Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

Automatic analysis of microRNA Microarray images using Mathematical Morphology

Felix Meyenhofer, Olivier Schaad, Patrick Descombes and Michel Kocher

***Abstract*—The micro array are an experimental technique for parallel determination of molecular concentration. The image analysis is an important, time consuming and error prone step of the process. We describe here an automatic procedure able to analyze the micro array data and to accurately provide the level of concentration for each microRNA (miRNA). The proposed method has the advantage, compared to commercial products, to minimize the user interaction, leading to a more reproducible data analysis**

I. Introduction

MICROARRAY is a technology that enables whole genome studies for gene expression and genotyping. Microarrays combine two technologies: miniaturization and parallelization. Miniaturization in the sense that the size of the individual experiment is on the order of 5 to 50 mm and parallelization in the sense that the same protocols are simultaneously applied to one thousand to one million individual experiments. This approach was first developed by Pat Brown [1] for DNA/ DNA hybridization. It was then extended to protein-protein interactions [2] and recently to microRNA [3]. The method requires the generation of arrays of thousands to millions of probes printed or synthesized on a solid surface.

The experiment consists in the recognition of a labeled target by a specific probe on the array. The intensity of the readout at a specific location will give information about the concentration of the target. For example, in order to assay the accumulation of transcripts, a typical microarray experiment will consist in extracting and purifying the total RNA of the samples, reverse transcribing the mRNA into cDNA, labeling the cDNA via in vitro transcription, then hybridizing the solution of targets to the arrays of probes. A direct comparison of two samples can be achieved either by hybridizing each sample onto individual arrays (single color arrays; e.g. Affymetrix, Illumina) or by labeling the two samples with two different dyes such as Cy3/Cy5 or Alexa3/ Alexa5 and cohybridizing them onto a single array (dual color arrays; e.g. Agilent, custom arrays). After hybridization, fluorescence measurements are made for each dye separately. The measurements are subsequently processed in order to evaluate the level of expression of a specific transcript on single color arrays, or the ratio of expression when dealing with a co-hybridization experiment on dual color arrays. The principle of producing an image of multiple, variable spots is one of the common features shared by many microarrays technologies. The acquisition of the fluorescent image can be obtained by different methods such as confocal laser or CCD camera scanning. The digitization process will produce an image that will contain the signal intensity representing the total fluorescent in a small region (a square of 5 x 5 μm for Affymetrix, a bead of 3 μm diameter for Illumina, a spot of 50 to 100 μm diameter for most dual color systems). The next step is to extract the information for each area where a probe has been printed, and then to evaluate the expression levels. A major issue is to accurately and reproducibly quantify spot intensities and shapes. This is particularly critical for arrays showing low degrees of reproducibility between batches. The segmentation of the raw image into individual probes is either completely automatic in large (more 10^6 probes) commercial arrays such as Affymetrix, Agilent or Illumina, or manual with the help of commercial softwares such as Imagene™, ScanArray™, GenePix™ for custom arrays or commercial low density arrays. The semi automatic procedure usually present in image analysis softwares for microarray feature extraction encounter problems such as being time-consuming and user-dependent. The variability of the results mostly depends on how much time the user will spend inspecting the data and the same user will often produce different results at different sittings.

Several methods using different segmentation algorithms have been developed to automatically extract signals from arrays [4-7], recently summarized in an evaluation study [8].

Here we present a signal extraction method that relies on mathematical morphology operators. This approach has been applied to extract signals from miRNA microarrays from Invitrogen (NCode™ Multi-Species miRNA Microarray). This particular type of microarrays are designed for the detection of microRNA in different species including human, rat, mouse, C. Elegans, drosophila, zebrafish on a single microarray. This design implies a high number of blank spots and cross-hybridizing spots, in addition to the usual problems of microarrays artifacts such as scratches, doughnut-shaped spots that are usually present. We have developed a program that requires as little human intervention as possible to limit subjective variability, using information on the general design of the array as input and in particular the presence of references spot (anchors) at each corner of the sub-array in order to initially position the grid. The output created is a tab delimited file with the signal and the background for each spot.

II. Micro array data

NCode™ miRNA arrays contains 16 grids arranged in a layout called meta grid. An artificial miRNA image is shown in figure 2 left. The upper left corner of each grid can be identified and are named Upper Left Corner (ULC) spot as visualized in figure 2 right. Each grid contains several

Manuscript received April 2, 2007. This work was supported in part by the swiss mational fund (NCCR Frontiers in Genetics), the University of Geneva and the University of Applied Sciences in Geneva.

F. M. and M. K. are members of the University of Applied Sciences in Geneva (corresponding author e-mail: Michel.Kocher@ hesge.ch).

