Microarray Image Pattern Analysis

Chaur-Chin Chen and Dali Lai *
Department of Computer Science
National Tsing Huaq University
Hsinchu, Taiwan 300
E-mail: cchen@cs.nthu.edu.tw
Fax: +886 3 572 3694
Web: www.cs.nthu.edu.tw/∼cchen

C.Y. Kao and C.N. Chen
Angiogenesis Research Center
National Taiwan University
Taipei, Taiwan 106
E-mail: cykao@csie.ntu.edu.tw
Fax: +886 2 2362 8167
Web: www.angio.bioinfo.ntu.edu.tw

Abstract

Microarray images used to study gene expression in cancer diseases has recently attracted a variety of researchers including medical doctors, computational biologists, and bioinformaticians. Most commercial softwares rely heavily on manual operations to obtain spot features from a microarry image consisting of several hundreds to tens of thousdands of spots with the image resolution of 3000 pixels per inch or 1000 pixels per centimeter, which is not practical. We propose a nearly automatic approach for computing spot features and test our method by 30 cDNA microarray images made from the tissues of patients of gastric cancer provided by the angiogenesis research center at National Taiwan University (ARCNTU). The average of Pearson correlation coefficients between spot features obtained by our method and by Array-Pro Analyzer on 30 microarray images is over 0.92, which encourages our work.

Keywords: M-A plot, Microarray, spot features, tissues.

1 Introduction

Microarrays are widely adopted for simultaneously investigating gene expression in a number of diseases such as adenocarcinoma, breast cancer, colon cancer, gastric cancer, hepatoma, lymphoma, and etc. [3]. A microarray is typically a glass or polymer slide, onto which DNA molecules are attached at fixed locations called spots. There may be up to many thousands of spots on an array, each containing tens of millions of identical DNA molecules or fragments of identical molecules of lengths from tens to hundreds of nucleotides [4]. The resolution of a microarray image is usually 3000 pixels per inch or 1000 pixels per centimeter. Given such a high resolution of microarray image, how to make the spot computations *fast, accurate, and repeatable* is one of the central issues of developing software for microarray image pattern analysis. Conventional softwares such as Array-Pro Analyzer 3.0 [10] and GenePix Pro 4.0 [11] are either expensive or requesting many user-specified parameters such that the work is infeasible or is not repeatable. This paper depicts a paradigm of microarray image analysis as shown in Figure 1. The purpose is to fast and accurately compute the repeatable statistics such as *spot features*, associated with the number of spot pixels and background pixels, standard deviations, and etc. for further statistical analysis.

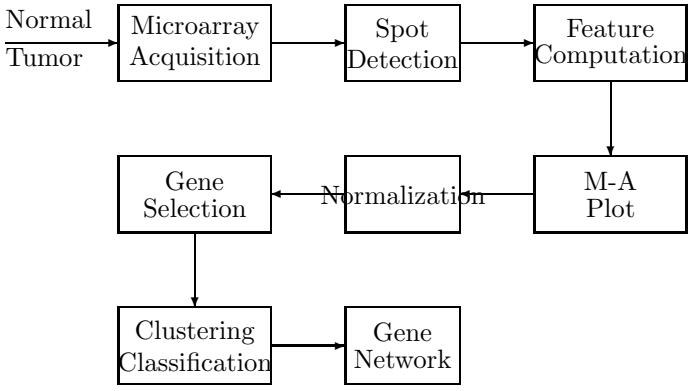

Figure 1. A Paradigm of Microarray Image Pattern Analysis.

A cDNA microarry [3] made of nylon membrane is acquired via a sequence of biological experiments and is scanned by a high resolution color

*This work was supported by NSC 92-2213-E-007-069 and 92-EC-17-A-19-S1-0016.

scanner Umax 6000 to get a color image represented by three primitive color signals of 24-bit (R,G,B) triple for each pixel. After applying mimor manual image operations such as rotation, cropping, and color to gray conversion by $g = 0.299R + 0.587G + 0.114B$, we obtain an image of size 600×896 as shown in Figure 2. The remaining of this paper is to describe, as characterized in Figure 1, step by step, how to get the spot features from Figure 2 which consists of $384 = 16 \times 24$ potential spots corresponding to genes under studies with darker spots regarded as more expressed ones. We demonstrate the intermediate results by images and data plots.

