IEEE International Conference on Image Processing (ICIP 2003), Barcelona, Sep. 2003

AN UNSUPERVISED ARTIFACT CORRECTION APPROACH FOR THE ANALYSIS OF DNA MICROARRAY IMAGES

Konstantinos Blekas[*], Nikolas P. Galatsanos[*] and Ioannis Georgiou[**]
[*] Department of Computer Science, University of Ioannina, Ioannina, Greece 45110
[**] Medical School, University of Ioannina, Ioannina, Greece 45110

ABSTRACT

Image processing for analysis of microarray images is an important and challenging problem because imperfections and fabrication artifacts often impair our ability to measure accurately the quantities of interest in these images. In this paper we propose a microarray image analysis framework that provides a new method that automatically addresses each spot area in the image. Then, a new unsupervised clustering method is used which is based on a Gaussian mixture model (GMM) and the minimum description length (MDL) criterion, that allows the automatic spot area segmentation and the image artifacts isolation and correction to obtain more accurate spot quantitative values. Experimental results demonstrates the advantages of the proposed scheme in efficiently analysing microarrays.

1. INTRODUCTION

The DNA microarrays [1, 2] are used to measure the expression levels of thousands of genes simultaneously over different time points, and different experiments. In microarray experiments, the two mRNA samples to be compared are reverse transcripted into cDNA, labeled using two different fluorophores (red (R) and green(G)) and then hybridized simultaneously to a glass slide. The fluorescence intensities of the R and G correspond to the level of hybridization of the two samples to the DNA sequences spotted on the slide. The microarray images are structured with intensity spots located on a grid. An example of such grid is shown in Figure 1. Gene expression data derived from arrays measure spots quantitatively and can be used further for analytical purposes [3].

It has been shown [1] that background correction is an important task in the analysis of microarrays. This is necessary in order to remove the contribution in intensity which is not due to the hybridization of the cDNA samples to the spotted DNA. The R and G intensities of a perfect microarray image depend only on the dye of interest. However, due to system imperfections the resulting images in addition to background fluorescence contain other types of undesired signals which are termed in the rest of this paper as *artifacts*. The correction of such artifacts is crucial to making accurate expression measurements, because unlike background fluorescence their spatial location is unknown and can lead to errors that propagate to all subsequent stages of analysis.

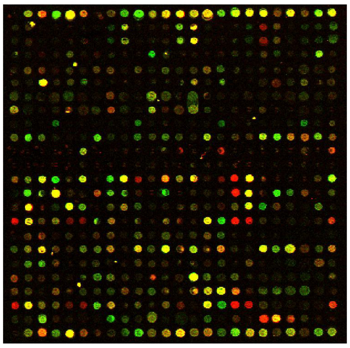


Figure 1. A 24x24 spots size grid of a microarray image.

Several spatial and histogram-based techniques have been introduced for the analysis of the microarray images [1, 2]. These methods correct only for background fluorescence and ignore the presence of artifacts [1, 2]. This paper presents a novel methodology for the microarray image analysis that addresses this problem. It includes three main steps. First, the borders of each spot of the microarray are determined by proposing a new algorithm which initially uses the global properties of the microarray image. Then, they are refined sequentially based on the local image properties. This task, that will be referred to in the rest of this paper as *addressing*, is automatic without human intervention. This is deemed necessary for the analysis of large microarray images with many spots. Following this first step, the pixels in the area of each spot are clustered either into two foreground (F) and background (B), or three, F, B and artifact (A) clusters. The clustering and the determination of the number of clusters to be used is based on a Gaussian mixture model (*GMM*) [4, 5], and the minimum description length (*MDL*) criterion [6]. This is the *segmentation* task of our method. Finally, based on this clustering results the means of the R and G components of the F and B clusters *only* are used to estimate the spot quantitative measurements (normalized R/G ratio). This last task represents the *reduction* step of our approach. To our knowledge it is the first time in microarray analysis that artifact isolation and correction is attempted. In what follows section 2 describes the proposed approach, while in section 3 experimental results are presented where our method is tested. Finally in section 4 we present our conclusions.

