Quantitative Analysis of Microarray Images

Leila Muresan, Bettina Heise, Erich Peter Klement
Department of Knowledge-based Mathematical Systems
Johannes Kepler University,
Linz, Austria
{leila.muresan, bettina.heise, ep.klement} @jku.at

Jan Kybic
Center for Machine Perception
Czech Technical University
Prague, Czech Republic
kybic@fel.cvut.cz

***Abstract*—The goal of quantitative analysis of microarrays is to determine the strength of the hybridization for every element of the array (spot). Since new technology allows the detection of the signal at single molecule level, new methods for analysis are necessary. A detection error of 10% is considered acceptable. In this paper we discuss three approaches to single peak detection inside the spots of the arrays, and compare the results we obtain on simulated and real images. These approaches are: global thresholding, an adaptive filter combined with local thresholding. We proposed a third algorithm, a statistical estimation of the background combined with clustering, which produces comparable results to well known algorithms, without having to perform the manual adjustment of the parameters.**

I. INTRODUCTION

From an image processing point of view, the task of quantitative microarray analysis consists of recognizing and counting single peaks in fluorescence images. (According to biological terminology, the elements of the microarray are called spots, and the bright signals inside them peaks. Peaks are diffraction limited point-like objects). Due to the development of an ultra-sensitive microarray platform, based on a combination of anti-adsorptive thin glass slides and the Cytoscout® (a scanning device with high throughput capabilities and high sensitivity, see [6], [11]) detection at single molecule level becomes possible. At this level, the current techniques [4, 5, 7, 8], based on the computation of the mean intensity for an array element, are obsolete. More sensitive and accurate methods are required.

The challenges of this task are robustness to noise, to different concentration of oligonucleotides (resulting in different densities of the peaks to be detected) and the total automation of the procedure. To one pixel of a commercial scanner correspond 400 pixels for the new technique. This leads to a considerable increase in the size of the images that have to be analyzed (the size of one image is approximately 8GB). The size of the images and the high density of peaks makes the peak counting task practically impossible for the human operator. Also, an efficient automatized method should not be computationally too expensive.

The results are tested on real as well as simulated data, since ground truth for since ground truth for this kind of images is not available.

This work was supported by the Austrian Genome Program GEN-AU and CEEPUS SK-42

0-7803-9134-9/05/$20.00 ©2005 IEEE

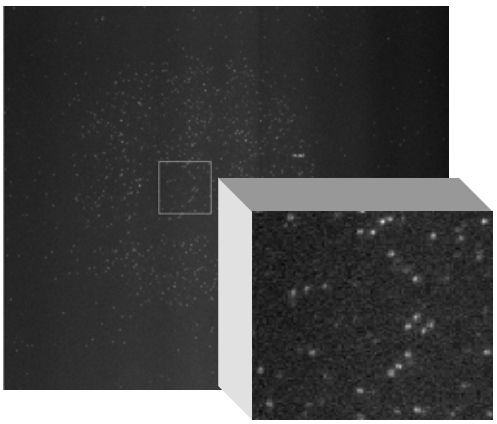


Fig 1. Oligonucleotide microarray peak with single molecule sensitivity (concentration: 0.8 amol/ 80 μl)

As real test images we used 16-bit scans of arrays hybridized with different concentrations of oligonucleotides (0.08, 0.8 and 8 amol/80 μl) resulting in different peak densities in the image.

For simulations, we generated images containing various numbers of peaks, and added a percentage of Gaussian and/or Poisson noise.

A pre-processing step in the scan analysis consists in detecting a region of interest, around the position of every single array element. Each such region is analyzed separately. In the following we discuss the use of two thresholding methods and one method based on robust estimation of the background.

II. PEAK DETECTION BY THRESHOLDING

As a first approach, a global threshold method was used to convert greyscale images into binary images and count subsequently the resulting blobs. As all the microarray images in our test dataset show a Poisson-like histogram, the triangle algorithm [14] seems suitable for a fast automated thresholding.

