Third Canadian Conference on Computer and Robot Vision
CRV 2006, June 7th~9th, Quebec City, Canada
http://www.computerrobotvision.org/

Simple Software for Microarray Image Analysis

Chaur-Chin Chen
Department of Computer Science
National Tsing Hua University
Hsinchu 30013, Taiwan
E-MAIL: cchen@cs.nthu.edu.tw
PHONE: + 886 3 573 1078
FAX: + 886 3 572 3694

Cheng-Yan Kao, Chun-Fan Chang, Hsueh-Ting Chu, Chiung-Nien Chen
Angiogenesis Research Center
National Taiwan University Hospital
Taipei 106, Taiwan

Abstract

A set of microarray images were acquired by a sequence of biological experiments which were scanned via a high resolution scanner. For each spot corresponding to a gene, the ratio of Cy3 and Cy5 fluorescent signal intensities was obtained and which may be normalied based on piecewise linear regression such as lowess method. In this study, we computed from 55 microarray images to get an $M \times N$ genematrix, A, with $N = 55$ patients and $M = 13574$ effected genes in each microarray. We start with our gene discovery from a genematrix $A \in R^{M\times N}$, $M = 13574$, $N = 55$, including $N_1 = 29$ patients of hepatitis B virus (HBV), $N_2 = 21$ patients of hepatitis C virus (HCV), 1 patient clinically diagnosed to be infected with HCV as well as HBV, and 4 patients were neither HCV nor HBV infected. Simple software was developed to solve the following problems: (i) Detect differentially expressed genes and (ii) Select a subset of genes which best distinguishes HBV patients from HCV ones.

Keywords: *cDNA microarray, dendrogram, Fisher criterion, M-A plot.*

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. [1, 2, 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 [12] and GenePix Pro 4.0 [13] 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. We attempt 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 clustering and classification, and gene discovery.

A cDNA microarray [3] made of glass slide is acquired via a sequence of biological experiments and is scanned by a high resolution color scanner like GenePix 4000 to get a color image, each pixel is represented by two 16-bit signals, corresponding to Cy3 and Cy5 channels. Our cDNA microarray im-

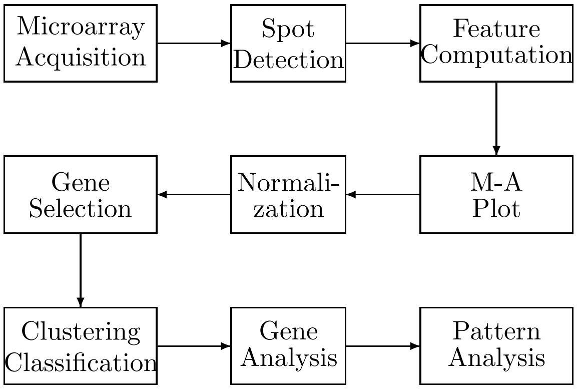

Figure 1. A Paradigm of Microarray Image Pattern Analysis.

age consists of 156×119 genes which is further divided into 60 effective blocks. Figure 2(a) shows a typical block of 270 genes with an ideal spot enclosed by a circular region of diameter $80\mu m$, each pixel has size 15 μm, the spacing between horizontal spots and vertical spots are $150\mu m$ and $145.161\mu m$, respectively. The remaining of this paper is to describe, as characterized in Figure 1, step by step, how to get the spot features from a microarray image and how to select a set of genes for clustering and classification. The genes in our study corresponding to the brighter spots are regarded as more expressed ones. We demonstrate the intermediate results by images and data plots.

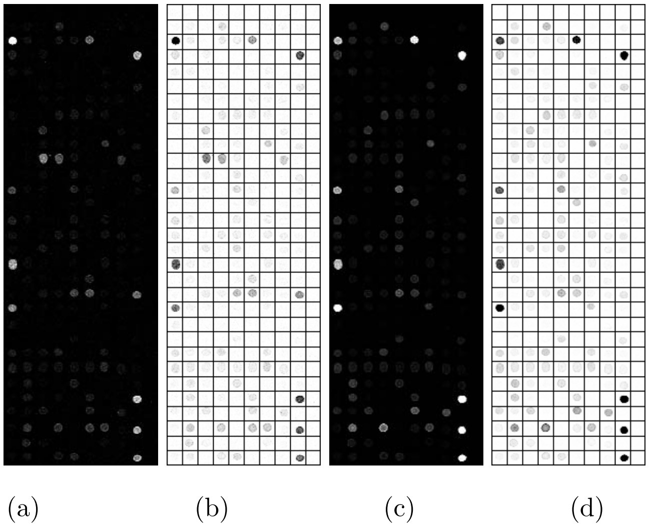

Figure 2. (a) Cy3 Block, (b) Segmentation of (a), (c) Cy5 Block, (d) Segmentation of (c).

