Chapter XIX
Image Processing and Machine Learning Techniques for the Segmentation of cDNA Microarray Images

Nikolaos Giannakeas
University of Ioannina, Greece

Dimitrios I. Fotiadis
University of Ioannina, Greece

ABSTRACT

Microarray technology allows the comprehensive measurement of the expression level of many genes simultaneously on a common substrate. Typical applications of microarrays include the quantification of expression profiles of a system under different experimental conditions, or expression profile comparisons of two systems for one or more conditions. Microarray image analysis is a crucial step in the analysis of microarray data. In this chapter an extensive overview of the segmentation of the microarray image is presented. Methods already presented in the literature are classified into two main categories: methods which are based on image processing techniques and those which are based on Machine learning techniques. A novel classification-based application for the segmentation is also presented to demonstrate efficiency.

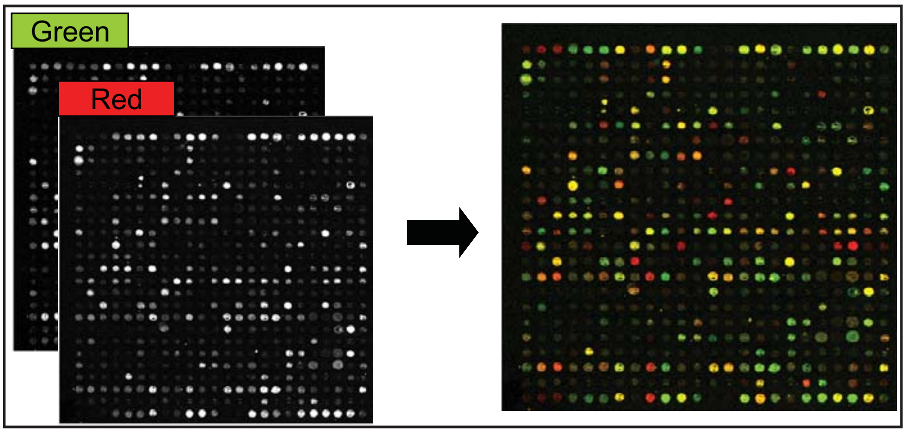
INTRODUCTION

Several types of microarrays have been developed to address different biological processes: (i) cDNA microarrays (Eisen, 1999) are used for the monitoring of the gene expression levels to study the effects of certain treatments, diseases, and developmental stages on gene expression. As a result, microarray gene expression profiling can be used to identify disease genes by comparing gene expression in diseased and normal cells. (ii) Comparative genomic hybridization application assesses genome content in different cells or closely related organisms (Pollack et al., 1999). (iii) SNP detection arrays identify single nucleotide polymorphism among alleles within or between populations (Moran & Whitney, 2004). (iv) Finally, Chromatin immunoprecipitation (chIP) technologies determine protein binding site occupancy throughout the genome, employing ChIP-on-chip technology (Buck & Lieb, 2004).

The experiment of cDNA microarrays typically starts by taking two biological tissues and extracting their mRNA. The mRNA samples are reverse transcribed into complementary DNA (cDNA) and labelled with fluorescent dyes resulting in a fluorescence-tagged cDNA. The most common dyes for tagging cDNA are the red fluorescent dye Cy5 (emission from 630-660 nm) and the green-fluorescent dye Cy3 (emission from 510-550 nm). Next, the tagged cDNA copy, called the sample probe, is hybridized on a slide containing a grid or array of single-stranded cDNAs called probes. Probes are usually known genes of interest which were printed on a glass microscope slide by a robotic arrayer. According to the hybridization principles, a sample probe will only hybridize with its complementary probe. The probe-sample hybridization process on a microarray typically occurs after several hours. All unhybridized sample probes are then washed off and the microarray is scanned twice, at different wavelengths corresponding to the different dyes used in the assay. The digital image scanner records the intensity level at each grid location producing two greyscale images. The intensity level is correlated with the absolute amount of RNA in the original sample, and thus, the expression level of the gene associated with this RNA.

