BIOINFORMATICS

Vol. 18 no. 5 2002
Pages 747–756



Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering

Daniel Bozinov [1,*,†] *and Jörg Rahnenführer* [2,†]

[1]*Center for Human Molecular Genetics, University of Nebraska Medical Center, Omaha, NE 68198, USA and* [2]*Department of Statistics, University of California, Berkeley, CA 94720, USA*

Received on August 9, 2001; revised on November 8, 2001 and December 4, 2001; accepted on December 12, 2001

ABSTRACT

Motivation: Microarray images challenge existing analytical methods in many ways given that gene spots are often comprised of characteristic imperfections. Irregular contours, donut shapes, artifacts, and low or heterogeneous expression impair corresponding values for red and green intensities as well as their ratio R/G. New approaches are needed to ensure accurate data extraction from these images.

Results: Herein we introduce a novel method for intensity assessment of gene spots. The technique is based on clustering pixels of a target area into foreground and background. For this purpose we implemented two clustering algorithms derived from k-means and Partitioning Around Medoids (PAM), respectively. Results from the analysis of real gene spots indicate that our approach performs superior to other existing analytical methods. This is particularly true for spots generally considered as problematic due to imperfections or almost absent expression. Both PX_{PAM} and PX_{KMEANS} prove to be highly robust against various types of artifacts through adaptive partitioning, which more correctly assesses expression intensity values.

Availability: The implementation of this method is a combination of two complementary tools Extractiff (Java) and Pixclust (free statistical language R), which are available upon request from the authors.

Contact: dbozinov@unmc.edu

INTRODUCTION

With the emergence of gene expression microarray technology, a unique and very powerful instrument has been introduced to bioscientists. Suddenly the expression of thousands of genes can be simultaneously captured in space and time. Such snapshots of expression patterns for a large gene library or yet a whole genome can provide a synopsis of metabolic events in a specific cytological sample (Bittner *et al.*, 2000; Eisen *et al.*, 1998; Golub *et al.*, 1999; Perou *et al.*, 2000; Ross *et al.*, 2000). At the same time they can also provide more elementary information about the behavior of single genes. This duality makes microarray technology an exciting tool to the field of Functional Genomics. Scientists, however, are also facing substantial challenges in analyzing the overwhelming amount of projected data. Data acquisition as well as data analysis generally require many interdependent steps, each in itself representing a multi-dimensional optimization problem (Bozinov, 2001). Insufficient processing at any phase during a series of analyses can impair or even jeopardize an entire experiment. Hence, all subsequent processing steps depend heavily on the reliability and especially the validity of the previous data manipulation. In particular, the accuracy of intensity assessment for gene spots hybridized with two samples that have been labeled with Cy3 and Cy5, respectively, has the highest impact on the final results of any inference of microarray studies. Technical errors, which are introduced during initial measurements of gene expression, are hard and often impossible to compensate for at a later stage. Therefore it is critical to capture the information that is embedded in two TIFF images containing the excitation over gene spots at 532 and 635 nm as accurately as possible in their corresponding expression value. Any method that claims to generate representative expression values for gene spots must resolve many obstacles in successful image analysis endeavors. A new and presumably superior method has to be robust whenever the potential for erroneous measurement due to artifacts, less but perfect printing or poor hybridization does exist. The generated intensity values must reflect the actual expression as precisely as possible. An unsupervised approach where no arbitrary cutoffs are necessary to distinguish between

*To whom correspondence should be addressed.
† Both authors contributed equally to this manuscript.

© Oxford University Press 2002

foreground and background is desirable. Furthermore, this method should be feasible for complete automation in high-throughput tasks. Lastly, such a method should be dynamic and flexible enough to adapt to known as well as unforeseen problems in spots satisfactorily. A new technique for image analysis that accomplishes all of the above criteria and also performs effectively in real world applications would qualify as a new standard technique for microarray image analysis. In this paper we address these objectives by proposing a sophisticated and intriguingly intuitive means of extracting representative quantifiers for conventional as well as poor gene spots. Herein we will concentrate on the criteria by which pixels within a target area can be categorized as foreground (f) or background (b). This information is then used to extract the optimal value for red and green intensities and the R/G ratio of a spot. In the next section we will describe a novel concept for improved assessment of gene expression spots and discuss its implementation in two algorithmic ways. We will illustrate the advantages of this procedure by examining application results using a variety of real microarray spots of diverse levels of quality to demonstrate the analytical power of our method. Based on our examples we will point out some shortcomings of existing approaches for microarray image analysis. In our conclusion we ultimately propose an improved method for extracting representative scalars for cDNA microarray spots.

IMAGE ANALYSIS FOR MICROARRAYS

Problem statement

In microarray experiments thousands of DNA sequences are printed in a high-density array on a solid surface, typically a coated glass slide. Two individual heterogeneous mRNA samples are labeled with either a red-fluorescent dye Cy5 or a green-fluorescent dye Cy3, respectively (Schena *et al.*, 1995; Shalon *et al.*, 1996). They are then mixed and hybridized to the arrayed cDNA sequences. The relative abundance of all these mRNA sequences in the two samples is measured by differential hybridization. Utilizing a laser scanner, images of those slides are created, generating two 16-bit TIFF files with spatial information about the fluorescence measurements for red and green dye. The ratio of those two intensities for a particular gene spot describes the relative abundance of the corresponding mRNA sequences in the two target samples. This is our starting point. Subsequently we will refer to the final results of this process for the red-labeled samples as R values and for the green-labeled samples as G values. The phase of image processing which attains two values for intensities R and G and one value R/G of relative abundance for a single spot can be divided into three basic steps:

(1) Addressing: identifying the areas in an image that belong to spots. The combined area of a spot and its background is called the *target area*.

