# Mathematical Morphology applied to Spot Segmentation and Quantification of Gene Microarray Images

Kashif I.Siddiqui*, Alfred O. Hero*†#, and Matheen M. Siddiqui‡
Dept. of Electrical Engineering and Computer Science*,
Dept. of Biomedical Engineering†, and Dept. of Statistics#
The University of Michigan Ann Arbor, MI 48109
Dept. of Computer Science‡, Boston University, Boston, MA 02215

*Abstract*—**DNA microarray technology is a very powerful technique used in modern biology which is extensively used for identification of sequence (gene/gene mutation) and determination of gene expression. A typical gene microarray image consists of a few hundred to several thousand spots and the extent of hybridization of these spots determines the level of gene expression (abundance) in the sample. The massive scale and variability of gene microarray data creates new challenging problems of gene clustering, feature extraction and data mining. A major issue in gene microarray data analysis is to accurately quantify spot shapes and intensities of microarray image. In this paper we propose a method for performing accurate spot segmentation of a microarray image, using morphological image analysis techniques, followed by quantification of the shapes of the segmented spots using B-Splines.**

*Keywords*—**Mathematical Morphology, Gene Microarray Images, Spot-Segmentation, Spot-Quantification, Watershed Transform, Feature Extraction, Bio-Informatics**

## I. Introduction

Gene microarrays, or chips, have revolutionized the field of experimental genetics because they permit estimation of the relative expression levels of thousands of genes simultaneously. Typically, a gene microarray consists of large number of known DNA probe sequences that are placed on distinct locations on a slide. The level of hybridization of an unknown target to probe gives estimate of the abundance of the probes in the unknown target [1].

In spotted arrays two mRNA (messenger RNA) samples, namely the control sample and the treatment sample, are reverse transcribed into cDNA (complementary DNA) samples and then tagged with two different dyes. Then these two samples are mixed and scanned to produce a spotted image depicting the variations in fluorescent intensities at each probe position. A sample microarray image is shown in Fig.1. The integrated intensity within each spot is a measure of the level of gene expression or equivalently the mRNA abundance in the sample [1]. As gene microarrays can suffer from a high-level background noise level, accurate spot segmentation is essential for quantifying this intensity. Mathematical morphology methods can be employed for lossless spot segmentation and to quantify spot shape variability. The main focus of this paper is the extraction of spot features from a gene microarray image, which along with the spot intensity can be used for statistical analysis of spot shape and intensity variations. For this purpose we have to segment the microarray image, which is just an intensity image and can be characterized by connected pixels of similar intensity values. For the segmentation of intensity images there are four main approaches: 1) Threshold techniques, which are based on the principle that all pixels whose value lie within a certain range belong to a specific class of hybrid levels, 2) Boundary-based techniques [2], which keep track of rapidly changing pixel values at the boundary between two regions, 3) Region-based methods [3], which are based on comparison of one pixel with its neighbor and if they all have similar values they are said to belong to the same class (an important special case is the algorithm called Seeded Region Growing (SRG) and 4) Hybrid techniques, which are a combination of boundary and region-based methods, and are very reliable in producing closed boundaries [13] (Morphological Watershed Segmentation belongs to this class). There is a variety of software available to perform segmentation of spotted microarray images e.g. Spot [15], which uses Seeded Region Growing. Here we propose Morphological Watershed Segmentation which has the advantages: 1) no seeding within spot boundaries is necessary, 2) the watershed region provide partition which is used to isolate local noise background for each spots, and 3)the implementation is fully automatic, elimination need for gridding or other manual pre-processing. The problem at hand can be viewed as reducing a database of genes $\mathcal{X} = \{X_i\}_{i=1}^{K}$ to $\mathcal{X} = \{X_i\}_{i=1}^{L}$ where $L \leq K$ are spots with strong responses and then applying further gene filtering techniques like Posterior Pareto Front Analysis [5] to find the most prominent genes.

The outline of the paper is as follows. In Sec. II we provide a brief overview of mathematical morphology, in Sec.III we provide experimental results of application of morphological techniques for microarray segmentation and finally in Sec. IV we describe applications to spot shape and intensity characterization over a population from microarray experiments.

[1]Email contact: {kisrar*,hero†}@eecs.umich.edu. The data for this work was provided by the Sensarray Microarray Node, Anand Swaroop (Director), The University of Michigan, Ann Arbor, MI, 48109.

