A SHORT STUDY ON MICROARRAY IMAGE PROCESSING

PREPARED BY:

SENIHA ESEN YUKSEL

CVIP Lab
University of Louisville
Electrical and Computer Engineering

Date: July 05, 2004

ABSTRACT

The ultimate goal of microarray imaging is to find the differences of the gene expressions under two separate conditions. Understanding how gene expressions change will allow us to better understand and cure the diseases. This technology has the advantage of collecting a large amount of data in a single experiment, but still some challenges exist to quantify and compare this data.

The process starts with extracting the mRNA of the cells, labeling them and hybridizing (base pair) them with their corresponding DNA's on a glass microscope slide. Then the glass is scanned and the spots are obtained from the emitting fluorescence. This forms the microarray image. The image should be processed for information. First, the boundaries of the spots are identified by gridding, then by segmentation procedures background is removed from the image and the intensity of each spot is calculated. Then the intensity data obtained from these spots should be processed by statistical methods to understand the gene expressions.

INTRODUCTION

Complementary DNA microarray imaging is a very powerful technology to analyze and understand the genetic expression. Understanding the genetic expression would help us solve the mechanisms of diseases and develop personalized medicines. Especially for genetically heterogeneous diseases such as cancers, heart diseases, multiple sclerosis and diabetes type 1 - caused by several genetic defects resulting in the same observable characteristics- DNA analysis would be very helpful to identify each gene defect of each individual and to invent effective therapies for each patient. However, this is a largely probabilistic science as it has been estimated that any two copies of the human genome differ from one another by 0.1%, i.e. three million variants over three billion that make up the human genome.[1]

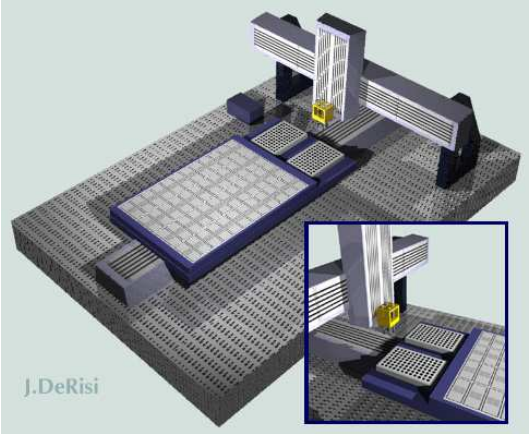
"A microarray experiment produces thousand of genes on a single slide, and all the genes can be globally viewed at the same time. This provides a systematic and comprehensive way to survey the DNA and RNA variations, which could become a standard tool for both molecular biology research and genomic clinical diagnosis, such as cancer diagnosis and type 1 and type 2 diabetes diagnoses." [2]

This short study is two fold; first, the process to form the microarray image is explained. Second, the current image processing techniques are introduced. The image processing task is divided into 4 main parts, namely the gridding, denoising, segmentation and information extraction. The methods used in the commercial packages and the recent approaches are discussed in these parts.

HOW IS A MICROARRAY IMAGE FORMED?

A microarray image is a means to compare the gene transcription of two or more different kinds of cells. The comparison is done via comparing the cDNA hybridization. But what is cDNA hybridization?

During transcription genes are coded into messenger RNA's (mRNA) in the cell nucleus. These mRNA are isolated from the RNA as they form only the 3% of all RNA in a cell. But these mRNAs can degrade very quickly, so they are reverse-transcribed back into the DNA form which is the complementary DNA (cDNA)[3]. Since the aim is to compare the two different cells, the general practice is to fluorescently label the cDNA's with rhodamine (Cy3 - red) and fluorescein (Cy5 - green) to be able to separate the two samples. Then these red and green dye tagged cDNA's are mixed and placed at the microarray.

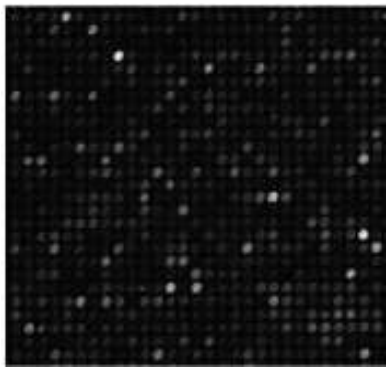


A microarray is a high-density array on a microscope slide consisting of thousands of spots, each spot representing one gene. On each spot, single strands of a large number of different DNA sequences are printed by an arraying machine (seen in Figure 1). The arraying machine produces a rectangular grid of thousands of spots. In this gridding, ideally, all spots should be circles with a constant diameter [4].