2. PROPOSED APPROACH

2.1. Automatic identification of spot areas (addressing)

The addressing procedure of our framework uses a scheme which combines a global and a local segmentation mechanism for assigning borders to each of the microarray spot areas $S(i,j)$, where (i,j) indicates the index of the spot. The proposed approach initially creates *global* borders, which are common to all spots in the same grid area. Following this step, the global borders are refined sequentially using the *local* characteristics of the microarray image.

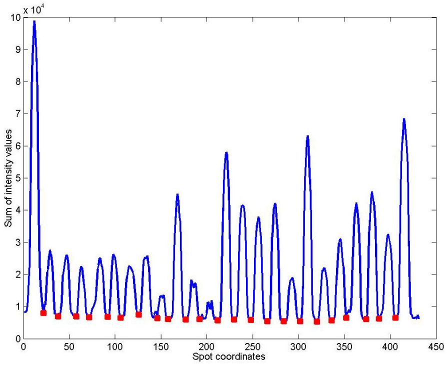


Figure 2. This signal is obtained by summing up the rows of both R and G channels of the microarray image.

During the global phase, we sum the combined R and G intensities along the rows and columns of the microarray image. The peaks of the resulting signals give the vertical and horizontal coordinates of the spot centers, respectively. We use the mid point of two successive peaks of the raw and column sums to define the global horizontal and vertical, respectively, borders between spots. In Figure 2 the row sum of the grid in Figure 1 is shown. The dots indicate the horizontal global borders.

After determing the global borders of the spots the next phase is the refinement step. Starting from the top left spot we redefine the border between the spots $S(i,j)$ and $S(i+1,j)$ ($S(i,j+1)$). This is achieved by finding the local minimum of the signal obtained as the sum of row (column) intensities of the R and G planes in the segment between these two spot areas. The region where we search for the local minimum is the neighborhood around the global border. Figure 3 illustrates an example of the global border refinement process. This procedure is repeated in a row-by-row or column-by-column fashion, scanning the entire microarray image. After addressing the spot areas the segmentation and reduction steps follow.

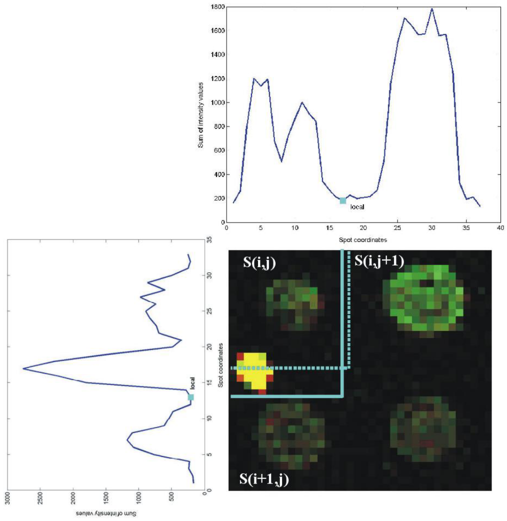


Figure 3. The global borders (dotted lines) are refined (solid lines) based on the local sums. The signals on the left and above the microarray image are the local row and column sums, respectively

2.2. Clustering based on *GMM* and *MDL*

The pixels in a spot are denoted $x_i \in R^2$ with $i = 1 \ldots N$ where $x_i = [R_i, G_i]$, for the R and G intensities respectively. In order to classify the pixel location i as F, B or A we formulate this task as a clustering problem of the two-dimensional input dataset $X = [x_1, x_2, \ldots, x_N]$. To solve this problem we use a Gaussian mixture model (*GMM*) [4]. For our approach each mixture component k models the F, the B or the A region of the spot area that follows the Normal density function $N(x_i; \mu_k, \Sigma_k)$, where μ_k and Σ_k are the 2x1 mean vector and the 2x2 covariance matrix, respectively.