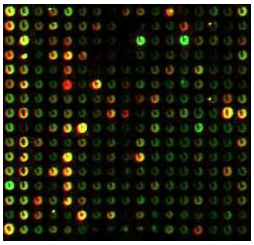

Fig. 1. *A sample gene microarray image*

## II. Morphological Methods For Spot Segmentation

In this section we briefly present the basic principles, definitions and notations used in mathematical morphology, for further details see [7], [11]. Let $f$ be function which defines a grayscale image defined on $\Re^2 \rightarrow \Re$, $B$ be a planar structuring element such that $B \subseteq \Re^2$ and $\psi$ be an image operator which transforms a grayscale image $f$ into another image according to some specific task .

An important morphological operator erosion (dilation) is defined as,

$$\psi_{erosion}(f) = \bigwedge_{(\varepsilon,\eta)\epsilon B} f(x+\varepsilon, y+\eta) = (f \ominus B) \tag{1}$$

$$\psi_{dilation}(f) = \bigvee_{(\varepsilon,\eta)\epsilon B} f(x-\varepsilon, y-\eta) = (f \oplus B) \tag{2}$$

Erosion (dilation) replaces the value of the image $f$ at a pixel $(x, y)$ by the infimum (supremum) of the values of $f$ over a structuring element $B$ ($B'$ - reflection of $B$ around the origin), which results in "shrinkage" ("expansion") of the image. Another very important morphological operation which is used extensively is called structural opening (structural closing) and is defined as,

$$f \circ b = (f \ominus b) \oplus b \tag{3}$$

$$f \bullet b = (f \oplus b) \ominus b \tag{4}$$

and is used to undo the effect of erosion (dilation) by applying the associated dilation (erosion).

Another useful operator is the so-called area opening, which is used to remove grains having area below a given value from the image. Mathematically it is defined as,

$$\psi_{aopn}(f, a) = \bigwedge_{t\epsilon\Re} \{(x, y)\epsilon \bigcup \{F_s(t) : |F_s(t)| \geq a\}\} \tag{5}$$

where, $F_s(t)$ is a cross-section of the image intensity $f$ and $F_s(t)|s = 1, 2, 3$ are grains of the cross-section $F(t)$ of the image $f$and $a$ is the threshold level [12]. On the other hand area closing is used to fill in the holes in the image, whose area is smaller than a given value. It is important to note that openings (closings) are increasing, anti-extensive (extensive) and idempotent. They both are smoothing filters and are used for smoothing contours of an image, suppressing small islands and cutting narrow isthmuses. The amount of smoothening is determined by the size and shape of the structuring element used. Note that supremum of openings is also an opening and infimum of closings is also a closing. This becomes very useful in practice since it allows us to develop larger openings (closings) using elementary openings (closings).

The opening (closing) of an image $f$ removes peaks (hollows) and ridges (ravines) from the topographic surface of the image $f$. Another operator, which produces such peaks (hollows) and ridges (ravines), called the opening top-hat operator (closing top-hat operator) is defined as,

$$\psi_{opnth}(f) = f - f \circ b \tag{6}$$

$$\psi_{closeth}(f) = f - f \bullet b \tag{7}$$

A morphological operator $\psi$ is said to be a morphological filter, if it is increasing and idempotent. The combination of different morphological filters also results in a morphological filter. Alternating filters are combination of closings and openings and are defined as,

$$\pi_k(f) = (f \circ kB) \bullet kB \tag{8}$$

$$\rho_k(f) = (f \bullet kB) \circ kB \tag{9}$$

where $kB$ represents $(k-1)$ dilations and $k$ is the size of the filter.

We can combine alternating filters to form an alternating sequential filter (ASF). This is combination of multiple closings and opening with decreasing number of dilations and it is given by,

$$\mu_k(f) = \pi_k \pi_{k-1} \ldots \pi_1(f) \tag{10}$$

$$\upsilon_k(f) = \rho_k \rho_{k-1} \ldots \rho_1(f) \tag{11}$$

These filters can be used to reduce noise or simplify variations in gene microarray images.

