spot, we shall discriminate between foreground and background via the concentration of the peaks in the two regions. The model we adopt is that of a mixture of two Poisson processes with piecewise constant intensities λ_1 and λ_2 (for foreground and background, respectively).

1) Estimation of concentrations based on method of moments (MOM)—A first approach for concentration estimation is to consider the count of the detected peaks inside non-overlapping, systematic quadrats, covering the subimage: y_i , where i is an index over the lattice structure. The counts are modeled as a mixture of two Poisson distributions, with constant concentrations λ_1, λ_2 (expressed in counts per quadrat):

$p(y_i | \lambda_1, \lambda_2, \eta_1) = \eta_1 \mathcal{P}(\lambda_1) + (1 - \eta_1) \mathcal{P}(\lambda_2)$, where η_1 denotes the weight of the first component. This simple model doesn't account for correlations between neighbouring quadrats.

The three parameters λ_1 , λ_2 and η_1 are determined via the method of moments for Poisson mixtures distribution discussed by Everitt, as described in [26].

Although inefficient compared to other estimators, this simple method offers a closed form solution in the case of mixture of two Poisson distributions, which is crucial for the speed of the analysis for such a large quantity of data.

Let Y_i be the random variables representing the counts in the quadrat i and y_i the measured value of these variables. The first three factorial moments $E(H_j(Y)|\lambda_1, \lambda_2, \eta_1)$ of the random

variable Y , $H_j(Y) = \frac{Y!}{(Y-j)!}$, $j = 1, 2, 3$ are matched with empirical moments obtained from y_i , $\bar{H}_j = \frac{1}{N} \sum_{i=1}^N H_j(y_i)$.

Since $E(H_j(Y)|\lambda_k) = \lambda_k$ the equation system for λ_1 , λ_2 and η_1 becomes:

$$\eta_1 (\lambda_1 - \lambda_2) + \lambda_2 = v_1$$

$$\eta_1 (\lambda_1^2 - \lambda_2^2) + \lambda_2^2 = v_2$$

$$\eta_1 (\lambda_1^3 - \lambda_2^3) + \lambda_2^3 = v_3$$

with $v_j = \frac{1}{N} \sum_{i: y_i \geq j} y_i (y_i - 1) \dots (y_i - (j - 1))$.

We tested several quadrat sizes, but the results obtained on real images were robust for quadrats of size 20×20 pixels and above. However further study is necessary to select the optimal quadrat size.

2) Concentration estimation based on Expectation-Maximization approach—

As a second approach we adopt the method of Byers and Raftery, used in minefield detection [27]. The location of the detected peaks are treated as a mixture of two spatial Poisson processes with different concentrations for foreground and background.

In the case of a single spatial Poisson processes with constant concentration λ the distribution D_k of the distance from a point of the Poisson process to its k -th nearest neighbor (kNN) can be written as

$$P(D_k \geq x) = \sum_{q=0}^{k-1} \frac{e^{-\lambda \pi x^2} (\lambda \pi x^2)^q}{q!} = 1 - F_{D_k}(x). \quad (10)$$

This leads to the density function:

$$f_{D_k}(x) = \frac{dF_{D_k}(x)}{dx} = \frac{e^{-\lambda \pi x^2} 2(\lambda \pi)^k x^{2k-1}}{(k-1)!}$$

meaning that the D_k^2 follows a transformed Gamma distribution, $D_k^2 \sim \Gamma(k, \lambda\pi)$. Here λ is measured in counts of single molecules per pixels squared.

The maximum likelihood estimate of the rate of the Poisson process is:

$$\widehat{\lambda} = \frac{k}{\pi \sum_{i=1}^n d_i^2} \quad (11)$$

where d_i , $i = 1, \dots, n$ are the realizations of the k -th nearest neighbour distances.