Assuming that a mixture contains K components, a mixture model M_K of Gaussian can be written as

$$f(x_i \mid M_K) = \sum_{k=1}^{K} \pi_k N(x_i; \mu_k, \Sigma_k), \tag{1}$$

defined by the mixture coefficients π_k and the means μ_k and the covariances Σ_k. The log-likelihood of the observed dataset X corresponding to the above model is

$$L(X \mid M_K) = \sum_{i=1}^{N} \log f(x_i \mid M_K) \cdot \tag{2}$$

When the model order K is known the expectation maximization (EM) algorithm [7] is a classical approach that can be applied for log-likelihood maximization with respect to the parameters π_k, μ_k and Σ_k. However, it is well known that the results of the EM algorithm for this application are very sensitive to the initialization used [5]. In order to ameliorate this problem we use a new incremental scheme for Gaussian mixture modeling that has been recently introduced and is termed as *greedy EM*

[5]. Starting with one component, this algorithm sequentially adds a new one by performing a global search over the input dataset. This is followed by a local search based on partial EM steps to fine tune the parameters of the new component. It was shown in [5] that the results of this algorithm for the *GMM* clustering problem are independent on the initial model parameter values.

Based on this formulation, the decision of whether or not artifacts are present in a spot area is equivalent to selecting the order *K* of the *GMM*. If a *GMM* with *K=2* (*GMM(2)*) is selected then the two clusters corresponds to F and B. In contrast a *GMM* with *K=3* (*GMM(3)*) implies the presence of the A cluster also. Thus, the crux of the proposed approach becomes a model order selection problem, i.e. to find the number of components in the mixture that best fit the input dataset *X*. One of the most popular model selection criteria is the Minimum Description Length (*MDL*) [6] defined as

$$MDL(K) = -L(X \mid M_K) + \frac{1}{2} C \log_2(N) \cdot \qquad (3)$$

The first term in Equation (3) gives a measure of how well the model fits the data. Obviously a complex model with more terms would provide a better fit. The second term of Equation (3) is a term that penalizes the complexity of the model and corresponds to the number of bits needed to encode the model parameters. *C* gives the number of free parameters and is given for a d-dimensional *GMM(K)* by $C = Kd\left(d+3\right)/2 + (K-1)$, assuming full covariances matrices.

Therefore, the number of components and thus the presence of an artifact within a spot area can be determined by applying the following procedure:

```
1. For K={2, 3}
   a. Estimate the model parameters ($\pi_k, \mu_k, \Sigma_k$) of the GMM(K)
      by applying the greedy EM algorithm and compute the log-
      likelihood $L(X \mid M_K)$.
   b. Calculate the value of MDL(K) according to Equation (3).
2. If $MDL(2)/MDL(3) \le T$ then use GMM(2)
3. else if $MDL(2)/MDL(3) > T$ then use the GMM(3).
```

We determined that a good value was $T = 1.01$. The last two steps can be seen as an *artifact correction* methodology that eliminates the effect of an artifact into quantitatively measuring any noisy spot area.

The last task of the proposed scheme is to define the criterion used to select the artifact component from a *GMM(3).* Since we assume as artifacts erroneous random samples that are neither foreground nor background, we propose a criterion based on the total variance of the class. Thus, we consider as artifact the component k $(k = 1,2,3)$ that has the greatest degree of scattering among all three classes, i.e $k = \arg\max_k \left(\sqrt{\sigma_{R,k}^2 + \sigma_{G,k}^2}\right)$. Once the A cluster is determined the remaining 2 clusters are labeled as F and B based on the relation $\|\mu_F\| > \|\mu_B\|$.