The distance function $d(\bullet, \bullet)$ is a map from $\Re^2$x$\Re^2$ into the set $\Re^+$. If $d(q, r)$ is defined as the distance between $q$ and $r$, then the distance transform has the properties that $d(u, u) = d(v, v) = 0$, $d(u, v) = d(v, u)$ and $d(u, w) \leq d(u, v) + d(v, w)$ for every $u, v, w \epsilon \Re^2$. Using the distance function we can define the distance transform $D_u(f)$ of $f$ at point $u \epsilon \Re^2$ as,

$$D_u(f) = \bigwedge_{v\epsilon\Re^2} d(u, v) \tag{12}$$

The non-zero values of $D_u(f)$ (distance transform of the foreground) gives the minimum distance of a pixel in background from the foreground boundary, while the non-zero values of $D_u^{'}(f)$ (distance transform of the background) gives the minimum distance of a pixel in foreground from the foreground boundary. This transform aggregates the distance information from a continuum of erosions and dilations into a single grayscale function.

A regional minimum (regional maximum), $Min_{reg}$ ($Max_{reg}$), of an image $f$ is a connected component of

pixels in $f$ with a given value $a$, such that every pixel in the neighborhood of $Min_{reg}$ ($Max_{reg}$) has a value strictly larger (smaller) than $a$. Every regional minimum $Min_{reg}$ has a catchment basin $C(Min_{reg})$ associated with it, which is a collection of all points of the topographic surface of $f$, such that a drop of water falling at any point slides along the surface until it reaches $Min_{reg}$ [10], [13].

Now, by flooding the topographic surface of an image from its regional minimum and preventing the merging of water coming from different sources, we partition the image into two different sets; the catchment basins and the watershed lines, where each catchment basin contains one and only one regional minimum [12]. Using the above analogy we can define the watershed transform as,

$$W(f) = D \bigcap (\bigcup_{s \epsilon \Re} C(Min_{reg_s}))^{'} \tag{13}$$

where, $D$ represents connected domain of the image $f$, see [13] for further details. In the next section we apply these morphological techniques to spot segmentation in gene microarray images.

## III. Spot Segmentation Of Gene Microarray Image

Image Segmentation is defined as the process of isolating objects in the image from the background i.e., partitioning the image into disjointed regions, such that each region is homogeneous with respect to some property [8]. Therefore, spot segmentation can be defined as the process of extracting the appropriate homogenous spots and the noise background, having the desired homogeneity property, from a microarray image. Estimation of noise background is important since it allows for the correction of the spot intensities.

In this section we apply morphological techniques, discussed briefly above, for spot segmentation of a gene microarray image. A portion of the original image's grayscale version is shown in Fig. 2. It can be seen, that there are bright regions inside the spots, which will cause faulty binarization of the image,but applying an area opening (5) solves this problem and the result is a much smoother image, which is depicted in Fig. 3. Thresholding the image in Fig. 3 produces the binary image shown in Fig. 4. The number of spots produced during thresholding is determined by the threshold level we select and thus can be used to filter those spots with weak hybridization levels and can also be used for multi-threshold extraction of spots of varying intensity levels.

Two iterations of the alternating sequential filter (ASF), characterized by a cross structuring element [12] having unit radius, are applied to the image in Fig. 4 while using the sequence of opening followed by closing operators (8),(9),(10),(11). Next we find the regional maxima of the image in Fig. 4, according to the connectivity defined by cross structuring element. These regional maximums act as markers for each cell, and can be seen as dark regions within spots in Fig. 5.

Now we apply the watershed transform (12) to the negation of the original image using the markers found previously and using the box-structuring element to define connectivity. These watershed lines are used to act as external markers, which mark the crest lines of the original image Fig. 1. Further we locate the regional minima of the original image and use them as internal markers. These external markers and internal markers are combined in to a joint marker, which is shown in Fig. 5 overlaid over the original image and it can be seen that the spot boundaries are well constrained between external and internal markers.

The watershed transform is applied to the gradient of the image [4] using the combined marker (Fig. 5) and the cross structuring element of unit radius. The resulting watershed lines are shown as green boundaries around spots and are overlaid over the original image in Fig. 6. Generally the gradient operator is overly sensitive to grayscale variation and noise and it can cause creation of a large number of irrelevant catchment basins, a problem called oversegmentaion. However, by using watershed transform techniques we can avoid overgsegmentation problems as seen in the final result in Fig. 6, at low computational cost. Another advantage of watershed segmentation is that we extract and characterize noise background features since the watershed provides regions in the neighborhood of each spot. Now by using boundaries of the extracted spots we can find foreground intensities for each spot and use them for statistical analysis. The plot of Cy5 vs. Cy3 intensities of the extracted spots is shown in Fig. 87.