(2) Segmentation: partitioning the target area of every spot into distinct regions, usually foreground f (the spot itself) and background b.

(3) Reduction: extracting two scalar values for the intensities of R and G and assigning one value R/G for the relative abundance to each spot.

These three steps are usually performed independently. In this paper we will present a new method that algorithmically combines the second and the third step in a coherent way. The first step can be accomplished by either manual grid alignment or fully automated, 'since each element of an array is printed automatically to a pre-defined location [and] detectable signals [can be safely assumed to] form a regular array' (see microarray image analysis at www.nhgri.nih.gov). To automate, we decided to use a technique in which the entire microarray image is clustered into foreground and background and the marginal distributions of the cluster membership indices are used to determine spot rows and columns. Later, however, it will be shown that in our approach the addressing step and the resulting target area size only have minor impact on the final spot assessments. The second and third steps are much more interesting problems as they hold characteristic impediments which have to be treated appropriately.

Ideally every spot on a microarray has the shape of a circle with a diameter consistent with all the other spots of that microarray. Furthermore, the intensities of pixels assigned to foreground or background should be invariant across all pixels of the same category. Unfortunately, such an ideal situation is almost never the case (Brown *et al.*, 2001). By examining images from various sources such as the microarray core facilities at the University of California, Berkeley (UCB) and the University of Nebraska Medical Center (UNMC), we found that deviations from that ideal can be divided into four main problem categories:

(1) Variable size: different diameters.

(2) Variable contour: sickle shape, donut shape, oval or pear shape, scratched or interrupted shape (see Figure 4k).

(3) Normalization: high background and/or low foreground.

(4) Spatial artifacts: smeared or incorrectly segmented areas, caused e.g. by dirt on the slide or slide treatment.

Existing methods

Many image analysis methods have been adapted towards microarrays to deal with its specific problems. In general

they can be classified into spatial and distributional methods. Spatial methods try to capture the shape of a spot; distributional methods treat the pixels of a spot target area separately as they use the histogram of the pixel intensity values to discriminate between foreground and background. Hybrid approaches combine both ideas. Typical representatives of those classes have to be described in more detail as they address the four problem categories in different ways.

Simple shape methods assume circular spots. The fixed circle method is implemented in most of the common software packages, such as ScanAlyze (Eisen, 1999), GenePix (Axon Instruments Inc., 1999) and QuantArray (GSI Lumonics, 1999) all of which are widely accepted analysis tools. GenePix also offers the possibility to estimate the diameter of each circle separately in order to address problem no. 1. In ScanAlyze this can be done manually, which proves to be very tedious. All of these basic implementations ignore the occurrence of irregular shapes as stated in problem no. 2. Well-known adaptive shape segmentation methods are watershed (Beucher and Meyer, 1993; Vincent and Soille, 1991) and seeded region growing (Adams and Bischof, 1994). They select starting points and enlarge the detected spot area step by step until some criterion is reached. Seeded region growing has been implemented in the Spot software (Buckley, 2000), written in the free statistical programming language R by members of the Terry Speed microarray analysis group (http://www.stat.berkeley.EDU/users/terry/zarray/). For more details, refer to Yang *et al.* (2000). In the context of microarray analysis these methods prove to be very effective as their natural weakness of finding a suitable starting point can be neglected because it is always located in the known theoretical center of each spot.

Two distributional methods are implemented in QuantArray. The first is a statistical approach that uses a threshold value based on a Mann–Whitney test. It is described in Chen *et al.* (1997): initially eight pixels are selected at random out of a predefined background region. Eight target group pixels with the lowest intensities in the target area are then compared to them based on a rank-sum statistic; more specifically a Mann–Whitney test is applied with a fixed critical value such as 0.05 or 0.01. Iteratively the darkest of the target group pixels is replaced with the darkest not yet chosen until the test becomes statistically significant. Finally the eight current target group pixels and all lighter ones are assigned to the spot area. This method deals well with many difficulties, but it relies heavily on selecting a good background region. The other histogram method defines the background and the foreground as pixels with intensities between the 5th and 20th percentile and between the 80th and 95th percentile, respectively. This is also not adaptive since it doesn't consider the variable size of the spot at all. GenePix also defines the regression ratio. This is the slope of a linear regression line that is fitted to the scatter plot of the R and G values. The regression ratio is a very sensitive and thus weak measure, since it can be highly influenced by some outliers or even by a group of saturated pixels. For all these methods it is crucial that the target areas of the spots do not include large sections of other neighboring spots.

Both, purely spatial and purely distributional methods show significant shortcomings in problem categories 3 and 4. Problem no. 3 is assessed by most of the above methods by finding a suitable background region, which guarantees the correction for systematic errors due to noise signals on the entire glass slide. This can become a difficult task, for which many different choices for the background region of a spot have been suggested in the literature and by commercial and non-commercial implementations. All these methods suffer from a major defect. As soon as an artifact falls into the background area, the background value is significantly overestimated. Moreover, the intensity normalization problem is one important reason why generally microarray experiments are designed as relative measurement procedures for two DNA samples. Problem no. 4 is addressed by current methods with varying success. Shape segmentation methods are unable to find the spot when it is divided into more than one section. Weak signals pose a greater problem because there is no marked transition between foreground and background and thus no consistent shape information is generated for the spot area. In such a case, the spot area is often maximally enlarged, causing small artifacts to negatively influence the foreground results. On the other hand, pixel based methods can again be disadvantageous when neighboring spots are included in the area for a spot and its background.

A good example of a hybrid solution is ImaGene (BioDiscovery Inc., 1997). An initial circular shape estimation is the basis for a separation of the pixels into foreground and background. In the distributional representation a fixed interval can be chosen to eliminate those pixels that have extreme intensity values. This takes care of outliers caused by artifacts. However, the interval in this approach is arbitrary and non-adaptive from gene spot to gene spot. Nevertheless, hybrid type approaches in general are agreed to be the most promising solutions for this task. In the next paragraph we describe a method which overcomes all of the previously described limitations.

