# DNA MICROARRAY IMAGE INTENSITY EXTRACTION USING EIGENSPOTS

*Sotirios A. Tsaftaris, Ramandeep Ahuja, Derek Shiell, and Aggelos K. Katsaggelos*

Department of Electrical Engineering and Computer Science
Email: {stsaft, rah111, djs740, aggk}@eecs.northwestern.edu
Northwestern University, 2145 Sheridan Rd., Evanston, IL 60208, USA

## ABSTRACT

DNA microarrays are commonly used in the rapid analysis of gene expression in organisms. Image analysis is used to measure the average intensity of circular image areas (spots), which correspond to the level of expression of the genes. A crucial aspect of image analysis is the estimation of the background noise. Currently, background subtraction algorithms are used to estimate the local background noise and subtract it from the signal. In this paper we use Principal Component Analysis (PCA) to de-correlate the signal from the noise, by projecting each spot on the space of eigenvectors, which we term eigenspots. PCA is well suited for such application due to the structural nature of the images. To compare the proposed method with other background estimation methods we use the industry standard signal-to-noise metric *xdev*.

**Index Terms—***DNA microarray, biochip, eigenspaces, noise, segmentation.*

## 1. INTRODUCTION

Microarrays (or biochips) allow for the simultaneous study of all the genes in an organism in a single experiment. This is made possible by spotting (placing) thousands of short DNA sequences on a surface. For microarrays manufactured using *in situ* synthesis (such as the ones studied in this work) each spot is fairly circular and the microarray image itself is very structured. Microarray technology relies on hybridization between the genes (messenger RNA or cDNA) and the DNA probes spotted on the array. Two gene pools are used; a test one and a control one. The genes bind to the probes on the array and become immobilized. The gene sequences are labeled with fluorescent dyes Cy3 and Cy5 (control and test, respectively) [2]. Thus, the level of gene expression, corresponding to the amount of gene sequences immobilized on a specific spot on the array, is proportional to its intensity.

The main objective from a biological standpoint is to determine the gene expression level in cells. The process of acquiring an image involves laser scanning at two different wavelengths, each corresponding to the excitation level of each dye. This process results in two 16-bit images labeled as red (Cy3) and green (Cy5). The gene expression is deduced from the ratio of the spot intensity of the two channels. Analysis of the microarray image involves finding the layout of the spots and superimposing a grid, the centers of the spots, segmenting the spots (determining background from foreground), spot intensity estimation, performing quality control to remove unreliable spots, performing dye normalization and gathering statistics to determine differentially expressed genes [2].

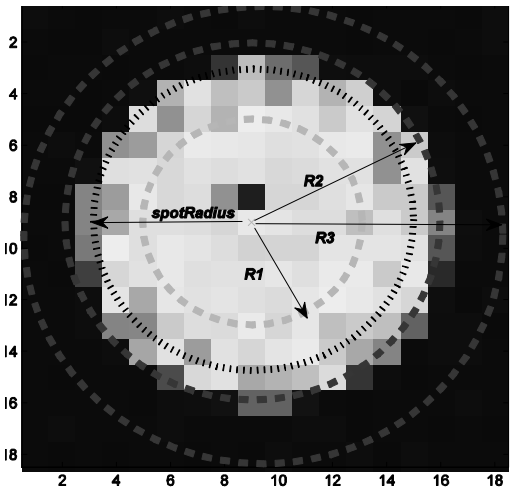


**Fig. 1.** A DNA microarray spot, (adopted from [1]).

In this work we assume that the center of the spots are known and we focus on the spot intensity estimation. Traditional spot intensity estimation techniques require the segmentation of each spot into foreground and background regions. Subsequently an average measurement of the foreground intensity is taken, followed by an estimate of the background noise, which is afterwards subtracted from the average foreground intensity. This provides the background adjusted spot intensity. This process is repeated for each channel and the ratio of the two channels is used to assess the level of gene expression.