Automated quantification of gene expression levels is realized analyzing the microarray images. Microarray images contain several blocks (or subgrids) which consist of a number of spots, placed in rows and columns (Fig. 1). The level of intensity of each spot represents the amount of sample which is hybridized with the corresponding gene. The processing of microarray images (Schena et al., 1995) includes three stages: initially, spots and blocks are preliminarily located from the images (gridding). Second, using the available gridding information, each microarray spot is individually segmented into foreground and background. Finally, intensity extraction,

Figure 1. A Block of a Typical Microarray Image



calculates the foreground fluorescence intensity, which represents each gene expression level, and the background intensities. Ideally, the image analysis would be a rather trivial process, if all the spots had circular shape, similar size, and the background was noise and artefact free. However, a scanned microarray image has none of the above characteristics, thus microarray image analysis becomes a difficult task. In this chapter, we describe several microarray segmentation algorithms based on image processing and machine learning techniques.

BACKGROUND

Resent studies in microarrays have been shown that segmentation methods can significantly influence microarray data precision (Ahmed et al., 2004). Several methods have been proposed for the segmentation of microarray images. These methods can be classified into four categories: (i) Fixed and adaptive circle, (ii) histogram-based, (iii) adaptive shape, (iv) clustering. The first three categories are based on image processing techniques, while the fourth is based on machine learning. Fig. 2 shows an overview of the already developed methods for microarray image segmentation. Earliest approaches fit a circle (with fixed or adaptive size) around each spot, characterizing the pixels in the circle as signal pixels and the pixels out of the circle as background pixels. Such an approach is used by Scanalyse (Eisen, 1999) and Dapple (Buhler et al., 2000). Histogram-based techniques estimate a threshold (GSI Lumonics, 1999; Chen et al., 1997), such that pixels with intensity lower than the calculated threshold are characterized as background pixels, whereas pixels with higher intensity as signal pixels. The adaptive shape segmentation methods are usually based on the Watershed Transform (Siddiqui et al., 2002) and the Seed Region Growing algorithm (Buckley, 2000; Wang et al., 2001). The most recent techniques employ clustering algorithms such as K-means (Bozinov & Rahnenführer, 2002; Ergüt et al., 2003; Wu & Yan, 2004), Fuzzy C-Means (FCM) (Ergüt et al., 2003), Expectation-Maximization (EM) (Blekas et al., 2005) and Partitioning Around Medoid (PAM) (Nagarajan, 2003). A hybrid method (Rahnenführer & Bozinov, 2004) which engages Image Processing and Machine learning techniques has been proposed. In this chapter we address a pixel by pixel classification approach for the segmentation of microarray images. The current application, which is presented in section 4, classifies the pixels of the image into two categories (foreground and background) using the Bayes classifier. Already developed clustering techniques generate groups of pixels, characterizing these pixels as signal or background using a set of rules, i.e. the group with the maximum mean intensity value is characterized as signal. Instead of this, the current approach directly classifies each pixel to the designated category.

MICROARRAY SEGMENTATION METHODS

Image Processing Techniques

Fixed or Adaptive Circle Segmentation

Fixed circle segmentation is the earliest method developed for microarray image analysis. This algorithm is implemented by Eisen et al, (Eisen, 1999) and it is included in the ScanAnalyze software tool. The method assumes that all the spots are circular with a constant radius. A circular mask of a fixed radius, called target mask, is placed on each spot location, considering all the pixels inside the mask as foreground pixels. On the other hand, background contains any external pixel which is close to the corresponding spot. Fig. 3 shows the way that masks are placed on each spot using the ScanAlyze software.

The elimination of the constant radius assumption was the next step in microarray image analysis studies, generating the adaptive circle algorithm. Assuming that the shape of all spots is circular, the radius for each spot is automatically estimated or manually adjusted by the user, for each spot. For instance, Dapple (Buhler et al., 2000) estimates the radius of the spot using the Laplacian-based edge detection. The manual approaches are extremely difficult and time consuming due to the large amount of microarray spots contained in a single image.

Histogram-Based Segmentation

The methods inn this category are based on the histogram of the image. Histogram-based approaches fix a circular mask on each spot, which is larger than the spot size, and then a threshold value of the pixel intensity is computed to seperate the foreground and background pixels within the mask. QuantArray software (GSI Lumonics, 1999) calculates the threshold globally from the

Figure 2. Current status of microarray image segmentation methods.