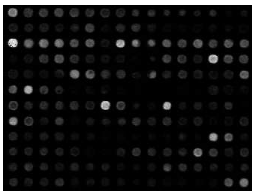

Fig. 2. *Grayscale version of original microarray image*

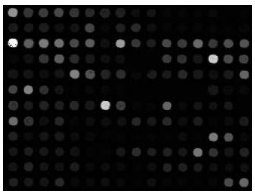

Fig. 3. *Result after application of area opening*

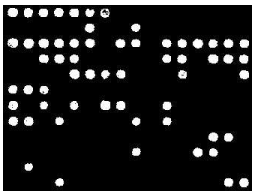

Fig. 4. *Result after application of thresholding*

## IV. Quantification Of Spot Shapes

Our accurate spot and noise segmentation permits quantification of spot shape and other characteristics e.g. noise,

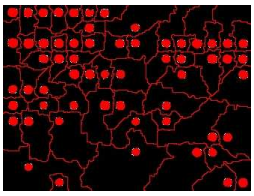

Fig. 5. *Combined internal and external marker*

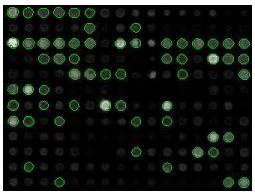

Fig. 6. *Watershed lines overlaid over the original image*

background averaging and subtraction. Here we illustrate the utility of segmentation for shape quantification. After segmentation of the image, the number $L$ of surviving spots in the image is determined (in our example $L$=60) and the co-ordinates of the centroid of each spot are calculated. Using the centroid of $i$-th spot, where $i$ = 1,2, ..., $L$ we calculate sample values of the boundary, which results in an $LxR$ matrix for the $x$ and $y$ coordinates of sample points of each spot boundary, where $R$ = the number of sample points along the boundary. Using the matrix of Cartesian coordinates of sample points of the boundary, we transform the boundary to polar coordinates with respect to the centroid of the $i$-th spot.

## A. B-Splines for Extracted Spots

To achieve a low dimensional parameterization of the spot shape we investigated planar curve model fitting based on the morphological segmentation described above. There are numerous methods available to represent closed boundaries as periodic planar curves, such as Fourier descriptors, Bezier curves, Beta-Splines and B-Splines. In this paper, we adopt a B-spline boundary model [14]. A B-spline consists of a set of spline coefficients and basis functions. Each spline coefficient is associated with one basis function. For a fixed centroid, each spot's boundary can be represented by a radial function $r(\theta)$, continuously indexed by polar angle $\theta \in (-\pi, \pi]$. This is given as,

$$r(\theta) = \sum_k B_k(\theta) c_k = \mathbf{B}^T(\theta)\mathbf{C} \qquad (14)$$

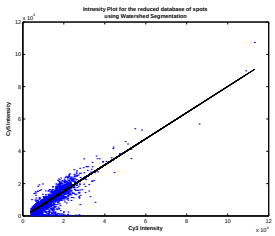


Fig. 7. *Scatter plot for foreground intensities of extracted spots using morphological watershed segmentation*

where $c_k = [x_k, y_k]^T$ is a 2D spline coefficient, $B_k(t)$ are the associated basis functions and $\mathbf{B}(\theta) = [B_1(\theta), ..., B_k(\theta)]^T$. From this we see spline boundary is a weighted combination of the spline coefficients, where the weights are given by the basis functions. Furthermore the basis functions are piecewise polynomial curves determined by $J$ fixed positions, called knots. For an m-th order B-spline these curves are specified by polynomial functions of degree m. Making use of the previously computed centroid matrix, which contains the centroid of each spot and locations of sample points on the detected spot boundary, we perform the B-spline fitting procedure to smoothly parameterize the spot shapes.