The feature that captures the properties of a spot area is the ratio of the difference of the fluorescence intensities (R and G) between two clusters (F and B), i.e. $Ratio = (\mu_F^R - \mu_B^R)/(\mu_F^G - \mu_B^G)$, where μ_F^R for example gives the mean of red component of the F cluster. This spot quality measurement is the most usually used [1, 2] since it reflects the transcript abundance for the red and green labeled mRNA samples. A ratio value $Ratio > 1$ ($Ratio < 1$) declares over-expression (under-expression) in the R labeled mRNA sample compared to the G.

3. EXPERIMENTAL RESULTS

The experiments described in this section were made using real datasets and have two objectives. First they test the effectiveness of the proposed method to automatically address spot areas of real microarray images. Second, to study the capabilities of the proposed artifact correction strategy in terms of segmenting properly spot areas and thus to extract more accurate gene expression quantitative values. In this spirit, we have selected real data used in [3]. In particular, we have used one such sample (a grid of which contains 24x24 spots as illustrated in Figure 1), where we have applied our addressing procedure for automatically partitioning it into 576 distinct spot areas. Table 1 represents six examples of spot areas $S(i, j)$ that have been addressed automatically. These results demonstrate that in all cases tested the proposed method created spot areas that contained the gene spot and the adjacent background, even in difficult cases of noisy spots with artifacts, see for example spots 3-6.

After addressing the spot areas in this microarray example, we used our clustering approach to estimate the corresponding fluorescence ratios. For comparison purposes, we have also applied two other known clustering approaches, the k-means and the partitioning around medoids (PAM) that have been proposed in [2]. Since these methods, as described in [2], do not provide a cluster selection criterion, in our study only two clusters (F and B) were assumed.

Table 1 summarizes the comparative results that have been provided for six spots. In each case, we illustrate the spot image segmentation, after classifying pixels according to the class label depicted when applying the greedy EM algorithm [5] to the mixture model with either two or three components. The segmentation map of the spot area is illustrated using light, dark and median gray values, which correspond to the F, B, and A pixels respectively. The *MDL* ratio (ratio $MDL(2)/MDL(3)$) is also shown and is used to decide the existence of an artifact cluster within the spot area, together with the calculated fluorescence ratios R/G for each of the four clustering approaches (*GMM(2), GMM(3)*, 2-means and PAM) respectively. Finally, the same feature is given as extracted by applying the ScanAlyze microarray image tool [8]. This is a spatial method that uses fixed circle segmentation. It must be noted that these values are published at the www at http://llmpp.nih.gov/lymphoma/.

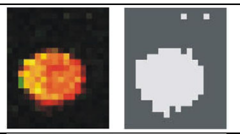
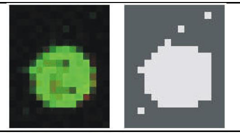
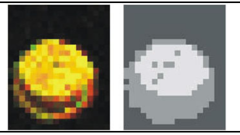
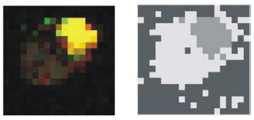
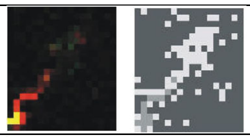
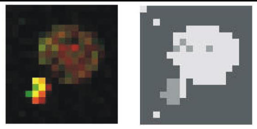
Spot examples		*Methods*	*Ratio*
1 S(1,3) MDL ratio 1.0025		GMM(3)	-
		GMM(2)	2.1254
		2-means	2.1005
		PAM	2.0431
		ScanAlyze	2.7570
2 S(10,7) MDL ratio 1.0002		GMM(3)	-
		GMM(2)	0.3501
		2-means	0.3319
		PAM	0.3411
		ScanAlyze	0.2233
3 S(1,15) MDL ratio 1.1081		GMM(3)	1.1190
		GMM(2)	1.2266
		2-means	1.1565
		PAM	1.1650
		ScanAlyze	0.9125
4 S(2,2) MDL ratio 1.0412		GMM(3)	1.4822
		GMM(2)	1.0816
		2-means	1.0173
		PAM	0.9637
		ScanAlyze	1.2360
5 S(11,15) MDL ratio 1.1059		GMM(3)	1.3350
		GMM(2)	2.4737
		2-means	2.6800
		PAM	7.4848
		ScanAlyze	0.9363
6 S(18,4) MDL ratio 1.0226		GMM(3)	1.6248
		GMM(2)	1.4737
		2-means	1.4878
		PAM	1.1061
		ScanAlyze	0.9189

Table1. Comparative results for six microarray spots. The left and right columns show the spot area and the segmentation, respectively. Light, dark and medium gray indicate the F, B and A pixel location, respectively.

