Biological Sequence Analysis Spring 2009 Dr. Richard Friedman
(212)851-4765 friedman@cancercenter.columbia.edu 824 ICRC

Lesson 12: Functional Genomics: I Analysis of Microarray Expression Data

In this class we are going to learn the basics of normalizing and analyzing microarray expression data.

Reading:

1. Pevsner, P. 189-203.
2. The powerpoint slides for this lesson available on the courseworks site (read notes as well as slides).
3. About AffylmGUI
http://bioinf.wehi.edu.au/affylmGUI/R/library/affylmGUI/doc/about.html
4. Running the Estrogen Dataset
http://bioinf.wehi.edu.au/affylmGUI/R/library/affylmGUI/doc/estrogen/estrogen.html

Theory

(For more detail see associated powerpoint slide and notes)

1. Concentrations 2 color experiment.

$$\frac{RNA(Experiment)}{RNA(Control)} = \frac{Intensity\,\mathrm{Re}\,d}{IntensityGreen}$$

2. Concentrations in 1 color experiment:

$$\frac{RNA_{\text{experiment}}}{RNA_{\text{control}}} = \frac{Intensity_{\text{experiment}}}{Intensity_{\text{control}}}$$

3. Probeset intensities as an average of probe instensities

$$I_{probeset} = \sum_{j=1,k} \frac{\log_2(PM_j - MM_j)}{k}$$

PM_j = perfect match in jth probe.
MM_j = mismatch in jth probe.

4. GCRMA Normalization

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}$$

Y_{ijn} = the background corrected intensity of the
jth probe on the
nth probeset on the
ith chip.

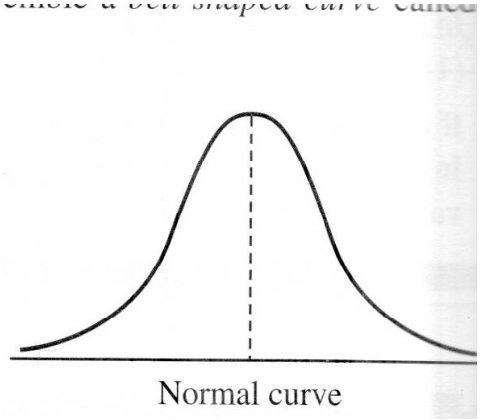
μ_{in} =is the intensity of the nth probeset on the ith chip

α_{jn} =is the intensity of the jth probe in the nth probeset.

Subject to the following constraint on probe effects:

$$\sum_{j=1,J} \alpha_{jn} \cong 0$$

5. Normal Distribution

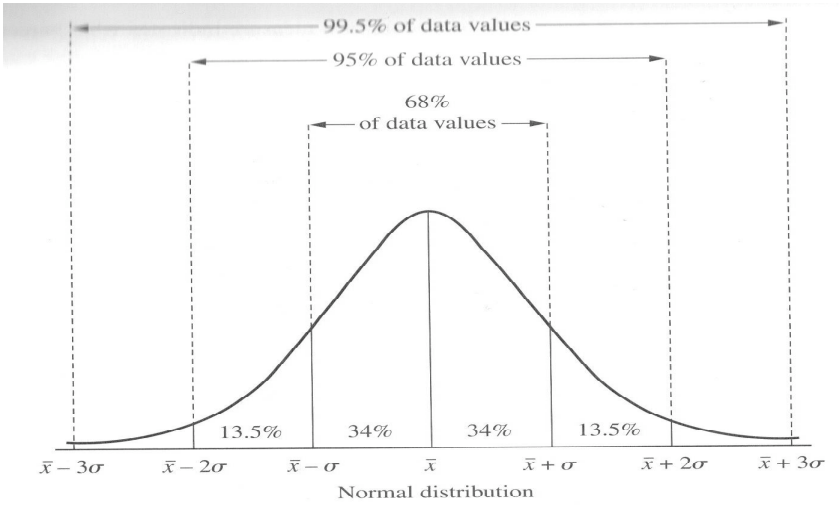

Normal curve

$$\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma}}$$

μ= mean - the center of the curve.