In particular, a B-spline is fit to the sample points by first sorting points by their angle about the centroid. Following this each sample point, $r_l$, is associated with a curve parameter $t_l$, which is computed using a measure of arc length: the sum of distances $|r_j - r_{j-1}|$ for $j = 2l$. Given $r_l$, and $t_l$, we can then find the control points $c_k$ that best fits the data while forming a closed loop via linear least squares[15]. Following this we can alternate between updating the $t_l$, and the control points so that the fitting error of the spline to the data is minimized. A typical result of our application of B-Splines for one of the $L$ extracted spots in the gene microarray image is shown in Fig. 8. Using this boundary model we can easily compute moments of the spot boundary, mean and standard deviation of the spot radii etc.

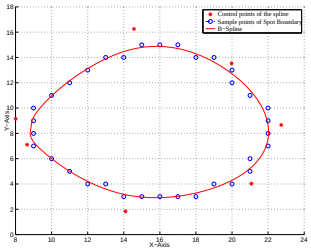


Fig. 8. *Estimated B-Spline around the boundary of one of the extracted spots*

## B. Circularity Coefficients

After determining the coefficient vectors of the B-splines, we developed a database of spot shapes in the microarray image. This database can be used to query for a possible correlation of spot shape to factors such as intensity, background noise and microarray print-head variation. We also constructed a database $\Gamma$ of shape statistics including the spot circularity coefficient $\tau_i$. The circularity coefficient is defined as the ratio of first moment squared of the splined boundary $r(\theta)$ to second moment of the splined boundary, which can be computed directly from spline coefficients,

$$\tau_i = \frac{(\int \theta r(\theta) d\theta)^2}{\int \theta^2 r(\theta) d\theta} \frac{1}{\int r(\theta) d\theta} = \frac{\left(\sum_k c_k \overline{B_k^1}\right)^2}{\left(\sum_k c_k \overline{B_k^2}\right)\left(\sum_k c_k \overline{B_k^0}\right)} \qquad (15)$$

where $\overline{B_k^m} = \int \theta^m B_k(\theta)$. This shape statistic tells how close the spot's shape is to a perfect circle. Note that the range of this coefficient is $0 \leq \tau_i \leq 1$, where $\tau_i = 1$ for a circular spot boundary. Since the surface of the microarray print head is disk

shaped the closer the $\tau_i$ for $i$-th spot is to 1, the higher the confidence in the accuracy of measured probe response. Circularity coefficients for a few spots are shown in Fig. 9, where columns of the table represent the column indexes and rows represent the row indexes of the microarray grid, and blank cells correspond to spot locations which were were not detected as spots of interest during the analysis. Note that upper left part of table in Fig. 9 corresponds to circularity coefficients of spots in Region A of the microarray image in the Fig. 10 and similarly upper right part,lower left and lower right part of the table corresponds to spots in Region B, Region C and Region D respectively. Circularity coefficients for each spot are analyzed with other spot characteristics, and one of these analysis is shown in Fig. 11, which depicts that majority of extracted spots have high circularity coefficient, with the exception of those which have very small radii or those with very large radii.

TABLE I

| | 1413 | 1414 | 1415 | ··· | 1426 | 1427 | 1428 |
|---|---|---|---|---|---|---|---|
| 2116 | 0.89 | 0.95 | 0.90 | ··· | | | |
| 2117 | | | | ··· | | | |
| 2118 | 0.93 | 0.95 | 0.84 | ··· | 0.88 | 0.89 | 0.91 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 2126 | | 0.89 | | ··· | 0.93 | | |
| 2127 | | | | ··· | | 0.93 | 0.98 |

Fig. 9. *Table for Circularity Coefficients, $\tau_i$, for spots extracted from gene microarray image at mn-th location on the microarray grid*

### C. Eigen Analysis of Extracted Spots

Using the shape parameters given by the B-spline, the statistics of spot shape can be computed and analyzed as a function of spot intensity level. Any correlation between intensity and shape can subsequently be used to improve estimates of overall hybridization levels, possibly leading to more accurate gene microarray analysis. After extracting the spot boundary for the $i$-th spot we normalize the spot's intensity so that it sums to one and this normalized intensity can be viewed as a probability distribution, $Q_i(x, y)$ for the $i$-th spot. Now using the centroid and the distribution of the $i$-th spot, the covariance matrix, $\Omega_i$, for that spot is constructed by using the relation,

$$\Omega_i = [\Delta x_i, \Delta y_i]^{'} Q_i(x, y)[\Delta x_i, \Delta y_i] \tag{16}$$

