A Pattern Classification Approach to DNA Microarray Image Segmentation

Luis Rueda[1] and Juan Carlos Rojas[2]

[1] School of Computer Science, University of Windsor,
401 Sunset Avenue, Windsor, Ontario, N9B 3P4, Canada
lrueda@uwindsor.ca
[2] Department of Computer Science, University of Concepción,
Edmundo Larenas 215, Concepción, Chile
correorojas@gmail.com

Abstract. A new method for DNA microarray image segmentation based on pattern recognition techniques is introduced. The method performs an unsupervised classification of pixels using a clustering algorithm, and a subsequent supervised classification of the resulting regions. Additional fine tuning includes detecting region edges and merging, and morphological operators to eliminate noise from the spots. The results obtained on various microarray images show that the proposed technique is quite promising for segmentation of DNA microarray images, obtaining a very high accuracy on background and noise separation.

Keywords: DNA microarray images, segmentation, clustering, classification.

1 Introduction

DNA microarrays are techniques used to evaluate the expression of thousands of genes simultaneously. This paper focuses on DNA microarrays, in which the spots are layered in sub-grids. Segmentation is one of the most important steps in microarray image processing, and consists of identifying the pixels that belong to the spot from the pixels of the background and noise. Various microarray image segmentation approaches have been proposed, which assume a particular shape for the spots, while some of them have more freedom in this regard. Fixed circle is a method that assumes a circular shape with the same diameter for all spots [1][2]. Adaptive circle is a method that allows to adjust the radius of the circle for each spot [1]. While this method solves the problem of the radius of the circle, it fails to find the proper shape when the spots have irregular shapes. Elliptic methods assume an elliptical shape for the spots, and can adapt to a more general shape than the adaptive circle method, but cannot recognize irregularly shaped spots [3]. Seeded region growing is a method that groups pixels in regions based on a certain criterion of similarity, starting from initial points, the seeds [4][5]. Histogram-based methods and mathematical morphology have also been applied to microarray image segmentation [6][7]. The application of clustering to DNA

V. Kadirkamanathan et al. (Eds.): PRIB 2009, LNBI 5780, pp. 319–330, 2009.

microarray image segmentation is based mainly on two algorithms: k-means and expectation maximization [8][9][10]. The advantage of clustering with respect to other techniques is that it is not restricted to any predetermined shape for the spots. However, the power of clustering and in general, pattern recognition techniques has not been exploited in a comprehensive way as we do here. This paper introduces the use of pattern recognition techniques to devise a method for DNA microarray image segmentation. These techniques combined propose a general structure, which is composed of many steps, and the main steps are implemented with classifiers, while the others are implemented with algorithms developed for fine tuning.

2 The Proposed Method

The proposed approach is divided in various steps, starting with a method that discards images that do not have spots, followed by a series of region detectors and classification, ending with the use of morphological operators to eliminate noise. Fig. 1 illustrates this structure, in which the boxes represent the steps, while the arrows the output of each block. A brief description of these steps follows.

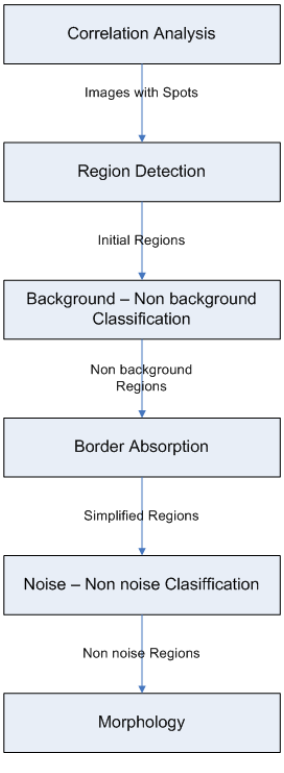

Fig. 1. General scheme of the proposed microarray segmentation technique

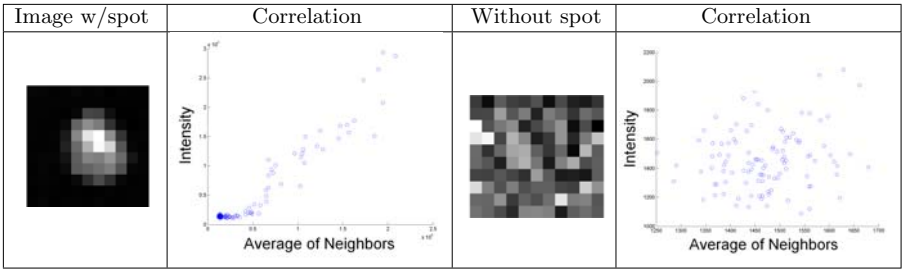

Fig. 2. Correlation plot for two images (with presence or absence of a spot). The intensities were increased to improve visualization.