However, as a drawback of all global threshold methods, changing illumination profile causes background variations, which result in an overlap of the intensity distributions for background and foreground, making accurate peak detection impossible. Noise may further deteriorate the results.

In order to minimize these effects, in the second approach, we combine a local thresholding method with a previously applied local adaptive smoothing filter [9]. This filter reduces the noise in the original image $x_O(i,j)$ without affecting the peaks, as described by Eq.1,

$$x_{AF}(i,j) = m_l(i,j) \quad + \quad \frac{\sigma_l^2(i,j)}{\sigma_l^2(i,j)+\sigma_n^2}[x_O(i,j)-m_l(i,j)] \tag{1}$$

where ml is the mean intensity value in a local neighbourhood l of pixel (i,j), σ_l^2 is the variance in the same neighbourhood, and σ_n^2 denotes the variance of the noise estimated by $\sigma_n{}^2 = \frac{1}{NM}\sum_{i,j}\sigma_l(i,j)^2$, the mean of all local variances in the image.

The original image $x_O(i,j)$ as well as the filtered image $x_{AF}(i,j)$ are binarized into the resulting images bO and bAF, respectively, according to the local adaptive threshold [13], in Eq. 2 :

$$b(i,j) = \begin{cases} 1, & for \quad x(i,j) \geq m_l(i,j) + k\,\sigma_n \\ 0, & otherwise \end{cases} \tag{2}$$

where k is a constant a adapted to the imaging conditions.

Finally, a valid peak is found at location (i_0,j_0) if the following conditions are fulfilled:

- $b_O(i_0,j_0) = 1$,
- $b_{AF}(i_0,j_0) = 1$,
- $x_{AF}(i_0,j_0)$ local maximum in x_O,
- $x_{AF}(i_0,j_0)$ local maximum in x_{AF} (see Fig. 2).

The results for different concentrations are summarized in Table II.*A*. Although the results of the local adaptive threshold are acceptable and the method has relatively low complexity, an appropriate estimation of the factor k is necessary, causing additional adaptation efforts for changing concentration and SNR. For this reason, the use of statistical methods for background estimation is justified.

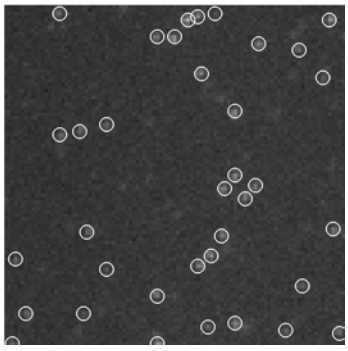


Fig 2. Results of the adaptive filter and local thresholding method for the detail image in fig.1 (k = 1.5, window size = 5)

III. Peak Detection Based on Statistical Background Estimation

The basic idea is to compute, for each image, certain features, which are sensitive to occurrence of peaks and subsequently, try to find a normally distributed model of the background, for each of these features. The peaks will represent outliers for these normal distributions, and by combination of outliers for different features, we try to eliminate false positives (assuming that noise is uncorrelated among the features). The method is suitable when sufficient background knowledge is available (the peaks represent less than a few percents of the background, which in real images usually is the case).

A. Features for peak detection

The features which are thought to be sensitive to occurrence of peaks are:

1. the variance of a 5 × 5 neighborhood,
2. the mean value of a 3 × 3 neighborhood (after applying the top-hat operation to the image, with a large structuring element, in order to eliminate the background),
3. the discrepancy and modified discrepancy value (described below),
4. the sum of the positive difference between the intensity value of the 4 (or 8 neighborhood) of a pixel and the mean value m of the 5 × 5 neighborhood: $\frac{1}{5}\sum_{|k+l|\leq 1}\left(x(i+k,j+l)-m\right)_+$, where
$$(a)_+ = \begin{cases} a, & if \quad a \geq 0 \\ 0, & otherwise \end{cases} \tag{3}$$
5. the result of Laplace filter.