In the case of a mixture of two Poisson processes with two intensity rates λ_1 and λ_2 , the model for D_k can be written as: $D_k = D_k(\lambda_1, \lambda_1) \propto p \Gamma^{\frac{1}{2}}(k, \lambda_1 \pi) + (1 - p) \Gamma^{\frac{1}{2}}(k, \lambda_2 \pi)$

As opposed to the method of moments' η , p represents a proportion of the samples D_k . The three unknown parameters that describe the distribution D_k : p , λ_1 , λ_2 , are estimated via the Expectation Maximization (EM) algorithm, together with the assignments to components ("missing data") $\delta_i \in \{0, 1\}$, where $\delta_i = 1$ if the i -th point belongs to the first component (signal), and $\delta_i = 0$ otherwise.

The expectation step is:

$$E(\widehat{\delta}_i^{(t+1)}) = \frac{p^{(t)} f_{D_k}(d_i, \widehat{\lambda}_1^{(t)})}{p^{(t)} f_{D_k}(d_i, \widehat{\lambda}_1^{(t)}) + (1 - p^{(t)}) f_{D_k}(d_i, \widehat{\lambda}_2^{(t)})}$$

and the maximization:

$$\widehat{\lambda}_1^{(t+1)} = \frac{k \sum_{i=1}^n \widehat{\delta}_i^{(t+1)}}{\pi \sum_{i=1}^n d_i^2 \widehat{\delta}_i^{(t+1)}}$$

$$\widehat{\lambda}_2^{(t+1)} = \frac{k \sum_{i=1}^n (1 - \widehat{\delta}_i^{(t+1)})}{\pi \sum_{i=1}^n d_i^2 (1 - \widehat{\delta}_i^{(t+1)})}$$

$$p^{(t+1)} = \sum_{i=1}^n \frac{\widehat{\delta}_i^{(t+1)}}{n}.$$

As initial values for the three parameters one can use the results obtained through the method of moments.

III. Results and discussion

The performance of the high resolution microarray image analysis was extensively tested on simulations as well as real images.

A. Evaluation of the detection method

The detection algorithm was tested on a set of simulation images with varied image quality parameters (measured by the SNR), as well as several molecule concentrations. Each image is of dimension 512×512 pixels and contains $N = 10, 50, 100, 500$ or 1000 randomly located simulated single molecules.

To a single molecule corresponds a diffraction limited spot, approximated by a two-dimensional Gaussian shape, with width s corresponding to the point spread function of the optical system (1.1 pixels [3] in our simulations). Both the constant background intensity and the peak intensity S were chosen on a logarithmic scale between 10 and 100. Noise is generated for each pixel as described in II-A3: the photon count noise was modeled by draws from Poisson distributions, and finally Gaussian noise was added to each pixel from $\mathcal{N}(0, \sigma)$, where σ was chosen as 0, 5, 10, 15 and 20% of the maximum peak intensity. For this special case of Poisson-Gaussian model described, we used the following dimensionless signal-to-noise (SNR) definition as in [28]: $SNR = S / \sqrt{B + \sigma^2}$, where S represents the maximum intensity of the single molecule profile, B the (local) background of the image and σ the standard deviation of the read-out noise.

The SNR for our simulations was between 0.9 and 31.6. For each set of parameters 10 images were generated and analyzed. The results are summarized in figure 3 (*a* and *b*). The fraction of correctly detected peaks over the number of simulated peaks gives the ratio of true positives, while the number of detected peaks that do not correspond to any simulated one represent the false positives of the method. For SNR above 10 (our fluorescence reader is at least 10, but more typically exceeds 15 [1]) the detection rate is higher than 80% true positives for less than 500 simulated molecules per image, and decreases to $\sim 60\%$ for high concentrations ($N = 1000$) due to overlapping peaks. Note that at this concentration regime one can revert to classical ensemble analysis.

The ratio of false positives is below 5% for very low concentrations ($N = 10$) and substantially lower for $N > 100$. The detection performance was similar for simulation with the same SNR using different weights of Poisson and Gaussian noise in the generation of the simulated image.