Correlation analysis discards regions that do not have spots by analyzing Pearson's index between the intensities of the pixels and the average intensity of its neighbors [21]. Fig. 2 shows an image with a spot and its correlation plot, where the x-axis represents the average neighbor intensity and the y-axis the pixel intensities, which tends to follow the $x = y$ line, reflecting the high correlation between the features. Fig. 2 also shows an image without a spot and its corresponding correlation plot, which has the shape of a cloud, reflecting the low correlation between the features for this case. Thus, the correlation index is a very good measure for the presence or absence of a spot. *Region detection* detects the initial regions of an image using k-means and different initial configurations for the number of clusters and centroids, generating a set of 303 different clusterings – we select the best clustering using the I-index [11]. *Background-non background classification* classifies the initials regions as regions that belong (or do not belong) to the background using a supervised classifier. *Border absorption* takes the regions that were classified as non background, and determines which are the main regions and which are the borders, and proceeds to merge the main regions with their borders. *Noise-non noise classification* classifies the initials regions as non noise regions using a supervised classifier. *Morphology* is finally used to eliminate noise that was not detected in the previous steps.

2.1 Unsupervised Classification

For the unsupervised classification of the pixels into different regions of the image, k-means was used and combined with the I-index to evaluate the quality of the clustering generated [11]. We used the Euclidean distance and the following features to represent each object (pixel) to be classified: pixel intensity, average of neighbor intensities (using an 8-vicinity), distance from the pixel to the center of the region, variance of the neighbor intensities (considering an 8-vicinity), and gradient (a vector that indicates the direction of maximum increment of intensities). The centroids of the clusters were initialized using the percentiles of the distribution of the feature values. We also used a random initialization of the centroids.

When clustering data, it is crucial to know the correct number of clusters. Since this is usually unknown, a difficult task is to find it automatically, i.e. without human intervention. The I-index is a coefficient used to measure the quality of the clustering and hence, it helps find the best number of clusters. The I-index aims to maximize $I(k) = \left\{ \frac{E_1}{kE_k} D_k \right\}^p$, where $E_k = \sum_{i=1}^{k} \sum_{j=1}^{|D_i|} u_{ij} \left\| x_j - \mu_i \right\|$ and $D_k = max_{i,j=1}^{k} \left\| \mu_i - \mu_j \right\|$, with u_{ij} being the membership of x_j to cluster D_i and $\mu_i = \frac{1}{n} \sum_{x_j \in D_i} x_j$. To avoid predominance of some feature over others, normalization is applied to each feature before using k-means.

The general strategy used to determine the final clustering consists of generating a large number of clusterings with different initializations and based on the quality of the clustering, to finally determine which one has the best performance. By following this procedure, the final implementation of the algorithm was configured to recognize between two and four clusters. For each number of clusters, 101 different clusterings were generated, 100 of the initial centroids were chosen at random, and one of them was predetermined, yielding a total number of 303 different clusterings. Each clustering was evaluated with the I-index and the one with the highest index value was selected.

2.2 Supervised Spot Classification

Once the regions have been identified, the next step is to identify those regions that belong to the spot and those that belong to the background. The first stage of this process consists of classifying a region as background or non background. The features for the background classification were, mainly, based on the average intensities and the spatial characteristics of the background, such as its distribution in the image. The features to represent an object (region) used for this classification are the following: average of intensities, percentage of the region perimeter that represents the border of the image, and the largest distance of a pixel that belongs to the region from the geometric center of the image.