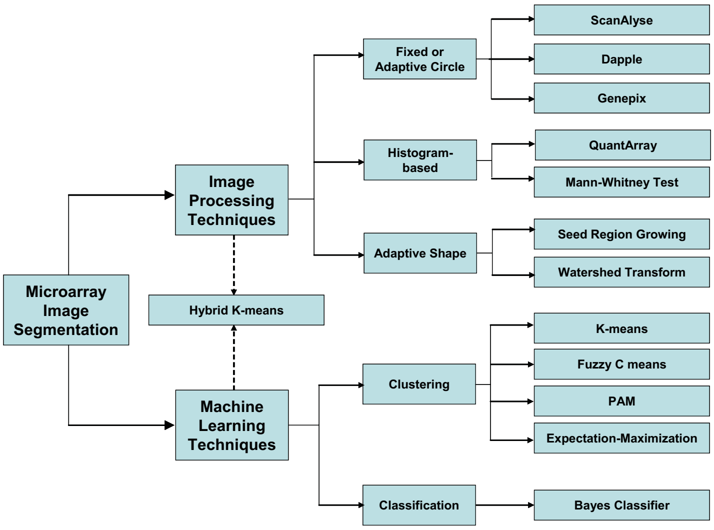
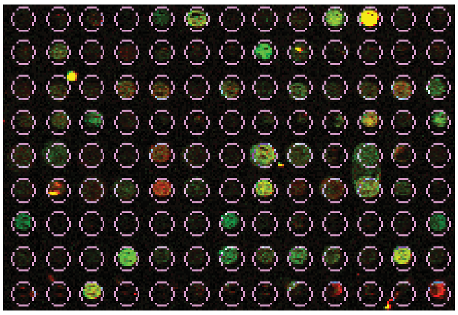


Figure 3. ScanAlyze image segmentation



histogram for all pixels within the masked area, while UCSF Spot software (Jain et al., 2002) estimates locally the threshold using the percentile cut-off. These methods can be unstable since the circular mask could be too large and it can cover neighbouring spots. In addition, adaptive thresholding segmentation methods do not provide the expected results when the signal is weak because there is no marked transition between foreground and background.

To overcome these difficulties, Chen et al. proposed a pixel selection method based on the Mann-Whitney test (Mann & Whitney, 1947). The Mann-Whitney test is a non-parametric statistical test for assessing the statistical significance of the difference between two distributions. The method associates a confidence level with every intensity measurement based on the significance level of the test. Each iteration of the algorithm calculates the threshold value between foreground and background pixels. At the first step, a circular target mask is placed in order to enclose all possible foreground pixels separating them from the background. A set of pixels from the background is randomly selected and compared against the pixels with the lowest intensity within the target mask. If the difference between the two sets is not significant, the algorithm discards some predetermined number of pixels from the target area and selects new pixels for investigation. Each iteration ends when the two sets significantly differ from each other, and the signal pixels are considered as the pixels remaining inside the target mask.

Adaptive Shape Segmentation

More sophisticated image processing techniques comprise the adaptive shape segmentation. These methods include no assumption on the size and the shape of the spot. The Seed Region Growing (SRG) algorithm (Adams & Bischof, 1994) selects randomly a small set of pixels, called seeds, as the initial points of a region in the area of each spot. At each iteration, the algorithm considers simultaneously the neighbouring pixels of every region grown from a seed. The neighbouring pixels are ordered under several criteria. The most common criterion uses only the intensity of the neighbouring pixels and the mean intensity of the growing region. This criterion C is defined as:

$$C(i) = \left| I(i) - \overline{I_s} \right|, \tag{1}$$

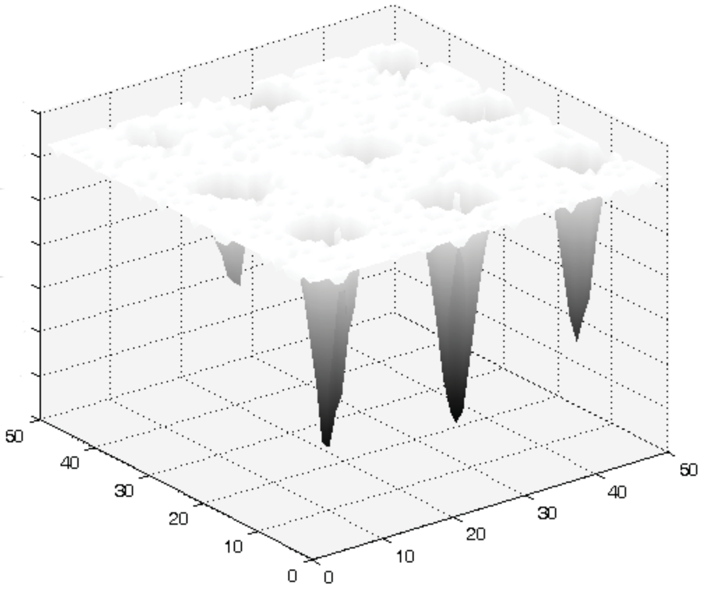
where S refers to the growing region, $I(i)$ is the intensity of the pixel i which is neighbour to the region S, and $\overline{I_s}$ is the mean intensity of the growing region S at the current iteration.

