Robust Microarray Image Segmentation and Quantification Using Extended Local Background

Kai Zhang[1] Marc Ma[1, 2] Hui-Yun Wang[3] Yu Wang[1] Tongsheng Wang[1, 4] Li Jia[1] I-Jen Yeh[1] Frank Shih[1] Patricia Soteropoulos[4]

[1]Dept. of Computer Science, [2]Center of Applied Mathematics and Statistics, New Jersey Institute of Technology
[3]Dept. of Molecular Genetics and Microbiology/The Cancer Institute of New Jersey, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School
[4]Center for Applied Genomics, Public Health Research Institute
Emails: {kxz6841, qma}@njit.edu, soteropoulos@phri.org

Abstract

DNA microarrays consist of thousands to hundreds of thousands of spots that must be scanned and digitized for further analysis. Thus, image processing is an important step for microarray data analysis. In general, three steps are carried out in image processing: gridding, segmentation and quantification. Errors in the first two steps will result in unreliable quantification in the last step. Since quantification is critical in the subsequent analysis, it is important that the first two steps are performed reliably and accurately. In this paper, we focus on a robust microarray image segmentation method based on a novel concept of using larger regions to estimate the local background noise values and local thresholds for segmentation. This larger region is termed as extended local background (ELB). We demonstrate that ELB is a viable method for robust image segmentation using a set of simulated Gaussian distributed noise data and real microarray images.

Keywords: DNA microarrays, robust image segmentation, extended local background model.

1. Introduction

Microarrays have become one of the most widely used functional genomics research tools and have revolutionized the way scientists approach biological problems [1-5]. Microarrays allow scientists to study the function, behavior or genetic variation of thousands of genes simultaneously. Studying gene expression on a genome-wide scale has played, and will continue to play, a major role in understanding complex biological processes. Microarrays have also been widely used for identification of single nucleotide polymorphisms (SNPs), for studying gene methylation and for studying protein abundance or function. Once a microarray has been scanned, the image must be digitized for further analysis. For spotted arrays in particular, image processing has been one of the bottlenecks in microarray data analysis [6].

In general, three steps are carried out in image processing: gridding, segmentation and quantification. Errors in the first two steps will result in unreliable quantification in the last step. Since quantification is critical in the subsequent analysis, it is important that the first two steps are performed reliably and accurately. Generally speaking, the goal of microarray image segmentation is to distinguish the foreground signals from the background signals, that is, to identify the intensity contribution due to the specific hybridization of the DNA samples.

There are several methods being used to estimate microarray background. Global background (GB) estimation calculates the average intensity level of all the pixels not belonging to signal regions. Thus, GB ignores the spatial background variation across the whole slide. Several other spatial or histogram-based techniques have been proposed for analyzing microarray images to overcome this limitation. Fixed circle (FC) segmentation fits circles with a constant diameter to all the spots in the image [7]. Adaptive circle (AC) segmentation estimates the circle's diameter individually for each spot [8] and may generate more reliable estimates. However, AC does not properly handle irregular shapes such as donuts. Adaptive shape (AS) segmentation offers a more flexible solution for dealing with irregular shapes, but it cannot give robust estimation for the foreground or background when large local variation of intensities exists. BlueFuse [6, 9], which is one of the histogram-based methods, uses a Bayesian model to generate a confidence measure for each spot. In comparison to the spatial-based approaches, histogram-based methods do not spend processing time in analyzing spatial distribution for each spot. Instead, they directly analyze the histogram distribution of local spot regions in which the pixels are categorized into

foreground and background based on some criteria. However, the quantization in histogram-based methods is unstable when a large target mask is set to compensate for spot size variation [10].

The density of spots on spotted microarrays continues to increase as more genes are identified and printing technology improves. With fewer pixels for local background noise estimation in the segmentation stage, the traditional approaches cannot make significant statistical sense. In this paper, we focus on a robust microarray image segmentation method based on a larger region to estimate the threshold of local noise for segmentation and quantification purposes. This larger region is termed as extended local background (ELB). We demonstrate that ELB is a viable method for robust image segmentation and quantification using a simulated Gaussian random noise image and real microarray images.

