Image analysis for detecting faulty spots from microarray images

Salla Ruosaari and Jaakko Hollmén

Helsinki University of Technology, Laboratory of Computer and
Information Science, P.O. Box 5400, 02015 HUT, Finland
Salla.Ruosaari@hut.fi, Jaakko.Hollmen@hut.fi

Abstract. Microarrays allow the monitoring of thousands of genes simultaneously. Before a measure of gene activity of an organism is obtained, however, many stages in the error-prone manual and automated process have to be performed. Without quality control, the resulting measures may, instead of being estimates of gene activity, be due to noise or systematic variation. We address the problem of detecting spots of low quality from the microarray images to prevent them to enter the subsequent analysis. We extract features describing spatial characteristics of the spots on the microarray image and train a classifier using a set of labeled spots. We assess the results for classification of individual spots using ROC analysis and for a compound classification using a non-symmetric cost structure for misclassifications.

1 Introduction

Microarray techniques have enabled the monitoring of thousands of genes simultaneously. These techniques have proven powerful in gene expression profiling for discovering new types of diseases and for predicting or diagnosing the type of a disease based on the gene expression measurements [1]. It is indeed an interesting possibility that we examine all genes of a given organism at the same time and possibly under different conditions. This opens up new ways of making discoveries, assuming that the large amounts of data can be reliably analyzed.

The rapidly increasing amount of gene expression data and the complex relationships about the function of the genes has made it more difficult to analyze and understand phenomena behind the data. For these reasons, functional genomics has become an interdisciplinary science involving both biologists and computer scientists.

Before estimates of the gene activities are obtained from an organism, a multi-phased process takes place allowing different sources of noise to enter the analysis. Noise is in fact a major issue with microarrays. Low quality measurements have to be detected before subsequent analysis such as clustering is performed and inferences are made. However, the detection of these poor quality spots has not been widely discussed. In this paper, we attempt to provide one solution to this problem.

2 Microarray technology

The microarray experiments are basically threefold involving the preparation of the samples of interest, the array construction and sample analysis, and the data handling and interpretation. The microarray itself is simply a glass slide onto which differing single-stranded DNA chains have been attached at fixed loci.

The phenomenon that microarrays exploit is the preferential binding of complementary single-stranded sequences. Popularly mRNA extracted from two different samples are brought into contact as they are washed over the microarray. Hybridization takes places at spots where complementary sequences meet. Therefore, hybridizations of certain nucleic acid sequences on the slide indicate the presence of the complementary chain in the samples of interest.

2.1 Two-sample competitive hybridization and dye separation

A popular experimental procedure is the monitoring of the mRNA abundance in two samples. When two samples are simultaneously allowed to hybridize with the sequences on the slide, the relative abundance of the hybridized mRNA in the samples can be measured. This measure is assumed to reflect the relative protein manufacturing activity in the cells. Often a common reference is used making further comparisons of gene activities e.g. between individuals possible.

The two samples are labeled with different fluorescent dyes allowing their separation when excited with the corresponding laser. When the whole slide is scanned, two 16-bit images looking as Fig. 1 are obtained, each reflecting the gene activities of the respective sample. The intensities of the image pixels correspond to the level of hybridization of the samples to the DNA sequences on the microarray slide.

2.2 From digitized images to intensity measures

To get an estimate of the gene activities, the pixels corresponding to the gene spots, and consequently the genes, must be found. The images are segmented or partitioned into foreground (i.e. belonging to a gene) and background regions. The gene activity estimates are then derived from the foreground regions. Many different methods exist including the average intensity of pixels inside some predefined area around the assumed spot center or within an area found by seeded region growing or histogram segmentation. Estimates of the background noise can also be obtained. The estimates can be global, i.e. all genes are assumed to include the same noise or local, i.e. the background estimate is determined individually for all genes or for some set of genes using a (predefined) combination of the pixel intensities outside the area used for gene activity estimation.

The gene activity estimation has an impact on the subsequent data analysis and interpretation. If the gene's measured activity is not due to the activity itself, subsequent analysis using this erroneous estimate will, of course, be misleading. To overcome this, background correction is often done, usually simply by subtracting the background intensity estimates from the gene activity estimates.

