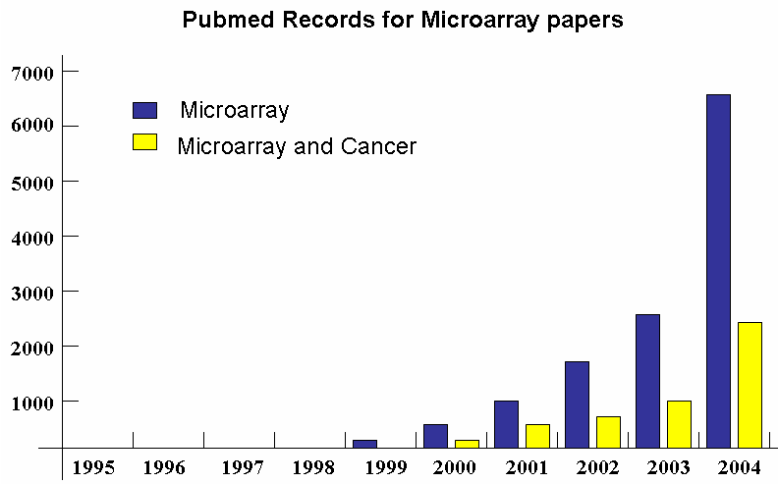
STATC141 Spring 2005
Lecture 10

Gene Expression Data Analysis (I)

Introduction to microarray technology
Since it innovation, Microarray technology has been widely used in biological and medical research.



In order to explain microarray data analysis, it is important to first have an understanding of microarray technology. I begin with an introduction to genes and why scientists are interested in them.

What is a gene?
Genes are specific sequences of DNA that determine our individual characteristics. Our genes are what determine our eye color, hair color, height, etc. A gene produces amino acids, which are the building blocks of proteins. Proteins regulate everything that happens in our body. So in the overall picture, genes regulate the actions of proteins.

A brief science background
Researchers evaluate the state of a cell based on what genes are expressed within it. Microarray technology allows researchers the ability to see what genes are being expressed in individual cells, and at what level. In other words, how much of each possible protein is the cell producing under particular circumstances? This is the information that microarray technology can provide.

Microarray technology allows us to take a 'photograph' of genes and catch them in action. For example, a photo generally shows who is wearing what type of clothes or shoes (normal DNA versus mutations) and captures the activity of people, such as sitting, sleeping or running (activity level of your genes, which ones are dormant and which ones have high/low levels of activity). By looking at the level of activity in different genes, scientists can narrow their research of high/low levels of activity. An example would be a comparison of genes from tumor tissue against ones that are non-cancerous. If there are fairly high/low levels of activity of genes in the tumor cells compared to the non-cancerous cells, scientists can then take a closer look into the roles these particular genes play.

What is a microarray?
Now that we have discussed what a gene is and what we gain from microarray technology, we can now explain what a microarray actually is.

In brief, microarray is a new powerful technology for biological exploration which enables one to simultaneously measure the level of activity of up to 60,000 genes. In particular, the amount of mRNA for each gene in a given sample (or a pair of samples) is measured.

The microarray technology is based on central dogma (figure 1) and complementary hybridization (figure 2). Microarrays exploit the preferential binding of complementary single-stranded nucleic acid sequences. The idea behind microarray is to measure the amount of mRNA to find which genes are being expressed (active). Measuring protein levels might be better, but is much harder.

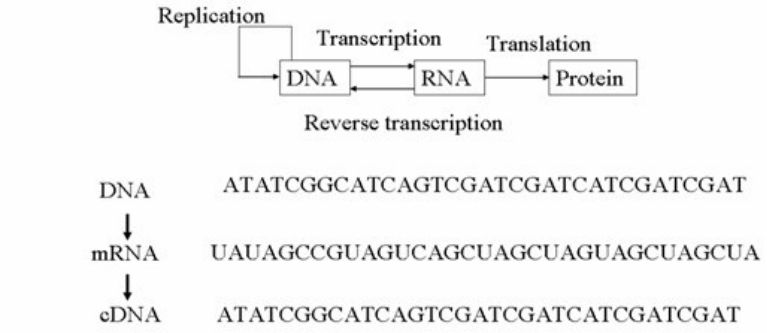


Figure 1. Sketch of central dogma

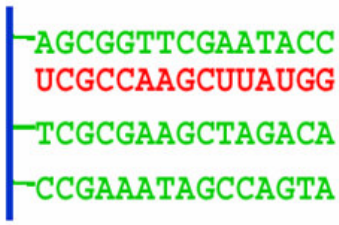


Figure 2. Due to Watson-Crick base pairing, an mRNA molecule will hybridize to a complementary DNA molecule

Figure 3. How complementary hybridization is usually done:
1). put the DNA sequence that can base pair on array
2). convert mRNA to cDNA using reverse transcriptase

Different types of microarrays

	Stanford/ Pat Brown	Affymetrix
How DNA sequences are laid down	Spotting	Photolithography
Length of DNA sequences	DNA or cDNA (Complete sequences)	Oligonucleotides