2. Flexible ELB Local Noise Model

In general, the variance of local background noise increases as the number of pixels used for background estimation decreases, which will result in very unreliable estimates of local noise. In this paper we introduce ELB, a novel concept for robust local noise estimation. By using a larger local region when compared with other spatial and histogram based methods, ELB overcomes many of the limitations associated with these methods. Unlike the spatial based methods such as FC, AC and AS, ELB does not assume any spot shape; neither does it spend any time in computing spot shapes, which makes ELB very efficient. Essentially, ELB is a histogram-based method. It uses local histogram information on a suitable (larger) population to generate better and more robust estimates of local noise. Fig. 1 shows one allowable ELB configuration. Pixels in the gray shaded area are initially tentatively treated as the candidate background region subject to later refinement.

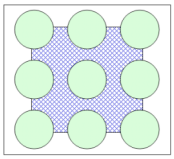


Fig. 1: Extended local background.

Fig. 2 illustrates one example of background histogram distribution based on the ELB for a grid point from a spotted microarray image (CAG human 19K oligo array). From this histogram, useful statistical values such as mean, median and variance can be estimated. In ELB, we use the following equation to compute the threshold value for segmentation, which is the sum of the mean value of intensities of candidate background pixels, $\bar{v}$, and twice the standard deviation, $\sigma_{background}$.

$$v_{cutoff} = \bar{v} + 2\sigma_{background}. \quad (1)$$

Within the proximity of the spot region, only pixels with intensities above this threshold will be classified as candidate foreground pixels from which we compute the median value, which is assumed to be the sum of true signal intensity and noise intensity, $v_{(signal+noise)}$. Finally, the true signal intensity is calculated using the following equation in which $v_{ELB} \equiv \bar{v}$.

$$v_{TrueSignal} = v_{(signal+noise)} - v_{ELB}. \quad (2)$$

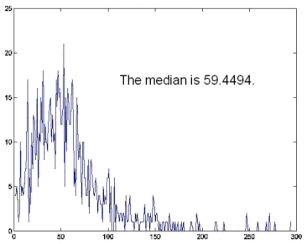
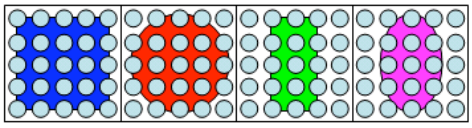


Fig.2: Background histogram distribution based on the ELB.

We allow a flexible ELB configuration in terms of shape and size to be customized by users. Fig. 3 shows different configurations of ELB including square, circle, rectangle and ellipse. The size of ELB can be measured by the length of the side of the square region, the length and width of rectangular region, the diameter of circular region, and the lengths of the long and short axes for elliptical regions. The length of sides or axes can be defined in terms of numbers of spots or pixels. For example, to set a rectangular ELB with size of 5 by 3 in terms number of spots, one will get the configuration as shown in Fig. 3 (c).



(a) (b) (c) (d)

Fig. 3: Different allowable configurations of ELB.

Fig. 4 shows the local background region configurations [11] of some popular microarray image processing software packages ScanAlyze, QuantArray, ImaGene and GenePix.

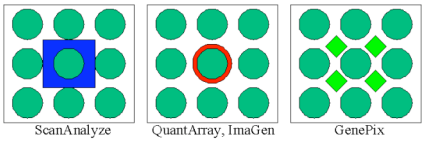


Fig. 4: Background region definitions of ScanAlyze, QuantArray/ImaGene and GenePix.

To estimate local background noise, ScanAlyze allows the user to define a size of a square region,

which is similar to our ELB method, however only square regions are allowed in ScanAlyze. QuantArray, ImaGene and GenePix use fixed regions. One obvious disadvantage for fixed regions is that if the specific spot is defective for any reason, the defect will significantly affect the background estimation.

For high-density DNA microarrays with dense grid layout (such as those with spot diameter 20 pixels and distance between two adjacent spots 26 pixels), the area of the ELB region is on the order of 2000 square pixels, significantly larger than the area used by QuantArray, ImaGene or GenePix in estimating local noise level for each spot. The ELB approach can also tolerate more errors in the gridding stage. Since it is a histogram-based method, it is not sensitive to the errors in spatial arrangement such as grid location. In terms of the size of the region used for local noise estimation, both the ELB and ScanAlyze approach are better choices in generating robust local noise estimates in microarrays with dense a grid layout.