Next, the algorithm finds the region's neighbouring pixel m which corresponds to the minimum criterion C and labels it as a region pixel or as a boundary pixel between two regions. If all the neighbours of m belong to a single region, m is also labelled to the region. Therefore if m has only one neighbour that belongs to another region, m is marked as boundary pixel. The algorithm iterates until all pixels have been assigned to a region or labelled as boundary pixels.

Another adaptive shape segmentation method is based on the Watershed Transform (WT) (Roerdink & Meijster, 2000). WT is a popular segmentation method which originates from mathematical morphology. The image is considered as a topographical relief, where the height of each point is related to its grey level, resulting to a topological relief with many wells in the position where spots are located, as it is shown in Fig. 4.

Imaginary rain falls on the terrain. The watersheds are the lines separating the catchment basins. The output of the watershed algorithm is a tessellation of the input image into its different catchment basins, each one characterized by a unique label. The pixels that belong to the watershed lines are assigned a special label. Siddiqui et al. (Siddiqui et al., 2002) had implemented a segmentation method where the watershed transform is applied to the gradient of the image (Serra, 1982) and not to the original one. The gradient operator is very sensitive to grayscale variation

Figure 4. A microarray image represented as a topographical relief



and noise and it can cause development of a large number of irrelevant catchment basins. However, these oversegmentation problems can be overcome using watershed transform techniques without any additional computational effort.

Machine Learning Based Segmentation

Nowadays, traditional image processing techniques have to cope with more complex problems in the fields of medical informatics and bioinformatics. The integration of machine learning in image processing is very likely to have a great benefit to the field, which will contribute to a better analysis of medical and biological data. Microarray image analysis methods have already employed several machine learning techniques in order to segment the image. Clustering is the most common technique that is used for the segmentation of the microarray images. The idea of the clustering application is to divide the pixels of the image into several clusters (usually two clusters) and then to characterize these clusters as signal or background. Clustering algorithms, such as K-means, Fuzzy C-Means, Expectation-Maximization etc. have been employed in several microarray imaging segmentation studies.

The K-means segmentation algorithm is based on the traditional K-means clustering (MacQueen, 1967). It employs a square-error criterion, which is calculated for each of the two clusters. The square error criterion for the two clusters is given as:

$$E^2 = \sum_{k=1}^{2} e_k^2, \qquad (2)$$

where

$$e_k^2 = \sum_{i=1}^{n_k} \left(x_{ik} - M_k\right)^2, \qquad k = 1,2 \qquad (3)$$

and x_{ik} is the feature vector of the i^{th} pixel , n_k is the number of pixels which belong to the k^{th} cluster and M_k are the centroids of the k^{th} cluster, respectively:

$$M_k = \left(\frac{1}{n_k}\right)\sum_{i=1}^{n_k} x_{ik}. \qquad (4)$$

K-means is commonly fed with the intensity of each pixel in the microarray image as features. However, there is already developed segmentation methods based on the K-means algorithm, which use more intensity features of each pixel (such as mean intensity of the neighbourhood of the pixel, or spatial features). For instance, Ergüt et al. (Ergüt et al., 2003) employed the K-means algorithm using only the intensity of the pixel as a feature, while Wu et al. (Wu & Yan, 2004) used three intensity-based features as well as the Euclidean distance between the pixel and the center of the spot, as the fourth feature. Both the channels of the microarray image are segmented simultaneously. Thus, for each pixel the intensities from both channels are combined to one feature vector. The number of cluster centres K is set usually to two, due to the fact that the segmentation is used for characterizing the pixels of the image as foreground or background pixels.

A number of studies employ the Fuzzy C-Means (FCM) (Bezdek, 1981), instead of the crisp K-means algorithm. FCM is a data clustering technique in which a dataset is grouped into K clusters with each data point in the dataset belonging to a cluster to a certain degree. For example, a certain pixel that lies close to the centroid of a signal cluster will have a high degree of belonging or membership to that cluster and another pixel that lies far away from the centroid of a cluster will have a low degree of membership to that cluster.