The second stage consists of separating noise from spots. This procedure takes the regions that were classified as non background and merged with its borders, and classifies them as noise or non noise regions. The features used in this step are the following, which are shown grouped in four categories:

Statistics of Intensities

- average of intensities
- variance of intensities (calculated as $\frac{1}{n-1} \sum_{i=1}^{n} \left(I_i - \overline{I} \right)^2$, where n is the total number of pixels that belong to the region, I_i is the intensity of pixel i, and $\overline{I}$ is the average of intensities of the region)
- standard deviation of intensities (calculated as $\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(I_i - \overline{I} \right)^2}$)
- average of neighbor regions that are non background

Geometric Features

- total area of a region
- ratio between the total area of a region and its perimeter (excluding its holes)
- distance between the geometric center of the region and the geometric center of the image
- percentage of the border of the region that limits with the background
- percentage of the region perimeter that represents the border of the image
- length of the skeleton of the region (defined as the medial axis of the region, where a point belongs to the medial axis, if it has more than one closest neighbor in the border of the region [20])

Comparison with the Edges of Neighboring Regions

- ratio between the border of the region and the border of its neighbors (average of pixel intensities)
- difference between the border of the region and the border of its neighbors (average of pixel intensities)
- ratio between the border of the region and the border of its neighbors that are non background (average of pixel intensities)

Comparison with the Average Intensities of Other Regions

- ratio between the region and the neighbor regions that are non background
- ratio between the region and the largest neighbor that is non background
- ratio between the region and the largest region that is non background
- ratio between the region and the background

For each of the supervised classifications, various classifiers were implemented, tested and compared: the logistic linear classifier (LG) [13], Fisher's lineal classifier (FISH) [12], the nearest mean classifier (NM) that uses the Euclidean distance [14], the k-nearest neighbor classifier (k-NN) using the Euclidean distance [15], a support vector machine (SVM) with a linear kernel [16], the naive Bayes classifier (NV) [17], a linear classifier using principal component analysis (PCA) [18], the quadratic classifier (QUAD) that assumes a normal distribution for each class [12], and the binary decision tree classifier (BTREE) [13][19].

2.3 Post-processing

In order to improve the final segmentation two post-processing stages are performed on the resulting regions classified as spots. *Border absorption* is applied to detect the borders of the regions, and to eliminate false edges. In a nutshell, this stage detects which regions, classified as non background, are borders of other regions, and then merge the main regions with their border regions, generating a new region. Two conditions are demanded for considering a region as a border of another region. Firstly, considering the pixels of both regions that conform the border between them, the average of intensities of the pixels from

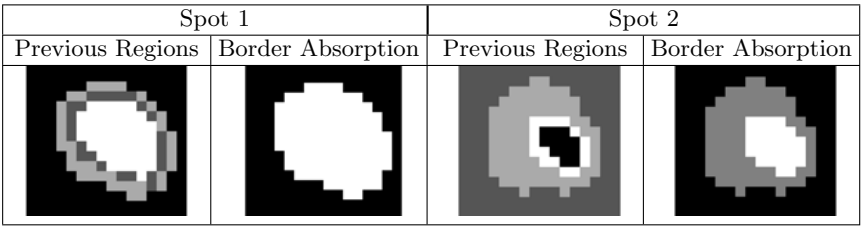

Fig. 3. Border absorption applied on two spots

the region that is a possible border is lower than the average of intensities of the pixels from the possible main region. Secondly, when the morphological operation dilation is applied over the main region [7,20], with an 8-vicinity and ignoring its holes, and this main extended region is overlapped with the possible border region, if it covers at least an 85% of its area. The algorithm also detects successive borders, i.e. if region B is a border of region A, and region C is a border of region B, then, the algorithm merges the tree regions into a single one. Fig. 3 shows two examples of border absorption, showing the simplification of the resulting regions.

The other stage involves applying *mathematical morphology*. The basic operations used in this work are erosion, which produces thinning of the objects, and dilation, which produces thickening of the objects. Both operations are controlled by a shape called structured element, which consists of a matrix with zeros and ones that is translated through the domain of the image. The combination of these two operations, a dilation and soon an erosion, are used as the last filter of noise on the regions that have been classified as non noise.

3 Experimental Results

The DNA microarrays images used in the experiments were obtained from the Stanford Microarray Database (SMD), publicly available at smd.stanford.edu. The images used here are mainly from experiments with *Arabidopsis Thaliana* and *Austrofundulus Limnaeus*. The images extracted from the database represent individual spots, based on its correlation value and a classification index that represents whether or not the image contains a spot. The regions were generated using the unsupervised classification, which were then classified as background or non background. The regions used for the second supervised classifier, which correspond to regions that were classified as nonbackground in the previous supervised step, were merged with their borders and classified as noise or non noise. Finally, morphological operators were applied to the regions classified as non noise, and the resulting regions were considered as the spots detected by the algorithm.