In the case of spots 1, 2 in Table 1 where no artifacts were found, the estimated fluorescence ratios are almost the same for *GMM(2)* or the two simple nearest-neighborhood techniques, 2-means and PAM. The four spots 3-6 represent cases where artifacts within the spot area are detected. Based on the *MDL* criterion in our methodology, the model with the three components is selected and the artifact component that holds the greater variability is removed. As observed in Table 1, the results are, in some cases, significantly different in terms of the calculated fluorescent ratios, when only two components are used for the clustering problem. For example, during the analysis of the spot 4, the *GMM(3)* model gives a greater by almost 50% gene expression value as compared to the one calculated for the other 2-cluster approaches. In this case a large artifact misleads the clustering algorithms when only two components are used since the artifact appears to be the foreground. This artifact explains the fluorescent ratio values (≈1.0) calculated by the *GMM(2),* 2-means and PAM method. The artifact region has almost equal R and G intensities as can be seen by its yellow color. Analogous observations can be made for spots 5, 6.

4. DISCUSSION AND CONCLUSIONS

In this paper we have proposed a new fully automated approach for the analysis of microarray images. Two are the main novelties of the proposed approach. First, the automatic method for finding the spot area based on the column and raw sums of the image intensities. Second, the *GMM*-based method for segmentation to F, B and A of the pixels in the spot area. This allows, for the first time to our knowledge, to identify and adjust the computation of the spots features to the artifacts which are present. The experiments demonstrated the ability of our approach to quantitatively measure the features of difficult spot areas, containing a large amount of artifacts. In addition, the proposed *GMM* clustering framework provides a very rich description of the spots properties and allow us to derive easily additional metrics which capture the morphology of the spot area. For example, the uniformity of the fluorescent intensities in the F or B areas is captured by the variances of the *GMM*. The overall color of the spot F or B areas are captured by the means and the covariances of the *GMM*. The evaluation of all these new metrics constitutes directions of our future research.

5. REFERENCES

[1] Y.H. Yang, M.J. Buckley, S. Duboit and T.P. Speed. "Comparison of Methods for Image Analysis on cDNA Microarray Data". *Technical Report* #584. Department of Statistics, University of California, Berkeley, 2000.

[2] D. Bozinov and J. Rahnenfuhrer, "Unsupervised Technique for Robust Target Separation and Analysis of DNA Microarray Spots Through Adaptive Pixel Clustering," *Bioinformatics*, vol.18 (5) pp.747-756, 2002.

[3] A.A. Alizadeh, M.B. Eisen, et. al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling". *Nature*, vol.403, pp.503-511, 2000.

[4] C.M. Bishop. *Neural Networks for Pattern Recognition*, Oxford Univ. Press Inc., New York, 1995.

[5] N. Vlassis and A. Likas, "A Greedy EM algorithm for Gaussian Mixture Learning". *Neural Processing Letters*, vol. 15, pp.77-87, 2002.

[6] J. Risannen. "Modelling by shortest data description". *Automatica*, vol.14, pp.465-471, 1978.

[7] N. Dempster, A.P. Laird and D. Rubin, "Maximum Likelihood from incomplete data via the EM algorithm". *J. R. Statist. Soc. B*, vol. 39, pp.1-38, 1977.

[8] M.B. Eisen. *ScanAlyze*. http://rana.Stanford.EDU/software, 1999.