The FCM algorithm is based on the minimization of the following objective function:

$$J_m = \sum_{i}^{n_k}\sum_{k=1}^{3}(u_{ik})^{\gamma}\left\|x_{ik} - M_k\right\|^2, \qquad (5)$$

where, x_{ik} is the feature vector of the i^{th} pixel, M_k is the centroid of each cluster, u_{ik} is the degree of membership of x_{ik} in each cluster, $\left\|x_{ik} - M_k\right\|^2$ is the Euclidean distance between x_{ik} and M_k, n_k is the number of the pixels that belong to the k^{th} cluster. The parameter γ is the weighting exponent for u_{ik} which controls the fuzziness of the resulting clusters. Each pixel is classified in the cluster with the maximum calculated membership function.

A more robust than K-means and FCM clustering technique is the Partition Around Medoids (PAM) (Kaufman & Rousseeuw, 1989) clustering. PAM minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances. The algorithm first computes a number of representative objects, called medoids. A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal. In the classification literature, such representative objects are called centrotypes. After finding the set of medoids, each object of the dataset is assigned to the nearest medoid. Nagarajan et al. (Nagarajan, 2003) developed a segmentation method based on the PAM to extract the target intensity of the spots. The distribution of the pixel intensity in a grid containing a spot is assumed to be the superposition of the foreground and the local background. The partitioning around medoids is used to generate a binary partition of the pixel intensity distribution. The medoid (PAM) of the cluster members are chosen as the cluster representatives.

According to the assumption that the pixels of the image could be grouped in several clusters in order to extract the signal and background of a microarray image, more sophisticated methods had been proposed in the literature. Blekas et al. (Blekas et al., 2005) proposed a Gaussian Mixture Model (GMM) approach for the analysis of the microarray images using the Expectation-Maximization (EM) (Dempster et al., 1977) algorithm. EM is an ideal candidate for solving parameter estimation problems for the GMM or other neural networks. This methodology provides modelling, flexibility and adaptability to the data, which are well-known strengths of GMM. The maximum likelihood and maximum a posteriori approaches are used to estimate the GMM parameters via the expectation-maximization algorithm. The approach has the ability to detect and compensate for artefacts that might occur in microarray images. This is accomplished by a model-based criterion that selects the number of the mixture components.

Hybrid K-Means Segmentation

Finally, it is meaningful to refer a work presented by Rahnenführer et al. (Rahnenführer & Bozinov, 2004) who tried to engage both the image processing and the machine learning techniques. The hybrid K-means algorithm is an extended version of the K-means segmentation approach (Bozinov & Rahnenführer, 2002). The machine learning contribution includes repeated clustering in order to increase the number of foreground pixels. As long as the minimum amount of foreground pixels is not reached, the remaining background pixels are clustered into two groups and the group with pixels of higher intensity is assigned as foreground. After the clustering, the number of outlier pixels in the segmentation result is reduced with mask matching.

CLASSIFICATION-BASED APPLICATION

Segmentation Using the Bayes Classifier

In this section a novel segmentation method that classifies the pixels of the image into two categories (foreground and background) using classification techniques is presented. This classification-based approach directly classifies each pixel to the designated category. More specifically, the Bayes classifier (Gonzalez et al., 2004) is employed to classify the pixels of the image into different classes. The Bayes Classifier is fed with an informative set of 11 features (Table 1) (Giannakeas & Fotiadis, 2007) to deal with the artefacts of the image. Thus, the method can classify the pixels of the image into signal, background and artefacts.

Table 1. Features for the classification segmentation

FEATURE TYPE	CHANNEL		DESCRIPTION
Intensity features	GREEN	1	Intensity of the pixel
		2	Mean intensity value of the 3x3 neighbourhood of the pixel
		3	Intensity standard deviation of the 3x3 neighbourhood of the pixel
	RED	4	Intensity of the pixel
		5	Mean intensity value of the 3x3 neighbourhood of the pixel
		6	Intensity standard deviation of the 3x3 neighbourhood of the pixel
Spatial Features		7	x-coordinate of the pixel in the image
		8	y-coordinate of the pixel in the image
		9	Euclidean distance between the pixel and the centre of the spot
Shape features	GREEN	10	Correlation of the neighbourhood of the pixel and the Gaussian template
	RED	11	Correlation of the neighbourhood of the pixel and the Gaussian template