3.1 Correlation

The experiments performed with the correlation coefficient between the pixel intensities and the average intensities of the neighbor pixels were done using Pearson's coefficient, and a threshold 0.7384, which was found experimentally, was used to eliminate images that do not have spots with a very high accuracy.

3.2 Unsupervised Classification

A series of experiments were then performed with the aim of determining the set of features that gives better results, and the range for the numbers of clusters using the I-index as a measure of the quality of the segmentation produced. To find the feature space that gives the best initial image segmentation with the k-means algorithm, a series of experiments with k-means were performed for testing the different configurations. This consists of comparing visually the segmentation generated with different feature spaces for k-means. The experiments performed showed that the feature space involving the intensities of the pixels and the average of intensities of the neighbors generates the best results among all possible feature spaces tested.

To determine an appropriate range for numbers of clusters to be used in the k-means algorithm, a test with two, three, four, five and six clusters was conducted, and the results were compared afterwards. These experiments allowed to conclude that the range for the numbers of clusters that gives the best results is between two and four, because a larger number of clusters generates an excessively large number of regions that are difficult to classify. Fig. 4 shows this scenario. The best number of clusters was found using the I-index, searching over 101 different clusterings for each number of clusters in the range (two, three and four), with a total of 303 clusterings. For each number of clusters, we generated 101 different clusterings: one of them has the initial configuration of the centroids pre-determined based on percentiles of the values obtained for each feature, and the remaining 100 initial configurations of centroids were selected at random in the range of values registered for each feature. Then, the algorithm obtains the I-index value for each clustering produced, and selects the one that delivers the largest value. This procedure is applied to each group of clusters, yielding three different clusterings with two, three, and four clusters respectively, and their corresponding image segmentations and I-index values. These experiments demonstrate the validity of the I index as an evaluator of the quality of the segmentation produced.

3.3 Background Classification

In this step, a set of supervised classifiers were tested in the classification of the regions into background and non background, and following a ten-fold cross-validation setup, obtaining the average error rate over the ten folds. Table 1 shows the results obtained in these experiments, where the error rate for each classifier is listed. These results indicate that the lowest error rates were obtained

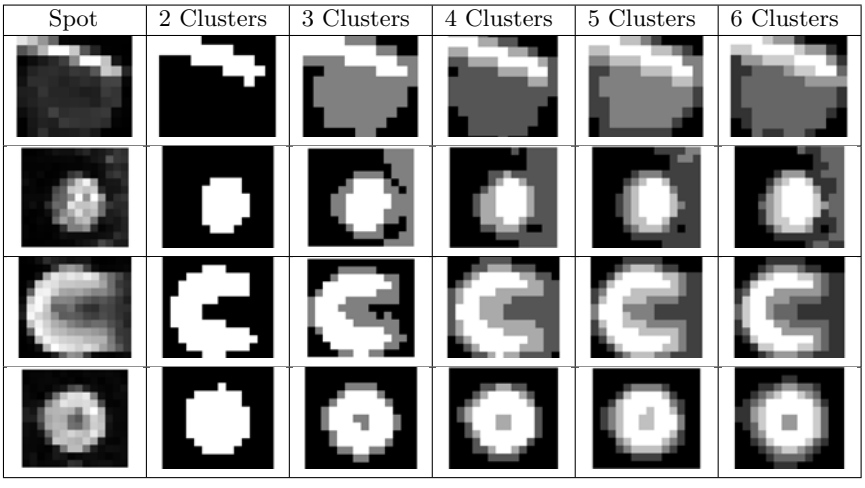

Fig. 4. Comparison of the segmentation generated with k-means using different numbers of clusters and the set of features pixel intensities with average of intensities of the neighbors

Table 1. Error rates for background vs non background classification

Classifier	**Error Rate**
LOG	3.47
PCA	3.72
FISH	3.47
NM	40.89
QUAD	8.43
KNN	13.51
BTREE	9.54
NV	4.58
SVM	4.37

with the logistic linear classifier and the minimum least square linear classifier, with an error rate of 3.47% for both cases. Both classifiers show a very low error rate, indicating that the schemes recognize background versus non background quite accurately. Border absorption was then applied to the regions that were classified as non background. The experiments show that the algorithm detects quite accurately when a region is a border of another region. This resulted in images simpler and easier to process for the next level using supervised classifiers.