This paper presents a new method for estimating spot intensities without subtracting a local background estimate. We use Principal Component Analysis (PCA) to de-correlate the signal from the noise by projecting each spot on the orthonormal space of eigenvectors, which we term eigenspots. PCA is well suited for this application due to the highly structural form of microarray images. We use images from Agilent Corporation for our experiments and Agilent's spot quality metric to assess the performance of our method.

This paper is organized as follows. In section 2 we describe how local background subtraction (LBS) algorithms work and the metric used to judge the quality of spots. In section 3 we present our work based on PCA. Section 4 presents our findings. Finally section 5 offers conclusions and future work.

## 2. SYSTEM MODEL

In this section we present Agilent's LBS algorithm, which is used to compare the proposed method [8]. Assuming known spot centers and based on parameters known from the manufacturer we can define a rectangular region of interest of *N*x*N* pixels, known in the industry as *vignette*. Figure 1 shows a vignette taken from an image of an Agilent microarray. There are four radii defined. *spotRadius* is the average radius of all spots in the microarray image after all spots have been identified and segmented. The radii *R1*, *R2* are defined as percentages of the *spotRadius* and are user selected parameters, while *R3* is equal to 0.5*N*.

The region enclosed by radii *R2* and *R3* according to Agilent defines the local background region, $\beta$. An estimate of the intensity of the spot is taken by averaging the signal from the circular area of radius *R1* denoted by *s*, defined as $\mu_{s,r}$ and $\mu_{s,g}$ for the red and green channel respectively. The signal from the area defined by the radii *R1* and *R2* is ignored due to a deficiency in the manufacturing of the array, which results in probes with smaller lengths at the perimeter of the spot compared to the length of the probes at the center. We will see later that our proposed method identifies those regions.

To estimate the local background according to Agilent's method the mean of $\beta$ is taken, defined as $\mu_{\beta,r}$ and $\mu_{\beta,g}$ for the red and green channel, respectively. Other LBS methods take the median of $\beta$ or the median or average of predefined spots in the vignette as an estimate of the background noise. Interested readers are referred to [10] for a review of such methods.

Finally, the background adjusted spot intensity for each channel is defined as:

$$\mu_r = \mu_{s,r} - \mu_{\beta,r} \quad (1)$$

$$\mu_g = \mu_{s,g} - \mu_{\beta,g} \quad (2)$$

A necessary step in microarray analysis is dye normalization, which is necessary due to the different excitation levels of the dyes. For simplicity, we will ignore this step, since also it has no effect in the following analysis.

We describe next the *xdev* metric which represents an industry accepted metric for evaluating the quality of the estimate of the average intensity of the spot and it is used in this paper. Let $N_{s,r}$, $N_{s,g}$ denote the number of pixels in the *s* region of the red and green channel, respectively. Let $N_{\beta,r}$, $N_{\beta,g}$ denote the total number of background pixels for a spot in the red and green channel respectively. The standard deviation per channel, are defined as

$$\sigma_r = \sqrt{\sigma_{s,r}^2 / N_{s,r} + \sigma_{\beta,r}^2 / N_{\beta,r}} \ , \quad (3)$$

$$\sigma_r = \sqrt{\sigma_{s,g}^2 / N_{s,g} + \sigma_{\beta,g}^2 / N_{\beta,g}} \ , \quad (4)$$

where $\sigma_{s,r}^2 (\sigma_{s,g}^2)$ and $\sigma_{\beta,r}^2 (\sigma_{\beta,g}^2)$ are respectively the variances of the pixels in the *s* and $\beta$ regions for the red (and green) channel. The log ratio ( *lRatio* ) of the dye normalized and background subtracted means of the Cookie signal is given by

$$lRatio = Log_{10}\left(\mu_r / \mu_g\right). \quad (5)$$

*lRatio* is the term used to define the expression of genes as we will see in the results section. The standard error ( *lRatioError* ) in calculating *lRatio* is shown below using the standard error propagation model [2],

