A Genetic Approach to cDNA Microarray Image Analysis

E. Zacharia, D. Maroulis

Dept. of Informatics and Telecommunications, University of Athens, Greece
rtsimage@di.uoa.gr

Abstract

Microarray image analysis is a significant tool for cDNA microarrays and it is divided in two main stages: Gridding and Spot-Segmentation. Most of the available microarray image analysis tools require human intervention to specify certain landmarks on the grid, or even to precisely locate individual spots. This paper focuses on the development of an original, fully automated gridding and spot-segmentation approach based on a genetic algorithm. This approach involves three main steps: a) Preprocessing of input images by wavelet-based noise reduction and Box-Cox transformation adjustment b) Gridding the preprocessed images by detecting the rectangular regions where individual spots are placed, c) Spot-segmenting together with model-based quantificating of individual spots using a genetic algorithm. The proposed genetic algorithm searches within a multidimentional-parameter space to determine, in parallel, the parameters of multiple diffusion models that optimally fit the characteristics of possible spots. Experiments with 16-bit microarray images show that the proposed method is effective and results in higher percentage of spot detection than that of existing method.

Keywords: cDNA Microarrays, Image Analysis, Gridding, Spot Segmentation, Parallel Genetic Algorithm.

1. Introduction

The cDNA microarrays [Leung et. Al. (2003)] is a powerful biotechnology tool with which the expression levels of thousands of genes over different samples are measured simultaneously. Because of its high throughput capacity and its ability to compare gene expression in normal and abnormal samples, this technique has a large impact on a variety of different application areas [Lobenhofer et. Al. (2001)]. In cDNA microarrays experiments [Cambell et. Al. (2006)], two distinct populations of mRNA are reversely transcribed to cDNAs, coloured with fluorescent dyes, then mixed and finally hybridized to a glass slide. The end product of the experiment is two digital images, one for each population of mRNA.

Ideally, the pixel intensity of each image varies in proportion to the hybridization level. However, in reality, the resulting images are contaminated with noise and artifacts. Moreover, they may contain inner holes, scratches and uneven background

[Chen et Al. (2006)]. Image-processing techniques [Smyth et. Al. (2002)], [Chen et. Al. (1997)] are applied on microarray images in order to remove artifacts, adjust image intensity, detect the grid structure (Gridding) and determine the boundaries of each individual spot (Spot Segmentation). These techniques are vital to microarray experiments due to the fact that the subsequent steps in experiments such as intensity extraction from each individual spot, clustering and identification of differentially expressed genes can be considerably affected. Since all the stages of microarray image analysis are challenging tasks, they have generally been separately dealt with in the literature.

Preprocessing applications which have recently been applied are based on stationary wavelet filtering techniques [Wang et. Al. (2003)] or on vector processing techniques [Steinfath et. Al. (2001)], which remove noise while preserving the structural cDNA image information.

Image adjustments are executed through transformations which are based either on a small set of available spot characteristics (i.e. mean, median, variance) or on the values of all the pixels of an image. As mentioned in Wit et. Al. (2003), more accurate adjustments can be obtained by transformations based on all pixel values. Such transformations are: the Box-Cox logarithmic transformation [Ekstrom et. Al. (2004)] and the Hyperbolic Sine transformation [Durbin et. Al. (2002)].

Gridding applications are achieved by a variety of different methods. Deng et. Al. (2004) use axis projections of image intensity. This method is not ideal in the case of misalignments and rotations of different grids. Angulo et. Al. (2003) propose morphological methods for grid segmentation. Since these methods use axis projections as a central component, misalignments and rotations of different grids can cause problems too. Yang et. Al. (2002) use template matching and seeded region growing methods while Katzer et. Al. (2003) use a Markov random field (MRF) for semiautomatic gridding. All aforementioned techniques require mandatory input parameters such as the number of rows and columns of the grid structure and, at times, manual intervention in order to locate precisely the grid.

Gridding is followed by spot-segmentation techniques. Several methods have been developed to segment microarrays spots. ScanAlyze software [Eisen (1999)] uses a fixed circle segmentation method which is based on a fixed radius circle of the spot. GenePix [Axon (2002)] uses an adaptive circle segmentation method which is based on a circle with adjustable radius so as to fit the spot. However, both of them are not optimal for non-circularly spots. Vesanen et. Al. (2002) suggested a method based on a generalized hit-or-miss transformation. A drawback of this method is the requirement of a training sample from which a typical spot shape can be learned. The Chen et. Al. (1997) method is based on the nonparametric Mann–Whitney test. This method does not make any assumptions on the spot shape. However, it requires a background sample, which makes it difficult to provide reliably in an automated system. Yang et. Al. (2002) use a seeded region growing method, however this method presents the weakness of relying heavily on a seed point.