The concept of the Bayes classifier is to estimate the a posteriori probability of a sample (pixel) to belong in a class. The a posteriori probability is given by the Bayes theorem:

$$P(w_i \mid x) = \frac{p(x \mid w_i)P(w_i)}{\sum_{i=1}^{2} p(x \mid w_i)P(w_i)} = \frac{p(x \mid w_i)P(w_i)}{p(x)}, \qquad (6)$$

where, $x \in R^{11}$ is the feature vector, w_i: $i = 1,2$ are the two classes, $P(w_i)$ is the a priori probability that an arbitrary sample belongs to class w_i, $P(w_i \mid x)$ is the a posteriori conditional probability that a specific sample belongs to a class, $p(x)$ is the density distribution of all samples, and $p(x|w_i)$ is the conditional density distribution of all samples belonging to w_i.

The Gaussian density function is often used to model the distribution of feature values of a particular class. The general multivariate Gaussian density function is given as:

$$p(x) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}, \qquad (7)$$

where D is the dimension of the feature vector (D=11 in our case). μ_i and Σ_i are the mean vector and the covariance matrix of the features of the corresponding class respectively:

$$\mu_i = \frac{1}{N_i}\sum_{x \in w_i} x, \qquad \mu_i \in R^{11}, \qquad (8)$$

$$\Sigma_i = \frac{1}{N_i}\sum_{x \in w_i} xx^T - \mu_i\mu_i^T, \qquad \Sigma_i \in R^{11\times 11}, \qquad (9)$$

where N_i is the number of pixels belonging to class w_i.

In the training stage, the proposed approach estimates the mean vector and the covariance matrix for each class. Given the mean vector, the covariance matrix and the gaussian density distribution, the a posteriori probability is estimated for each sample and each class in the testing stage. The pixel is classified to the class with the maximum a posteriori probability.

Results

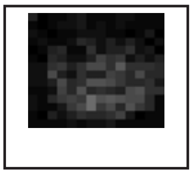
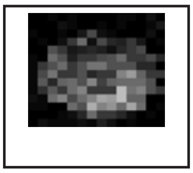
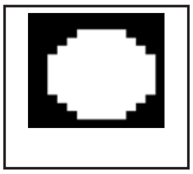
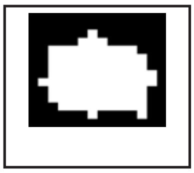
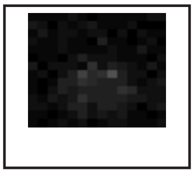
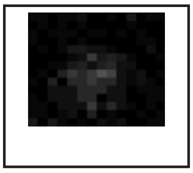
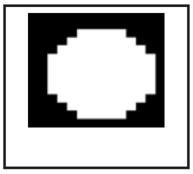
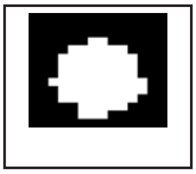
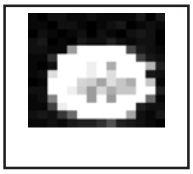
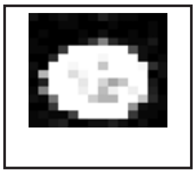
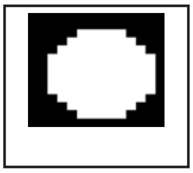
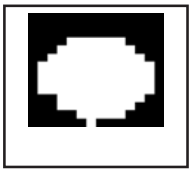
To quantify the effectiveness of the classification-based approach an image from the Stanford Microarray Database (SMD) (Gollub et al., 2003) is used. This image includes 16 blocks and each block consists of 576 spots, forming 24x24 rows and columns. Two of the blocks are used for the training (1152 spots and ~350000 pixels) and 14 blocks for the testing (5184 spots and ~3500000 pixels). In order to extract pixel by pixel information from the annotation, we simulate the fixed circle segmentation that is used by Scanalyse. For this task, the known radius of the fixed circle and the coordinates of the centres of each spot are used. Thus, a binary map is generated for the whole block, characterizing the pixels inside the circle as signal pixels and the pixel outside of the circle as background.

