Active Contours and a Background Intensity Estimator for Analysis of Microarray Spots

M. Katzer [1], K. Horvay [1], H. Küster [2], J. Landgrebe[1], S. Loop [1],
B. Spielbauer[1], E. Brunner[2] and T. Pieler [1]

[1] Zentrum Biochemie, Medizinische Fakultät,
Universität Göttingen, 37073 Göttingen
[2] Zentrum für Genomforschung, Fakultät für Biologie,
Universität Bielefeld 33615 Bielefeld
[3]Abteilung Medizinische Statistik, Medizinische Fakultät,
Universität Göttingen, 37073 Göttingen
Email: mkatzer@gwdg.de

Abstract. Differential gene expression experiments using DNA microarray technology yield raw data in the form of fluorescence images which must be segmented and quantified to extract the expression data.

This work proposes an active contour method for the precise segmentation of DNA spots in microarray images as well as an estimator for the unspecific background intensity.

The proposed methods are compared to commonly employed methods using data from experiments where some knowledge of the studied biological systems exists.

1 Problem definition

1.1 Introduction

Printed cDNA microarrays have become a widespread tool to screen many genes for differences in their expression under two or more different conditions by a concurrent hybridisation assay. Typical applications of this technique in medical research are the characterization of normal and pathological tissues by (global) expression profiles, gene regulation studies e. g. in metabolic diseases or genotyping for personalized treatments.

RNA from two different samples is differently labelled usin fluorescent dyes and hybridised to immobilised cDNA probes (the microarray). The relative intensities of the two dyes contain information on the relative quantities of different RNA sequences in the two samples. In overlaid fluorescence images of the microarray, the colour of the probe spots resembles the ratio of fluorescence intensities. Typical microarrays contain several thousands to several ten thousands of different probes, such that automatic image analysis methods are necessary.

Quantitative analysis of microarray images faces two major problems: The irregular shape of the mechanically printed spots of probe material on the array and the background signal, which is strongly influenced by the array surface processing before hybridisation.

1.2 Microarray technology

The cDNA probes of Microarrays are printed on glass slides which have a DNA-binding surface coating. After coupling the probes to the surface, the remaining area must be blocked against further binding of DNA to avoid background signal. This step is critical, since large quantities of blocking agent also produce a fluorescent film on the array. This primarily affects the DNA-free surface, since especially in the presence of blocking problems, there are spots with lower intensity than their surrounding ('black holes'). A spot is the surface area carrying probe DNA for one sequence.

1.3 Image analysis problem

The primary aim of quantitative microarray image analysis is the estimation of the two dye fluorescence intensities for each spot. Besides the desired signal, the image intensity contains fluorescence signal of surface contaminations and the slide itself.

It is common practice to estimate the dye fluorescence intensity by subtracting the mean or median intensity of a sample of image pixels on a circle or box surrounding a spot from the mean or median intensity inside the spot [1]. This local background estimation approach is not robust to contamination by e. g. dust particles and ignores the different surface properties inside and outside of the spot, however. Obviously the accurate segmentation of the spot regions is a closely related problem. Various algorithms have been proposed for this purpose which are based on geometric models of the spot shape (Axon GenePix$^{\mathrm{TM}}$, [2]), nonparametric testing [3], seeded region growing [4] and clustering of pixel intensities [5]. These approaches impose strong assumptions e. g. on the spot shape or require prespecified background samples or seed points, which may be difficult to provide robustly in an automated system.

2 Progress reported in this work

We describe a background intensity estimator with improved robustness to particle contaminations as well as an active contour model for the segmentation of DNA spots that is adaptive to variable shape and size of the spots and easy to initialize.

2.1 Modeling and Estimating Dye Fluorescence and Background Intensity

According to the insight that surface blocking causes different background signal in- and outside of the spots, we use two different background intensity models on the printed and DNA-free parts of the array surface. There is one component for e. g. glass fluorescence that acts on the complete array surface. For the DNA free surface area, a second component describes the more variable fluorescence of the blocking agent and dust particles.

Instead of looking at a small sample of local background intensity values, we use a larger window enclosing several adjacent spots (but without their signal regions). Typically, a histogram on this background region shows a prominent background peak with a strong right skew. The peak is associated with the common background component, such that we use the mode (the most frequent intensity value) as our estimator of background intensity. The enlarged background region and smoothing techniques improve the robustness of this estimator.

2.2 Semi-continuous Active Contour Model

Active contour models [6] have been applied successfully to many image segmentation problems. The approach describes object contours as continuous curves under an energy functional, which models curve flexibility and the fit to object edges. For efficiency reasons, discretised versions of the model (polygonal curve with integer vertex coordinates) are applied in practice.

Since the small size (5-50 Pixels in diameter) of microarray spots causes difficulties with the discretised approach, we have developed a more scale-independent semi-continuous active contour model with a polygonal structure and continuous vertex coordinates. The image energy component includes the relative orientation of image gradient direction and contour normal. This makes contour initialization less critical, since edges of spots next to the target spot will not be fit.