In this paper, we present an original approach to unsupervised gridding and spot-segmentation in microarray images, based on a genetic algorithm. This algorithm searches within a multidimentional parameter space to determine, in parallel, the parameters of multiple diffusion models that optimally fit the characteristics of possible spots. The spot-segmentation is achieved by modeling spots using diffusion modeling functions. To the best of our knowledge genetic algorithms have not been previously applied to microarray spot-segmentation.

The rest of this paper is structured in four sessions. Section 2 describes the diffusion modeling of microarray spots used in this study. The proposed approach to gridding and spot-segmenting is described in Section 3. The results of its application in microarray images are presented in section 4, whereas the conclusions of this study are summarized in section 5.

2. Diffusion modeling of Spots

Microarray spots share with each other some common characteristics, such as having an approximately elliptical shape which can be depicted to a ‘volcano’ or to ‘plateaus’ in 3-dimensional space. These simple characteristics can be captured by tuning the parameters of a mathematical model so that it fits an image region containing a spot.

The diffusion model proposed by [Bettens et. Al. (1997)] suggests that the spots are modeled by a mathematical function representing the actual diffusion process of a protein into a polyacrylamide gel. The assumptions for this process are the following: a) the medium of the diffusion is two-dimensional and anisotropic, i.e. there are two main directions of diffusion (*x* and *y*) with different diffusion constants D_x and D_y, b) the diffusing substance is initially distributed uniformly on a disc with radius *a*.

The cDNA strands are hybridized by a diffusion process too [Gadgil et Al. (2004)]. However, their diffusion process differs from that of the proteins because it is isotropic. Therefore, spots can be modeled on the aforementioned diffusion model only if it is properly modified from an anisotropic to an isotropic one. We assume that the constant D_x cannot differ from the constant D_y more than a threshold T_D.

$$\left|D_x - D_y\right| <= \mathrm{T_D} \quad \textbf{(1)}$$

3. A Genetic Approach to Gridding and Spot-Segmentation

The proposed approach to microarray gridding and spot-segmentation consists of three main steps: a) Preprocessing of input images by wavelet-based noise reduction and Box-Cox transformation adjustment, b) Gridding the preprocessed images by detecting the rectangular regions where individual spots are placed, and c) Spot-segmenting together with model-based quantificating of individual spots using a genetic algorithm.

3.1 Preprocessing of microarray images

The quality of microarray images suffers from the existence of noise (i.e. dust on the slide), artifacts (i.e. inner holes and scratches) and uneven background [Chen et. Al. (2006)]. Therefore, stationary wavelet-base filtering of the input images [Wang et. Al. (2003)] is used as a de-noised step, which involves the following three steps: Firstly, the image is decomposed by applying the biorthogonal wavelet procedure. After the decomposition procedure at two levels, the SureShrink thresholding algorithm [Donoho et. Al. (1995)] is applied to the detail-images in order to eliminate the noise source. Finally, the denoised detail-images and the approximation image are reconstructed.

Moreover, lowly expressed genes in microarray experiments are depicted by spots that are poorly contrasted and ill-defined. Box-Cox transformation, [Ekstrom et. Al. (2004)] is employed in order to adjust properly the intensities of microarray images. The equation of the Box-Cox transformation as a function of (x, y, λ_1, λ_2) is the following:

$$Y(x,y,\lambda_1,\lambda_2)=\begin{cases}\dfrac{k((Z(x,y)+\lambda_1)^{\lambda_2}-1)}{\lambda_2},\lambda_2\neq 0\\ k\log(Z(x,y)+\lambda_1),\lambda_2=0\end{cases} \quad \textbf{(2)}$$

where Z stands for the denoised microarray image, Y is the transformed microarray image, and (x,y) are pixel coordinates of the aforementioned images. λ_1 is a positive, non-zero constant ($\lambda_1>0$), λ_2 is a positive or zero constant ($\lambda_2\geq 0$), and k is a constant used to scale the transformed pixel intensities so that a saturated pixel (for a 16-bit image, a saturated pixel has intensity equal to 65535) corresponds to a value of $Y(x, y, \lambda_1, \lambda_2)=1$.