Table 2 presents the overall accuracy, specificity and selectivity of classification based methodology while the Table 3 shows the results of the Bayes classification in each individual spot. The accuracy, the specificity and the sensitivity of the proposed method is defined as:

Table 2. Accuracy results for the classification-based segmentation

Images	Training	Training	Acc %	Sp %	Se %
Experiment lc4b007rex2	2 Blocks (1152 spots)	14 Blocks (8050 spots)	86.28	90.21	82.21

Table 3. Results of the classification-based segmentation in individual spots

Description	Red	Green	Annotation	Result
Expressed spot				
Low expressed spot				
High expressed spot				

$$Acc = \frac{\#\ of\ \ correctly\ detected\ \ pixels}{total\ \#\ of\ \ pixels\ in\ the\ image}, \tag{10}$$

$$Sp = \frac{\#of\ \ correctly\ identified\ signal\ \ pixels}{total\ \#\ of\ \ signal\ \ pixels}, \tag{11}$$

$$Se = \frac{\#of\ \ correctly\ identified\ background\ \ pixels}{total\ \#\ of\ \ background\ \ pixels}. \tag{12}$$

In Table 3, three different types of spot are selected for illustration, expressed, low expressed and high expressed spot.

The current classification-based method detects efficiently the signal and the background pixels as it is shown in Table 2. We also have to stress that better results are reported for specificity as it is shown in Table 3. The main reason for this is the imbalanced dataset, i.e. it contains a large number of background pixels compared to the signal ones.

To compare the clustering-based techniques versus the classification based one four clustering approaches based on the K-means and the FCM algorithms are employed. Initially, both of these algorithms are fed with only the intensity of the two channels of the image. Then all the features of Table 1 are used. The accuracy results are shown in Table IV. As it is shown in this Table, the reported accuracy of the Bayes classifier is quite better than the other four approaches. In addition, the specificity (Sp) of the background class is extremely increased using the supervised classification-based segmentation.

FUTURE TRENDS

Future trends in the field of microarray technology could include: (1) The manufacturing and use of simpler arrays for quick and more-accurate-clinical diagnoses. (2) The acceptance of national (if not international) standards for array manufacturing, scanning, and analysis, and (3) The emergence and increasing use of smaller nano-arrays.

Accordingly, investigation and future challenges are generated for the computer-based

Table 4. Comparison of clustering and classification-based methods

Method	**Acc %**	**Signal Sp %**	**Background Sp %**
K-means (2 features)	65.49	88.14	41.95
K-means (11 features)	73.85	91.36	55.66
FCM (2 features)	65.88	87.21	43.71
FCM (11 features)	74.20	**90.94**	56.79
Bayes (11 features)	**86.28**	90.21	**82.21**

analysis of the microarrays. The development of new intelligent image processing techniques to eliminate the noise sources inherent in the DNA microarray process becomes more challenging. Additionally, the development of advanced image processing methodologies is significant to speed up the real-time diagnosis and implementation procedures of the next generation of system-on-a-chip devices. The extension of the machine leaning applications in the field of microarray image processing could provide more robust and effective tools for these purposes. Those methods can ultimately provide a new generation of diagnostic systems that can help to unlock the unknown patterns of complex diseases and their molecular phenotypes and allow rapid and responsive treatment mechanisms for these diseases.

CONCLUSION

An overview of the already developed methods for microarray is presented in this chapter. We categorized all the methods into two main categories, earlier approaches which use image processing techniques and approaches which use Machine learning techniques such as clustering. Image processing is an important stage in the circle of life of a microarray experiment. Reliability of this stage strongly influences the results of data analysis performed on extracted gene expressions. Several methods related to image processing or Machine learning techniques have been developed in this area. In this chapter we emphasized to the significance of the classification-based techniques for the segmentation of microarray image analysis. A Bayes classifier is presented to demonstrate the effectiveness of the classification techniques. According to the promising accuracy results, the precision of the microarray data during the next steps of the experiment might be significantly influenced.

REFERENCES

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

Ahmed, A. A., Vias, M., Iyer, N. G., Caldas, C., & Brenton J. D. (2004). Microarray segmentation methods signifficantly influence data precision. *Nucleic Acids Research*, *32*, e50.

Bezdek, J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. *Plenum Press.* New York.

Blekas, K., Galatsanos, N., Likas, A., & Lagaris, I. E. (2005). Mixture Model Analysis of DNA

Microarray Images. *IEEE Transactions on Medical Imaging*, *24*(7), 901-909.

Bozinov, D., & Rahnenführer, J. (2002). Unsupervised Technique for Robust Target Separation and Analysis of DNA Microarray Spots Through Adaptive Pixel Clustering. *Bioinformatics, 18*(5), 747-756.

Buck, M. J., & Lieb, J. D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics, 83*, 349-360.