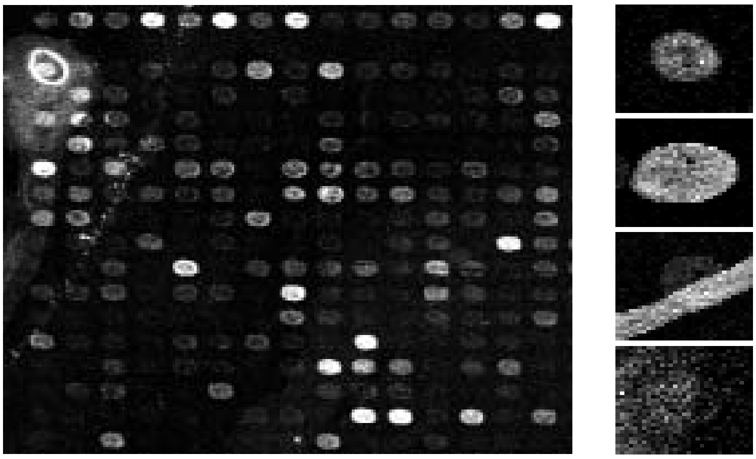


Fig. 1. A scanned microarray image and four example spots, which demonstrate possible problems, i.e. spots of varying sizes, scratches, and noise.

Depending on how the gene activity estimate and the background estimate have been derived, the resulting measures may be largely deviant.

Image analysis methods using predefined regions, histogram segmentation or region growing essentially all lead to biased results, even if background correction is used, if the data quality is not taken into consideration. This can be understood by observing Fig. 1. The spots may be of various sizes or contaminated and can therefore have an effect on the activity estimation when no attention to the spatial information is given. The Mann-Whitney segmentation algorithm may provide better results as it associates a confidence level with every intensity measurement based on significance [2]. If the noise level on the slides is not constant, non gene activity due measures may start dominating the results as most of the genes on typical slides are silent. Background estimations may be even more affected by contamination. In order for the background correction to be effective, the background estimates should be derived iteratively and not by using the same pixels for each spot. Moreover, the most contaminated spots should be excluded from the analysis as the measure does not reflect the gene activity at all. Replicate measurements may be of help [3] especially when the median of the measures is used in the analysis.

To this day, little has been published on data quality related issues. Previously, the effect of the choice of image analysis method has been assessed. It has been shown that the background adjustment can substantially reduce the precision of low-intensity spot values whilst the choice of segmentation procedure has a smaller impact [4]. Measures based on spot size irregularity, signal-to-noise ratio, local background level and variation, and intensity saturation have been used to evaluate spot quality [5]. Experiments on error models for gene array data and expression level estimation from noisy data have been carried out [6]. The intrinsic noise of cells has also been researched [7,8].

3 Detection of faulty spots

Our work is based on analyzing real-valued raw 16-bit images with the approximate gene loci known. Each gene spot is searched from a 31 x 31 environment defined by the gene center locus obtained as a result of previous image segmentation with QuantArray software. The sizes of these blocks were chosen to allow some non-exactness in gene loci and to be large enough to be able to include valid spot pixels. We apply image analysis techniques in extracting spatial features describing relevant properties of microarray spots [9].

3.1 Defining the spot area

The spot area is defined on the basis of raw pixel intensity values and their spatial distribution. We assume that the intensity of the spot pixels deviates from the background intensity in the positive direction. At the initial step, the raw pixels are judged to belong to the spot if their raw intensity is more than 12.5 percent of the maximum pixel intensity found in the 31 x 31 image. This is how the histogram segmentation methods work. Here, however, the histogram segmentation forms only the initial step of the segmentation procedure.

From these regions, the largest connected block of pixels is picked using eight-connectivity, and pixels inside the area are joined to the area using four-connectivity. This way, we obtain a binary image in which the spot area is differentiated from the background. Examples of these images, which can be regarded as masks for the original intensity images, are shown in Fig. 2.

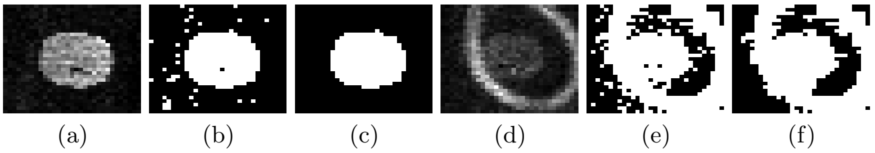


Fig. 2. The search for the spot area is presented using a non-faulty spot (a-c) and a faulty spot (d-e). The 31 x 31 pixel block around the spot centers (a and d), the corresponding binary image obtained using threshold 12.5 percent of the maximum intensity found within this block (b and e), and the largest connected region of the binary image with holes filled (c and f).