3. ELB on Simulated Noise Data

We performed a simple numerical experiment to establish the statistical basis of ELB and validate that a larger population of pixels yields more robust statistical estimates. We generated a square Gaussian random noise image with 500 pixels on each side. The probability distribution function is defined in Eq. 3.

$$f(x \mid \mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad (3)$$

in which we set the mean value as $\mu = 300$ and the standard deviation as $\sigma = 200$. The random image and the noise levels are shown in Fig. 5.

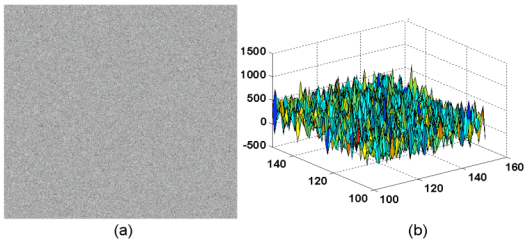


Fig. 5: (a) Gaussian noise image (500 by 500 pixels) for ideal experiment; (b) Intensity level of a sub-image from (a).

We select 225 grid points evenly spaced on the random noise image, Fig. 5 (a) and calculate the mean value and standard deviation of the pixel intensities at each one of these locations with varying size of the local region. The size of each local region is determined such that no overlapping of adjacent local regions will occur to help isolate factors that may influence the statistics. Fig. 6 (a) shows the distributions of mean values for varying size of local region used for calculating the means. We can observe that as the local region becomes larger, the variance becomes smaller (the bell curve gets "thinner"). This indicates that the estimates approach the expected value (the preset mean value) more robustly. Fig. 6 (b) shows the details of histogram distribution corresponding to a local region of 10 by 10 pixels, from which we can observe that a large number of estimates of mean values are far off from the expected mean value, which renders these estimates less robust. Fig. 7 shows that the larger the population used in local noise estimation, the more robust these estimates become.

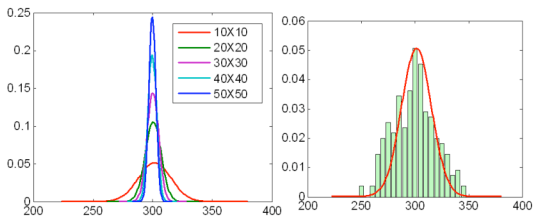


Fig. 6: (a) Distributions of mean values for varying size of local regions used for calculating the means. (b) the histogram and fitted distribution for a 10 by 10 local region.

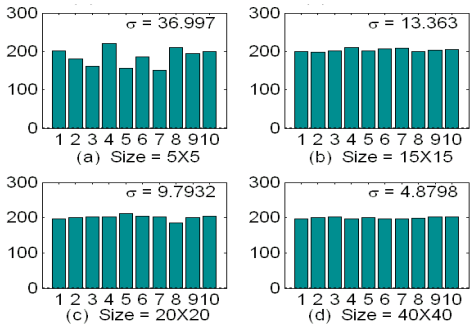


Fig. 7: The standard deviation of pixel intensities.

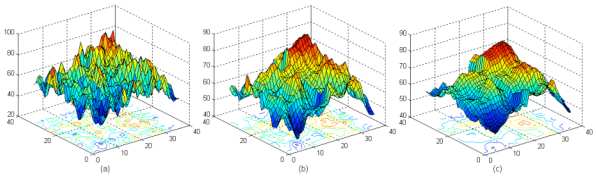
While computing the mean value for a local region, we also obtain the standard deviation of the pixel intensities. For each size of the local region to be examined, we pick 10 grid points and plot the standard deviation across these grid points. As shown in Fig. 7, we see the standard deviation estimates becomes more stable as the size of the local region becomes larger, which implies that the threshold value, v_{cutoff} for each local region becomes more reliable.

We can use the above procedure to fine-tune the size of the local region such that the standard deviation of a random subset of grid points becomes stable and approaches the preset value. Fig. 7 (d) shows that a local region size of 40×40 is already reasonably good in terms of the standard deviation (4.8798) of the standard deviations computed for all 225 local regions.