New approach: adaptive pixel clustering

A method that combines the shape independent advantages of histogram methods with some kind of attention towards foreground and background discrimination is ideal for this purpose. Thus we believe the most natural way to address the problem at hand is to use clustering algo-

rithms (Jain *et al.*, 1999). They allow for optimal detection of groups without the necessity of supervision. The input for the cluster algorithm are two-dimensional values (R_i, G_i) for pixel x_i. A suitable choice of the number of clusters K can be the most challenging part for other type of applications. Here we already know $K = 2$, due to the particular structure of the spot analysis. This task can be regarded as a discrete 2-center problem with the objective to identify two subsets with the smallest possible expansion (Agarwal *et al.*, 1997). For our purpose, the objective is rather to minimize the dispersion within the two subsets. In any case, the very constrained nature of microarray images favors successful image segmentation. General image segmentation techniques have been discussed by Haralick and Shapiro (1985).

Most mathematical cluster algorithms try to optimize an objective function. K-means for example, which has been extensively studied in the literature (MacQueen, 1967), looks for minimum variance partitions. The result of k-means are k cluster centers $m_1, \ldots, m_k$. Its objective function is the sum of the squared Euclidean distances of all data points to their closest of these centers. Like many algorithms k-means finds a local minimum of its objective. Bad local optima can be avoided through iterative application with random starting points. For details regarding statistical properties see e.g. Bock (1992). More robust multidimensional generalizations of k-means have recently been investigated by Pötzelberger and Strasser (2001). An extensive simulation study of Rahnenführer (2002) shows that these algorithms do not have the outliers' sensitivity of k-means, where single data points can strongly affect the outcome.

For the purpose of microarray image analysis, such an insensitive distance measure is necessary. Furthermore it would be advantageous if the method returned a real, representative pixel, whereas a mean value could be located in an area with no representatives at all. Also, the method should enable a quality measure to confirm that the target area truly contains two clusters.

Therefore, an eminent choice for clustering microarray images is Partitioning Around Medoids (PAM; Kaufman and Rousseeuw, 1990). PAM allows the choice of an arbitrary dissimilarity matrix for the data points and searches for k so-called medoids. These representatives of the clusters are data points themselves. Single bad pixels from regions with artifacts or imprecise fits of red and green area have minimal impact on the final result. Additionally, PAM provides a silhouette plot and a corresponding quality index allowing verification of the number of clusters (Struyf *et al.*, 1997).

The aim of PAM is to minimize the sum of distances of all observations to their closest medoid.

When $d_{ij} = d(x_i, x_j)$ denotes the distance between the observations x_i and x_j, PAM finds a local minimum of the objective function (O1)

$$\sum_{i=1}^{n} \min_{k=1\ldots K} d(x_i, m_k) \qquad \text{(O1)}$$

with the medoids $m_1, \ldots, m_k$. An outlier-insensitive choice for the dissimilarity matrix is based on the Manhattan distance, which is the sum of the absolute differences over all coordinates. By choosing the Euclidean distance instead, PAM and k-means would have the same objective. Thus, PAM with Manhattan distance is predisposed for the particular task of microarray image clustering.

Specification

The pixel clustering methods are defined as follows. Let R_i and G_i be the expression values for the ith pixel of a fixed spot and for the red and green dye, respectively.

DEFINITION. PX_{PAM} (pixel extraction with PAM).

0. *Calculation of dissimilarity matrix of spot pixels.* Calculate the Manhattan distances between all pairs of pixels. The distance d_{ij} between the pixels x_i and x_j is the sum of the absolute differences $d_{ij} = d(x_i, x_j) = |R_i - R_j| + |G_i - G_j|$.

1. *Construct two initial representatives (Build phase in PAM).* Define m_1 as the object with smallest $\Sigma_{i=1\ldots n} d(x_i, m_1)$ and m_2 as the object that decreases the objective (O1) with fixed m_1 as much as possible.

2. *Find local optimum of cluster problem (Swap phase in PAM).* Repeat the following step until convergence.

Consider all pairs of objects (i, j) with $i \in \{m_1, m_2\}$ and $j \notin \{m_1, m_2\}$ and perform the swap $i \leftrightarrow j$ (if any) which decreases the objective (O1) most. The outcome are two medoid pixels m_1 and m_2.

3. *Reduction.* Let R_{fg} and G_{fg} be the values of the medoid pixel with higher intensity values and R_{bg} and G_{bg} those of the other one. Calculate $(R_{\text{fg}} - R_{\text{bg}})/(G_{\text{fg}} - G_{\text{bg}})$ as the final relative abundance estimate.

DEFINITION. $\text{PX}_{\text{KMEANS}}$ (pixel extraction with k-means).

1. *Construct initial representatives.* Define starting midpoints $m_1 = (R_{\text{fg}}, G_{\text{fg}})$ and $m_2 = (R_{\text{bg}}, G_{\text{bg}})$, where R_{fg} and G_{fg} are values of the pixels with highest intensity values red and green and R_{bg} and G_{bg} the lowest ones.

2. *Find local optimum of cluster problem (k-means).* Repeat alternating the following two steps until convergence;

—assign each data point to its closest of the two midpoints;

—calculate two new midpoints as the means of all points assigned to the old midpoints, respectively.

The outcome are two cluster means m_1 and m_2.

3. *Reduction.* Let R_{fg} and G_{fg} be the values of the cluster mean with higher intensity values and R_{bg} and G_{bg} those of the other one. Calculate $(R_{fg} - R_{bg})/(G_{fg} - G_{bg})$ as the final relative abundance estimate.