3.2 Gridding microarray images

The pre-processed microarray images are segmented into rectangular regions each one ideally containing one individual microarray spot, as follows:

For a line l of the pre-processed image Y we define the real function $P^{(Y)}_{line}(l)$ which shows the possibility of the line containing spots. The $P^{(Y)}_{line}(l)$ function is calculated by the following equation:

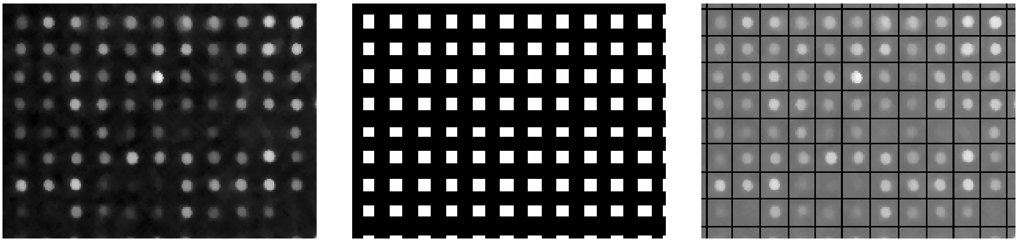
$$P^{(Y)}{}_{line}(l) = \frac{\sum\limits_{\substack{[(x,y)\in Y]\wedge \\ [x=l]\wedge \\ [(x,y):Y[x][y]>Ts]}} 1}{\sum\limits_{\substack{[(x,y)\in Y]\wedge \\ [x=l]}} 1} \qquad \textbf{(3)}$$

where *(x,y)* are the coordinates of the image *Y*, and T_s is a threshold.

Respectively, for a column *c* of the image *Y* we define the real function $P^{(Y)}{}_{column}(c)$ which shows the possibility of the column containing spots.

The algorithm proceeds by calculating the real functions $P^{(Y)}{}_{line}(l)$ and $P^{(Y)}{}_{column}(c)$ for each line *l* and column *c*. If the line l_{xo} and the column c_{yo} of the pixel (x_o,y_o) have values of $P^{(Y)}{}_{line}(l_{xo})$ and $P^{(Y)}{}_{column}(c_{yo})$ larger than the threshold T_p ($P^{(Y)}{}_{line}(l_{xo})>T_p$ and $P^{(Y)}{}_{column}(c_{yo})>T_p$), the pixel is inside a region. Otherwise, the pixel is situated between two different regions. The T_p and T_s thresholds are defined through experimentation.

The algorithm continues by drawing a grid separating in half the pixels (black) which are found between two distinct neighboring regions (white). A segmentation example of a microarray sub-image is illustrated in Figure 1.



(a) (b) (c)

Figure 1. *Microarray image transformation and gridding: (a) Pre-processed image, (b) Image where the pixels which are inside a region are depicted in white color while the pixels which are between two distinct neighboring regions are depicted in black color, (c) Output image where gridding is depicted.*

3.4 Spot Segmentation in microarray images

This step aims to determine the optimal diffusion model for each microarray spot, in the pre-processed microarray image. Finding the optimal model parameters is not a straightforward process due to inner holes, scratches, uneven background and spot overlapping. In order to tune automatically the parameters of the diffusion models so that they optimally fit the microarray spots, we developed an original method based on a genetic algorithm capable of dealing with the afore-mentioned situations.

Genetic algorithms are stochastic non-linear optimization algorithms based on the theory of natural selection and evolution [Goldberg et. Al. (1989)]. Compared to traditional search and optimization procedures, genetic algorithms are parallel, robust optimizers, suitable for solving problems for which there is little or no a priori knowledge of the underlying processes.

The genetic approach to spot-segmentation, proposed in this paper, assumes that the procedure of gridding may give smaller regions or even sifted regions as compared to ideal regions. Therefore, for each rectangular region *R* we create another one, *R',* which has the same center as *R* but its length and width are n times larger than the ones of *R*.

The developed genetic algorithm performs a parallel search for the optimal model parameters of an N^2 number of different microarray spots. Each spot model has its centre inside region *R* but a part of the spot can appear in region *R'*.