4. ELB on Real Microarray Images

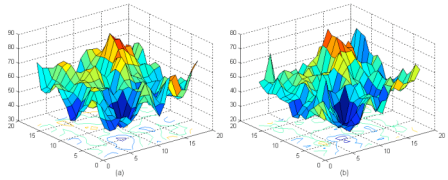
We computed the local noise values of images of the CAG human 19K array using ELB for segmentation and quantification and compared the results against

those by GenePix Pro 5.0. Fig. 8 illustrates the local background noise values using different sizes of ELB with square configuration. The sizes are 3 by 3, 5 by 5 and 7 by 7 in terms of the number of spots on each side of a square shape for ELB configuration. Results in Fig. 8 confirm that a larger ELB region yields more reliable values of local background noise level across the whole chip while still capturing the intrinsic variation of local noise values due to spatial difference. Fig. 9 shows comparison results on a fraction of a microarray image, from which we see that ELB gives smoother distribution of local background noise estimates than GenePix. GenePix uses only a very small number of pixels for background noise estimation and the statistical and numerical errors are considerably larger than those with ELB.



(a) (b) (c)

Fig. 8: Local background noise variation over whole chip based on different sizes of ELB: (a) 3 by 3 (b) 5 by 5 and (c) 7 by 7.



(a) (b)

Fig. 9: (a) Local background noise distribution using (a) ELB and (b) GenePix 5.0 software.

5. Summary and Future Work

We have presented ELB, a novel and viable method for robust microarray image segmentation and quantification based on using a larger local region. We have done preliminary validation of ELB approach using a numerical experiment and the comparisons against GenePix results. Meanwhile, our ELB-based image quantification still gives signal values in the traditional sense via Eq. (2). In contrast, the "true signal values" estimated by the BlueFuse approach can neither be easily associated with mean nor median of pixel intensities.

What we have not discussed is the role of different shapes of ELB configuration in segmentation and quantification. The shape of the ELB may also be an important factor in capturing the spatial difference in localization and quantification of local noise levels. For example, for a rectangular subarray, it might be best to use shape configurations like those in Fig. 3 (c) or (d).

Our future work will include optimizing the size of the ELB region on-the-fly for low computational cost and reasonable variation of variance in intensity estimates for all spots as mentioned in the end of Section 3. We will also perform a more systematic comparison of results by ELB with different shape configurations and size definitions against those generated by other popular microarray image processing software.

References

[1] D. Amararunga and J. Cabrera, "Exploration and Analysis of DNA Microarray and Protein Array Data", Wiley-Interscience-John Wiley and Sons, Inc., 2004.

[2] S. Dudoit, Y. H. Yang, M. J. Callow and T. P. Speed, "Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments", *Statistica Sinica* 12: 111-139, 2002.

[3] H.-Y. Wang, M. Luo, H. Li, "A Genotyping System Capable of Simultaneously Analyzing >1,000 Single Nucleotide Polymorphisms in a Haploid Genome," *Genome Research,* 2005 (in press).

[4] M. Schena, D. Shalon, R.W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a cDNA microarray," *Science* 270: 467–470, 1995.

[5] J. L. DeRisi, V. R. Lyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science* 278: 680–686, 1997.

[6] N. Haan and G. Snudden, "Microarrays in the real world: Image analysis," *European Biopharmaceutical Review*, pp. 44–47, 2004.

[7] Y.H. Yang, M.J. Buckley, S. Dudoit and T.P. Speed. "Comparison of methods for image analysis on cDNA microarray data," *Journal of Computational and Graphical Statistics* 11:108-136, 2002.

[8] L. Stefano and Y. Lu, "Gridding and Compression of Microarray Images," *IEEE Computational Systems Bioinformatics Conference*, 2004.

[9] http://www.cambridgebluegnome.co.uk/

[10] X. H. Wang, R. S. H. Istepanian and Y. H. Song, "Application of wavelet modulus maxima in microarray spots recognition," *IEEE Trans. Nanobios* 2(4): 190-192, 2003.

[11] http://www.stat.berkeley.edu/~sandrine.