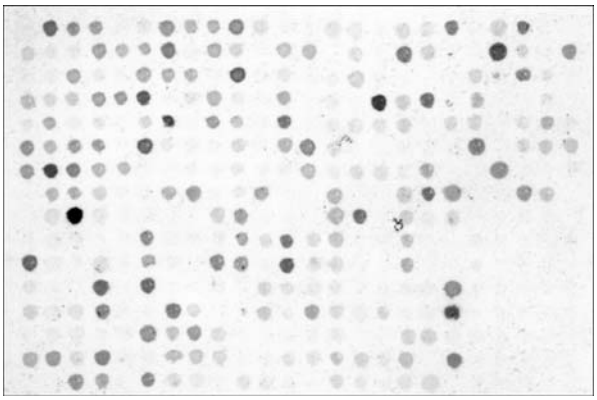

Figure 2. An input 600×896 gray-level microarray image.

2 Gridding and Spot Detection

The purpose of microarry image processing is to locate the spots, a set of local darker pixels as shown in Figure 2 and compute the mean or median of these pixel values associated with the mean or median of surrounding background pixels. Conventional software used to let users manually move pre-defined circular shapes of the same radius or elliptic shapes with known x-intercepts and y-intercepts to fit the spots [11] which may not be appropriate since a microarray image usually contains a lot of noise during the processes of biological experiments and the shape of each spot need not be circular or elliptic. We propose to put a grid of h by k pixels ($h = k = 36$ for our images) according to previous studies to cover each spot as shown in Figure 3 and do local segmentation based on a simple thresholding algorithm [5] [8] for the area of each grid. Consider an image block consisting of n pixels, with g_i representing the gray level for pixel i, where $0 \le g_i \le 255$. A segmentation algorithm is to find a threshold $T \in [0, 255]$ such that a pixel i being classified as "spot pixel" if $g_i \le T$ and "background pixel" if $g_i > T$ such that the following criterion is maximized.

$$C_T = |m_1 - m_0|/(p_1 s_1^2 + p_0 s_0^2), \qquad (1)$$

where n_0 and n_1 are the number of "spot pixels" and "background pixels", respectively with $p_0 = n_0/n$, $p_1 = n_1/n$ and

$$m_0 = \frac{1}{n_0} \sum_{j:\ g_j \le T} g_j,\ \ m_1 = \frac{1}{n_1} \sum_{j:\ g_j > T} g_j,\ \ n_0 + n_1 = n.$$

$$s_0^2 = \frac{1}{n_0} \sum_{j:\ g_j \le T} (g_j - m_0)^2,\ \ s_1^2 = \frac{1}{n_1} \sum_{j:\ g_j > T} (g_j - m_1)^2.$$

The segmentation result for the image in Figure 2 is shown in Figure 4. Note that the shapes of spots are generally not circular with the same diameter and not all of them have circular or elliptic shapes. Furthermore, a sequence of microarray experiments might introduce a lot of noise before the image is acquired, a smoothing operation like mean or median filtering [7] might be applied to further reducing noise before a local segmentation is adopted. Here, we adopted a 5×5 mean filter to get rid of noise and smooth the image.

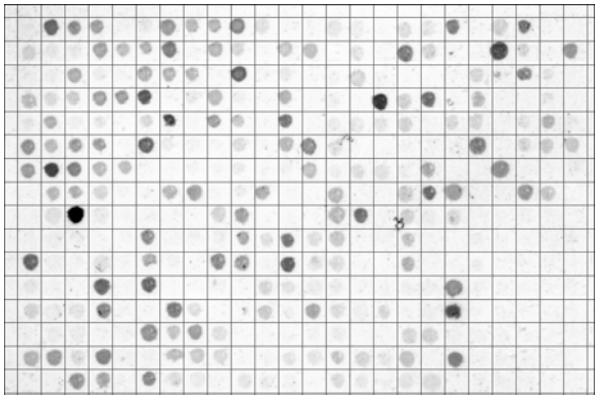

Figure 3. Result of Gridding.