Figure 1: arraying machine [6]

Upon placing the cDNA's on the microarray, the cDNA's are hybridized to the spots i.e. the single strand cDNA's find their base DNA's on the array and bind to form the helix structure. Because of this hybridization process, the labeled cDNA samples are called as **probes** in literature.

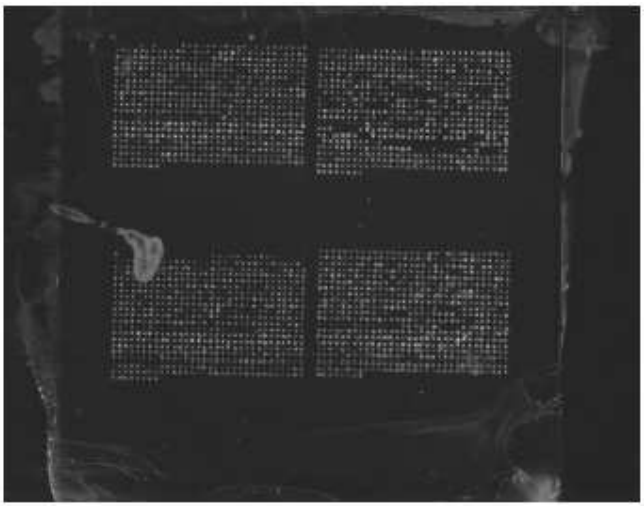
The task now is to determine how many cDNA's have combined with which DNA. Therefore the microscope slide is scanned with a red and a green laser to detect the amount of cDNA in each spot. A detector, either a charge coupled device (CCD) or a confocal microscope, records the intensity of emitted light as an image. High intensity spots on the image correspond to the spots with more bound probes (cDNA's).



As in Figure 2, the microarray image will have brighter red & green parts and also yellow spots. The yellow spots correspond to the equal amounts of cDNA from each cell (equal red and green combination makes yellow). The quantification of DNA in each spots is done by comparing the intensities of spots. The brighter the spot is the more mRNA is coming from the cell dyed with that color[3].

Figure 2: A microarray image [3]

This procedure above mentioned is a two channel microarray experiment. If instead of two dyes, bitoin is used, then it is called a one channel microarray experiment. In a one channel microarray experiment, the image appears only black and white and the black spots correspond to the more hybridized gene expressions. And two slides are used to compare the different conditions. An example of a single channel microarray image is given in Figure 3.



"The resolution of such an image is typically 10μ per pixel with intensity values of 16 bit accuray, resulting in about 50MB of raw data for each array when the complete slide surface is used." [7]

These images are generally stored in TIFF format.

Figure 3: Single channel microarray image [7]

Now that an image is obtained, the problem boils to an image processing problem where the background should be removed, spots should be identified, intensities should be calculated and quantification should be obtained. Actually, these tasks are not trivial either, because of the noise and other problems introduced during the image producing. These problems will be discussed in the next part of this survey. A brief, very explanatory flash animation of this part on obtaining the image from cells can be found in [4].

PROBLEMS IN MICROARRAY IMAGING

- Noise problems:

"Noise is introduced from the source and the detector. Source noise includes the photon noise, dust on the slides and treatment of the glass slides. Detector noise includes features of the amplification and digitization process"[7].

- Printing Problems:

"The robot coordinate system may be slightly rotated against the image coordinate system. The print tips are not necessarily aligned to the row and column distances to the print grid. Therefore, the global spot pattern is not periodic in general. The print tips are not attached rigidly to the print head such that they do not break or scratch the array surface when set down. Therefore, the needles may vibrate slightly, causing random deviations from the ideal print positions."[8]

- Background problems:

"There can be a few molecules that would hybridize to a wrong spot or to the glass slide. The extra light from these molecules would form the background of the image."[3]

- Reproducibility problems:

The amount of DNA that bound to the glass depends on the amount of the DNA printed on the glass. Therefore the size and shape of spots are variable in each image. However, this technology has reproduction and quality issues due to the variation in the biological systems and due to variation of the amount of DNA printed on the glass during the printing process (fabrication problems). These issues limit the application of this technology to complex biological systems. [9]

- Protein production:

The final product of this technique is the proportional changes in the RNA. However, some RNA is involved in the protein production and some is not. Therefore the only knowledge of RNA in the replication process without the knowledge of its existence in the proteins may not be enough to detect the patterns in diseases like cancer.