PX_{PAM} as well as PX_{KMEANS} address both spot analysis steps, segmentation (2) and reduction (3). PX_{PAM} clusters the pixels in a robust way in foreground f and background b and takes the resulting medoids as best representatives for their clusters. For one-dimensional data it is well known that the median minimizes the sum of absolute distances to one point. If we took medians in step (3) instead of medoids after applying PAM we expect to get similar midpoints. In PX_{PAM} the third step of reduction is already resolved by addressing step 2. This is one of the advantages of using PAM for our application. The method provides an algorithmic combination of the two usually independent image analysis steps 'segmentation' and 'reduction'.

Refinement

Both algorithms perform well in most gene spot cases. However, in order to cover more spots and become more reliable and practical, certain adjustments are still necessary. Especially PX_{KMEANS} can be tuned such that it performs similar to PX_{PAM}. The reason for favoring k-means over PAM is essentially the time complexity of PAM. PAM provides all attributes for clustering gene spot pixels. It intrinsically returns realistic medoids rather than potentially hypothetical values. It is very robust against outliers and against collapses into one cluster. It also offers a silhouette plot as a tool to verify the actually existent number of clusters. But the price for this scientific clarity of PAM is its $O(n^2)$ execution time, due to the tremendous amount of spots on a microarray. An implementation of the underlying concept with a very low execution time can be achieved by using the PX_{KMEANS} algorithm, which has a time complexity of $O(rkn)$. For appropriate robustness a significant adjustment is necessary though. Spatial medians instead of means are chosen as cluster representatives. We recognize that it is somehow counterintuitive to first minimize an objective based on squared distances and then calculate the midpoints within the resulting groups as spatial medians. However, we found that for microarray spot data PX_{PAM} and the refined PX_{KMEANS} yield very similar results, as indicated in Figure 4a–l. This stability is a good indicator of the validity of the outcome. In a direct comparison on a dataset of 15 000 gene spots PX_{PAM} required 5.4 h, while PX_{KMEANS} needed only 27.7 s. Another attempt to maintain the features of PX_{PAM} and also reduce the time complexity replaces PAM with CLARA (Struyf *et al.*, 1997). Although very intriguing, this approach lacks stability. The crucial common ingredient in all of these methods is the adaptive step of discriminating between foreground f and background b by clustering the pixels into two groups.

In PX_{KMEANS}, instead of random starting points for the two clusters we use the maximal and minimal observed values in the target area for red and green intensities, respectively. This provides a more meaningful initial representation of foreground and background and assures the convergence to an adequate optimum, usually the global minimum. This adjustment further avoids the immediate termination of the algorithm by drawing the same values twice. It has to be stressed that in PX_{PAM} and PX_{KMEANS} convergence is guaranteed, as the objective function decreases in every step until no further progress terminates the algorithm.

For the rarely observed but not unrealistic case of 'black holes' an additional refinement can be implemented. Here, the spot intensities are below background intensities, thus our shape independent method could swap foreground and background. Consequently, an appropriate measure is to flag all spots where dispersion of pixels in the xy-space of the presumed foreground area is higher than the one of the presumed background area.

The pixel clustering approach can be considered as a distributional method with some attention towards the identification of a confined shape, the gene spot. In order to further emphasize the influence of a presumed gene spot region, PX_{PAM} and PX_{KMEANS} can be incorporated into a hybrid solution, where the accumulation of all target clustering results can be used as a subsequent 'Bayesian'-filter for each target area separately. The two identified clusters in a target area can be further refined by subsequently applying another spatial convolution. Smoothing would result in reassigning those pixels that belong to a different cluster than the majority of their adjacent pixels.

RESULTS

Every target area that contains the pixel information for a single gene spot and its adjacent background, undergoes a series of transformations in order to extract final values for red and green intensities. The compartmentalization of a candidate spot including nearby background pixels is illustrated in Figure 1. All pixels within that target area are then clustered and plotted by their R and G values into a two dimensional cluster-plot (Figure 2). It is important to point out that each pixel keeps a label with its original x and y coordinates. Once the pixels are separated into two classes f and b, their representative medians are chosen to draw a linear connection. Its slope visualizes the overall ratio of R and G for this particular gene spot. This strategy appears to be a very effective and robust

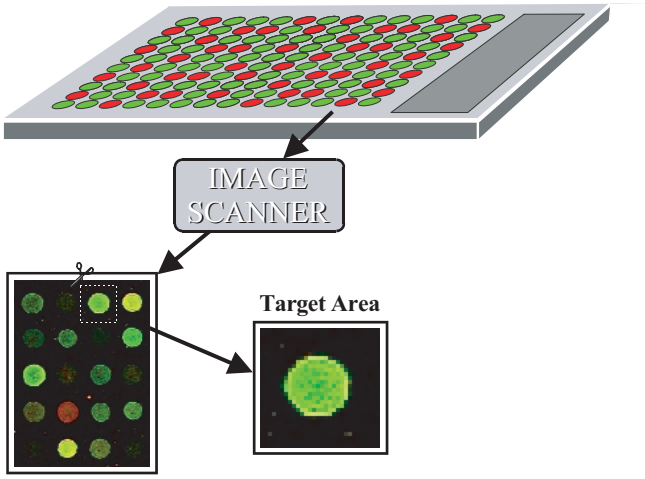


Fig. 1. Addressing of a target area that consists of a candidate gene spot and surrounding background pixels.