3 Feature Computation

The most important output statistics from a microarray image processing are the spot features

which measure the gene expression level corresponding to the quantity of molecules during a certain period. A spot feature is usually computed by the difference

$$G = |F_\mu - B_\mu|, \tag{2}$$

where F_μ is the mean spot intensity and B_μ is the mean background intensity defined as the mean of gray levels of pixels surrounding the detected spots (complementary to the inner 30×30 pixels in the grid). One of the microarray applications is to search for a set of genes in a certain disease which are differentially expressed in tumor tissues but not expressed in normal tissues or vice versa. The image shown in Figure 2 is a microarray image acquired from a normal tissue of a gastric patient whose tumor tissues are also experimented to get the spot detection and feature computation with the segmentation result shown in Figure 5 and the corresponding spot feature R can be similarly computed by Equation 1.

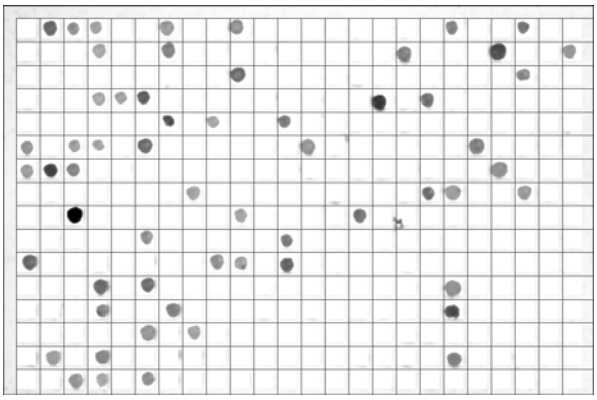

Figure 4. Spot Detection for a Microarray with Normal Tissues.

4 M-A Plot of Gene Expression Levels

A pair of microarray images made of normal and tumor tissues, respectively, were acquired via a sequence of biological experiments. An image of size 600×896 containing 16×24 potential spots corresponding to 384 genes in this study was obtained after manual operations of minor rotation and cropping on the original image. We have

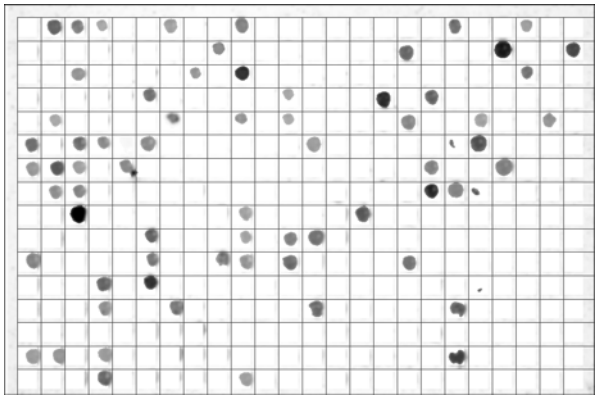

Figure 5. Spot Detection for a Microarray with Tumor Tissues.

demonstrated that we can obtain 384 pairs of spot features $\{(G_j, R_j) \mid 1 \le j \le 384\}$ corresponding to the target genes under studies. Microarray analysts attempt to determine which genes are differentially expressed by analyzing the plot of vectors $\{(G_j, R_j)\}$, $1 \le j \le 384$. Since some unavoidable noise of biological unknowns or phenomena, the experiment designers usually put some control spots such as plant genes to reduce the difference crossing microarrays and leave some spots *blank* regarded as "unused" to normalize the intensity levels of each spot. Our microarray images used 16 plant genes (spots, "*"), marked 39 unused spots (blanks, "+"), and one positive control gene located in the 11th row and the 12th column for quality control (it must be "dark" enough if it is in effect). Table 1 lists first 16 pairs of spot features derived from normal and tumor tissues, respectively. An M-A plot [4] of these spot features are given in Figure 6, where the coordinates on x-axis (A values) and y-axis (M values) are defined as

$$\begin{aligned} A_j &= \tfrac{1}{2}[\log_2(G_j) + \log_2(R_j)] \\ M_j &= [\log_2(G_j) - \log_2(R_j)] \end{aligned} \tag{3}$$

The differentially expressed genes are defined as those feature vectors in the M-A plot that are far above or below the horizontal line corresponding to $M = 0$. Due to the uncertainty and distortion of experiments, a variety of statistical approaches have been proposed to make the interpretations more practical, whereas no universally best method has come out yet [3][4].