Chromosome: The parameters of the diffusion models that correspond to microarray spots which are contained in an N^2 number of different *R'* regions are encoded into a single three-dimensional (3D) chromosome *m*. The chromosome consists of N^2 segments m_{ij}, i=1,2,3,…N, j=1,2,3,…,N. Each segment is a string of real values representing the diffusion-model parameters of a spot which belongs to a region *R'*.

Genetic Operations: At the beginning, an initial population of randomly generated chromosomes is selected. This population evolves then, through the genetic algorithm, by creating new generations of population. In each generation of the genetic algorithm, the P_r% of the best chromosomes is maintained in the next generation of population. The rest are reproduced by applying: 1) the joint application of the BLX-*a* crossover and of the Dynamic Heuristic one [Herrera et. Al. (2005)] and 2) the Wavelet mutation [Ling et. Al. (2006)]. The joint application of the BLX-*a* and the Dynamic Heuristic crossover is the most promising crossover application, [Herrera et. Al. (2005)] while wavelet mutation exhibits a fine-tune ability, [Ling et. Al. (2006)] .

Fitness Function: The fitness of a chromosome *m* as a solution to the particular optimization problem is defined by the following equation:

$$f(m) = \sum_{i=1}^{i=N} \sum_{j=1}^{j=N} f_L(m_{ij}) \tag{4}$$

where the real valued function $f_L(m_{ij})$ is defined as the local fitness of a chromosome segment m_{ij}, i=1,2,3,…,N , j=1,2,3,…,N.

The local fitness $f_L(m_{ij})$ of a chromosome segment m_{ij} is computed by the following equation:

$$f_L(m_{ij}) = \frac{\sum\limits_{\substack{[(x,y)\in Y]\ \wedge \\ [(x,y):C(x,y|m_{ij})>B(m_{ij})]}} d_L(x,y\,|\,m_{ij})}{\sum\limits_{\substack{[(x,y)\in Y]\ \wedge \\ [(x,y):C(x,y|m_{ij})>B(m_{ij})}} 1} \tag{5}$$

where

$$d_L(x,y\,|\,m_{ij}) = \begin{cases} 1, & \text{if } \left|C(x,y\,|\,m_{ij}) - Y(x,y)\right| < b \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

(x, y) are pixel coordinates in the transformed image *Y*

$C(x\ y|m_{ij})$ is the value of the diffusion model encoded by the chromosome segment m_{ij}

$b = p\ Y(x, y)$ and $0 < p \leq 1$ is a constant.

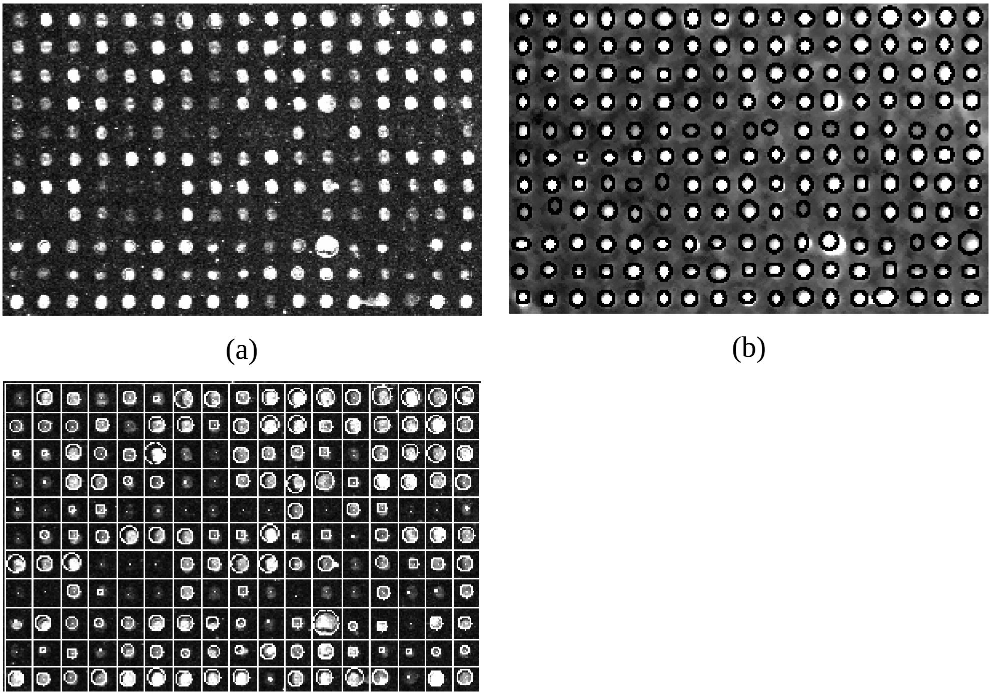
The local fitness expresses the percentage of pixels of the model spot for which $C(x,y|m_{ij})$ differs from $Y(x,y)$ less than b. If $|C(x,y|m_{ij})$-$Y(x,y)|<b$ and $C(x,y|m_{ij})>B(m_{ij})$ then $f_L(m_{ij})$ =1. The parameter b controls the tolerance level of the local fitness to include as fittest solutions, models that approximate spots with irregularities, and asymmetries, or even spots which have volcano or plateau shape, with arbitrary precision.