σ= standard deviation of the curve - width of the curve.

6. Standard deviation and percent



7. Estimates of the mean and the standard deviation of the mean:

The mean:

$$\bar{x} = \frac{\sum_{i=1,N} x_i}{N}$$

The standard deviation of the mean:

$$s_x = \sqrt{\frac{\sum_{i=1,N} (x_i - \bar{x})^2}{N(N-1)}}$$

8. The z distribution:

$$z = \frac{\bar{x} - \mu}{s_x}$$

z= number of standard deviations from the mean.

9. Example of use of z distribution:

Let's say that we want to know if the amount of carbon dioxide in air is equal to 10.00(mg/m^3). We take a 1000 measurements or carbon dioxide in a sample and find it to be 10.43(mg/m^3). The standard deviation of our measurement is .24(mg/m^3). We want to calculate the probability that the concentration of CO_2 is our sample is 10.00(mg/m^3).

$$z = \frac{\bar{x} - \mu}{s_x}$$

$$\bar{x} = 10.43 mg/m^3$$

$$\mu = 10.00 mg/m^3$$

$$s_x = .24 mg/m^3$$

$$N = 1000$$

$$z = \frac{10.43 - 10.00}{.24} = 1.79$$

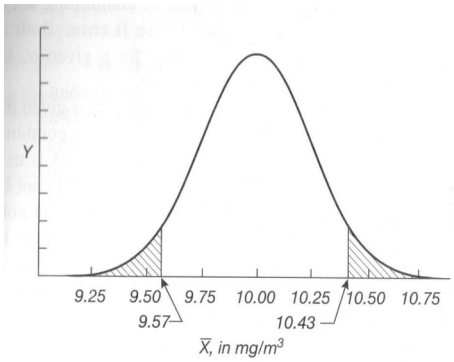
From the normal distribution:

$$p(\bar{x} \geq 10.43 mg/m^3) = p(z \geq 1.79) = 0.0367$$

We also have to take into account the probability that it is less than 1.79 standard deviation. Since the function is symmetrical:

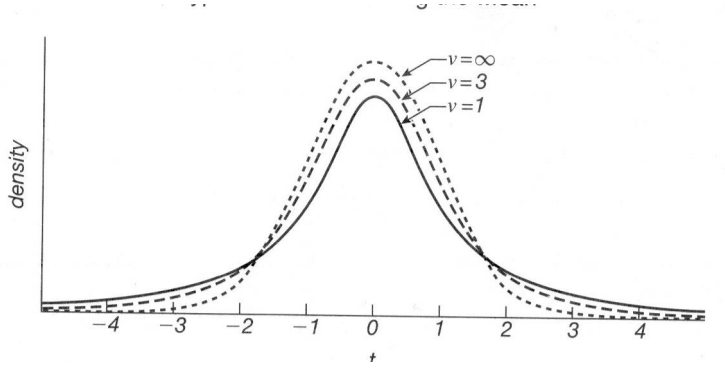
$$p(z \leq -1.79) = 0.0367$$

The total probability that the measured mean can be .43(mg/m^3) away from 10.00 by chance is the area under the curve .43(mg/m^3) away from 10.00 (mg/m^3) in either direction.



$$p(z \geq 1.79 \text{ OR } z \leq -1.79) = 0.0367 + 0.0367 = 0.0734$$

10. For small sample sizes use t not z.



$$t = \frac{\bar{x} - \mu}{s_x}$$

$N \to \infty,\ t \to z$

10. The t distribution of the difference between 2 means.

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}}}$$

s_p= pooled standard deviation of samples.

$$s_p^2 = \frac{\sum_{i=1,N_1} (x_{i1} - \bar{x}_1)^2}{N_1 + N_2 - 2} + \frac{\sum_{i=1,N_2} (x_{i2} - \bar{x}_2)^2}{N_1 + N_2 - 2}$$