$$lRatioError \cong \sqrt{(\ln 10)^{-2}(A - B)} \ , \quad (6)$$

where $A = \sigma^2{}_r / \mu^2{}_r + \sigma^2{}_g / \mu^2{}_g$ and $B = 2\sigma^2{}_{r,g} / \left(\mu_r \cdot \mu_g\right)$, where $\sigma_{r,g}^2$ is the covariance between the two channels. The term *B* is typically neglected since it approaches zero assuming there is no cross correlation between the background subtracted signals in the red and green Channels.

The quality metric *xdev* is defined as the signal to noise metric used in this paper and is a function of the ratio of intensities and the error in estimating the ratio of intensities [1], [9]:

$$xdev = \frac{lRatio}{lRatioErro\,r} \ . \quad (7)$$

This parameter has been proposed by Agilent (the manufacturer of the arrays used in this experiment. A higher |*xdev*| implies greater confidence in the spot intensity ratio estimate. This confidence metric is analogous to the signal-to-noise metric for images.

## 2. EIGENSPOT METHOD

Principal Component Analysis (PCA) is a statistical method used to map a dataset to a new coordinate system such that the variation of the data set is largest along the first principal component, second largest along the second principal component, and so on [5]. The principal components are orthogonal to each other. To remove the noise from the image, we utilize only the Principal Components that contain most of the energy in the signal.

To analyze an image using PCA we scan (by rows or columns) each vignette to convert into a vector of size $N^2 \times 1$ pixels. The collection of column vectors from all the spots in the image form the observation dataset. If the image has *L* spots, the observation matrix *x* has dimensions $N^2 \times L$ pixels, for each of the red and green channels. Alternatively both channels can be utilized in forming an $N^2 \times 2L$ matrix *x*.

We obtain the covariance matrix of *x* is obtained as

$$C_x = E\{(x - u_x) \cdot (x - u_x)^T\} \ , \quad (8)$$

where

$$u_x = E\{x\} \ . \quad (9)$$

We arrange the eigenvectors (henceforth referred to as eigenspots) of the covariance matrix $C_x$ in descending order based on the value of the variance of the corresponding eigenvalues. We stack each eigenspot as a horizontal vector to form the matrix of eigenvectors *V*. We can obtain the co-ordinates of the dataset $x$ in the new principal component space as

$$W = V \cdot (x - u_x) \ , \quad (10)$$

and the original dataset $x$ by re-writing Eq. (10) as

$$x = V^T \cdot W + u_x \ , \quad (11)$$

since $V^T = V^{-1}$ holds for orthogonal matrices.

We reduce the dimensionality of the dataset by keeping only the first *K* eigenvectors, thus forming the matrix $V_K$. The new dataset can then be written as

$$x_k = V_K{}^T \cdot W + u_x \quad (12)$$

Each column *of* $x_k$ is transformed back to a vignette, in order to calculate |*xdev*|. For our method we do not subtract the background signal (as in Eqs (1) and (2)) and therefore $\mu_r = \mu_{s,r}$, $\mu_g = \mu_{s,g}$. Consequently, the 2nd term in Eqs. (3) and (4) is omitted from the standard deviations. |*xdev*| is calculated from Eq. (7), using the above means and standard deviations, and Eqs. (5) and (6).