O. S. and P. D. are members of the University of Geneva

hundreds of quasi circular spots. The intensity of each spot are very diverse and reflect the concentration of miRNA. Due to the experimental process, the image suffers two major drawbacks. First a geometrical distortion due to the mechanical printing process. Second, the background noise intensity is not homogeneous and has to be estimated locally for each spot. Note: the method has been developed using NCode arrays as a template. However, it can be modified at whish for the analysis of other formats of microarrays.

III. Methodology

The methodology developed to identify each spot and to compute their average internal and external intensity is divided into three steps. First, the rough localization of the meta grid (this grid contains 16 grids). Second, the precise detection of each grid. Finally, within each grid, the detection of each spot and the computation of the average inner and outer intensity.

A. Meta grid detection

This section describes the algorithms used to roughly align the data to an artificial meta-grid. To do so, a global rotation is applied to the data in order to align as well as possible the ULC spots to an artificial Cartesian grid. The idea used here is to project the data on a vertical axis and to iteratively apply a rotation to the data, using a step of 0.1[deg], until the projected variance is minimum. This optimization algorithms is depicted in figure 1

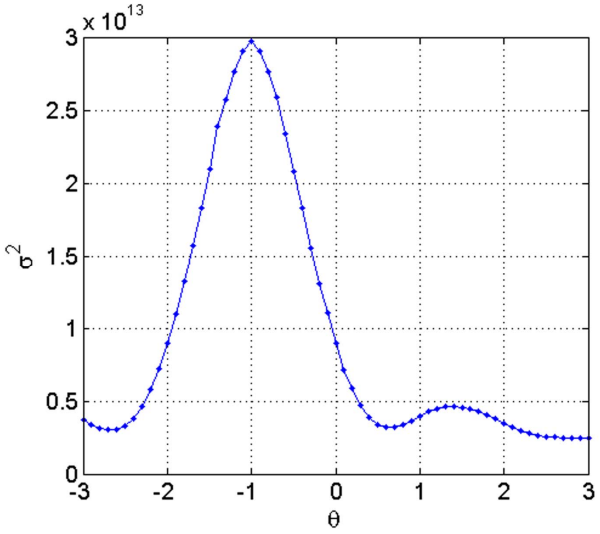

Fig. 1. Evolution of the projected variance as a function of the rotation angle. The maximum is obtained for θ = -1 [deg].

To have a complete localization of the meta-grid, the ULC spots have to be individually detected. This is achieved using four criteria, the intensity, the shape, the size and the position. First an open-close [9] operation is performed to get rid of the noise (shape criterion). Second, a non critical manual threshold is applied to detect the very bright spots (intensity criterion). Third, a binary area opening [10] is applied to get rid of the too small clear noise (size criterion). Finally, based on the information contained in a file regarding the ULC spots position and on the horizontal and vertical projections of the processed data, the Region Of Interest (ROI) in which the ULC spots have to be located is computed. This file, called GAL describes block and feature-indicator positions and geometry.

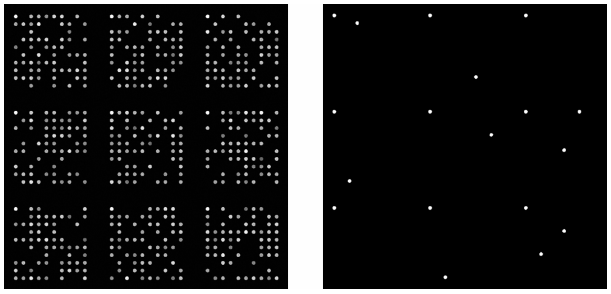

Fig. 2 Artificial image made of 9 blocs before and after the three first steps. More than nine spots have been detected by the image analysis method and the information contained in the GAL file is required in order to select the ULC spots.

B. Single grid detection

The goal of this procedure is to create an artificial grid adapted to each bloc. This final grid will be used in the following section as a marker to precisely locate each spot. This procedure is made of 4 steps which are respectively the edge preserving noise cleaning, the binarization of the numeric de-noised image, the matching between an artificial grid and the detected spots and finally, an affine mapping of the artificial grid on the real data.

First, the noise has to be decreased by using a leveling filter. This auto-dual filter [11] consists in applying an auto-dual reconstruction from a marker image into a mask image. The marker image is a filtered version of the original image and has the advantage of having a very small noise and the disadvantage of having mislocalized boundaries. This image is obtained by applying a very strong median filter (20x20), about the size of a spot, to the original data. The mask image is, in this case, the original image itself.

Second, the denoised numeric image is binarized by applying a Rank Hit Or Miss Transform (RHMT) [9]. This filter uses both shape and intensity description of the spot and of the background to individually detect each spot.