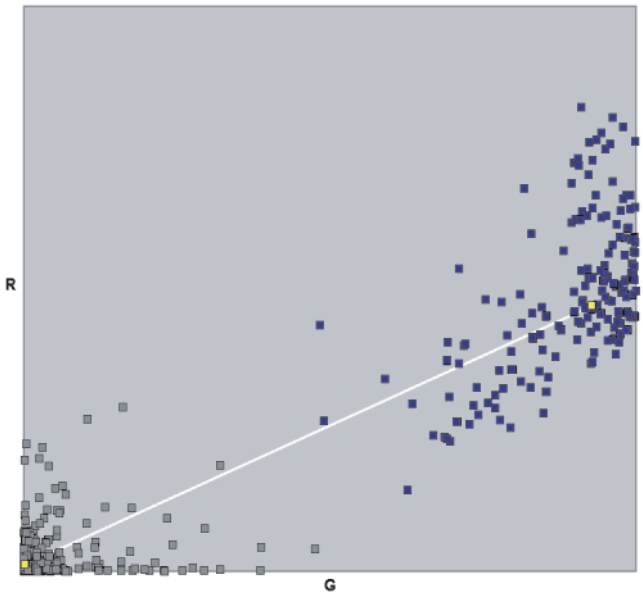


Fig. 2. R versus G plot of all pixels after clustering. Pixels with high intensity levels are separated from background pixels. The slope of the connection between the representative medoids of the two clusters reflects the total gene spot ratio R/G.

way to quantify total gene spot expression. The spatial bivalence plot as a visual feedback of the clustering results is strong support for the accuracy of this method. It is undoubtedly useful to give consideration to the assessment of classification results (Gordon, 1996). The bivalence plot (Figure 3) essentially reorganizes the pixels based on their xy-labels. The shade in the bivalence plot can alter only between white and black. A pixel that has been categorized to belong to the foreground cluster will be displayed as a white dot, while a pixel that the clustering algorithm considered as background will here become a black dot. It is very evident that the clustering approach divides the pixels in a meaningful manner, as the bivalent distribution of the pixels between the two clusters clearly resembles the shape of the gene spot. Similar success can be demonstrated for other cases in the following section.

Case study

A favorable way of concept validation for microarray image analysis is to process realistic data and review the outcome. Therefore twelve real microarray spots were chosen, which best illustrate the analytical power of our approach. It is important to point out that these spots are neither special nor obscure cases but rather represent one or more typical problem categories that are often encountered in microarray images. Every spot consists of a target window of 30×30 pixels. This is an arbitrary size that in our case best fit the grid of the spotted arrays. Later, it will be shown that the window size does not influence the final result for R, G and R/G.

The first spot in the list of Figure 4 is a simple well-formed spot with predominance in the red spectrum. All of the existing methods mentioned previously should deliver acceptable results in the range $R/G = 3.5$, meaning that the measured intensity for red is about 3.5 times larger than the intensity for green. On the R/G-cluster plot [ii] one can easily see that the pixels of the target area are clearly separated into two reasonable groups with separation into high and low intensity pixels. The spatial biva-

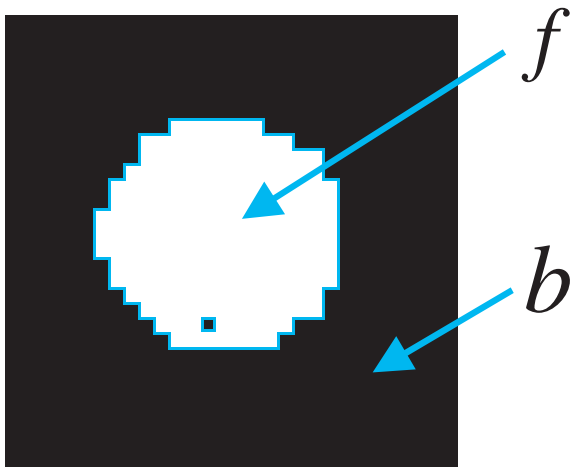


Fig. 3. Spatial bivalence plot based on the xy-labels of each pixel. White dots represents the foreground cluster, while black dots represent the background cluster.