Table 1. Spot Features G_j and R_j for Normal and Tumor Tissues.

Index	1	2	3	4	5	6	7	8
G_j	8	102	74	64	5	12	71	47
R_j	5	106	88	59	6	8	67	21
Index	9	10	11	12	13	14	15	16
G_j	53	87	18	3	4	12	8	4
R_j	24	94	13	3	4	17	13	3

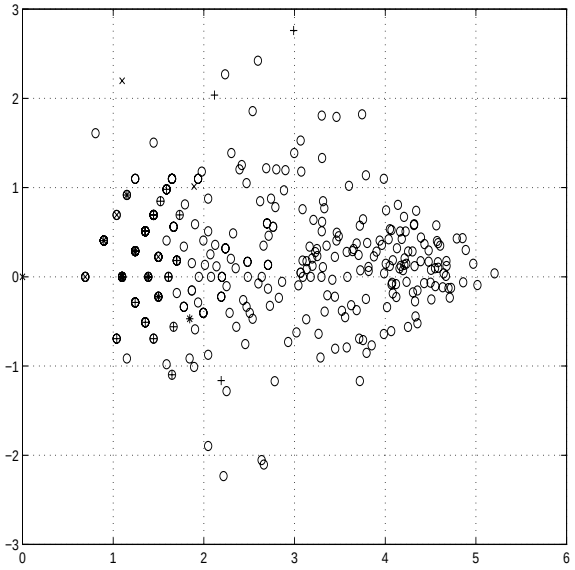

Figure 6. M-A Plot of 384 Spot Features.

5 Discussion and Conclusion

We propsed a nearly automatic microarray image processing system which takes within less than 15 minutes running on a Linux based system with Pentinum 4 CPU to compute spot features and M-A plot for a pair of cDNA microarray images of size 600×896 containing 16×24 spots with each expressed gene consisting in average of 300∼600 pixels out of 900 pixels in a grid. We compared our system with a commercial software by ArrayPro Analyzer 3.0 on 30 cDNA microarray images from patients of gastric cancer provided by ARC-NTU [9]. The high Pearson correlation coefficients (0.92 in average) encourages our approach. If the locations of control genes or spots can be further re-arranged or specifically designed, an automatic system to report statistics for spot features and their associated statistics will be available which merits further studies.

6 Acknowledgments

This work was supported by NSC Grant 92-2213-E-007-069 and 92-EC-17-A-19-S1-0016.

References

[1] A.A. Alizadeh et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", *Nature*, vol. 403, 503-511, 2000.

[2] U. Alon et al., "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucletide Arrays", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, 6475-6750, 1999.

[3] A. Brazma and etc., "Minimum 3nformation about Microarray Experiment (MIAME) - toward Standards for Micorarray Data," *Nature Genetics*, vol. 29, 365-371, 2001.

[4] H.C. Causton, J. Quackenbush, and A. Brazma, *Microarray Gene Expression Analysis*, Blackwell Publishing, 2003.

[5] C.C. Chen and R.C. Dubes, Environmental Studies of ICM Segmentation Algorithm, *Journal of Information Science and Engineering*, vol. 6, 325-337, 1990.

[6] H.Y. Chuang, H. Liu, S. Brown, C. McMunn-Coffran, C.Y. Kao, D.F. Hsu, Identifying Significant Genes from Microarray Data, *Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering*, BIBE2004, 358-365, May 19-21, 2004.

[7] R.C. Gonzalez and R.E. Woods, *Digital Image Processing*, Prentice-Hall, Inc., 2002.

[8] N. Otsu. "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. SMC-9, 62-66, 1979.

[9] http://www.angio.bioinfo.ntu.edu.tw

[10] http://www.mediacy.com/arraypro.htm

[11] http://www.axon.com/gn_GenePixSoftware.html