3.2 Spatial features of the spots

We assume that features extracted from the spot area can be used to describe the quality of the measurement. The features are collected into a feature vector $\mathbf{x} = [x_1, \ldots, x_6]$ and are later used to discard redundant low quality data from subsequent analysis. Through the choice of the features, an implicit model for the spots is defined. The image pixel coordinates are denoted as (h, v) pairs and the individual pixel coordinates with h_i and v_i, $i = 1, \ldots, n$, n being the number of pixels belonging to the spot in this context. The features we extract are:

The horizontal range of the spot	$x_1 = \max(\lvert h_i - h_j \rvert), i \neq j$
The vertical range of the spot	$x_2 = \max(\lvert v_i - v_j \rvert), i \neq j$
The elongation of the spot as the ratio of the eigenvalues	$x_3 = \lambda_1 / \lambda_2$
The circularity of the spot as the ratio between the area of the estimated spot and an ideal circle with the same perimeter	$x_4 = 4\pi Area/(Perimeter)^2$
The uniformity of the spot expressed as the difference between the Euclidean distance of the mass centers between the binary image and the intensity image masked with the binary image	$x_5 = \lVert 1/n \sum_{i=1}^{n} (h_i, v_i) - 1/n \sum_{i=1}^{n} \mathrm{int}_i(h_i, v_i) \rVert$
The Euclidean distance between the assumed spot center and the binary image	$x_6 = \lVert 1/n \sum_{i=1}^{n} (h_i, v_i) - (h_c, v_c) \rVert$

3.3 Classification based on the spatial features

As stated earlier, our primary task is to classify microarray spots to classes *faulty* and *good*. This binary class variable c_i is predicted on the basis of six features, or input variables, describing relevant properties of the objects to be classified. Having access to n labeled training data, that is, pairs $(\mathbf{x}_i, c_i), i = 1, \ldots, n$, we can train a classification model in order to classify future cases where label information is not available.

Based on the assumption that classes have differing statistical properties in terms of the distributions of the feature variables, we may use the class-conditional approach [10,11]. Suppose we already have a classification model, we may assign a spot to the class to which it is most likely to belong to, i.e. whose posterior probability is the largest. Using Bayes rule, this is equivalent to assigning the spot, i.e. the feature vector $\mathbf{x}$ derived from it, to the class c_i for which the discriminant function g_i is the largest, as in $c_j = \arg\max_k \, g_k(\mathbf{x}_j)$, where $g_i(\mathbf{x}) = \log p(\mathbf{x}|c_i)p(c_i)$. The underlying distributions $p(\mathbf{x}|c_i)$ are assumed to be Gaussian. The parameters of the class-conditional distributions, i.e. mean vectors and covariance matrices, are estimated from pre-labeled training data. The prior probabilities are not of concern because the optimal bias is found by observing the misclassification costs.

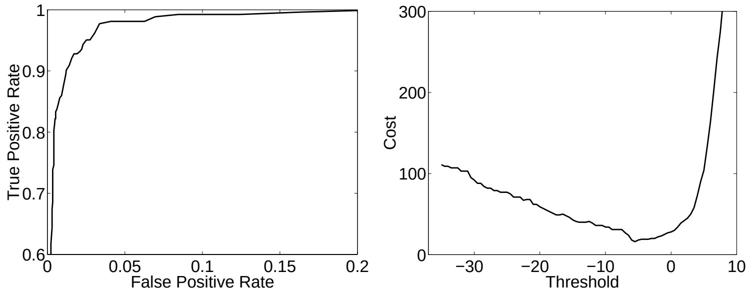
3.4 Assessment of classification results

Before put to practice, it is important to assess the accuracy of the proposed scheme to detect faulty spots. We are interested in the following two aspects: first, how well the individual spots are classified correctly and how often the spots are misclassified in the two possible directions (good as faulty and faulty as good) and second, combining the results for the three classifications of replicate spot measurements, what is the most beneficial compound result that fulfills our goal. In both approaches, we have the problem of choosing an optimal decision function.

Receiver Operating Characteristics (ROC) Curve [12,13] visualizes the tradeoff between false alarms and detection, helping the user in choosing an optimal decision function. With the ROC curve, we can assess the errors made in the classification of individual spots. However, we are in fact faced by the need to classify three spots that are repetitive measurements of the same gene expression, two of which are possibly redundant.

We are fundamentally interested in correct classification of good spots as good (true negative, tn) and faulty spots as faulty (true positive, tp), but the situation is complicated by our consideration of classifying good spots as faulty (false positive, fp) not being so harmful as long as at least one of the replicate good spots is classified correctly. On the contrary, classifying faulty spots as good (false negative, fn) is considered harmful, since possible measurements of the faulty spots may enter the subsequent analysis. Formulating the above as a matrix for misclassification costs, we get $\Lambda = (\lambda_{ij}) = \Sigma\mathrm{fn}/\Sigma(\mathrm{tn}+\mathrm{fn})$, with the exception $\lambda_{i4} = 1$, when $i = 1, 2, 3$.

The entries in the cost matrix λ_{ij} signify how much cost is incurred when the compound configuration of three spots i is chosen when j is in fact the right choice. For instance, the entry λ_{41} signifies the cost of classifying the compound classification *faulty faulty faulty* as *good good good*, and therefore a cost of 1 units is incurred. The order of the outcomes is irrelevant as long as the classification-label pairs match. The cost matrix contains off-diagonal zeros to allow misclassifications of some good spots if at least one good spot is classified as good. If a good spot finally enters the subsequent analysis, our goal is fulfilled.