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\left(\frac{\sum_{i=1,N_1} (x_{i1} - \bar{x}_1)^2}{N_1 + N_2 - 2} + \frac{\sum_{i=1,N_2} (x_{i2} - \bar{x}_2)^2}{N_1 + N_2 - 2}\right)}}$$

11. Log transformation to stabilize variance and log fold change, m.

$x -> \log_2(x)$

$$m = \log_2 x_2 - \log_2 x_1 = \log_2\left(\frac{x_2}{x_1}\right)$$

12. Problems with applying –test to microarray experiments:

A. Multiple Tests- 1,000 of genes.
B. Multiple comparisons more than 2 experiments.

LIMMA: LInear Model for Microarray Analysis

13. How LIMMA solves multiple test problem:

A. Variance stabilization by addition of fudge factor $s_0(m)$ reduces false positives.

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}} + s_0(m)}$$

B. Benjamini-Hochberg false discovery rate correction:

Uncorrected p-value= rate of false discovery if only 1 test.

Corrected p-value= rate of false discovery if all of the genes above it were tested.

14. How LIMMA solves multiple comparison problem:

A. Multiple samples: Include variablility of all samples in standard error.
Example: Comparing means of 2 samples when 3 samples taken:

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\left(\frac{\sum_{i=1,N_1}(x_{i1} - \bar{x}_1)^2}{N_1 + N_2 + N_3 - 3} + \frac{\sum_{i=1,N_2}(x_{i2} - \bar{x}_2)^2}{N_1 + N_2 + N_3 - 3} + \frac{\sum_{i=1,N_3}(x_{i3} - \bar{x}_3)^2}{N_1 + N_2 + N_3 - 3}\right)}}$$

B. Multiple Comparison test (done by user-rarely used).
Example: Bonferroni Correction:

$P_{Corrected}$ =# comparisons x P_{Raw}

15. Bayesian t:

$B = \log(\text{odds of differential expression})$

$\text{odds} \geq 1$

Tentative cutoff:

$B = \log(odds) \geq 0$

or

$p_{FDR} \leq .05$

Caveat Experimentalist: Results must be validated by PCR.

Summary of Commands:

Note: In this document different fonts have different meanings:

Times is used to explain commands.

`Courier is used to indicate commands and command options.`

`Courier italics are used to indicate command parameters, for example, filenames.`

`Courier bold is used to indicate commands that are not displayed.`

`Courier bold italics are used to indicate computer-generated output.`

Helvetica is used to indicate menu items.

I. Installing Software

Bioconductor Software is installed in the computer classroom. The following sites are for installing Bioconductor software in your own labs.

http://www.bioconductor.org	Web-site to download Bioconductor- a microarray analysis suite that runs in the R statistical programming environment.
http://bioinf.wehi.edu.au/affylmGUI/	Web-site to download AfflmGUI, a graphic user interface for normalization and signifcance analysis of Affymetrix data.

II. Running AffylmGUI.

`R`	Click on R icon on desktop.

Package->Load Package->AffylmGUI	Get AffylmGUI graphic user interface. Follow instructions in the estrogen test set Tutorial Except:
GCRMA	Recommended normalization procedure.
All Genes/B Statistic/FDR/	Recommended toptable options. (Save differences at 10hr and 48 hr).

Lab

1. Use affylmGUI to find a list of genes that are significantly differentially expressed in the estrogen experiment between estrogen present and estrogen absent at 10 hours. The estrogen files are in:
   C:\Program Files\R\R-2.6.1\library\estrogen\extdata.
   Your output should be include:

   A. A density plot of on of the chips.
   B. An image of one of the chips.
   C. A boxplot of the quantile distribution of the unnormalized data.
   (Normalise with PLM)
   D. A signed residual plot of all of the PLM normalized arrays.
   E. An RLE plot of all of the PLM normalized arrays.
   F. An NUSE plot of all of the PLM normalized arrays.
   (Close AffylmGUI and reload. Normalize with GCRMA.
   G. A boxplot of the quantile distribution of the GCRMAnormalized data
   H. An Excel file of the normalized expression values
   I. An Excel file of all of the genes sorted by decreasing B value at 10 hours.