The discrepancy norm $\| x \|_D$ on R^n is defined [3] as a mapping

$$\| x \|_D := \max_{1\leq\alpha\leq\beta\leq n} \left| \sum_{i=\alpha}^{\beta} x_i \right| \tag{4}$$

Considering that ideal peaks show a radial symmetric appearance and are more or less size restricted in our images by 5 x 5 pixels, the sequence X of the regarded pixels over which we calculate the discrepancy $D(i,j)$ was chosen in the following way:

$$\begin{aligned} X = (x_0, x_1, x_2, \ldots x_{23}) = \\ (x(i,j), x(i+1,j), x(i+2,j), \\ x(i,j), x(i+1,j+1), x(i+2,j+2), \\ \ldots, x(i,j), x(i+1,j-1), x(i+2,j-2)) \end{aligned} \tag{5}$$

starting repeatedly in the central point (i, j),. Subtracting the mean value m of X , the modified radial discrepancy DR. is:

$$D_R = \max_{1\leq\alpha\leq\beta\leq n} \left| \sum_{i=\alpha}^{\beta} (x_i - m) \right|, \tag{6}$$

For reduction of computing time, only local maxima were considered as central points. An analysis of "sensitivity" of

each feature (the sensitivity to peaks, to the dimension of the chosen neighborhood, to noise) is necessary.

B. Outlier detection

In order to detect outliers in every computed feature, we shall use the "*modified z-score method*", for normal distributions. Let z_{ij} be the value of a feature at pixel *(i,j)* (described in section *A*). The distribution of the feature for the background is standardized, using robust estimators, and then outliers are detected. The method consists of the following steps [10, 12]:

1. Compute the median *med* of the z_{ij}
2. Compute the median *MAD* of $\left|z_{ij} - med\right|$
3. If the following inequality holds:

$$0.6745 \cdot \frac{z_{ij} - med}{MAD} > t \qquad (1.)$$

then z_{ij} is an outlier with respect to the feature z.

The effect of the threshold t in Eq. 7 will be discussed below. However one should keep in mind that, in case of normal distribution, 99.7 % of the data is closer to the mean than 3σ, where σ is the standard deviation of the distribution and for outlier detection, usually a value of $t = 3.5$ is chosen.

For variance, the distribution is χ^2 and not the normal one. However, for large sample sizes (as in or case), small deviations from normality will not compromise the result of the test ([10]).

C. Counting peaks

Since, in practice, the result still contains a lot of background data, some further steps are necessary. In order to select only the data related to the peaks, clustering algorithms are applied.

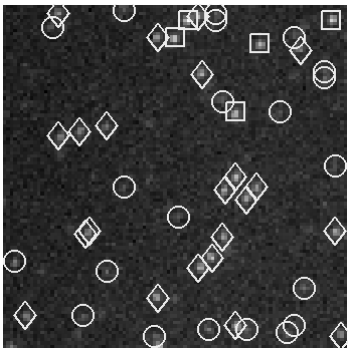
Since the appearance of the peaks shows a great variability, (due to background variance, focus etc.) it seems more reasonable to divide the candidates, instead of two, into three categories: rejected candidates, weak candidates, strong candidates. We tested two clustering algorithms: fuzzy c-means and Gustafson-Kessel. The Gustafson-Kessel method (see [1]) provides better results, since it is sensitive to the size and shape of the clusters.

IV. RESULTS

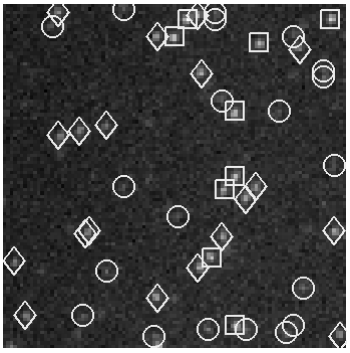
In Table 1, we present the number of peak candidates, found as a result of outlier detection for the features 1 and 2 (variance and mean). The test images contain 100, 500 and 1000 peaks, respectively. The background intensity is a tenth of the mean peak intensity, and Poisson noise was added.