3 Results

The active contour segmentation and mode background estimation was evaluated using two datasets from a pilot study on root nodule development in the plant *Medicago truncatula* [7] (6 Arrays of 13824 spots) and from an experiment on vegetal localization of mRNA in *Xenopus laevis* oocytes (unpublished, 12 arrays of 9216 spots), which were previously analysed using the commercial image analysis system Imagene$^{\mathrm{TM}}$. These datasets were used for evaluation because they are comparably easy to interpret (*Medicago*) and have a large number of biologically relevant controls (65 spots representing the *Velo1*, *Dead end* and *XWNT11* genes of *Xenopus laevis*). These controls were confirmed by specific hybridisation experiments and are therefore more reliable than the homology based annotation of the *Medicago* data. We repeated the statistical analysis (Loess normalisation, t-test) of both experiments using raw intensity data extracted with our methods as implemented in the AIM (Automatic Image processing for Microarray experiments) system [8] and found the following results:

- Compared to local median background estimation, the mode background estimator lowers the proportion of spots having higher background than foreground estimation (i. e. missing values in the statistical analysis) from 8% to 0.6% for the Medicago data and from 7% to 2% with the Xenopus data.

Table 1. Sums of the p-value ranks of positive controls in the *Xenopus* experiment with different spot segmentation algorithms, signal (Fg) and background (Bg) intensity estimators. 'Zero' means no background correction was applied.

Program	Spot segmentation	Fg Estimator	Bg Estimator	Control ranksums
Imagene	Mann-Whitney	Median	Median	87849
Imagene	Mann-Whitney	Median	Zero	79015
AIM	Active Contour	Median	Mode	66658
AIM	Active Contour	Median	Zero	66327
AIM	Active Contour	Mean	Mode	64246
AIM	Active Contour	Mean	Zero	65824

- For the *Medicago* experiment, the ranking of candidate genes was almost reproduced. Ranking criterion was the estimated degree of differential expression. There was a correlation of 0.854 between the differential expression estimates in both analyses.
- For the *Xenopus* experiment, the ranking of the known positive controls was improved. Ranking criterion was the FDR-adjusted p-value calculated from a one sided t-test. Among the first 200 candidates there are 13 positive controls in the original analysis and 23 positive controls with our image analysis methods (first 300: 25/35, first 400: 36/42, first 800: 43/51, mode background and mean signal estimators were used). Table 1 states detailed results for different spot segmentation methods as well as background and foreground estimators.

For the comparison of spot segmentation methods, we did not subtract background intensity. Signal intensity of the spots is either estimated by the median or mean intensity within the signal region identified by the Mann-Whitney or active contour algorithm. The positive controls of the *Xenopus* experiment appear at better rank places when the signal is segmented using the active contour approach.

Background correction has counterproductive effects when the simple local median background estimator is used. This result is in accordance with recent findings of Qin and Kerr [9]. Our proposed mode background estimator does not clearly change the ranking when the median foreground intensity estimator is applied and slightly improves the ranking with mean foreground intensity estimation.

The rank sums in Table 1 are relatively high because there are candidates with stronger effect than the positive controls, and a small number of positive controls failed for technical reasons.

4 Discussion

We approximately reproduced the candidate ranking in the *Medicago* experiment, while there is clearly an improvement from the Mann-Whitney segmentation based analysis to the active contour segmentation based analysis of the *Xenopus* data.

The results on background estimation show that our assumption of different background intensity models inside and outside of the printed spots is useful in practice, although the choice of the segmentation algorithm seems to have more influence than background correction. The combination of mean signal and mode background estimators is recommended, since it improves the ranking of the positive controls.

The generally bad performance of background adjustment shows that this issue requires more research work on the sources of variance of the background estimators.

Altogether the results show that the accuracy of our methods is comparable or better than accepted methods as measured by biologically meaningful criteria. Extending the study to more different datasets is desirable.

5 Acknowledgements

Part of this work was funded by the *V*olkswagenstiftung.

References

1. Brown CS, Goodwin PC, Sorger PK. Image metrics in the statistical analysis of DNA microarray data. Proc Natl Acad Sci USA 2001;98(16):8944–8949.
2. Ekstrøm CT, Bak S, Kristensen C, Rudemo M. Spot shape modelling and data transformations for microarrays. Bioinformatics 2004;20(14):2270–2278.
3. Chen Y, Dougherty ER, Bittner ML. Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images. Journal of Biomedical Optics 1997;2(4):364–374.
4. Yang YH, Buckley MJ, Dudoit S, Speed TP. Comparison of Methods for Image Analysis on cDNA Microarray Data. Journal of Computational and Graphical Statistics 2002;11:108–136.
5. Bozinov D, Rahnenführer J. Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering. Bioinformatics 2002;18(5):747–756.
6. Kass M, Witkin A, Terzopoulos D. Snakes: Active Contour Models. International Journal of Computer Vision 1988;p. 321–331.
7. El Yahyaoui F, Küster H, Amor BBen, Hohnjec N, Pühler A, Becker A, et al. Expression profiling in Medicago truncatula identifies more than 750 genes differentially expressed during nodulation, including many potential regulators of the symbiotic program. Plant Physiol 2004;p. in press.
8. Katzer M, Kummert F, Sagerer G. Methods for Automatic Microarray Image Segmentation. IEEE Transactions on Nano-Bioscience 2003;2(4):202–214.
9. Qin L, Kerr KF. Empirical evaluation of data transformations and ranking statistics for microarray analysis. Nucleic Acids Res 2004;32(18):5471–5479.