STAT 6560
Graphical Methods

Spring Semester 2009

Project 2

Bill Welbourn

Utah State University

Department of Mathematics and Statistics

3900 Old Main Hill

Logan, UT 84322–3900

Graphics for Biomedical Informatics

R Packages Discussed in this Project: *affyPLM*, *gplots*, and *hopach*.

Project Objective: Provide an example of a bioinformatics analysis, particularly the importance of graphical representation of data and results.

Motivation: We would like to examine the underlying biological mechanisms which produce a phenotypic trait. In particular, it is desired to understand the proteins involved in some underlying disease causative pathway. Preventing – or perhaps modifying – the biological function of the protein, could reduce the likelihood of disease. For example, suppose that in the presence of a certain protein, the odds of cancer increases substantially. If we could somehow hinder the protein's function, it is likely we would see a decrease in the incidence of cancer.

Background: As it turns out, proteins are very complex, three-dimensional molecules, possessing four levels to their structure. So, rather than studying proteins directly, we instead examine a much simpler molecular structure, DNA. The idea is that when expressed, DNA molecules give rise to proteins. Thus, in studying the mechanics of DNA, we are indirectly also studying the underlying mechanics of proteins.

An Example (based on the study conducted by Obst et al. (2007);
data available at `http://www.ncbi.nlm.nih.gov/projects/geo/`, GEO Accession: GSE5245)

Study Design: Suppose that we are interested in how T receptor cells respond, depending on the duration of exposure to an antigen. This could be important, for example, in determining how the immune system responds within HIV+ patients. To carry out the investigation, we collect a random sample of sixteen (16) mice from a breeding farm. The mice are randomly assigned to one of three antigen exposure groups: (*i*) five (5) mice are randomly assigned to receive a short duration of exposure to antigen; (*ii*) six (6) mice are randomly assigned to receive a long duration of exposure to antigen; and (*iii*) the remaining five (5) mice serve as a control, and as such are not exposed to antigen. Gene expression, measured on the continuum of mRNA (the product obtained when DNA expresses) present, are the responses of interest to the investigators.

Figure 1: Affymetrix microarrays: U133 Plus 2.0 (left), array for the Human Genome; 430 2.0 Array (right), array for the Mouse Genome. Image taken from `http://en.wikipedia.org/wiki/File:Affymetrix-microarray.jpg`, April 16, 2009.

Materials: A tissue sample is drawn from each of the sixteen mice, and mRNA extracted. To ready the mRNA for hybridization to a microarray, the mRNA material for each mouse is stained with the Cy5 dye. The fluorescent mRNA product for each mouse is then hybridized to a unique Affymetrix GeneChip Mouse Genome 430 2.0 Array (see `http://www.affymetrix.com/products_services/arrays/specific/mouse430_2.affx`), so that sixteen biological replicates are obtained. Figure 1, displays two examples of the structure to a microarray.

Methods: (*i*) Quality Control: To assess the veracity of the microarray samples, we create MAPlots and Probe Level Model (PLM) plots; (*ii*) Statistical Analysis: The two mice groups exposed to antigen are collapsed into a single group. Gene expression profiles for the merged group are compared to those of the control group, applying (essentially) a two-sample t-test to each gene represented on the microarrays.

Quality Control – The affyPLM R Package

- MAplot() function: We could create scatterplots of gene expression profiles, for each of the pairwise array combinations (for our example, we would have 120 scatterplots). However, this leads to a cumbersome visual task for quality control. Instead, we create a "pseudo" microarray, utilizing the median gene expression values from the collected microarray samples. The gene expression profile for this pseudo array is compared to those for each of the sixteen arrays, by way of "M-A" scatterplots (MAplots). Let X and Y represent the respective gene expression profiles (on the $\log_2$ scale) for the pseudo microarray and array i ($i = 1, \ldots, 16$). For each of the sixteen microarrays, the MAplot is a scatterplot of M plotted against A, where

$$M = X - Y \qquad and \qquad A = 0.5(X + Y). \tag{1}$$

  Essentially, the MAplot is a rotated and scaled X-Y scatterplot. A locally weighted polynomial regression (loess) curve is fit to the MAplot. Quality problems exist when the loess curve tends to oscillate and/or possess greater variability (when compared to other array MAplots) about the reference line, $M = 0$.

  The MAplot() function takes as its main argument, the "processed" gene chip data. Options to the function include: (i) displaying a subset of the array MAplots; and (ii) scatterplot smoothers (requires geneplotter R package). Figure 2, displays MAplots for six of the mouse microarrays. Based on this figure, the MAplots for the microarrays with labels, GSM118666.CEL and GSM118669.CEL, seem to deviate from the line, $M = 0$ (shown in blue). This suggests a potential quality control issue among the sixteen arrays.

- image() function: For each gene, we model its expression (response) value by a linear regression model. The residuals (or functions thereof) are plotted as an overlay on the respective microarray image. The resulting plot is called a probe level model (PLM) image. Potential quality control issues exist, when the presence of a pattern(s) is(are) apparent.

  The image() function takes a similar main argument as that of the MAplot() function – slightly modified version of the processed gene chip data. Options to the function include: (i) displaying a subset of the array overlayed residual plots;

and (*ii*) the type (e.g., raw residual values, the sign (i.e., positive or negative) of the residuals) of residuals to display. Figure 3, displays the PLM image for array three, where we have chosen to display the sign of the residuals. The plot suggests an artifact (e.g., a fingerprint) may be present on the array, as indicated by the blue "bean" shaped mass toward the lower right corner of the figure.