Third, an artificial grid is computed based on the spatial information contained in the GAL file and on the location of the ULC of each grid. The center of mass of each detected spot is computed. Based on the fact that most of the spots have a very low intensity (noise level), there is not a one to one match. The candidates are obtained by performing the intersection between the artificial grid and the dilated version of the centers of mass computed on the real data. The radius of the structuring element is based on the inter-spot distance given in the GAL file.

Fourth, an affine transform is applied to the artificial grid to best match, in the least square sense, the position of the detected spots in the original data.

Figure 3 illustrates the describe aligment procedure.

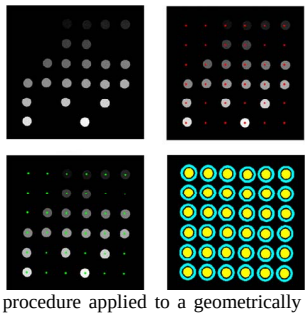

Fig. 3 Alignment procedure applied to a geometrically distorted grid. (A) original image, (B) overlaid by the grid dots, (C) overlaid by affine transformed grid dots, (D) measure masks for the object and the background

C. Spot detection

Each single grid having being detected, the last task to be achieved is to detect, every spot in order to compute the mean intensity within the spot and to compare it with the mean intensity of the environment (noise). To do so, two alternatives have been investigated. The first, using the HMT Opening [9], which is based on the *a priori* shape knowledge. The second, using the watershed transform [12] does not use this a priori shape knowledge, and relies only on the real data information. The marker image of the watershed transform is obtained by the centers of the detected spots after grid alignment. The segmentation function is computed by applying a morphological gradient [9] to the de-noised image obtained by the leveling described in the previous section. The de-noising step is useful in order to prevent any leakage in the watershed process as illustrated in figure 4.

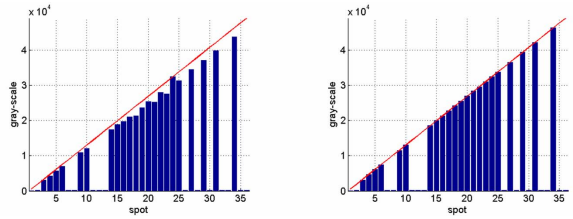

Fig 4. Illustration of the advantage of using a data-based detection (watershed), on the left, compared to a method based on an *a-priori* knowledge of the shape (opening HMT), on the right. In this figure, artificial data made of spots with linear growing intensity are segmented using the two approaches.

D. Evaluation of the performances by using artificial data

In order to evaluate the performance of the above-mentioned algorithms, the artificial micro array image shown in figure 5 containing four blocks, each block containing 100 spots which intensities varying linearly between 5000 and 50000 (16 bit) has been created. This image was rotated by 0.6 [deg] in the clockwise orientation, and a separate affine transform was applied to each bloc. Furthermore, noise was added to the image so that 30% of the spots fall in the noise range.

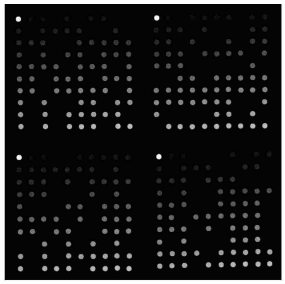

Fig 5. Test image used as ground truth to evaluate the performance of the proposed method.

The detected rotation angle is 0.7 [deg], the average error position for the ULC spots is 1 pixel and, after having performed the affine transform on each grid, the average error position between the artificial grid and the real one is of 1.28 pixel. Figure 6 describes the very good agreement between the theoretical and detected mean inner intensity.

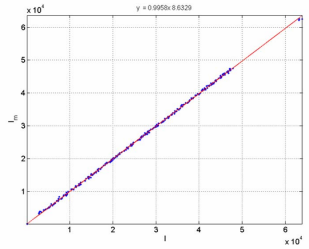

Fig. 6 Scatter plot obtained by analyzing the data presented in figure 5. The x-axis is the true value intensity whereas the y-axis is the measured intensity provided by our algorithm.

E. Evaluation of the performances by using real data

The algorithms described in this paper have been applied to real data and a comparison was made between the inner average value of the spots computed by the ImaGene© commercial software and by our own software (Fig. 7). Globally we observed a very good correlation between the two methods for the vast majority of the data.

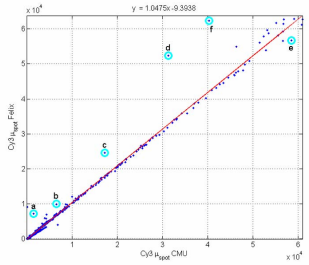

Fig. 7 Scatter plot comparing the results produced by the two softwares. The x-axis is the estimated intensities measured by ImaGene© software and the y-axis is the estimated intensities measured by our algorithm.