lence plot [iii] of the two clusters identifies a white zone as the correct foreground f and a black zone as the remaining background b. The minor artifact in the lower left corner is relatively weak and therefore still categorized as background as one can verify in the corresponding bivalence plot. Figure 4b shows a green spot with a yellow artifact that potentially could bias the background intensity in methods with a predefined background area. The clustering algorithm considers the artifact as foreground where its influence is negligible, see Figure 4b[ii]. Figure 4c[i]–[iii] illustrates how the clustering successfully manages so-called donut shaped spots. The dark area located in the center, which occasionally appears due to retrieval of the microarrayer's spotting pin, is correctly categorized as background b. If those black pixels were classified as foreground, they would 'darken' the final intensity value of that spot. This would be the case for both fixed circle methods as well as seeded region growing. In the following three examples 4d–f the contours are not round but rather pear-shaped. Additionally, the pixels of spot 4e are heterogeneous in their intensity. All three shapes are correctly identified by PX_{PAM} as well as PX_{KMEANS}. The heterogeneous pixel intensities in 4e are inherently resolved by PX_{PAM} as it returns only the medoid of the two clusters f and b. In PX_{KMEANS} this is accomplished by returning the medians of R and G for both clusters. Irregular shapes like these would most likely impair the results acquired with fixed circle methods. Figure 4g displays a spot with an artificial hole within its structure. This black region as well as the missing regions in the sickle-shaped spot 4h and the missing 'bite' in 4i are categorized as background and therefore do not weaken the analysis of the foreground region at all. Figures 4j and k are particularly interesting examples. Figure 4j shows a large artifact attached to the spot which mimics expression in the equilibrium of R and G. Without any pre-assumptions of where to expect the foreground region, our clustering algorithms will consequently determine the foreground region as the combination of the green spot and the yellow artifact. But as they return the medoid or the median, respectively, the artifact is considered as a large section of outliers and therefore doesn't influence the final result. This is exactly the strength of PX_{PAM} and PX_{KMEANS} where the previously mentioned step 2 (segmentation) and step 3 (reduction) are addressed in a robust manner. Figure 4k shows a spot with a scratch dividing it into several parts. To our best knowledge neither of the existing methods will handle spots 4j and k effectively. They are either influenced by the large yellow artifact or will not successfully identify the region of interest within the target area of both 4j and especially k. The last spot 4l essentially illustrates that the clustering algorithm PX_{PAM} (or PX_{KMEANS}) also works well for spots with very low overall expression. The two clusters f and b are certainly closer to each other but still easily separated by our clustering approach. This is shown in the R/G plot of Figure 4l[ii]. As previously indicated the size of the target area does not significantly influence the final outcome of our method. To demonstrate this we have selected one gene spot and applied our clustering algorithms PX_{PAM} and PX_{KMEANS} to examine the possible effect of variable image size on the intensities R and G. Figures 5a–c show identical results although the pixel matrices for the spot are 25×25, 30×30 and 35×35, respectively. Moreover, in Figure 5d one can see that even fractions of neighboring spots that accidentally entered a 40×40 target area have minimal influence on the final result, as long as the real spot is significantly larger than the artifact and of sufficient contrast. Analysis based on the same three window sizes has been applied to 65 additional gene spots to generate an 'M versus A' plot, which illustrates the effect of variable target areas. Here, logged R/G ratios (M) are plotted against the corresponding overall gene intensities (A). Each data point represents a gene at a target size of 30×30, whereas potentially two appendices visualize the deviations for 25×25 and 35×35 pixels, respectively. The clear observation that gene spots with higher signal-to-noise ratio are conceivably resilient to inconsistent values resulting from target size variation draws the attention to the lower intensity spectrum of this dataset. Even at this spectral end the extracted values remain surprisingly stable. Furthermore the resulting ratios R/G are barely affected by the size variation. In this context it has to be stressed that in real experiments those spots are more likely to be flagged because of other reasons such as their insufficient quality due to the presence of substantial artifacts. For the pixel clustering approach the choice of a good target size is advisable but not critical, to generate reliable intensity values.

CONCLUSION

A novel technique for microarray image analysis has been described, based on the idea of discriminating a target area of one gene spot into foreground pixels and background pixels by clustering. For this purpose we are proposing a highly robust pixel-extracting version of PAMs that intrinsically combines the two independent steps of 'segmentation' and 'reduction' in one algorithmic procedure. A more time efficient version based on k-means was also developed. In both PX_{PAM} and PX_{KMEANS} the number of clusters is known to be $k = 2$ reflecting foreground f and background b, simplifying the clustering procedure. Both clustering algorithms were able to cope with exemplary spots that contain one or more commonly encountered imperfections impeding analysis. The clustering results have been displayed in R/G plots as well as spatial bivalence plots to demonstrate the analytical power of the technique. Variability regarding shape or pixel intensity and even artifacts encapsulated within the spots are dynamically iso-

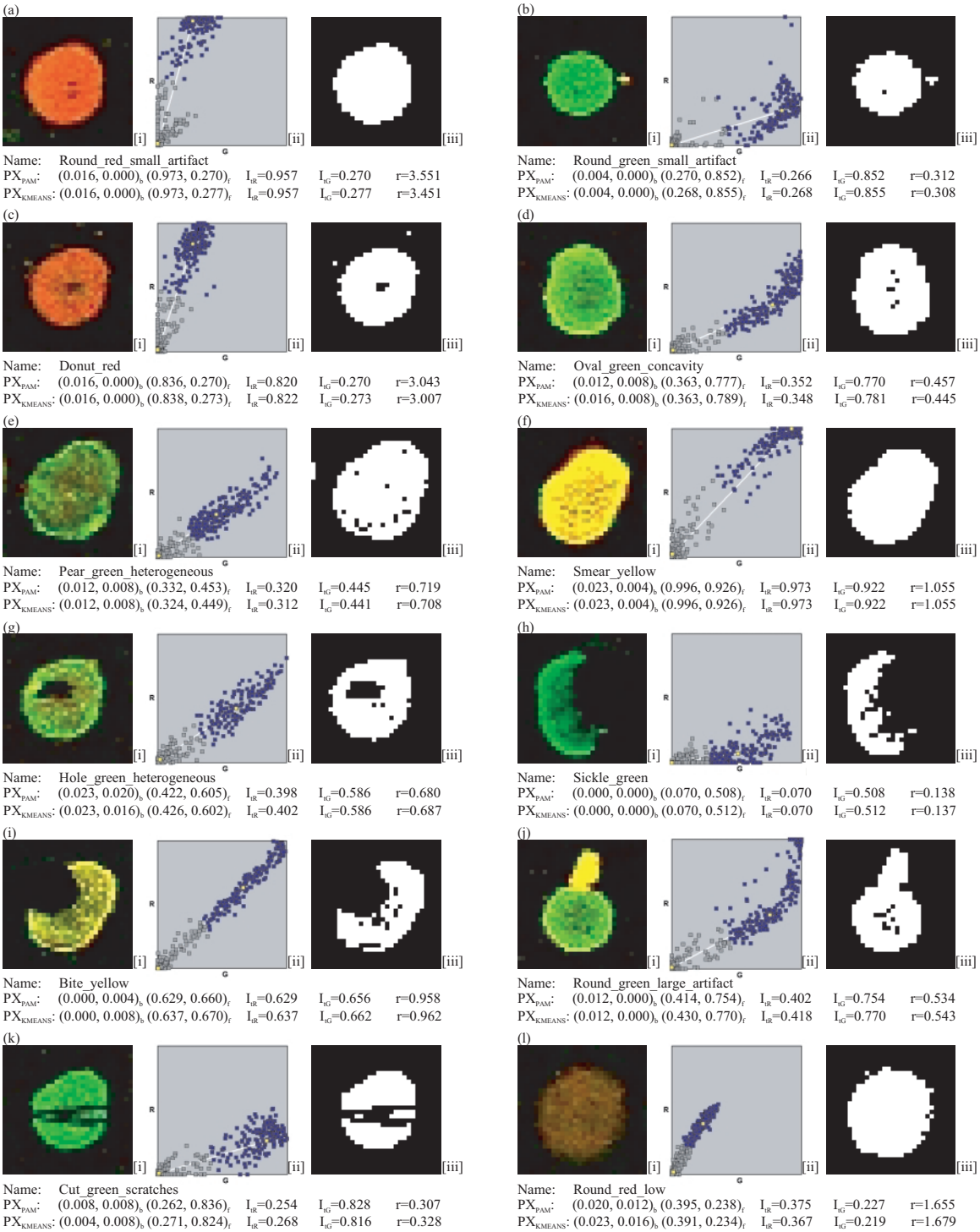
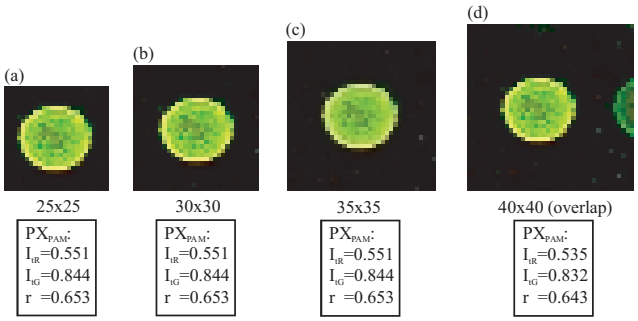


