Improving Microarray Image Acquisition and Quantification

Rashi Gupta
Department of Mathematics and Statistics
and
Institute of Biotechnology
University of Helsinki
Finland

1 INTRODUCTION

DNA microarrays provide an independent platform that offers the ability to analyze extremely large amount of information in parallel [1]. An important application is the gene expression analysis which gives an insight into an organism's metabolism and its regulation.

Thousands of DNA sequences relevant to the biological questions are selected and printed, with identified genes, on the glass microscope slide with a robotic arrayer. The DNA sequences are spotted in a rectangular layout with one spot for each sequence. A popular experimental procedure is monitoring the mRNA abundance in two samples. From the two biological samples in question, mRNA are extracted and labeled, one usually with Cy5 and the other with Cy3, which emit at 630-660 nm and 510-550 nm respectively [2]. When two labeled samples are simultaneously allowed to hybridize with the probes printed on the slide, the relative abundance of the hybridized mRNA in the samples can be measured [3].

After hybridization, the slide is scanned with lasers and two digital TIFF images are produced, where intensities reflect the abundance of the labeled samples. The digital images are then analyzed and the gene activities in the two samples are determined separately. These measurements are used to determine the relative expressions of the genes, which is commonly presented as the ratio of the sample activities [3].

Many sources of variation are associated with each step of the experimental procedure of microarray. Efforts have been made to optimize the laboratory protocols and image analysis to increase the accuracy of the results. Normalization methods have been developed to account for the asymmetry in data[4]. Little attention is however paid to the scanning procedures, although it has been proposed that errors in microarray data may be introduced during scanning.

In all microarray experiments, the chips involved in the experiment are scanned once and intensities of gene extracted from these scans by using any commercial or freely available image analysis software. Care is taken by designing sophisticated image analysis tools that capture the true picture of what is spotted on the chip but absolutely no precautions are made of what quantity is being captured in the scans. In subsequent sections, the issues involved in scanning are mentioned and methods to improve the acquisition and hence improving the data quality are discussed.

2 DATA GENERATION AND PROBLEMS

2.1 Image acquisition protocols

Microarray images are acquired by using laser scanner. The scanner performs an area scan of the slide and produces for each dye a digital map or an image, of the fluorescent intensities for each pixel. For a typical microarray experiment, the scanner produces two 16-bit tagged image file format (TIFF files), one for each fluorescent dye. Different dyes absorb and emit light at different wavelengths. Inorder to measure the abundance of the two fluorescent dyes for each spot, the scanners are designed to generate excitation light at different wavelengths and detect different emission wavelength. The dyes used usually are Cy3 and Cy5 having emission in 510-550nm and 630-660nm ranges respectively.

A number of settings (eg: scan rate, laser power, PMT voltage) can be adjusted by the user at the time of scanning. In general, a higher laser power excites more photons and generates more source signal and more source noise. A higher PMT voltage amplifies more electrons per photon and generates more detector noise and more signals.

It is desirable to use a higher laser power rather than a higher PMT voltage, as this would excite more photons for the signal rather than produce more signals per photon. However, high laser power can damage the hybridized samples by photobleaching, and depending on the number of scans to be performed on each sample, the laser power will need to be adjusted accordingly.

The process of scanning an array is known as image acquisition, whereas the process of converting images to numerical data is referred to as image quantification or processing. A wide variety of different scanning instruments are available, and a number of different image acquisition and quantification packages are associated with them. In general, selection of image quantification parameters (e.g. ‘adaptive’, ‘fixed circle’, ‘spot distance’) should be carefully assessed and decided for each project as a whole, and this also depends upon the array design, slide type and spot morphology. It should be noted that image quantification method should be identical for all slides constituting a project, whereas image acquisition parameters, for instance laser power and/or photo multiplier can be optimized from slide to slide.