3.4 Noise Classification

In this step, a set of supervised classifiers were tested in the classification of the regions, which were classified in the previous step as non background regions

Table 2. Error rates for noise vs non noise classification

Classifier	Error Rate
LOG	18.88
PCA	21.53
FISH	19.76
NM	33.92
QUAD	23.64
KNN	26.84
BTREE	20.35
NV	24.19
SVM	23.60

and then merged with their borders, into two classes, noise or non noise. The classifiers were tested using a ten-fold cross-validation procedure. Table 2 shows the results obtained in these experiments, where the error rate of each classifier is listed. These results indicate that the lowest error rate was obtained for the logistic linear classifier with an error rate of 18.88%. These results imply that the hardest task of the proposed approach is to recognize the signals including noise, and thus it justifies the use of a specific level of supervised classification for detecting them. It also suggests the need of using additional filters to detect and remove noise, which are implemented in the next step by using morphological operators.

3.5 The Complete Segmentation

We tested the complete segmentation method over a set of images of spots selected and classified based on general features. The aim is to compare visually the original image with the segmentation that the algorithm outputs. The supervised classifiers used in the implementation of the algorithm were selected based on their performance in the tests. In this series of experiments, the same classifier was used to implement the background and noise classification steps, the logistic linear classifier, which was shown to be the most accurate among all classifiers tested. The images included in the experiments are of various characteristics:

Regular Spots: This set groups spots that show a circular-like shape, and the image does not present signals of noise. The variations between the images are given by the size, intensity and location of the spot in the image.

Irregular Spots: This set groups spots that do not have a circular-like shape, and the image does not present noise signals. Some of the shapes considered in this set are elliptic and half-a-moon. The variations between the images are given by the size, intensity and location of the spot in the image.

Noisy Spots: This set groups images of spots that present different levels of noise. In addition to the level of noise, other variations between the images are given by the shape, size, intensity and location of the spot in the image.

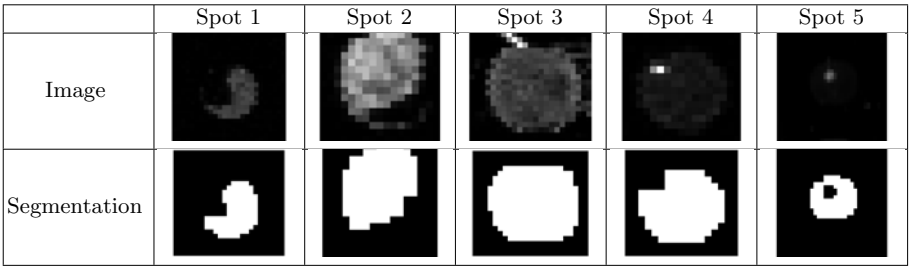

Fig. 5. Experiments performed with the complete algorithm

Fig. 5 shows the results of the experiments over a set of spot images using the complete algorithm. The set of spot images considers different configurations, from nearly perfect spots to quite irregular spots and noisy images. The results show the power of the algorithm to produce accurate segmentation and to easily adapt to all these different configurations. When dealing with regular spots, the algorithm segments accurately the images with these characteristics, independently of the size, location or brightness of the spots. The results of the experiments with irregular spots show that the algorithm detects the main features of the spots accurately, and the quality of the segmentation depends on the smoothness of the spot borders – smoother spot borders produce a better segmentation. The results of the experiments with noisy spots show that the performance of the algorithm depends on the level of noise present in the image. If the magnitude of noise compared to the spot is low, the algorithm generates a segmentation that is arbitrarily close to the real spot, while if the magnitude of noise is high, the segmentation will differ substantially from the real spot. In conclusion, the main factors that affect the quality of the segmentation are the level of noise and the smoothness of the borders of the spots. In general, the proposed approach is able to deal with different types of images, capturing different shapes and eliminating noise accordingly.

4 Conclusions

A combination of techniques from the field of pattern recognition is shown to be a very powerful scheme for segmentation of DNA microarray images. Supervised and unsupervised classification techniques have been shown to be effective in the segmentation of real-life images, when performed in sequence and complemented with fine tuning that includes border absorption and morphology. Experiments have been performed in real-life images from the Stanford microarray database, which show that the system is highly accurate in identifying the pixels belonging to the spots, and separating them from background and noise. The proposed approach is a framework for the development of such a system that encourages future research on variations of the different parameters of the system, including the number and selection of the features, the unsupervised and supervised classification schemes, the evaluation of the quality of the clusters, among others.