Fig. 4. Various gene spots (plots [i]) with one or more imperfections for analysis. Plot [ii] displays all pixels of a target area after clustering them into foreground f (blue) and background b (gray). The bivalence plot [iii] shows the two clusters f and b in xy-alignment. For the algorithms $\mathrm{PX_{PAM}}$ and $\mathrm{PX_{KMEANS}}$ the relative values of the cluster midpoints as well as the total intensities for red (I_{tR}) and green (I_{tG}) are given. The final relative abundance r is the ratio of $I_{\mathrm{tR}}/I_{\mathrm{tG}}$.

I.



II.

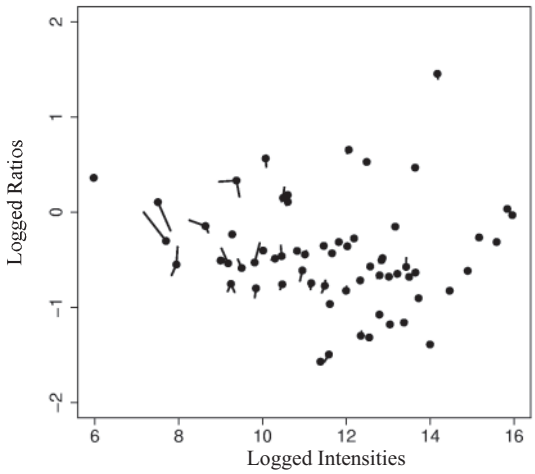


Fig. 5. (I) Different window sizes for the target area have been selected (a–c). The final results for intensities and ratio remain constant. Even overlapping of a neighboring spot (d) only minimally changes the outcome. (II) M versus A—plot of 65 gene spots. The results for window size 30 × 30 are plotted as data points, whereas two possible appendices visualize the deviations for 25 × 25 and 35 × 35 pixels.

lated and their potential impact on final results is eliminated. As has been illustrated, the physical size of a target area does not influence the final results for intensities R and G of a spot, more evidence for the robustness of our technique. PX_{PAM} and PX_{KMEANS} have proven superiority over other existing approaches in many commonly known problem areas. They produce reliable values for weak gene spots, where other approaches often perform insufficiently.

Future work

As with any method, our approach to cluster pixels of a target area has revealed certain shortcomings. Spots with very low expression and considerably large artifacts in the background region could mislead the clustering algorithm. In this case the artifact would appear to be the foreground, while the actual gene spot in combination with the real background region would be considered as background, as is illustrated in Figure 6. Two possible strategies are anticipated for future endeavors. The first is to allow the clustering algorithm to partition the same target area in parallel into two and three classes. The number of clusters will be determined based on the silhouette plot. In the case of three clusters the non-background cluster with a lower amount of elements will be considered as an artifact while the other non-background cluster would represent the foreground region. Another strategy would be to develop the previously proposed hybrid version based on the accumulation of all target clustering results to generate a bivalence mask. This bivalence mask would represent a meaningful filter of the presumed location of the foreground region within the target area. Both possibilities are currently being considered for further investigation.

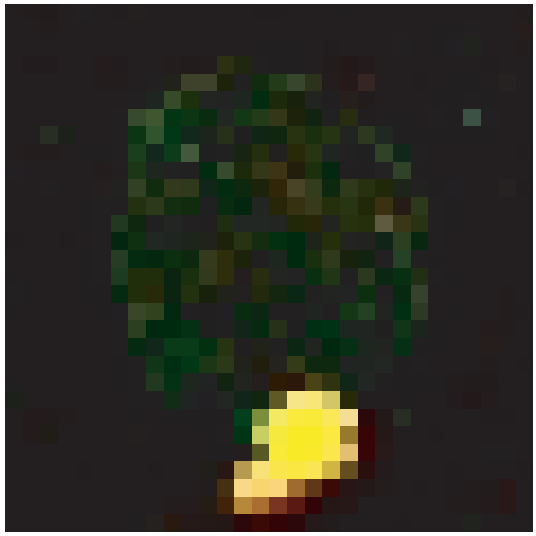


Fig. 6. Low intensity spot with high intensity artifact.