2 Gridding and Spot Detection

The purpose of microarray image processing is to locate the spots, a set of local brighter pixels as shown in Figure 2(a) 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-intercept and y-intercept to fit the spots [13] 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 ($h = k = 15$ for our images) according to previous studies to cover each spot and do local segmentation based on a simple thresholding algorithm [7] 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 $0 \leq g_i \leq L-1$. A segmentation algorithm is to find a threshold $T \in [0, L-1]$ such that a pixel i being classified as "spot pixel" if $g_i \geq T$ and "background pixel" if $g_i < T$ such that the following criterion is maximized.

$$C_T = (m_1 - m_0)^2/(p_1 s_1^2 + p_0 s_0^2), \tag{1}$$

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

$$m_0 = \frac{1}{n_0} \sum_{g_j \geq T} g_j, \quad m_1 = \frac{1}{n_1} \sum_{g_j < T} g_j.$$

$$s_0^2 = \frac{1}{n_0} \sum_{g_j \geq T} (g_j - m_0)^2, \quad s_1^2 = \frac{1}{n_1} \sum_{g_j < T} (g_j - m_1)^2.$$

The gridding and segmentation results for the image block of Figure 2(a) is shown in Figure 2(b). 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 [6] 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. Furthermore, a microarray slide might not be placed such that spots in located in the region of interest is well aligned by user-specified vertical and horizontal lines. A minor rotation may be justified before the spot feature compution [9].

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 [9]. 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 15×15 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(a) is a microarray image acquired from a tumor tissue of a patient of Hepatoma. The normal part from common reference is also experimented to get the spot detection and feature computation with the segmentation result shown in Figures 2(c)(d), where the corresponding spot feature R can be similarly computed by Equation (2).

4 M-A Plot of Gene Expression Levels

A microarray image in our experiment, hybridized by normal (dyed with cy5 fluorescence) and tumor (dyed with cy3 fluorescence) tissues, were acquired via a sequence of biological experiments. An image of size 2350×2020 containing $156 \times 119 = 18564$ potential spots was obtained after semi-manual operations. We have demonstrated that we can obtain 18564 pairs of spot features $\{(G_j, R_j) \mid 1 \leq j \leq 18564\}$ corresponding to the target genes under studies. Microarray analysts attempt to determine which genes are differentially expressed by analyzing the plot of vectors $\{(G_j, R_j)\}$, $1 \leq j \leq 18564$. 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* to normalize the intensity levels of each spot. Each of our microarray images contains only 13574 effected genes, the remaining 4990 spots either served as "control" or "unused". An M-A plot [4][8] of these spot features is given in Figure 3, 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 features 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]. We apply a simple linear least squares fit to adjust the ratio of Cy3/Cy5 such that the shape of normalized M-A plot tends to be horizontal. Figure 4 shows a linearly normalized M-A plot of gene expression. Other normalization methods can be applied to better fitting the design issue [10].

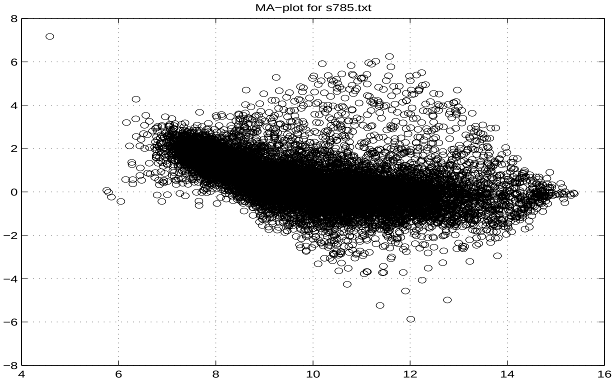

Figure 3. M-A Plot of 13574 Spot Features.

5 Seeking for Differentially Expressed Genes

Let A[1:Gn, 1:N] with Gn=13574, N=55, be the genematrix with each entry being the normalized $Cy3/Cy5$ *ratio*, where the tumor tissues dyed with the fluorescence light of wavelength 532 nanometers and the normal tissues dyed with the fluorescence light of wavelength 635 nanometers. A gene k is said to be *up regulated* if

$$\log_2(A[k,j]) \;\geq\; T \quad for\ 1 \leq j \leq N \tag{4}$$

and *down regulated* if

$$\log_2(A[k,j]) \;\leq\; -T \quad for\ 1 \leq j \leq N \tag{5}$$

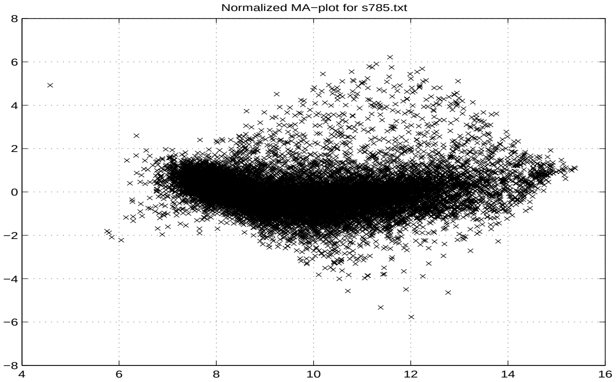

Figure 4. Normalized M-A Plot of 13574 Spot Features.

In our experiment of 55 patients of Hepatoma, the following genes are detected as *differentially expressed* when the threshold $T = 3.0$ is chosen.