B. Evaluation of concentration estimation

The concentration estimation algorithms were tested on data representing the position of single molecules and clutter, respectively. We assumed that signal (molecules' position) has a higher concentration than clutter. For each data set, two spatial Poisson processes are simulated: one of intensity λ_1 inside a disk of radius R (150 pixels in our case) and a second one, of intensity λ_2 , independent of the first one, in a rectangle excluding this area. The parameter values were chosen such that $(\lambda_1, \lambda_2) \in [0.01, 0.05] \times [0.005, 0.05]$, $\lambda_2 < \lambda_1$. For each (λ_1, λ_2) parameter pair, ten data sets were generated. The results of the estimation of the signal concentration λ_1 are presented in figure 4. The results in the case of MOM estimation are biased to lower intensities λ_1 than the true value. When the two concentrations are close together one can see a stronger bias, due to the failure of separating the two components of the mixture. In the case of EM estimation there is almost no bias. The true value belongs to the estimated confidence interval. Although more computationally expensive, the increased accuracy of the EM method makes it preferable to the MOM. The method is insensitive to the shape that has to be detected, allowing for the analysis of anomalous spots due to uneven drying, spotting errors etc. An example of unusual shape that could be analyzed is shown in fig. 5.

C. Correlation tests

Finally we have compared the results of our analysis to those obtained by classical low-resolution ensemble analysis on the downsampled images of the same data. Since our method results in peaks/pixel concentrations, while the ensemble approach gives mean pixel intensity values we compare the correlation coefficients of the estimated hybridization measures computed by each method with the ground truth concentrations used in the simulations.

For this purpose, 60 sets of images were generated with SNR between 2.85 and 31.6. Each set of image, characterized by a SNR value, contains 15 images. In each image, single molecules were simulated with concentrations $\lambda_1 \in \{0.003, 0.005, 0.007, 0.009, 0.01\}$ peaks/pixels inside a disk of radius 150 pixels. The concentration λ_2 of peaks corresponding to clutter outside this disk was also varied from 0.001 to $\lambda_1 - 0.002$ by a step of 0.002.

In fig. 1 are shown the images corresponding to background value $\lambda = 0.003$, $SNR = 12.52$. The single molecule simulation procedure is the same as the one described in section III-A.

We have created the corresponding low resolution images, by integrating the intensities of the high resolution images over 20×20 pixel patches. The pixel intensities of the downsampled images were modeled as a Gaussian mixture of foreground (signal) and background pixels respectively and the parameters of the mixture were estimated via a maximum likelihood (ML) approach. Moreover, we have applied the same downsampling and ML estimating procedure to the original high-resolution image after a wavelet denoising step. Thus we eliminated the effect of the background on the hybridization measure.

We have compared our analysis also with two state-of-the-art microarray spot segmentation and intensity estimation methods. One segmentation method is based on the Mann-Whitney rank-sum statistic (M-W) and is described in [29]. The second one uses a hierarchical Bayes approach to model the microarray spot parameters (spot shape and position, and signal intensity parameters) and uses a Markov Chain Monte Carlo (MCMC) algorithm to sample these parameters. The parameter estimates are based on the samples obtained via Gibbs sampling (see [13] for details).

The results are visualized in fig. 6. Each point in fig. 6 corresponds to a correlation coefficient computed between the hybridization measure (the estimated λ_1 values or mean foreground pixel intensities, respectively corresponding to one of the five methods enumerated above) and the true λ_1 values used in simulations. For the whole range of different SNR corresponding to the test data, our approach (with EM concentration estimation) has higher correlation coefficients than any of the alternative four algorithms.

The lowest correlation value for single molecule analysis, 0.77, was obtained at $SNR = 3.99$. Since some of the low-resolution spot analysis have failed for low SNR image data, we show the results for the data sets with $SNR \geq 5$. We mention that the analysis of a low-resolution microarray spot via the MCMC segmentation method takes approximately 10 minutes [13] for low SNR, this time complexity being a prohibitive factor for use on chips with several thousand spots. Our approach takes approximately 90 seconds/spot, and significantly less for low concentration spots. The mean of the correlation coefficients for these datasets are as follows:

- Our approach: 0.969
- Low resolution MLE: 0.858
- MLE on wavelet denoised images: 0.809. However on datasets with $SNR > 16$ the method outperforms the low resolution MLE.