where $\Delta x_i = x - mx_i$ and $\Delta y_i = x - my_i$, with $mx_i$ and $my_i$ being $x$ and $y$ coordinates respectively of the centroid of the $i$-th spot. This is followed by eigen-analysis of $\Omega_i$ of each spot to find eigen-vectors $\zeta_{i1}$ and $\zeta_{i2}$ for the covariance matrix of the $i$-th spot. The first eigen-vector $\zeta_{i1}$ lies along the axis that has the most 'mass' concentrated and the second eigen-vector $\zeta_{i2}$ is orthogonal to this. Thus we can see that the eigen-vectors line up with the distribution (intensity) of the spot as depicted in Fig. 10.

## V. Conclusion

Spot extraction of a gene microarray image has been achieved using the watershed segmentation and other morphological techniques for image analysis. This method is robust to noise problems leading to oversegmentation. The computational requirements of our procedure are very low. The detected boundaries of the extracted spots are used to obtain B-spline coefficients of the shape, which are further stored in a database for quantification of spot variations. Using morphological segmentation permits us to perform shape and intensity analysis.

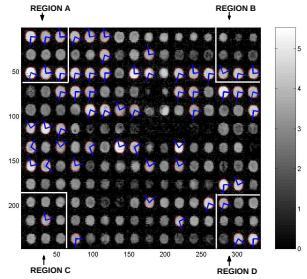


Fig. 10. *A gene microarray image with B-Spline boundaries for each spot and eigen axis of covariance matrix of each spot*

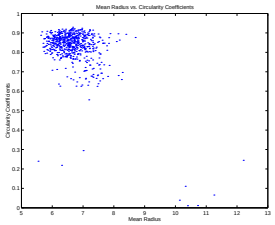


Fig. 11. *The majority of spots have high circularity with the exception of spots that have very small radii or those with very large radii*

## References


[1] M. B. Eisen and P. O. Brown, *"DNA Arrays for Analysis of Gene Expression"*, Methods Enzymol 303, 179-205 (1999).

[2] L. S. Davis, *"A survey of edge detection techniques"*, Comput. Graphics and Image Processing, vol. 4, 179-205 (1975).

[3] S. W. Zucker, *"Region growing: Childhood and adolescence"*, Comput. Graphics and Image Processing, vol. 5, 382-399 (1976).

[4] J. Serra, *"Image Analysis and Mathematical Morphology"*, London, England: Academic Press (1982).

[5] A. O. Hero and G. Fluery, *"Pareto-optimal methods for gene filtering"*, submitted to Journ. Amer. Statist. Soc. (JASA) (2002).

[6] C. R. Giardina and E. R. Dougherty, *"Morphological Methods in Image and Signal Processing"*, New Jersey: Prentice Hall (1988).

[7] H. J. A. M. Heijmans, *"Morphological Image Operators"*, Boston: Academic Press (1994).

[8] J. Serra and P. Soille, *"Mathematical Morphology and its Applications to Image and Signal Processing"*, Dordrecht, The Netherlands (1994).

[9] L. Vincent, *"Morphological Grayscale reconstruction in image analysis: Applications and efficient algorithms"*, IEEE Transactions on Image Processing, vol. 2, pp 176-201 (1993).

[10] F. Meyer and S. Beucher, *"Morphological segmentation"*, Journal of Visual Communications and Image Processing, vol. 1, pp. 21-46 (1990).

[11] Dougherty, E. R., *"An Introduction to Morphological Image Processing"*, SPIE Press, Bellingham (1992).

[12] J. Goutsias and S. Batman, *"Morphological Methods for Biomedical Image Analysis Handbook of Medical Imaging"*, Volume 2, Medical Image Processing and Analysis M. Sonka and J. M. Fitzpatrick (2000).

[13] S. Beucher, *"The watershed transformation applied to image segmentation"*, 10th Conf. on Signal and Image Processing in Microscopy and Microanalysis , Cambridge, UK (1991).

[14] Elaine Cohen, Richard F. Riesenfeld and Gershon Elber, *"Geometric Modeling with Splines: An Introduction"*, Natick Mass., Peters (2001).

[15] Yee Hwa Yang, Michael Buckley, Sandrine Dudoit and Terry Speed, *"Comparison of methods for image analysis on cDNA microarray data"*, University of California, Berkeley, Technical Report # 584 (2002).