Because of the absence of ground truth, it is impossible to decide which solution performs the best. Nevertheless, by carefully examining each spot for which the inner average value is sensibly difference, it is possible to get an insight about the reason of this difference (Fig 8).

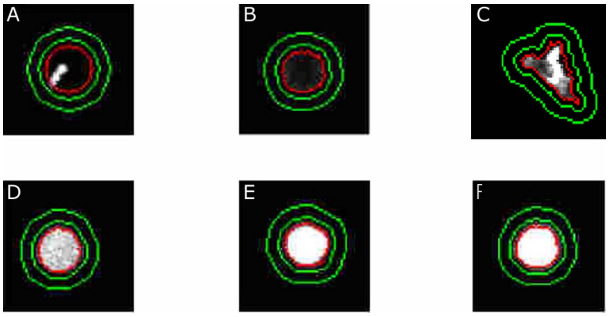

Fig. 8 Illustration of some strangely shaped spots observed on real data images.

To explain these differences, our hypothesis is that the ImaGene© software is using a fixed template to compute the inner average value whereas, in our solution, the watershed, a more shape adaptive method, has been used.

IV. Conclusion

The use of Mathematical Morphology as the paradigm for data analysis provides the following advantages. First, because of the strong mathematical foundation, it leads to reproducible results. Second, the choice of every parameter used in the algorithms is motivated by physical measurements such as shape and dimensions. The last point allows easy customization of the program for each commercial micro array.

In our approach, we choose to analyze each spot by segmenting them instead of assuming a perfect circular shape. We believe that this improves the quality of the results because of the high variability of the shapes we have observed.

We have applied the method to miRNA microarrays image (4000x4000 16 bits pixels) and compared the data obtained with a commercial microarrays analysis software (ImaGene©). Our initial evaluation demonstrates that the software developed performs extremely well for estimating the signal intensity. In addition, the whole analysis can be conducted with minor input from the user and most importantly, with very good control of subjective variability. We thus feel confident that this new software will be a valuable tool for extraction of signals from microarrays images.

Acknowledgment

The Authors thank Prof Walter Reith, Dr Isabelle Dunant, and Dr Mylene Doquier for providing microarrays data.

References

[1] P. O. Brown and D. Botstein, "Exploring the new world of the genome with DNA microarrays," *Nat Genet*, vol. 21, pp. 33-7, Jan 1999.

[2] L. A. Kung and M. Snyder, "Proteome chips for whole-organism assays," *Nat Rev Mol Cell Biol*, vol. 7, pp. 617-22, Aug 2006.

[3] L. A. Goff, M. Yang, J. Bowers, R. C. Getts, R. W. Padgett, and R. P. Hart, "Rational Probe Optimization and Enhanced Detection Strategy for MicroRNAs Using Microarrays," *RNA Biol*, vol. 2, Jul 20 2005.

[4] J. Angulo and J. Serra, "Automatic analysis of DNA microarray images using mathematical morphology," *Bioinformatics*, vol. 19, pp. 553-62, Mar 22 2003.

[5] R. J. Laws, T. L. Bergemann, F. Quiaoit, and L. P. Zhao, "SignalViewer: analyzing microarray images," *Bioinformatics*, vol. 19, pp. 1716-7, Sep 1 2003.

[6] E. Novikov and E. Barillot, "Software Package for Automatic Microarray Image Analysis (MAIA)," *Bioinformatics*, Jan 19 2007.

[7] A. M. White, D. S. Daly, A. R. Willse, M. Protic, and D. P. Chandler, "Automated Microarray Image Analysis Toolbox for MATLAB," *Bioinformatics*, vol. 21, pp. 3578-9, Sep 1 2005.

[8] A. Lehmussola, P. Ruusuvuori, and O. Yli-Harja, "Evaluating the performance of microarray segmentation algorithms," *Bioinformatics*, vol. 22, pp. 2910-7, Dec 1 2006.

[9] P. Soille, "Morphological Image Analysis," *Springer Verlag, 2nd edition, corrected second printing*, 2004.

[10] L. Vincent, "Morphological Area Opening and Closing for Grayscale Images,," *Proc. NATO Shape in Picture Workshop, Driebergen, The Netherlands, Springer-Verlag*, pp. 197-208, 1992.

[11] F. Meyer, " The levelings.," *In Heijmans H.J.A.M. and Roerdink J.B.T.M. (eds.), /Mathematical Morphology and its Applications to Image and Signal Processing, /(Proc. ISMM'98, Amsterdam, June 1998), Kluwer*, pp. 199-206, 1998.

[12] S. Beucher and F. Meyer, "The morphological approach to segmentation: the watershed transformation," *in: E. Dougherty (Ed.), Mathematical Morphology in Image Processing, Vol. 34 of Optical Engineering, Marcel Dekker, New York*, pp. 433--481., 1993.