TABLE I. NUMBER OF PEAKS SELECTED AS OUTLIERS FOR DIFFERENT THRESHOLD VALUES (FEATURES: VARIANCE AND MEAN).THE RESULTS VARY LESS THAN 10%

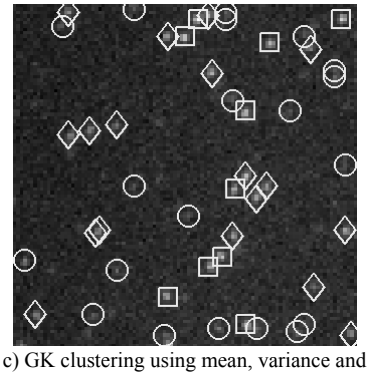
Generated peaks	**Detectedpeak candidates**		
	t = 3.5	**t = *3.0***	**t= *2.5***
100	95	99	109
500	503	521	546
1000	961	988	1021



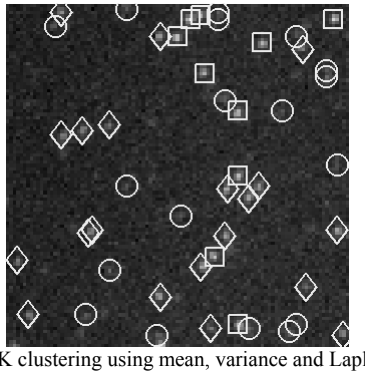
a) GK clustering using mean and variance as features



b) GK clustering using mean, variance and volume estimation as features



c) GK clustering using mean, variance and discrepancy as features



d) GK clustering using mean, variance and Laplace filter value as features

Fig 3. Detail of clustering result (GK clustering with different features) for microarray image having concentration 0.8 amol/80 µl. The three resulting clusters are:□ – clear peak, ◊ - peak out of focus, ○ - rejected peak candidate

Since the variation is less than 10%, no further processing is necessary. In real images, peaks show a great variability, the features change if the peak is defocused, and the SNR of the recorded images is low. In order to minimize these negative effects, we lower the outlier detection threshold to t = 3, introducing some background elements.

In order to reflect the characteristics of wide-field fluorescence microscopy imaging, we perform a clustering in three classes (focused objects, defocused objects and noise). The class of candidates least resembling to the assumed model is discarded as the elements introduced by noise. The weak candidates represent mainly defocused objects, while the set C_3 is the set of peaks in focus. The results of the Gustafson-Kessel clustering for three different concentrations are summarized in Table II.*B*. The three values for each t represent the cardinality of each cluster (the first is the number of rejected peaks, while the second and the third represent the cluster of defocused and focused peaks, respectively). The final result of detected peaks is $C_2 \bigcup C_3$. In the "*Features*" column, the features used for clustering are enumerated (denoted as in III.*A*). The set of peaks corresponding to focused and defocused objects varies little with the features used for clustering, the most significant differences occurring at very high concentrations.

TABLE I. A) NUMBER OF PEAKS IN MICROARRAY IMAGES USING ADAPTIVE FILTER AND LOCAL THRESHOLDING (K = 1.5)

Concentration (amol/80µl)	**Detectedpeak s**
0.08	101
0.8	875
8	3808

B) NUMBER OF PEAKS IN MICROARRAY USING DIFFERENT FEATURE COMBINATIONS AND GK CLUSTERING (C1 REJECTED, C2 DEFOCUSED, C3 FOCUSED OBJECTS)

Concentration (amol/80µl)	**Features**	**Detected peaks**					
		t = 3.5			t = 3.0		
		C_1	C_2	C_3	C_1	C_2	C_3
0.08	1, 2, 3	104	44	27	141	58	30
	1, 2, 4	98	54	23	109	74	46
	1, 2, 5	105	46	24	145	56	28
	1, 2	107	48	20	152	56	21
0.8	1, 2, 3	503	560	158	553	610	175
	1, 2, 4	503	530	188	585	561	192
	1, 2, 5	532	493	196	596	543	199
	1, 2	578	531	112	645	566	127
8	1, 2, 3	4409	1032	1040	4566	1132	1081
	1, 2, 4	3365	1267	1849	3048	2647	1084
	1, 2, 5	3118	2408	955	3218	2573	988
	1, 2	3258	2556	667	3535	2628	616

