

Summer Institute for Training in Biostatistics - 2005

Lecture III: A brief overview of microarray gene expression analysis

Sündüz Keleş

Department of Statistics and of Biostatistics & Medical Informatics

Affymetrix data files

As a result of a microarray gene expression experiment with Affymetrix high density oligonucleotide arrays, we acquire the following files:

CEL files contain information about the expression levels of the individual probes.

CDF file contains information about which probes belong to which probe set. The probe set information in the CEL file by itself is not particularly useful as there is no indication in the file as to which probe set a probe belongs. This information is stored in the CDF library file associated with a chip type. All the arrays belonging to a given type will share this same information.

⇒ These two files are the starting point of a (bio)statistician.

So, the experiments are done. What next?

- Low level analysis: image processing, normalization, quantification of expression levels.
- High level analysis: statistical test for differential expression, clustering/classification.

Microarray Analysis

1. Sample Generation, experimental design, hybridization ⇒ RAW DATA.
2. Data analysis.
 - (a) Scanning and image analysis ⇒ PROBE INTENSITIES. [**low level analysis**]
 - (b) Background correction, normalization and quantifying expression measures ⇒ NORMALIZED EXPRESSION MEASURES FOR EACH GENE (PROBE SET). [**low level analysis**]
 - (c) Ranking of genes, clustering/classification of genes ⇒ INTERESTING GENE SET. [**high level analysis**]
3. Validation of the interesting genes: using additional biological data and performing targeted experiments.

Case study

”The investigators in this experiment were interested in the effect of estrogen on the genes in ER+ breast cancer cells over time. After serum starvation of all eight samples, they exposed four samples to estrogen, and then measured mRNA transcript abundance after 10 hours for two samples and 48 hours for the other two. They left the remaining four samples untreated, and measured mRNA transcript abundance at 10 hours for two samples, and 48 hours for the other two. Since there are two factors in this experiment (estrogen and time), each at two levels (present or absent, 10 hours or 48 hours), this experiment is said to have a 2x2 factorial design.”

Bioconductor package Affy

Provides a platform for low level analysis of gene expression data from Affymetrix chips.

To do:

- Reading the CEL files into R. [Note: For spotted cDNA arrays, the first step would be image analysis – recall that one could get spots in variety of shapes from cDNA arrays and image analysis is very important to extract intensity signals efficiently.]
- Diagnostic plots: image plots to check for any abnormalities, histograms, M-A plots using raw unnormalized data.
- Normalization and quantification of expression levels.
- More diagnostic plots using normalized data.
- Simple summary statistics.

Application with the case study in R. [AffyEg.S]

MA plots

Measurement of relative expression versus Average log intensity plot.

Also known as the RI (Ratio versus Intensity).

MA plots can show the intensity-dependant ratio of raw microarray data. Let R represent raw intensity for one chip and G represent raw intensity for another chip. Define

$$M = \log_2(R/G) \quad A = \log_2 \sqrt{RG}.$$

Normalization

The purpose of normalization is to adjust for effects which arise from variation in the microarray technology rather than the biological differences between the RNA samples or printed probes.

Imbalances between chips may arise from hybridization on different days (temperature/humidity) or by different people, different scanner settings etc.

MA plots

In spotted cDNA arrays, R and G represent the green and red channel intensities measured on a single chip. For Affymetrix chips, there is only one channel on each array, so so M and A are defined based on two different chips, e.g., low10 versus low48.

What type of MA plot would we expect for intensities from two experiments that are exact replicates of each other?

Case study 2: Golub *et al.* (1999) data

Study of gene expression in two types of acute leukemias: acute lymphoblastic (ALL) and acute myeloid leukemia (AML). Gene expression levels were measured using HU6800 chip containing probes for approximately 6800 human genes. The data comprise 47 cases of ALL and 25 cases of AML. Samples are divided into a *learning set* with 38 observations and a *test set* with 34 observations.

Acknowledgements

Bioconductor exercises: Working with Affymetrix data: estrogen, a 2x2 factorial design example. R. Gentleman and Wolfgang Huber (2004).