Acknowledgements. This research work was partially supported by NSERC, the Natural Sciences and Research Council of Canada, grant No. RGPIN 261360, and the Chilean National Council for Technological and Scientific Research, FONDECYT grant No. 1060904.

References

1. Yang, Y., Buckley, M., Speed, T.: Analysis of cDNA Microarray Images. Briefings in Bioinformatics 2(4), 341–349 (2001)
2. Eisen, M.: ScanAlyze User Manual. Stanford University (1999)
3. Rueda, L., Qin, L.: A New Method for DNA Microarray Image Segmentation. In: Kamel, M.S., Campilho, A.C. (eds.) ICIAR 2005. LNCS, vol. 3656, pp. 886–893. Springer, Heidelberg (2005)
4. Adams, R., Bishof, L.: Seeded Region Growing. IEEE Trans. on Pattern Analysis and Machine Intelligence 16(6), 641–647 (1994)
5. Talbot, B.: Regularized Seeded Region Growing. In: Proc. of the 6th International Symposium ISMM 2002, pp. 91–99 (2002)
6. Ahmed, A., Vias, M., Iyer, N., Caldas, C., Brenton, J.: Microarray Segmentation Methods Significantly Influence Data Precision. Nucleic Acids Research 32(5), e50 (2004)
7. Angulo, J., Serra, J.: Automatic Analysis of DNA Microarray Images using Mathematical Morphology. Bioinformatics 19(5), 553–562 (2003)
8. Rueda, L., Qin, L.: An Improved Clustering-Based Approach for DNA Microarray Image Segmentation. In: Campilho, A.C., Kamel, M.S. (eds.) ICIAR 2004. LNCS, vol. 3212, pp. 17–24. Springer, Heidelberg (2004)
9. Wu, S., Yan, H.: Microarray Image Processing Based on Clustering and Morphological Analysis. In: Proc. of the First Asia-Pacific Bioinformatics Conference on Bioinformatics, pp. 111–118 (2003)
10. Li, Q., Fraley, C., Bumgarner, R., Yeung, K., Raftery, A.: Donuts, Scratches and Blanks: Robust Model-Based Segmentation of Microarray Images. Technical Report No. 473, Department of Statistics, University of Washington (2005)
11. Maulik, U., Bandyopadhyay, S.: Performance Evaluation of Some Clustering Algorithms and Validity Indices. IEEE Trans. on Pattern Anal. Mach. Intell. 24(12), 1650–1654 (2002)
12. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 3rd edn. Academic Press, London (2006)
13. Mitchell, T.: Machine Learning. McGraw-Hill, New York (1997)
14. Veenman, C., Tax, D.: A Weighted Nearest Mean Classifier for Sparce Subspaces. Computer Vision and Pattern Recognition 2, 1171–1176 (2005)
15. Song, Y., Huang, J., Zhou, D., Zha, H., Lee, C.: IKNN: Informative K-Nearest Neighbor Pattern Classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 248–264. Springer, Heidelberg (2007)
16. Abe, S.: Support Vector Machines for Pattern Classification. Springer, Heidelberg (2005)
17. Dash, D., Cooper, G.: Exact Model Averaging with Naive Bayesian Classifiers. In: Proc. of the 19th International Conference on Machine Learning, pp. 91–98 (2002)
18. Baek, K., Draper, B., Beveridge, J., She, K.: PCA vs. ICA: A Comparison on the FERET Data Set. In: Proc. of the 4th Int. Conference on Computer Vision, Pattern Recognition and Image Processing, Durham, NC, pp. 824–827 (2002)

19. Safavian, S., Landgrebe, D.: A Survey of Decision Tree Classifier Methodology. IEEE Trans. on Systems, Man, Cybernetics, 660–674 (1991)
20. Gonzalez, R., Woods, R., Eddins, S.: Digital Image Processing Using Matlab. Prentice-Hall, Englewood Cliffs (2003)
21. Draghici, S.: Data Analysis Tools for DNA Microarrays. Chapman and Hall/CRC, Boca Raton (2003)