4. Results

Several experiments were executed so as to evaluate the performance of the proposed algorithm on a set of microarray images at 16-bit grey level depth. An example of these images is the one illustrated in figure 2a which is a sub-image of “Array1.tif” that is used in [Leonardi et. Al. (2004)]. A population of 100 chromosomes was used. In each generation of the genetic algorithm, 20% of the best chromosomes were maintained in the population, whereas the remaining 80% were reproduced by crossover and mutation operations.

The best results were achieved using a high crossover probability of 0.8 and a high mutation probability of 0.8 which are in accordance with [Miller et. Al. (2003)] and [Janikow et. Al. (1991)] respectively. In particular, Janikow et. Al. suggest that the real-coded genetic algorithm may take advantage of such high mutation probability rates. The reason is that the real-coded genetic algorithm does not provide enough diversity via the crossover operation alone. Mutation on the other hand can select a new real value within the allowable range of each designed gene of the chromosome. The values of the thresholds which gave adequate spot segmentation results were: T_D=0.25, T_S =4000, and T_p=0.5.

Using the proposed approach, the accuracy of locating spots is considerably higher than that of the MicroZip software package. Our method segmented 84.6% of real spots instead of 77% of the MicroZip software package. Examples of output images containing indicative spot detection results are illustrated in Figure 2. This figure contains 187 real microarray spots. As we can see, the proposed method found and segmented all 187 real microarray spots while the MicroZip software program could not properly segmented 52 of them.



(a) (b)

(c)

***Figure 2**. Microarray spot segmentation results. (a) Input microarray subimage, (b) Output of the proposed method, (c) Output of the MicroZip software package.*

5. Conclusions

In this paper, an original method to detect the grid and segment microarray spots in microarray images based on a genetic algorithm has been presented. The genetic algorithm searches within a multidimensional-parameter space to determine, in parallel, the parameters of multiple diffusion models that optimally fit the characteristics of possible spots.

The proposed method has the following advantages: a) it does not require a training phase, b) it is capable of segmenting spots which have volcano or plateaus shape, c) it is capable of segmenting spots distorted by imperfect diffusion, d) its spot-segmentation rate is higher than that of the MicroZip software package.

Future development at this project includes further experimentation, optimization and parallelization of the proposed method, and its integration into a complete user-friendly software application.

Acknowledgments

This work was supported by the Greek General Secretariat of Research and Technology and the European Social Fund, through the PENED 2003 program (grant no. 03-ED-332). Moreover, it was partially funded by National and Kapodestrian University of Athens, Special Account of Research Grants.

References

Angulo, J., Serra, J. (2003), *Automatic analysis of DNA microarray images using mathematical morphology*, Bioinformatics, vol. 19, pp. 553–562.

Axon Instruments (2002), GenePix Pro documentation, http://www.axon.com.

Bettens, E., Scheunders, P., Dyck, D.V, Moens, L., Osta, P.V. (1997), *Computer analysis of two dimensional electrophoresis gels: A new segmentation and modelling algorithm*, Eletrophoresis, vol. 5, pp. 792-798

Campbell A.M., Heyer L. J. (2006), *Discovering Genomics, Proteomics & Bioinformatics*, 2nd edn, Pearson Benjamin Cummings, Printed in the USA, ISBN 0-8053-8219-4

Chen W.B., Zhang C., Liu W.L. (2006), *An Automated Gridding and Segmentation Method for cDNA Microarray Image Analysis*, IEEE Symposium on Computer based Medical Systems, Maribor, Slovenia.