Buckley, M. J. (2000). Spot User's Guide. *CSIRO Mathematical and Information Sciences,* Australia,. from http://www.cmis.csiro.au/iap/Spot/spotmanual.htm

Buhler, J., Ideker, T., & Haynor D. (2000). Dapple: improved techniques for finding spots on DNA microarrays. (UWCSE Tech Rep). Washington*: UWTR Dept. of Computer Science and Eng.,* University of Washington.

Chen, Y., Dougherty, E. R., & Bittner, M. L. (1997). Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images. *Journal Of Biomedical Optics*, *2*(4), 364–374.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*(1), 1-38.

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

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

Ergüt, E., Yardimci, Y., Mumcuoglu, E., & Konu, O. (2003). Analysis of microarray images using FCM and K-means clustering algorithm. *In IJCI 2003*, (pp. 116-121).

Giannakeas, N., & Fotiadis, D. I. (2007). Multichannel Segmentation of cDNA Microarray Images using the Bayes Classifier. *In 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* Lyon, France.

Gollub, J., Ball, C. A., Binkley, G., Demeter, K., Finkelstein, D. B., Hebert, J. M., Hernandez-Boussard, T., Jin, H., Kaplper, M., Matese, J. C., Schroeder, M., Brown, P. O., Botstein, D., & Sherlock, G. (2003). The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res*., *31*, 94-96.

Gonzalez, R. C., Woods, R. E., & Eddins, S. L. (2004). Digital image processing using MATLAB. *Prentice Hall, Upper Saddle River*, NJ.

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

Jain, A. N., Tokuyasu, T. A., Snijders, A. M., Segraves, R., Albertson, D. G., & Pinkel, D. (2002). Fully automatic quantification of microarray image data. *Genome Research*, *12*, 325–332.

Kaufman, L., & Rousseeuw, P. J. (1989). Finding Groups in Data - An Introduction to Cluster Analysis. *Wiley,* NY.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281-297. Berkeley: University of California Press.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*, 50-60.

Moran, G., Stokes, C., Thewes, S., Hube, B., Coleman, D. C., & Sullivan, D. (2004). Comparative genomics using Candida albicans DNA microarrays reveals absence and divergence of virulence-associated genes in Candida dubliniensis. *Microbiology*, *150*, 3363-3382.

Nagarajan, R. (2003). Intensity-Based Segmentation of Microarray Images. *IEEE Trans. On Medical Imaging, 22*(7), 882–889.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., & Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet, 23*, 41-46.

Rahnenführer, J., & Bozinov, V. (2004). Hybrid clustering for microarray image analysis combining intensity and shape features. *BMC Bioinformatics, 5*, 47.

Roerdink, J. B. T. M., & Meijster, A. (2000). The watershed transform: definitions, algorithms and parallelization strategies. *Fundamenta Informaticae, 41*(1-2), 187-228.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative motoring of gene expression patterns with a complementary DNA Microarray. *Science, 270*, 467-470.

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

Siddiqui, K. I., Hero, A., & Siddiqui, M. (2002). Mathematical Morphology applied to Spot Segmentation and Quantification of Gene Microarray Images. *In. Asilomar Conference on Signals and Systems.*

Wang, X., Ghosh, S., & Guo, S. W. (2001). Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, *29*(15), e75.

Wu, H., & Yan, H. (2003). Microarray Image Processing Based on Clustering and Morphological Analysis. *In First Asia Pacific Bioinformatics Conference*, (pp. 111-118).

KEY TERMS

Block: Blocks are also known as grids or subgrids. These are areas of the microarray slide (and relatively of the microarray image) in which a number of spots are located.

Classification: It is a procedure in which individual items are placed into groups based on quantitative information on one or more characteristics inherent in the items and based on a training set of previously labelled items.

Clustering: It is the task of decomposing or partitioning a dataset into groups so that the points in one group are similar to each other and are as different as possible from the points in the other groups.

Image Processing: The analysis of an image using techniques that can identify shades, colours and relationships that cannot be perceived by the human eye. In the biomedical field, image processing is used to produce medical diagnosis or to extract data for further analysis.

Machine Learning: It refers to the design and development of algorithms and techniques that allow computers to "learn". The purpose of machine learning is to extract information from several types of data automatically, using computational and statistical methods.

Microarray: Sets of miniaturized chemical reaction areas that may also be used to test DNA fragments, antibodies, or proteins, by using a chip having immobilised target and hybridising them with a probed sample.

Spot: It is a small and almost circular area in the microarray image whose mean intensity represents the expression level of the corresponding gene.