- MCMC segmentation: 0.809
- M-W segmentation: 0.709.

IV. Conclusion

In this paper we have presented the analysis of microarray images at single molecule resolution. The concentration of single molecules as a new measure of hybridization offers attractive advantages in the case of low abundance of target molecules, for background suppression, photon count fluctuation etc. We have shown that the single molecule detection algorithm performs well across a wide range of concentrations and image SNRs. The separation of the specifically bound molecules from clutter was also thoroughly tested. Furthermore we have shown that our approach provides good correlation results for concentrations and SNR values where the low resolution methods fail. The algorithms presented in the paper provide validated tools for other techniques based on the observation of individual fluorescent molecules.

Acknowledgments

This work was supported by the GEN-AU program of the Austrian Federal Ministry of Education, Science and Culture, by the Austrian Research Fund (FWF Project L422-N20), the European Commission (FP6 Project Autoscreen) and the state of Upper Austria.

References

- [1]. Hesse J, Jacak J, Kasper M, Regl G, Eichberger T, Winklmayr M, Aberger F, Sonnleitner M, Schlapak R, Howorka S, Muresan L, Frischauf A, Schütz GJ. RNA expression profiling at the single molecule level. *Genome Research*. 2006; 16:1041–1045. [PubMed: 16809670]
- [2]. Nygaard V, Holden M, Løland A, Langaas M, Myklebost O, Hovig E. Limitations of mRNA amplification from small-size cell samples. *BMC Genomics*. 2005; 6:147. [PubMed: 16253144]
- [3]. Hesse J, Sonnleitner M, Sonnleitner A, Freudenthaler G, Jacak J, Höglinger O, Schindler H, Schütz G. Single-molecule reader for high-throughput bioanalysis. *Anal. Chem*. 2004; 76:5960–5964. [PubMed: 15456321]
- [4]. Sonnleitner, M.; Freudenthaler, G.; Hesse, J.; Schütz, GJ. High-throughput scanning with single-molecule sensitivity; *SPIE Proceedings*; 2005; p. 202-210.
- [5]. Bajcsy P. An overview of DNA microarray grid alignment and foreground separation approaches. *EURASIP Journ. on Applied Signal Processing*. 2006; 2006:1–13.
- [6]. Yang YH, Buckley MJ, Speed T. Analysis of cDNA microarray images. *Briefings in Bioinformatics*. 2001; 2:341–349. [PubMed: 11808746]
- [7]. Yang YH, Buckley J, Dudoit S, Speed TP. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*. 2002; 11:108–136.
- [8]. Balagurunathan Y, Wang N, Dougherty ER, Nguyen D, Chen Y, Bittner M, Tent J, Carroll R. Noise factor analysis for cDNA microarrays. *Journal of Biomedical Optics*. 2004; 9(4):663–678. [PubMed: 15250753]
- [9]. Angulo J. Polar modelling and segmentation of genomic microarray spots using mathematical morphology. *Image Anal. Stereol*. 2008; 27:107–124.
- [10]. Chudin E, Kruglyak S, Baker S, Oeser S, Barker D, McDaniel T. A model of technical variation of microarray signals. *Journal of Computational Biology*. 2006; 13(4):996–1003. [PubMed: 16761924]
- [11]. Korn E, Habermann J, Upender M, Ried T, McShane L. Objective method of comparing DNA microarray image analysis systems. *Bioimaging*. 2004; 36(6):960–967.
- [12]. Li Q, Fraley C, Bumgarner R, Yeung K, Raftery A. Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics*. 2005; 21(12):2875–2882. [PubMed: 15845656]