IMAGE PROCESSING

"In the image processing phase, the aim is to extract each spotted DNA sequence as well as to obtain background estimates and quality measures." [2]

Image processing is formed of three steps: gridding, segmentation and information extraction. In the gridding process, the coordinates of each spot are determined. In the segmentation process, the pixels are segmented as background or foreground, and in the third step the intensities are extracted and the related genes are clustered. Then confidence measures are applied to the results to facilitate accurate microarray analysis which involves data normalization, filtering and data mining.

GRIDDING:

The arrayer determines the microarray image's structure, so the general form is known priori. So how many rows and columns appear in an image –so called grids- and how many spots exist in each grid is already known and they generally have a highly regular form. However, the exact places should be determined and each spot should be exactly located in the image for accurate calculation. This process of locating the grids and spots is called grid alignment or spot detection and it is an important step as an important number of spots can be misaligned and there can be some irregularities in the shape and size of the spots[10]. Individual translation of grids or spots, separation between row and columns of spots or grids should be taken into consideration[7]. "A good grid alignment algorithm should take full advantage of the regularity of the array design yet be flexible with the inherent inconsistencies." [10]

The current spot detection techniques used in the commercial software packages include manual intervention, fixed template (rectangles, ovals) and adaptive polygon (circles, snake) algorithms. These will be discussed more in the segmentation part of this report.

A novel approach has been given by Wang *et al.* [10] using three-color cDNA microarray platform where the printed platform is **fluorescein** labeled prior to hybridization. Labeling the glass slide prior to hybridization allows the evaluation of the array/spot morphology, DNA deposition and background levels. By the help of the prior knowledge of array/spot morphology, gridding can be done more accurately. Similarly, by the help of the prior knowledge of DNA on each spot, the pixels that don't have strong enough third-dye signals can be assigned to background in the segmentation phase. Moreover, since amount of DNA bound to the glass slide is dependent on the amount of DNA printed on the glass slide, more accurate quality measures can be performed [9,[10]. However, the common lack of the use of hardware for three-dye technology becomes the main downside of this algorithm.

DENOISING:

The noise in the microarray images occur due to photon noise, electronic noise, laser light reflection, dust on slide and so on. Before further gene expression analysis, the noise in the image should be removed. "Traditionally, statistical models are employed for noise reduction like variance analysis, ratio distribution, gamma distribution, empirical Bayes model and Bayesian estimation of array measurements. But these methods deal mainly with measurement error, such as preparation of the sample, cross hybridization, and fluctuation of fluorescence value from gene to gene. But none deals particularly with the effect of the noise"[11]. Therefore, two new approaches have been introduced and they will be discussed briefly to handle the inherent noise problem.

1) X. H. Wang [12][11] introduces a recent approach by using the wavelet transforms methods to denoise the image as wavelet is a powerful tool that can recognize the discontinuities and edges of the images as seen in Figure 4 and is very sensitive to weak signals[2].

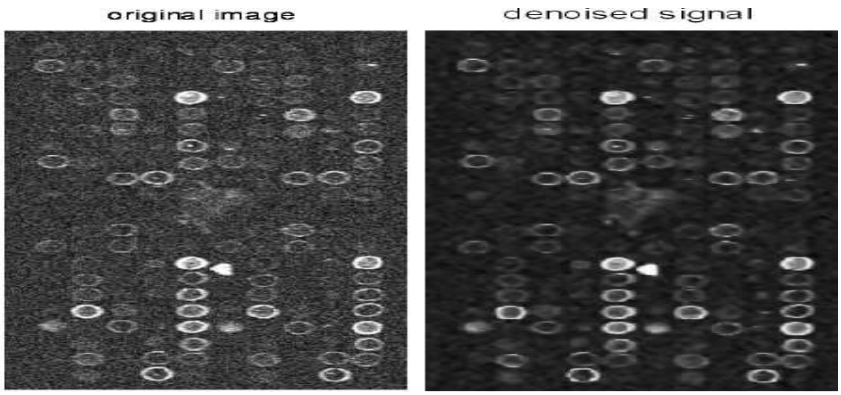


Figure 4: Denoising using wavelet transforms (Adapted from [11])