2.2 Need for more scans

Usually for each slide a certain combination of laser power and PMT is chosen while scanning. The choice of these two parameters is made so that almost all expression on the chip could be captured. But it has been observed that not all the genes spotted on the chip get accurately measured under a single scanner setting. There are genes with expression ranging from 50,000 or more to genes with expression as low as 200 or even less. Such wide range of expression is just impossible to be captured in just one scan, with a fixed setting, accurately. Surely a single scan with some PMT and Laser setting is suitable for most of the intensity range but surely not all. Thus there is a need to capture the various ranges of gene expression value and then combine the information from all the scans before any further analysis is carried out.

In addition to the above, quite a number of times a few genes do not get measured properly due to scanning issues. The expressions of these genes are lost or in- correctly captured. The possibility of making more scans and utilizing the expression value for gene from all the 3 scans help to even cross check for incorrectly measured spots.

A single scan attempts to capture the whole range of the expression. This may not give the true picture of the expression of the whole set of genes. Also there are problems of saturation because of the acquisition devices. As a result it is impossible to measure beyond 65,536. It is believed that above 50,000 erroneous ratios occur due to the saturation of pixels within the spots. As a result there is a need to be able to accurately measure the expression of gene.

2.3 Problems with scanning

Producibility of microarray data is dependent on a number of factors like RNA quality, the choice of reverse transcriptase, choice of genes, concentration at which the genes are spotted on to the microarray chip and many more. One problem that is observed in almost all labs is the instability of the fluorescent dye once it is incorporated into cDNA and hybridized to the array. As a result of instability of the dye, there is a loss of signal from one/both fluorescent channels. Also it was observed that dye not only degrades between the last wash and the first scan but also degrades after the first scan. Few factors that affect signal quality are summarized below:

1) RNA Degradation
2) Capture reagent formed aggregates
3) RNA contamination with genomic DNA
4) Arrays are covered with a plastic coverslip
5) Dye oxidation (Cy5)
6) Dye photobleached

3 PROPOSED APPROACHES

One of the solutions to the above problem has been made by Heidi Lyng et al. [5]. In their article, they proposed one such method to improve the expression of gene. The procedure aimed at making two scans of each channel. The primary scan was made taking care of the low expressed gene and the secondary scan was made such that the intensities of the brightest spot was just below the level of saturation. The primary scan forms the basis of analysis whereas the secondary scan was used to correct the intensities of spots with saturation in the primary scan. The PMT for the first scan was maintained so as to avoid intensities of the weakest spot below 200, allowing the brightest spots to reach the level of saturation. The second scan was acquired with a lower PMT setting such that no pixels are getting saturated. Reliable data for spots with saturation in the first set of image was extracted from the second set of image by their algorithm. However, the algorithm tried to correct the intensities above 50,000 where some pixels or the whole spot was getting saturated. They used the intensities from 20,000-30,000 to find the scaling factor and this factor was used to correct the intensities from 50,000 above. This correction is made for only a small range and no care is taken if scanning has created problems in the other ranges of intensities.

Another method is proposed by L.E.Dodd et al. [6]. The method is based on a censored regression model. The model does not require scanning the array at multiple PMT voltages.

4 DISCUSSIONS

It is extremely important to note that the gene expression is just an approximation to the true expression of the gene and additional measurements of the same gene though under different setting would help give better estimate of the expression for each gene. Thus good care should be taken to get the best measure of the expression from the chip before any analysis trying to infer the regulatory networks, pathways etc are made.

REFERENCES

1. Kohane, I.S., Kho, A.T. and Butte A.J. (2002) Microarrays for an Integrative Genomics. MIT Press.

2. Shalon, D., Smith, S.J. and Brown, P.O. (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*., **6**, 639-645.

3. Chen, Y., Dougherty, E.R. and Bittner, M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.,* **2**, 363-374.

4. Quackenbush, J. (2002) Microarray data normalization and transformation. *Nature Gen.,* **32** Suppl., 496-501.

5. Lyng, H., Badiee, A., Svendsrud, D.H., Hoving, E., Myklebost, O. and Stokke, T. (2004) Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction. BMC Genomics 2004, 5:10

6. Hsiao LL, Jensen RV, Yoshida T, Clark KE, Blumenstock JE, Gullans SR: Correcting for signal saturation errors in the analysis of microarray data. *BioTechniques* 2002, 32:330-336