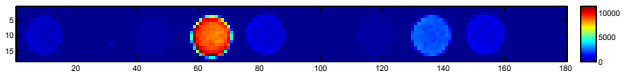

**Fig. 2.** First 10 Spots of the red channel of a microarray image.

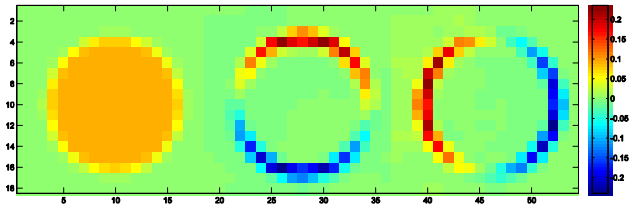

**Fig. 3.** Eigenspots obtained using Method I on the red channel of a microarray image.

## 3. RESULTS

We used six microarray images for our experiments obtained an Agilent system. We used *xdev* as a metric to compare the performance of the eigenspot method to Agilent's LBS method. As mentioned in Sec. 2, there are three choices for deriving the eigenspots used in the PCA method. Method I results by finding the eigenspots for the red and green channels independently and thus transforming them independently. Method S results by finding the eigenspots for one channel but transforming both channels using the same eigenspots set. Finally Method C results by performing PCA once on the dataset consisting of both the red and green channels. We should note here that we are training on the complete set of data since we are not interested in the classification properties of PCA, but on its de-correlation properties.

The results section is divided into two sub-sections. The first one discusses the performance of the PCA approach comparing all three eigenspot derivation methods to Agilent's LBS method, while the second one discusses the effects of the proposed method to gene expression analysis. All methods and metrics are implemented and tested in Matlab.

### 3.1 *xdev* Performance

We tested our method on 19061 spots of a microarray image. Fig. 2 shows the first 10 spots from the red channel of the image. The first three eigenspots obtained by PCA on the red channel are shown in Fig. 3, leftmost is the eigenspot with the highest energy. The energy in the first three eigenspots is 95.6%, 1.3%, and 1.11% respectively. After extensive experimentation with the parameter $K$, we decided to use $K$=3 for all the results presented here, since this value offered the best performance.

We can see from the shape and values of the eigenspots the effect of the manufacturing deficiency. Specifically, the leftmost eigenspot is a disk with same values at the center but smaller values at the perimeter of the spot, while the other two eigenspots have zero values at the center and large and positive values at the perimeter.

In Fig. 2, note that spots 2, 3, 7 and 10 have intensities close to zero, and are henceforth referred as 'inactive spots'. Figures 4 and 5 show the corresponding |*xdev*| and *lRatioError* for the 10 spots for all approaches. |*xdev*| for spots 6, 10 and 12 is not shown since their centers were not found, and were therefore excluded

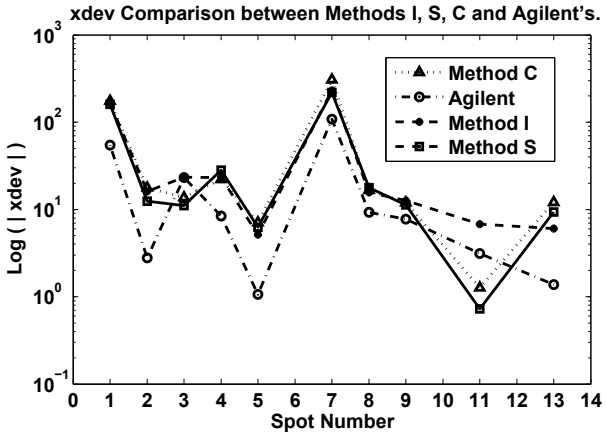


**Fig. 4.** *xdev* for 10 spots of a microarray image.

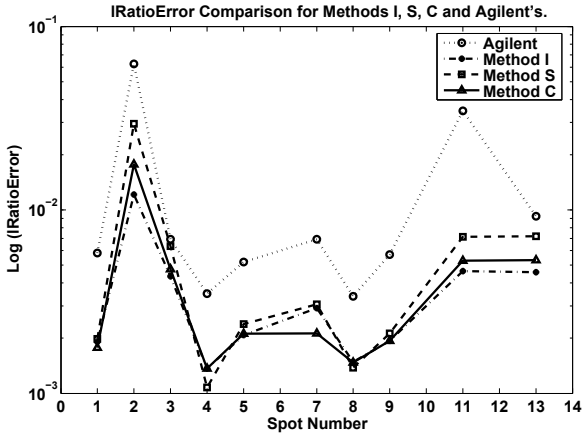


**Fig. 5.** *lRatioError* for 10 spots of an image.

from our analysis. Only the first 10 spots are shown to show the relative performance of the methods.