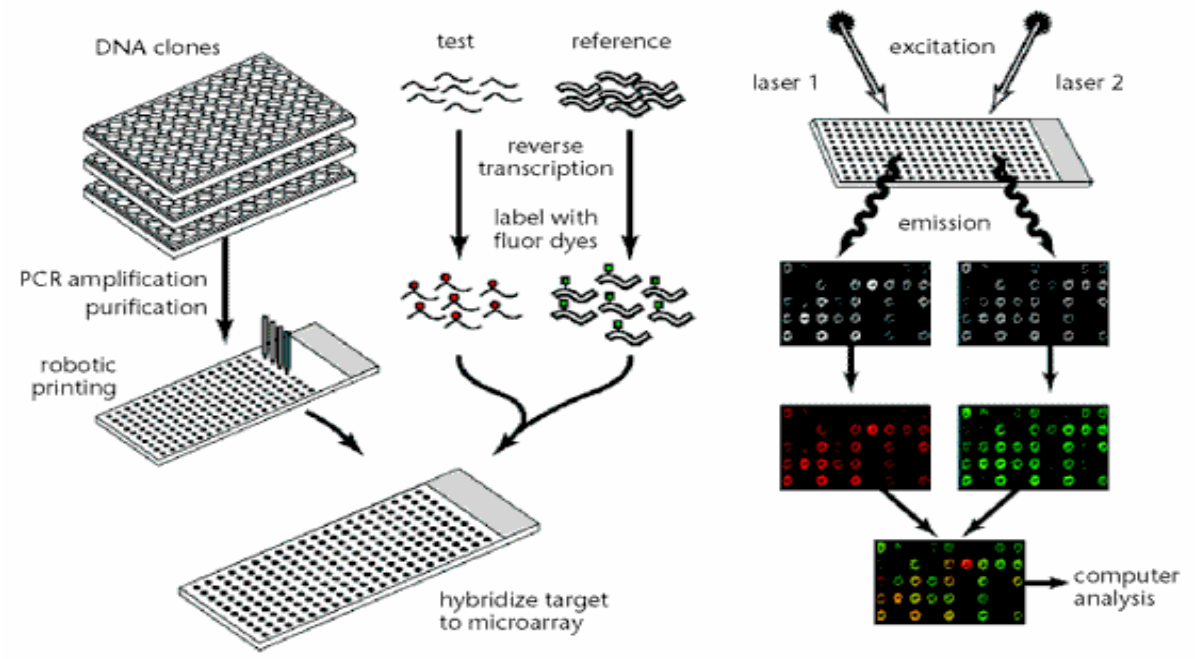
DNA (spotted) microarrays
Below link provides an animation which demonstrates how DNA (spotted) microarray experiments are performed (HIGHLY RECOMMENDED!).
http://www.bio.davidson.edu/courses/genomics/chip/chip.html

A DNA microarray is typically a glass (or some other material) slide, on to which DNA molecules are attached at fixed locations (spots). There may be tens of thousands of spots on an array, each containing a huge number of identical DNA molecules (or fragments of identical molecules), of lengths from twenty to hundreds of nucleotides. (According to quick napkin calculations by Wilhelm Ansorge and John Quackenbush in Schnookeloch in Heidelberg on 4 October, 2001, the number of DNA molecules in a microarry spot is 10^7-10^8). For gene expression studies, each of these molecules ideally should identify one gene or one exon in the genome; however, in practice this is not always so simple and may not even be generally possible due to families of similar genes in a genome. Microarrays that contain all of the about 6000 genes of the yeast genome have been available since 1997. The spots are either printed on the microarrays by a robot, or synthesized by photo-lithography (similarly as in computer chip productions) or by ink-jet printing.

There are different ways how spotted microarrays can be used to measure the gene expression levels. One of the most popular micorarray applications allows comparing gene expression levels in two different samples, e.g., the same cell type in a healthy and diseased state.

The total mRNA from the cells in two different conditions is extracted and labeled with two different fluorescent labels: for example a green dye for cells at condition 1 and a red dye for cells at condition 2 (to be more accurate, the labeling is typically done by synthesizing single stranded DNAs that are complementary to the extracted mRNA by a enzyme called reverse transcriptase). Both extracts are washed over the microarray. Labeled gene products from the extracts hybridise to their complementary sequences in the spots due to the preferential binding - complementary single stranded nucleic acid sequences tend to attract to each other and the longer the complementary parts, the stronger the attraction.

The dyes enable the amount of sample bound to a spot to be measured by the level of fluorescence emitted when it is excited by a laser. If the RNA from the sample in condition 1 is in abundance, the spot will be green, if the RNA from the sample in condition 2 is in abundance, it will be red. If both are equal, the spot will be yellow, while if neither is present it will not fluoresce and appear black. Thus, from the fluorescence intensities and colors for each spot, the relative expression levels of the genes in both samples can be estimated.



The raw data that are produced from microarray experiments are the hybridized microarray images. To obtain information about gene expression levels, these images should be analyzed, each spot on the array identified, its intensity measured and compared to the background. This is called image quantitation.

Image quantiation is done by image analysis software. To obtain the final gene expression matrix from spot quantiations, all the quantities related to some gene (either on the same array or on arrays measuring the same conditions in repeated experiments) have to be combined and the entire matrix has to be scaled to make different arrays comparable.

Summary of spotted microarrays

Two-dye design : Measures comparative expression level by competitive hybridization. On the contrary, Affymetrix measures absolute expression level.

Advantage:

- Cheaper
- Flexibility of custom-made array : Many labs can build their own cDNA/DNA library and print their own arrays.

Disadvantage:

Big variability introduced : Each gene usually has only one corresponding cDNA/DNA on the array.

Statistical issues in spotted microarray analysis

1. Image analysis: identify spot area and extract intensities for each spot
2. Normalization: normalizing dye effects, slide effects, etc
3. downstream analysis: finding differentially expressed genes; clustering and classification

Image Analysis:

- Identify spot area: Each spot contains around 100×100 pixels. Spot image may not be uniformly and roundly distributed. Some software (like ScanAlyze or ImaGene) have algorithms to help placing the grids and identify spot and background area locally. Sometimes still have to adjust manually.
- Extract intensities (data reduction): Aim to extract the minimum most informative statistics for further analysis. Usually use signal minus background and some spot quality indexes.