- [13]. Sarder P, Nehorai A, Davis PH, Stanley SLJ. Estimating gene signals from noisy microarray images. *IEEE Transactions on Nanobioscience*. 2008; 7(2)
- [14]. Starck, J-L.; Murtagh, F.; Bijaoui, A. *Image and Data Analysis: The Multiscale Approach*. Cambridge University Press; 1998.
- [15]. Olivo-Marin J-C. Extraction of spots in biological images using multiscale products. *Pattern Recognition*. 2002; 35:1989–1996.
- [16]. Mallat, S. *A wavelet tour of signal processing*. Academic Press; 1999.
- [17]. Starck JL, Fadili J, Murtagh F. The undecimated wavelet decomposition and its reconstruction. *IEEE Trans. on Image Processing*. 2007; 16(2)
- [18]. Donoho D, Johnstone I. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*. 1994; 81:425–455.
- [19]. Donoho D. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*. 1995; 41(3):613–627.
- [20]. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*. 1995; 57(1):289–300.
- [21]. Abramovich F, Benjamini Y. Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis*. 1996; 22:351–361.
- [22]. Johnstone I, Silverman B. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*. 2004; 32(4):1594–1649.
- [23]. Johnstone I, Silverman B. Empirical Bayes selection of wavelet thresholds. *The Annals of Statistics*. 2005; 33(4):1700–1752.
- [24]. Boulanger, J.; Sibarita, J-B.; Kervrann, C.; Bouthemy, P. Non-parametric regression for patch-based fluorescence microscopy image sequence denoising; *Proc. IEEE Int. Symp. on Biomedical Imaging: from nano to macro (ISBI)*; 2008;
- [25]. Blekas K, Galatsanos NP, Likas A, Lagaris IE. Mixture model analysis of DNA microarray images. *IEEE Transactions on Medical Imaging*. 2005; 24:901–909. [PubMed: 16011320]
- [26]. Frühwirth-Schnatter, S. *Finite Mixture and Markov Switching Models*. Springer; New York: 2006.
- [27]. Byers S, Raftery AE. Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association*. 1998; 93:577–584.
- [28]. Murphy, DB. *Fundamentals of light microscopy and electronic imaging*. Wiley-LISS; 2001.
- [29]. Chen Y, Dougherty E, Bittner M. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*. 1997; 2(4):364–374. [PubMed: 23014960]

Fig. 1. Simulation of microarrays spots. Upper row: spots at single molecule resolution (200nm pixel size) with different peak concentrations (λ) inside each spot. From left to right: $\lambda = 0.005, 0.007, 0.009, 0.011$ peaks per pixel. (Background concentration representing dirt, unspecific binding etc.: 0.003 peaks per pixel). Lower row: the same spots downsampled to $4\mu m$, the size used by existing commercial microarray systems.

Fig. 2. Analysis of a spot in a high-resolution microarray image. (a) Original image, bright features correspond to molecules bound to the chip. (b) Detection after undecimated wavelet thresholding. (c) Selection of single molecule locations (local maxima on denoised image inside the detection support in (b)), (d) Separation of hybridization signal from clutter.

Fig. 3. Results of detection on simulations. Left: ratio of true positives, right: ratio of false positives with respect to the true number of simulated single molecules. The size of simulated images is 512×512 pixels. The typical SNR is higher than 10, indicated in the figures by a vertical line.

Fig. 4.

Background/foreground separation for three different concentrations via Method of Moments (MOM) and Expectation Maximization (EM) methods (see II-B). The true λ values are represented as a stair-case function and for better visibility, the estimation results were slightly shifted on the abscissa.

Fig. 5. Anomalous shape detection. A high concentration region in donut shape was simulated on a background formed of low concentration clutter. The proposed approach is able to separate signal from clutter.

Fig. 6. Correlations between the estimated and the true signal concentrations. The single molecule analysis performs better than the analysis on the downsampled data (original and denoised via wavelet thresholding). The arrow indicates the correlation coefficient corresponding to the images in figure 1.