For Method I, the average |*xdev*| for the 10 spots is 50.69 whereas it is 22.01 for Agilent's LBS. On average the |*lRatio*| of Method I deviates from Agilent's |*lRatio*| by 0.03 for the same spots. Therefore, we can see that the increase in |*xdev*| of Method A is not due to the difference in |*lRatio*| but to the decrease in the *lRatioError* term of *xdev*, instead.

Subsequently, for Method S, eigenspots were derived from the red channel and the red and green channels were projected and reconstructed based on the first three eigenvectors. The corresponding average |*xdev*| is 47.62. Thus the average |*xdev*| dropped 6% compared to Method I.

Using Method C the average |*xdev*| is 58.49, which is 15.3% higher than the average |*xdev*| obtained using Method I.

The |*xdev*| of spots 3 and 11, derived using Agilent's LBS is larger than the corresponding |*xdev*| for Method S and C. This is due to the fact that spots 3 and 11 are inactive spots (low mean intensity. However, these spots are not differentially expressed.

From Fig. 5 we see that *lRatioError* of Method I is smaller than Agilent's for all spots. The proposed method reduces the error in estimating the Log Ratio of intensities compared to LBS methods. Furthermore, we can conclude that Method I is the best ap-

| | IMAGE NUMBER | | | | | |
|---|---|---|---|---|---|---|
| Performance Criteria | Im1 | Im2 | Im3 | Im4 | Im5 | Im6 |
| IN_AGI (Count) | 18473 | 18987 | 18747 | 18738 | 18811 | 21201 |
| OUT_AGI (Count) | 553 | 72 | 310 | 321 | 241 | 30 |
| IN_PCA (Count) | 18543 | 19004 | 18941 | 18967 | 19041 | 21224 |
| OUT_PCA (Count) | 518 | 57 | 120 | 94 | 20 | 12 |
| IN_AGI_OUT_PCA (Count) | 1 | 1 | 0 | 0 | 0 | 0 |
| IN_PCA_OUT_AGI (Count) | 36 | 16 | 190 | 227 | 221 | 18 |
| Mean \| *xdev*\| PCA | 29.18 | 54.81 | 38.15 | 26.10 | 19.37 | 79.14 |
| Mean \| *xdev*\| AGI | 24.48 | 31.40 | 24.58 | 20.70 | 10.43 | 11.30 |
| Total Spots (Count) | 19061 | 19061 | 19061 | 19061 | 19061 | 21236 |

**Fig. 6.** PCA performance for six microarray images.

proach in deriving eigenspots and applying the PCA-based spot intensity estimation when compare to the other methods.

### 3.2 Differential Expression

We evaluate now the performance of the proposed method in terms of the differential expression of a gene. A spot is considered to represent a differentially expressed (DE) gene if

$$|lRatio| \geq T, \quad (17)$$

where $1 \leq T \leq 3$, is chosen by the user and depends on the nature of the experiment [7]. Recall from Eq. (5) $lRatio<0$ represents under-expression, while $lRatio>0$ represents over-expression [10]. The following performance criteria are used to assess how the proposed method affects the analysis of gene expression:

1. *IN_AGI, IN_PCA*: Total number of spots characterized as not DE using Agilent's LBS ($|lRatio_A|<1$) and Method I ($|lRatio_I|<1$), respectively.
2. *OUT_AGI, OUT_PCA*: Total number of spots characterized as DE using Agilent's LBS ($|lRatio_A|\geq 1$) and Method I ($|lRatio_I|\geq 1$), respectively.
3. *IN_AGI_OUT_PCA*: Spots classified as not DE by Agilent (*IN_AGI*) but as DE by Method I ($|lRatio_I|\geq 1$).
4. *IN_PCA_OUT_AGI*: Spots classified as not DE using Method I (*IN_PCA*) but as DE using Agilent's method ($|lRatio_A|\geq 1$).