2) A second approach to remove the noise is introduced by O'Neill *et al.* [13] using the image reconstruction techniques. Instead of trying to find the contours of the spots and taking the rest as the noise, O'Neill tries to use a modified image tiling and hole filling method to obtain the noise first. Then the noise is subtracted from the image to obtain the original image. The advantage of the algorithm is that it doesn't necessitate the exact fitting of a boundary around the spot, which is an important step for automation.

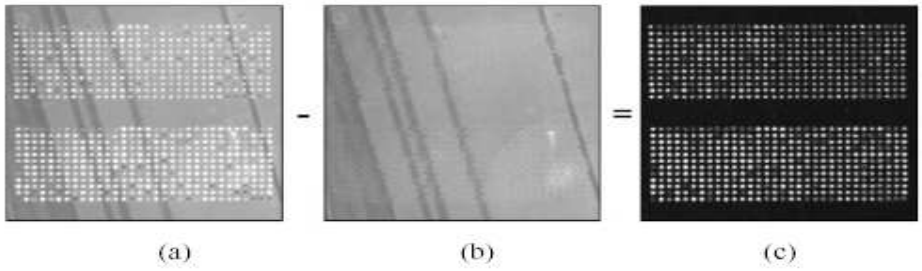


Figure 5: An example reconstruction: (a) the original image minus (b) the reconstructed background image leave (c) the final gene image (Adapted from [13])

SEGMENTATION:

Segmentation methods used in the commercial packages can be summarized as below in the first 4 points [7][14]. The 5th and 6th points are the new approaches to the segmentation problem in microarray images.

1) Fixed circle segmentation:

In fixed circle segmentation, a circle with constant diameter is fixed to all the spots. This method is easy, but the circles are not always circles or of the same size.

2) Adaptive circle segmentation

In this method the circle diameter is varying, but this method also misses the spots that are not of circular shape.

3) Histogram segmentation

In histogram segmentation, a mask is chosen to be larger than any spot, and the intensity is determined from the histogram of the pixels in the masked area. However, the selection of a large mask to compensate for the spot size variation makes the quantitation unstable.

4) Adaptive shape segmentation:

The most common adaptive shape segmentation methods include seeded region growing and watershed algorithms. In general, the flaw of these algorithms is that they need initial seed points, and determining the number and location of these seeds can be a problem. But in microarray imaging, this algorithm proves to be very effective since the number and the location of the spots (which are to be the seed points) are pretty much known beforehand.

After segmentation, most commercial software packages like ScanAlyze , GenePix, QuantArray, select the foreground intensity as the mean or the median of the pixel values within the segmented spot mask, and the background as the mean or median of the pixels surrounding the spot mask.

5) Wavelet Transforms:

In contrast to the approaches used in the commercial packages (1-2-3-4), noises do occur in the image and also spots can be rotated and translated. Therefore, more precise spot recognition algorithms are needed. The recent, novel approach by X. H. Wang[12][11] uses the wavelet transform method for spot recognition.

Wang *et al*. [12] applies wavelet transform to find the contours of the spots. Actually, this method also finds the noise in the spot as seen in Figure 6, but the error is as small as 4.5%.