The complexity of the proposed algorithm is dependent mainly on the computation of the z-statistic for each feature – section III.*B* (an approximate value of the median can be found as in [2], in $O(n)$ steps, where n is the size of data) and on the relatively high complexity of the Gustafson-Kessel algorithm (since it implies a matrix inversion step) – section III.*C*.

CONCLUSIONS

The task of peak counting for very large images, with low SNR cannot accommodate human intervention. The last method presented in the paper produces comparable results to well known algorithms, without having to perform the usual adjustment of the parameters. The results of this method can be used in a further analysis of peaks. Nevertheless, additional tests are necessary as well as an analysis of the sensitivity to different peak concentration, SNR and varying imaging conditions. An improvement of the algorithm time-complexity can be achieved by using a sub-sample of the original data and by replacing the Gustafson-Kessel clustering with a less computationally expensive one.

ACKNOWLEDGMENT

The authors thank the members of the Biophysics Institute, at Johannes Kepler University (Linz) and the Upper Austrian Research for interesting discussion and suggestions.

REFERENCES

[1] Babuska R.- Fuzzy modeling for Control, Kluwer, 1998

[2] Battiato S., Cantone D., Catalano D., Cincotti G., Hofri M. - An efficient algorithm for the approximate median selection problem, in: G. Bongiovanni, G. Gambosi, R. Petreschi (Eds.), *Proceedings of the Fourth Italian Conference*, CIAC 2000

[3] Bauer P., Bodenhofer B., Klement E.P., A fuzzy system for image pixel classification and its genetic optimization. In Trappl R. (Ed.), *Cybernetics and Systems* '96. Austrian Society for Cybernetic Studies, Wien, Vol. 1, pp. 285--290. 1996

[4] Braendle N., Bischof H., Lapp H. – Robust DNA microarray image analysis, *Machine Vision and Applications*, 15, pp. 11-28, 2003

[5] Gwynne P., Page G. - Microarray Analysis: the next revolution in molecular biology, *Science*, 1999

[6] Hesse J., Sonnleitner M., Sonnleitner A., Freudenthaler G., Jacak J., Höglinger O., Schindler H., Schütz G.- Single molecule reader for high-throughput bioanalysis, *Anal. Chem.* , 76, pp. 5960-5964, 2004

[7] Jain A. N., Tokuyasu T. A., Snijders A.M., Segraves R., Albertson D, G. Pinkel D.- Fully Automatic Quantification of Microarray Image Data, *Genome Researc*h, Vol. 12, Issue 2, 325-332, 2002

[8] Katzer M., Kummert F., Sagerer. G. - A Markov Random Field Model of Microarray Gridding., *Proc. 18th ACM Symposium on Applied Computing*, 2003

[9] Kuan DT, Sawchuk AA, Strand TC, Chavel P. Adaptive noise smoothing filter for images with signal-dependent noise, *IEEE Transaction on Pattern Analysis and Machine Intelligenc*e, 7(2), 165-177, 1985

[10] Moore D., McCabe G. – Introduction to the practice of statistics, W.H. Freeman and Co., New York, 492-569, 2002

[11] Schindler H.. Vorrichtung zur Visualisierung von Molekülen. PCT/AT99/00257, international patent, 2000

[12] Staudte R.G, Sheather S. J. - Robust estimation and testing, Wiley, 1990

[13] Trier OD, Jain AK.. Goal-Directed Evaluation of Binarization Methods, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995, 17 (12): pp. 1191-1201

[14] Zack G.W., Rogers W.E., Latt S.A.. Automatic Measurement of Sister Chromatid Exchange Frequency. *J. of Histochemistry and Cytochemistry*, 25(7): pp. 741-753, 1977