Chen, Y., Dougherty, E.R., Bittner, M.L. (1997), *Ratio-based decisions and the quantitative analysis of cDNA microarray images*, Journal of Biomedical Optics, vol. 2, pp. 364–367.

Deng N., Duan H. (2004), *The Automatic Gridding Algorithm based on projection for Microarray Image*, International conference on Intelligent Mechatronics and Automation, Chengdu, China

Donoho D.L.,Johnstone I.M. (1995), *Adapting to unknown smoothness via wavelet shrinkage*, Journal of American Statistical Association, vol. 90, no. 432, pp 1200-1224.

Durbin, B.P., Hardin, J.S., Hawkins, D.M., Rocke, D.M. (2002), *A variance-stabilizing transformation for gene-expression microarray data*, Bioinformatics, vol. 18, pp. 105–110.

Eisen M.B. (1999), ScanAlyze documentation, http://rana.lbl.gov/EisenSoftware.htm

Ekstrom, C.T., Bak, S., Kristensen, C., Rudemo, M. (2004), *Spot shape modelling and data transformations for microarrays*, Bioinformatics, vol. 20, pp 2270-2278.

Gadgil, C., Yeckel, A., Derby, J. J,. Hu, W.S. (2004), *A diffusion-reaction model for DNA microarray assays*, Journal of Biotechnology, vol. 114, pp. 31-45

Goldberg D.E. (1989), *Genetic Algorithms in Search, Optimization & Machine Learning*, Addison-Wesley, printed in the USA, ISBN 0-2011-5767-5.

Herrera, F., Lozano, M., Sanchez, A.M., (2005), *Hybrid crossover operators for real-coded genetic algorithms: An experimental study*, Soft Computing - A Fusion of Foundations, Methodologies and Applications, vol. 9 pp. 280-298.

Janikow C.Z., Michalewicz Z. (1991), *An experimental comparison of binary and floating point representations in genetic algorithms*, *4th International Conference on Genetic Algorithms*, San Diego, California.

Katzer M., Kummert F., Sagerer G. (2003), *A Markov Random Field Model of microarray gridding*, ACM Symposium on Applied computing, Florida, USA.

Leonardi S., Luo Y. (2004), *Gridding and Compression of Microarray images*, IEEE Computational Systems Bioinformatics Conference, Standford, USA.

Leung Y.F., Cavalierim D. (2003), *Fundamentals of cDNA microarray data analysis*, TRENDS in Genetics, vol. 19 no. 11, pp. 649-659.

Ling, S.H., Leung, F.H.F (2006), *An improved genetic algorithm with average-bound crossover and wavelet mutation operations*. Soft Computing - A Fusion of Foundations, Methodologies and Applications, vol. 11, pp. 7-31.

Lobenhofer, E.K., Bushel, P.R., Afshari, C.A., Hamadeh, H.K. (2001), *Progress in the Application of DNA Microarrays*, Environmental Health Perspectives, vol. 109, pp. 881-889.

Miller, M.T., Jerebko, A.K.,. Malley, J.D, Summers, R.M. (2003), *Feature selection for computer-aided polyp detection using genetic algorithms*, Proceedings of SPIE, vol. 5031, pp. 102-110.

Smyth, G.K., Yang, Y.H., Speed, T. (2002), *Statistical issues in cDNA microarray data analysis*, Functional Genomics: Methods and Protocols, vol. 224, pp. 111-136.

Steinfath, M., Wruck, W., Seidel, H. (2001), *Automated image analysis for array hybridization experiments*, Bioinformatics, vol. 17, pp. 634–641.

Vesanen P. (2002), *Calibration-free methods in segmentation of cDNA microarray images*, 12th Symp. Electronic Imaging Science and Technology, San Jose, California.

Wang, X.H., Istepanian, R.S.H., Song, Y.H. (2003), *Microarray Image Enhancement by denoising using stationary wavelet transform*, IEEE Transaction on nanobioscience, vol. 2, no 4, pp 184-189.

Wit, E., McClure, J. (2003), *Statistical adjustment of signal censoring in gene expression experiments*, Bioinformatics, vol. 19, pp. 1055–1060.

Yang, Y.H., Buckley, M.J., Duboit, S., Speed, T.P. (2002), *Comparison of methods for image analysis on cDNA microarray data*, Journal of Computational and Graphical Statistics, vol. 11, pp. 108–136.