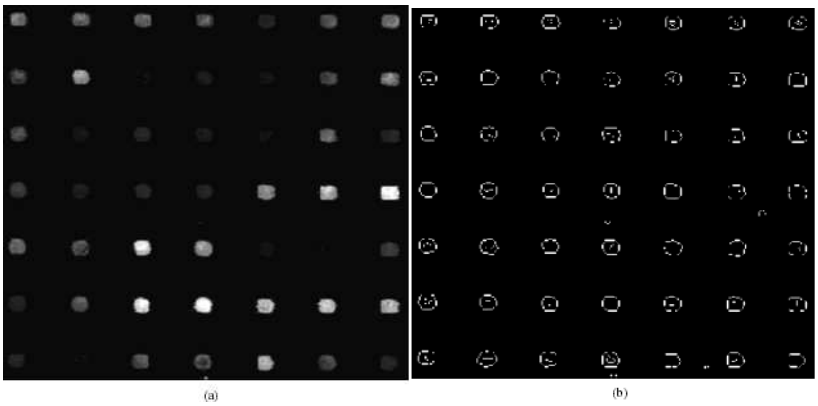


Figure 6: Microarray image and its modulus maxima (Adapted from [12])

The fact that this algorithm can catch even very weak spots is also the advantage of the algorithm that it can be very helpful in detecting the contours of the spots.

6) Markov Random Fields:

Katzer et al.[8] use Markov Random Fields for high-level image segmentation and active contours for single spots. The method is robust to rotation, noise and misaligned grids and it does not require calibration. Katzer[8] developed a partly continuous contour model and a robust energy minimization method, which increases the segmentation correctness for images with high intensity and spot shape variation to 94%.

INFORMATION EXTRACTION:

After the detection of the location, size and shape of each spot, it is necessary to calculate spot quality measures and partition the results into groups that have similar expression patterns. Grouping the genes based on their similarity is called clustering. Cluster information would help the biologists understand the functions of the previously unknown genes, understand the interactions between genes and possibly form gene pathways.

Two common clustering methods are k-means clustering and hierarchical clustering. In k-means clustering, the inter and intraclass distances are minimized iteratively, in hierarchical clustering a gene similarity matrix is formed and the highest scored genes are joined. If the clustering process is used, it would help to understand which expression levels change under a given condition or when a significant change has been made [16]. The final result will look like Figure 7.

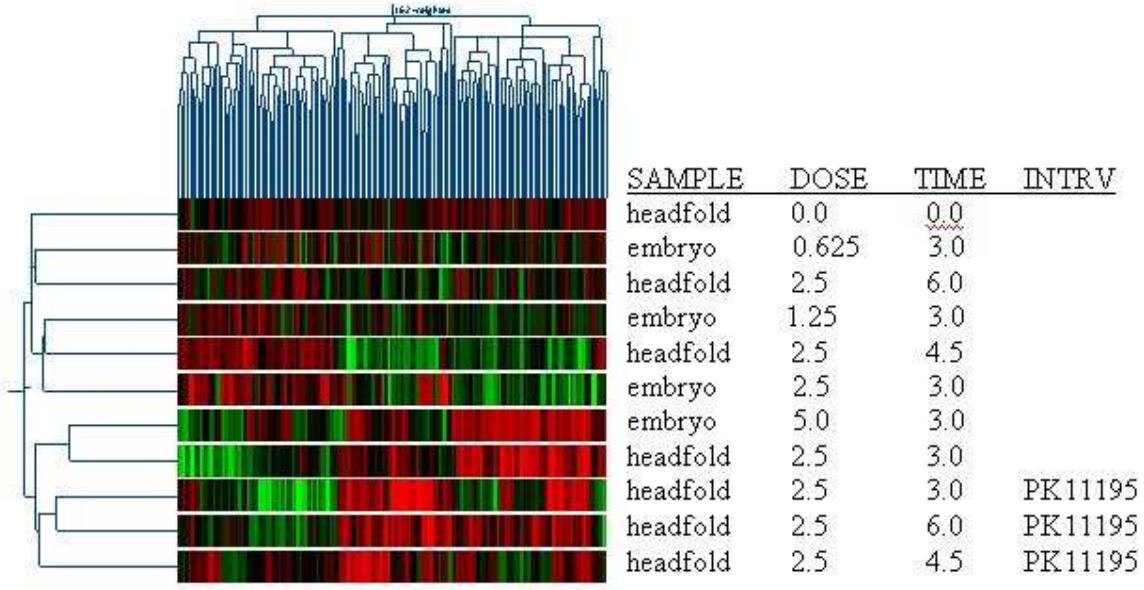


Figure 7: Hierarchical clustering example: Genes were color-coded for up regulation (red) or down regulation (green). Color intensity gives the magnitude relative to normal expression (black). Adapted from [15].