Table 1. Differentially Expressed Genes.

Index ↑	Feature#	Accession#
6078	17265	BC007058
10186	7559	AI133162
3182	4966	L32179
5942	1355	AL532086
8693	17711	X06290
8653	5388	BC008983
8857	10690	BI834172
5150	13529	AW609791

Index ↓	Feature#	Accession#
2855	289	M24173
10177	8963	BG766355
2044	15427	AI133196

6 Distinguishing HCV from HBV

Let $X[1 : K, 1 : N]$ be derived from $A[1 : 13574, 1 : 55]$ with $N = N_1 + N_2 = 29 + 21 = 50$ patients including $N_1 = 29$ HBV patients and $N_2 = 21$ HCV patients with K genes being selected to distinguish HCV from HBV. The selection of K genes are based on the condition $C_k > T_c$ for each gene k, where the Fisher's ratio C_k is defined below: the larger, the more separable.

$$C_k = (\mu_1(k) - \mu_2(k))^2/(p_1 s_1^2(k) + p_2 s_2^2(k)),$$

$$where\ p_1 = N_1/N, \quad p_2 = N_2/N$$

$$\mu_1(k) = \frac{1}{N_1}\sum_{j=1}^{N_1} X[k,j]$$

$$\mu_2(k) = \frac{1}{N_2}\sum_{j=N_1+1}^{N} X[k,j]$$

$$s_1^2(k) = \frac{1}{N_1}\sum_{j=1}^{N_1}(X[k,j] - \mu_1(k))^2$$

$$s_2^2(k) = \frac{1}{N_2}\sum_{j=N_1+1}^{N}(X[k,j] - \mu_2(k))^2$$

With $T_c = 1.3$, there are 26 genes are detected and are used to distinguish patients of HBV infection from HCV ones.

The dendrogram of complete-linkage, as shown in Figure 5, based on the 26 most discriminative genes selected by Fisher's criterion, matches the result of applying K-means clustering algorithm. There are only one HBV and one HCV clustered into *unexpected* groups.

7 Discussion and Conclusion

We have proposed a near automatic microarray image processing system which takes less than 10 minutes running on a Linux based system with Pentinum 4 3.00GHz CPU to compute spot features and M-A plot for a pair of cy3 and cy5 cDNA microarray images of size 2350 × 2020 containing 156×119 spots with most expressed genes consisting of 40∼120 pixels out of 225 pixels inside a grid. Gene expression was normalized based on a linear least squares fit on an M-A plot and is used for classification and clustering. In our study, cDNA microarray images from patients of Hepatoma are provided by ARCNTU [11]. We have applied our simple software to seek 11 differentially expressions with 3-fold of log ratio. By the usage of Matlab tools, we also demonstrate that the 26 most discriminative genes can distinguish patients with HBV from HCV fairly well. If the locations of control genes or spots can be further re-arranged or specifically designed, an

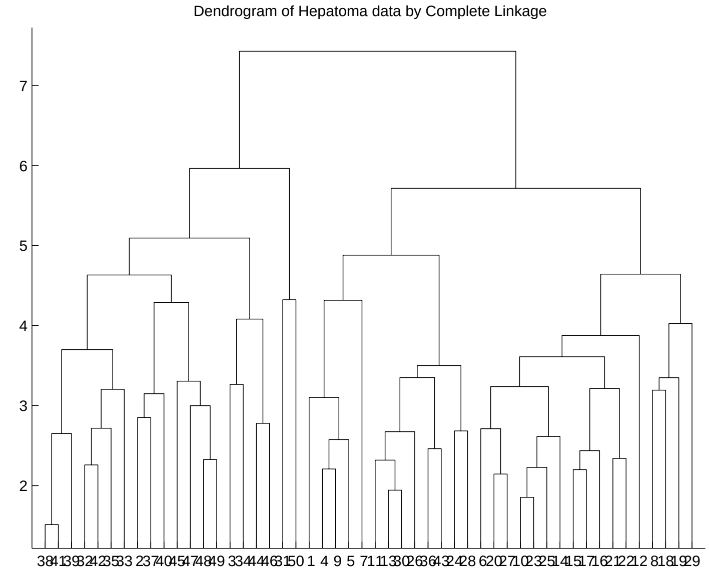

Figure 5. Dendrogram of 29 HBV and 21 HCV patients.

automatic system to report statistics for spot features and their associated statistics will be available which merits further studies. The genes discovered by our simple software need biological verifications before further usages such as gene therapy or drug design.

8 Acknowledgments

This work was supported by NSC 94-2213-E-007-089 and MOEA 93-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 Oligonucleotide 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 Information 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] 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.

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

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

[8] D. Stekel, Microarray Bioinformatics, *Cambridge University Press*, 2003.

[9] M.Y. Tsai, Gene Expression Computation on Microarray Image Data, *M.S. Thesis, National Tsing Hua University, Taiwan*, January 2006.

[10] Y.H. Yang et al., "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation", *Nucleic Acids Research*, Vol. 30, No. 4, e15, 2002.

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

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

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