ACKNOWLEDGEMENTS

We would like to thank Cathy Burson and Jim Eudy for carefully editing the manuscript. This work is partially supported by the German Research Foundation (DFG) (JR, RA 870/2-1), by Nebraska Research Initiative grant 31-3205-0502 (DB) and NSF grant #0091900 (Nebraska Informatics Center for the Life Sciences).

REFERENCES

Adams,R. and Bischof,L. (1994) Seeded region growing. *IEEE Trans. Pattern Anal. Mach. Intell.*, **16**, 641–647.

Agarwal,P., Sharir,M. and Welzl,E. (1997) The discrete 2-center problem. In *Proceedings of the 13th ACM Symposium on Computational Geometry*. pp. 147–155.

Axon Instruments Inc. (1999) *GenePix 4000A User's Guide*.

Beucher,S. and Meyer,F. (1993) The morphological approach to segmentation: the watershed transformation. In *Mathematical Morphology in Image Processing, Volume thirty fourth of Optics Engineering*, chapter 12, Marcel Dekker, New York, pp. 433–481.

BioDiscovery Inc. (1997) *ImaGene*. http://www.biodiscovery.com/imagene.asp.

Bittner,M., Meltzer,P., Chen,Y., Jiang,Y., Seftor,E., Hendrix,M., Radmacher,M., Simon,R., Yakhini,Z., Ben-Dor,A., Sampas,N., Dougherty,E., Wang,E., Marincola,F., Gooden,C., Lueders,J., Glatfelter,A., Pollock,P., Carpten,J., Gillanders,E., Leja,D., Dietrich,K., Beaudry,C., Berens,M., Alberts,D. and Sondak,V. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Bock,H.H. (1992) A clustering technique for maximizing ϕ-divergence, noncentrality and discriminating power. In Schader,M. (ed.), *Analyzing and Modeling Data and Knowledge*. Springer, New York, pp. 19–36.

Bozinov,D. (2001) Development of a novel software tool for high-throughput analysis and interpretation of gene expression microarray studies. *Beyond Genome 2001, Bioinformatics and Genome Research Conference*. San Francisco, CA.

Brown,C.S., Goodwin,P.C. and Sorger,P.K. (2001) Image metrics in the statistical analysis of DNA microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 8944–8949.

Buckley,M.J. (2000) The spot user's guide. *CSIRO Mathematical and Information Sciences*, http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm.

Chen,Y., Dougherty,E.R. and Bittner,M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biom. Opt.*, **2**, 364–374.

Duggan,D.J., Bittner,M., Chen,Y., Meltzer,P. and Trent,J.M. (1999) Expression profiling using cDNA microarrays. *Nature Genet.*, **21** (Suppl.).

Eisen,M.B. (1999) *ScanAlyze*. http://rana.Stanford.EDU/software/.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Gordon,A.D. (1996) A survey of constrained classification. *Comput. Stat. Data Anal.*, **21**, 17–29.

GSI Lumonics (1999) *QuantArray Analysis Software, Operator's Manual*.

Haralick,R.M. and Shapiro,L.G. (1985) Image segmentation techniques. *Computer Vision, Graphics and Image Processing*, **29**, 100–132.

Jain,A.K., Murty,M.N. and Flynn,P.J. (1999) Data clustering: a review. *ACM Computing Surveys*, **31**, 264–323.

Kaufman,L. and Rousseeuw,P.J. (1990) *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley, New York.

MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations. *Proc. 5th Berkeley Sumposium on Math. Stat. and Prob*.

Perou,C.M., Sorlie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H., Akslen,L.A., Fluge,O., Pergamenschikov,A., Williams,C.F., Zhu,S.X., Lonning,P.E., Borresen-Dale,A.L., Brown,P.O. and Botstein,D. (2000) Molecular portraits of human breast tumors. *Nature*, **406**, 747–752.

Pötzelberger,K. and Strasser,H. (2001) Clustering and quantization by MSP-partitions. *Stat. Decisions*, **19**, 331–372.

Rahnenführer,J. (2002) Multivariate permutation tests for the k-sample problem with clustered data. Report Series 35, SFB 010, Vienna University of Economics and Business Administration, Statistical Institute, to appear in *Computational Statistics* 2/2002.

Ross,D.T., Scherf,U., Eisen,M.B., Perou,C.M., Rees,C., Spellman,P., Iyer,V., Jeffrey,S.S., van de Rijn,M., Waltham,M., Pergamenschikov,A., Lee,J.C., Lashkari,D., Shalon,D., Myers,T.G., Weinstein,J.N., Botstein,D. and Brown,P.O. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.*, **24**, 227–235.

Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a DNA microarray. *Science*, **210**, 467–470.

Shalon,D., Smith,S.J. and Brown,P.O. (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.*, **6**, 639–645.

Soille,P. (1999) *Morphological Image Analysis: Principles and Applications*. Springer, New York.

Strasser,H. (2000) Reduction of complexity. *Technical Report*. Department of Statistics, Vienna University for Economics and Business Administration.

Struyf,A., Hubert,M. and Rousseeuw,P.J. (1997) Integrating robust clustering techniques in S-PLUS. *Comput. Stat. Data Anal.*, **26**, 17–37.

Vincent,L. and Soille,P. (1991) Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans. Patt. Anal. Mach. Intel.*, **13**, 583–598.

Yang,Y.H., Buckley,M.J., Dudoit,S. and Speed,T.P. (2000) Comparison of methods for image analysis on cDNA microarray data. *Technical Report #*584. Department of Statistics, University of California, Berkeley.