CONCLUSION

Microarray imaging is a very powerful technique to study the gene expressions, understand the interactions between the genes, and to discover the functions of the unknown genes. This knowledge would help to diagnose several diseases such as cancer or diabetes in the early stages and it would help to find person-specific cures.

However, no matter how carefully the experiment is conducted, some errors are introduced to the image. These errors can come from the nature of biology like the inconsistent hybridization of cDNA, as well as from technical issues like the noise from several sources. Furthermore, microarray imaging is still a very expensive technology so duplicating the experiments is not that easy. Therefore to be cost effective, each experiment should be analyzed at best, to obtain the best possible results.

REFERANCES:

[1] S. Bennett. (2004, Feb). Array of hope for personalized medicine, Current Drug Discovery. [online]. Available: http://www.currentdrugdiscovery.com/pdf/2004/520515.pdf

[2] R.S.H Istepanian, "Microarray image processing: current status and future directions," *IEEE Transactions on NanoBioscience*, Vol. 2, issue 4, pp. 173- 175, Dec. 2003

[3] Jeremy Buhler, (2002, August 27). [Online] Anatomy of a Comparative Gene Expression Study. Available: http://www.cs.wustl.edu/~jbuhler//research/array/ June 20, 2004 [date accessed]

[4] A. M. Campbell.(2001). DNA Microarray Methodology - Flash Animation. Available: http://www.bio.davidson.edu/courses/genomics/chip/chip.html June 21, 2004 [date accessed]

[5] A.Kuklin. (2000, May). Spot Checks. [online]. 3(5), pp. 52-54. Available http://pubs.acs.org/hotartcl/mdd/00/may/spotcheck.html

[6] Arrayer Overview. Available : http://cmgm.stanford.edu/pbrown/arrayer.html. June 21, 2004 [date accessed]

[7] M. Katzer, F. Kummert, G. Sagerer, "Methods for automatic microarray image segmentation;" *IEEE Transactions on NanoBioscience*, vol.2 issue 4, pp. 202-214, Dec. 2003

[8] M. Katzer, F. Kummert,G. Sagerer, "Methods for automatic microarray image segmentation;" *IEEE Transactions on NanoBioscience*, vol.2 issue 4, pp. 202-214, Dec. 2003

[9] M. Hessner, X. Wang, K. Hulse, L. Meyer, Y. Wu, S. Nye, S.-W. Guo,and S. Ghosh, "Three color cDNA microarrays: Quantitative assessment through the use of fluorescein-labeled probes," *Nucleic Acids Res.*, vol.31, pp. e14, 2003.

[10] X. Wang, N. Jiang, X. Feng, "A novel approach for high-quality microarray processing using third-dye array visualization technology ," *IEEE Transactions on NanoBioscience*, vol.2, pp.193-201, Dec.2003

[11] X.H Wang, R.S.H. Istepanian, Y. H. Song, "Microarray image enhancement by denoising using stationary wavelet transform;" *IEEE Transactions on NanoBioscience*, vol.2 issue 4, pp.184-189, Dec.2003

[12] X.H Wang, R.S.H. Istepanian, Y. H. Song, "Application of wavelet modulus maxima in microarray spots recognition;" *IEEE Transactions on NanoBioscience,* vol.2 issue 4, pp. 190- 192, Dec. 2003

[13] P. O'Neill, G.D. Magoulas, X. Liu," Improved processing of microarray data using image reconstruction techniques;" *IEEE Transactions on NanoBioscience*, vol.2 issue 4, pp. 176-183, Dec. 2003

[14] Y. H. Yang, M. J. Buckley, S. Dudoit and T. P. Speed, "Comparison of methods for image analysis on cDNA microarray data," University of California, Berkeley, Tech. Rep. #584, November 2000.

[15] 2001 Progress Report: Molecular Characterization of a Biological Threshold in Developmental Toxicity [July 02, 2004] Available:

http://cfpub.epa.gov/ncer_abstracts/index.cfm/fuseaction/display.abstractDetail/abstract/975/report/2001 July 2, 2004 [date accessed]

[16] E. Rouchka, CECS 694-02. Class Lecture "Microarray Image Analysis", University of Louisville, Louisville, Spring 2003.