(a) (b)

Fig. 3. Classification results presented with a ROC curve (a) and as a function of classification cost with a varying boundary threshold (b).

4 Experimental results

The covariance matrices and mean vectors of the class descriptive Normal distributions were estimated from data consisting of 7488 spots. The spots were visually determined to be either valid or faulty enabling the derivation of the class separating discriminant functions. Data consisting of 2881 spots, of which 2617 were valid and 264 faulty, was used to test the classifier. Each test spot was considered to be an independent sample. The results are presented with a ROC curve in Fig. 3 a.

The ROC curve characterizes the diagnostic accuracy of the classifier. The false positive rate is the probability of incorrectly classifying a valid spot and describes thus the specificity of our classifier. Equally, the true positive rate is the probability of correctly classifying a faulty spot. As random guessing would result in a linear curve connecting the points (0,0) and (1,1), our performance is much improved. Fig. 3 a shows, the true positive rate of our classifier is high even with rather low false positive rates indicating high sensitivity. However, perfect classifiers would have true positive rates equal to 1.0. Note that the false positive axis has been scaled from 0 to 0.2.

Attaining true positive rates close to one is difficult due to the various source and type of noise on the array. However, the optimal working point of the classifier can be found by associating costs with the different possible errors that can be made. This was done to assess the quality of replicate spot classification. The spots were considered in triplets with costs incurring each time a invalid spot is labeled as valid or with all valid spots being classified as faulty. The resulting curve is shown in Fig. 3 b.

The observing of Fig. 3 b shows that the location of the curve minimum is shifted from 0. The costs assigned to misclassifications introduces a bias into the class separating boundaries as the cost matrix is asymmetric. The classification costs are therefore minimal when the threshold equals ca. -6. With our data, this is the optimum working point. If more negative threshold is chosen, more faulty spots become labeled as valid reducing the sensitivity of the classifier. On the other hand, a more positive threshold reduces the specificity. However, costs also incur when threshold equal to -6 is chosen because the classifier is imperfect.

The nonsymmetric slopes of Fig. 3 b are due to the different variances of the features derived from valid and faulty spots. As the variance between the valid spots is small, the specificity decreases faster with increasing threshold than the sensitivity with decreasing threshold introducing costs. The features derived form high intensity noise are well separated from those derived from valid spots whereas the resemblance between valid spots and spot-like dirt is smaller. The noise spots that are very different from the valid ones become classified as valid only when the threshold is shifted very far away from the unbiased boundary. Thus, the slope is very gentle when moving in the reduced sensitivity direction.

5 Summary

Microarray technology offers new ways to explore the functions of the genome. For making reliable analyzes, the quality aspects of the data have to taken into account. In this paper, we proposed an automated classification of microarray image spots to classes faulty and good based on a on features derived form the spatial characteristics of the individual spots on the microarray. Assessment was presented for classification of individual spots using ROC analysis and for compound classification of replicate measurements using a non-symmetric misclassification cost matrix.

References

1. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.H. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
2. Yidong Chen, Edward R. Dougherty, and Michael L. Bittner. Ratio-based decisions and the quantitative analysis of cdna microarray images. *Journal of Biomedical Optics*, 1997.
3. Mei-Ling Ting Lee, Frank C. Kuo, G.A. Whitmore, and Jeffrey Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetetive cdna hybridizations. *Proc. Natl Acad. Sci. USA*, 2000.
4. Yee Hwa Yang, Michael J. Buckley, Sandrine Dudoit, and Terence P. Speed. Comparison of methods for image analysis on cdna microarray data. Technical Report 584, Department of Statistics, University of California, Berkeley, December 2000.
5. Xujing Wang, Soumitra Ghosh, and Sun-Wei Guo. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, 29(15), 2001.
6. Ron Dror. Noise models in gene array analysis. *Report in fulfillment of the area exam requirement in the MIT Department of Electrical Engineering and Computer Science*, 2001.
7. Mukund Thattai and Alexander van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proc. Natl Acad. Sci. USA*, 2001.
8. Ertugrul M. Ozbudak, Iren Kurtser Mukund Thattai, Alan D. Grossman, and Alexander van Ouderaarden. Regulation of noise in the expression of a single gene. *Nature Genetics*, 2002.
9. Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image Processing, Analysis and Machine Vision.* Chapman & Hall Computing, 1993.
10. David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining.* Adaptive Computation and Machine Learning Series. MIT Press, 2001.
11. Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification.* John Wiley & Sons, second edition, 2001.
12. J.P. Egan. *Signal Detection Theory and ROC Analysis.* New York: Academic Press, 1975.
13. John A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293, 1988.