The table in Fig. 6 shows the results for six DNA images. We observe ($6^{th}$ row) that there are a number of spots flagged as DE by Agilent but as not DE by Method I. Thus the proposed eigenspot method has the potential to reduce the review time and effort required by the experimenter by rejecting a spot based on its *lRatio*. This is a critical element since it reduces the number of follow-up experiments needed on selected spots.

In addition, we see that for Images 1 and 2, there is a single spot, classified as not DE by Agilent but classified as DE by PCA (Fig. 6, row 5). Thus our method identified a spot as DE, which Agilent's method missed.

Furthermore, we see that for all images the average |*xdev*| (row 7) of the proposed method is greater than the one of Agilent's (row 8). Finally, from inspecting the values of |*xdev*| for all the spots in the images, we observed that the proposed method has larger |*xdev*| compared to Agilent's for more than 75% of spots.

## 4. CONCLUSION

In this paper we showed that PCA and the proposed eigenspot method can be used successfully to de-correlate the spot intensity signal from the local background noise. It outperformed the local background subtraction method used by Agilent (a major provider of microarrays and image analysis software). As a comparison metric we used the industry accepted signal to noise metric |*xdev*|. We tested our method on six microarray images and showed that the average |*xdev*| is greater for all images. We also showed that our method improves the classification of differential expression of genes from spot data, by reducing the error in the ratio of intensities.

The proposed technique provides to the experimenter more freedom in deciding the significance of each spot. To that extend we are in collaboration with biologists to deploy and test our method. We are in the process of using the eigenspots as templates for finding the centers of spots and for classifying spot quality. Thus far, our preliminary results show great promise in that direction.

## 5. ACKNOWLEDGMENTS

The authors would like to thank Prof. Papoutsakis and Dr. Paredes in the Dept. of Chemical and Biological Engineering at Northwestern University, for providing the images used in this paper.

## 6. REFERENCES

[1] Agilent G2567AA Feature Extraction Software (v 7.5) User Manual, Agilent Technologies Inc., 2004.

[2] P. R. Bevington, *Data reduction and error analysis for the physical sciences*, McGraw-Hill, Inc. 1969.

[3] J. P. Brody, B. A. Williams, B. J. Wold, and S. R. Quake, "Significance and statistical errors in the analysis of DNA microarray data," *Proc. of Nat. Ac. of Sci. of United States of America*, vol. 99, no. 20., pp 12975-12978, Oct. 1999.

[4] S. W. Davies and D. A. Seale, "DNA microarray stochastic model," *IEEE Trans. on Nanobioscience*, vol. 4, no. 3, , pp 248-255, Sept. 2005.

[5] J.E. Jackson, *A user's guide to principal components,* John Wiley & Sons, Inc. 1991.

[6] G. Kamberova and S. Shah (Eds.), *DNA array image analysis: Nuts and Bolts,* DNA Press, 2002.

[7] C. Y. Kao, Chang, C. F., Chu, H., and Chen, C., "Simple software for microarray image analysis," Angiogenesis Research Center, National Taiwan University Hospital, 2006.

[8] U. Truong, M. Hartnett, C. G. Delenstarr, and A. Lucas, "A method for evaluating the performance of microarray image analysis software packages," Agilent Technologies, Inc., 2004.

[9] Z. Yakhini, C. Y. Enderwick, C. G. Delenstarr, P. K. Wolber, N. M. Sampas, "Method and system of extracting data from surface array deposited features," U. S. Patent 6,591,196 B1, July 8, 2003.

[10] Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed, "Comparisons of methods for image analysis in cDNA microarray data," *Journal of Computational & Graphical Statistics*, vol. 11, no. 1, pp. 108-136(29), Mar. 2002.