Results – The gplots and hopach R Packages

- heatmap.2() function (gplots): A heatmap is essentially a graphical representation of a quantitative variable over a two-dimensional "matrix map", where color is used to indicate the magnitude of the variable. The function takes a numeric matrix as its main argument. A number of options are available for the function, and these will be discussed in more detail when we examine the R code (below). A couple of options available within this function, not available in the heatmap() function, are: (*i*) displaying a legend, detailing the numerical values corresponding to the colors depicted in the heatmap; and (*ii*) displaying a "trace" line, which could aid in locating patterns within the data.

- dplot() function (hopach): Pursuant to the R documentation,

  > "Function to make a pseudo-color image of a distance matrix with the row and column ordering based on HOPACH clustering results."

  Essentially, the function takes a matrix of gene expression measurements and a matrix of pairwise gene (or array) distances (within the R code (below), we will utilize the Pearson Correlation Coefficient as a measure of distance) as its main arguments, and displays a heatmap, indicating cluster memberships with dashed lines.

- bootplot() function (hopach): Pursuant to the R documentation,

  > "After clustering, the boothopach or bootmedoids function can be used to estimated the membership of each element being clustered in each of the identified clusters (fuzzy clustering). The proportion of bootstrap resampled data sets in which each element is assigned to each cluster is called the "reappearance proportion" for the element and that cluster. This function plots these proportions in a colored barplot."

  Essentially, another (in addition to the dplot() function) graphical representation of cluster membership, either by gene or array distance.

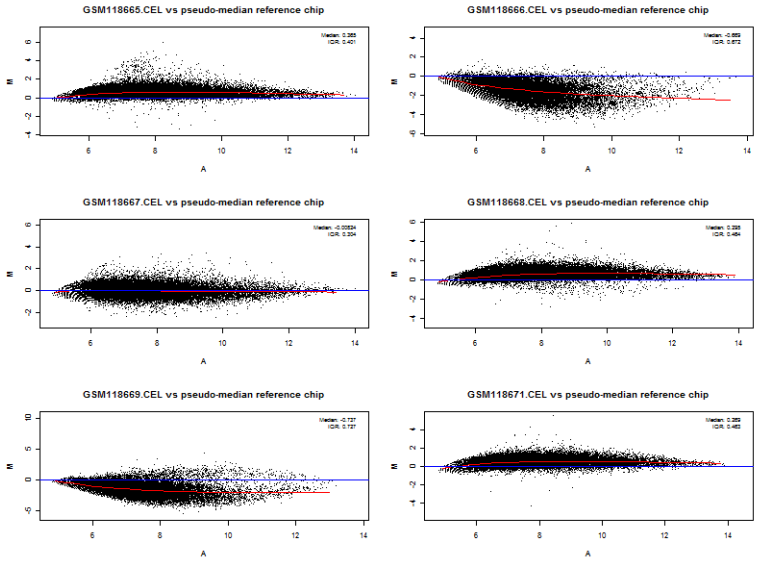

Figure 2: MAplots for six microarrays. The loess curves (shown in red) for the two arrays with respective labels, GSM118666.CEL and GSM118669.CEL, deviate from the line, $M = 0$ (shown in blue). This suggests a potential quality control issue among the arrays.

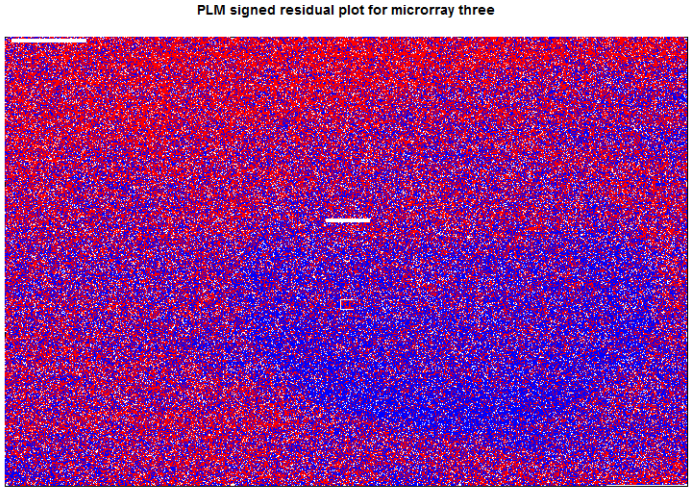

Figure 3: Probe Level Model (PLM) signed residual plot for the microarray with label, GSM118667.CEL. The blue "bean" shaped mass toward the bottom-right corner of the plot, suggests an artifact (e.g., fingerprint) is present on the array.

Links to the R code:

- http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/welbourn_william_project2_genetics.R.

- http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/welbourn_william_project2_bootplot.R.

Link to the data file:

- http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/welbourn_william_project2_genetics.txt

References

Obst, R., Van Santen, H., Melamed, R., Kamphorst, A. & et al. (2007), 'Sustained Antigen Presentation can Promote an Immunogenic T Cell Response, like Dendritic Cell Activation', *Proc Natl Acad Sci USA* **104**(39), 15460–15465.