Microarray Analysis: A Tutorial

Jesse Mecham

1 Introduction

Microarray analysis is becoming an increasingly integral part of biological research. Analysis of cell expression that would have previously taken months to perform can now be carried out in a matter of hours with the use of these miraculous chips. This paper is written as a brief introduction to the world of microarrays and the methods used to analyze them. Furthermore, a tutorial will be presented for readers who wish to perform a simple analysis. This is by no means meant as an all-inclusive source for array analysis. Much of the data used in this tutorial can be found in [7, 5, 8, 6, 1]. Interested readers should direct their further inquiries to those and other applicable sources.

2 Background

In 1995, the first microarray expression profiling was published in Science, ushering in a new era of gene expression analysis [10]. However, as with all new technologies, there is still much confusion about what constitutes a microarray analysis. This section will attempt to give a brief overview about the fundamentals upon which microarray analysis is built.

2.1 Microarrays

A microarray is device that allows for fast and precise analysis of messenger ribonucleic acid (mRNA) directly from a cell. It consists of two parts: the chip and the optical reader. The chip is constructed from a plate of glass to which tens of thousands of cDNA genes are chemically attached in specific locations called “spots” as shown in Figure 1 [7]. An optical reader is used to analyze which of the spots have sample mRNA bound to them and return the results in a format that researchers can use for further analysis.

An experiment is carried out by tagging all mRNA prepared from a sample (cell, tissue, or other biological source) with a fluorescent tag. These tags can be a single color (used in the Affymetrix GeneChip) or more colors (such as the complimentary Cy3 red and Cy5 green dies used in this tutorial). The sample is then allowed to hybridize with the genes on the microarray chip and then washed to remove all of the unbound mRNA’s. The chip is then run through the optical reader which records the location and intensities of the fluorescent tags.
There are three major steps involved in microarray analysis: experimental design, preprocessing, and data analysis [4]. They will be discussed in minor detail below.

2.2 Experimental Design

The first step of conducting a microarray analysis is the experimental design. This may well be the most important step since all decisions made here will drastically effect the results of all subsequent steps. It is important to have a clear understanding of exactly what the experiment is attempting to analyze from the start. If a clear plan is not made, it is likely that unanticipated technological and biological confounding factors arise and drastically alter the outcome of the experiment. Examples of such confounding factors include:

1. Hybridization variation between different transcripts
2. Unanticipated cell-cycle or developmental differences between cell lines
3. Variation in exposure levels to chemicals among cell lines
4. Amplification differences between samples
3. **TUTORIAL**

2.3 **Preprocessing**

Preprocessing is concerned with connecting the chip with final analysis. Preprocessing includes the following tasks:

1. **Data import** is involved in incorporating various file formats into a desired data object. This can be a challenge because different vendors sometimes utilize different data representations.

2. **Background adjustment** all comes down to one word—noise. The noise can be introduced from various sources such as optical distortion, non-specific hybridization, or equipment damage.

3. **Normalization** between samples needs to be established for a variety of reasons. Some of these reasons may include different reverse transcription efficiency levels or hybridization inequalities between samples. In order to properly summarize the results between samples, reasonable normalization algorithms must be applied to the data. Furthermore, normalization can accommodate for variations in spacial localities between chips (especially those that are done in-house).

4. **Summarization of data** is the process of reducing the various samples into an analysis which can infer biological properties. This is often considered the crux of microarray analysis by many. It involves applying various linear and non-linear models to a variety of learning techniques including (but by no means limited to): support vector machines (SVM), neural network, and Empirical Bayes algorithms.

2.4 **Quality Control**

Quality control is concerned with accuracy and reproducibility. Recently Dr. Piatetsky-Shapiro gave a speech at the university colloquium on such a topic. He focused on using various data mining techniques to verify and validate the results of microarray analysis. Some areas he covered included cross validation, identification of significant results, and corrective behavior. His slides can be found at http://www.KDnuggets.com.

3 **Tutorial**

Below is the a brief tutorial which will walk you through a simple microarray analysis. Note that any items in *italics* are optional and will not effect the outcome of your final results. Some commands are self-explanatory (or will be when you try them). Others require some explanation, which will be provided as deemed necessary. It should be noted that I took much of the information for this tutorial from various web sites and references. If you require further explanation, I found that my best sources were Google and the Bioconductor web page.

1. **Install R** (http://www.r-project.org)

   R is an “S”-like GNU project language and environment for statistical computing. It was chosen for this tutorial because it contains many packages for both linear and non-linear statistical modeling. It also includes an effective data handling and storage facility, a suite of operators for calculations on matrices, an integrated collection of
intermediate tools for data analysis, graphical facilities for data analysis, and display functions for either on-screen or hard copy.

2. Install Bioconductor (http://bioconductor.org)

Bioconductor is an Open source package for statistical analysis of genomic data. It includes both statistical and graphical tools essential and specific to microarray analysis. Due to the activity in the project, there is a constant influx of new packages. Unfortunately, there are some advanced microarray analysis components that are not yet available for Bioconductor. One example of this is the Granular Support Vector Machine that was required for the reproduction of the experiment mentioned above. However, because of the open source nature of the project, it is entirely possible to take the current SVM capabilities found in some Bioconductor implementations and extend them to accomplish the desired task.

3. Download the Tutorial Dataset (ApoAI.RData)

You should either be able to obtain this file from Dr. Clement or searching the web with Google for it. The apolipoprotein AI (ApoAI) gene is known to play a pivotal role in high density lipoprotein (HDL) metabolism. This means that mice which have the ApoAI gene knocked out (KO) have very low HDL cholesterol levels. The purpose of this experiment will bet to determine how ApoAI deficiency affects the action of other genes in the liver. This will help us gain some understanding as to which molecular pathways ApoAI operates on, which in turn may be used to create drugs to help lower cholesterol.

Data from both the knockout mice and wild-type mice were pooled together and marked with a green tag. This pooled data was included on all chips. A red tag was used to indicate either a knockout or wild-type set of genes depending on which sample was being run. Every chip contained both a red and a green sample and is illustrated in Figure 2.


5. >library(limma)

This is done in order to import the linear model microarray library we will use for this tutorial.

6. >load(“ApoAI.RData”)

This command reads in the microarray data.

7. >ls()

8. >names(RG)
3. TUTORIAL

9. \ Someone\ targets
10. \ Someone\ 
11. \ dim(RG)\ 
12. \ ncol(RG)\ 
13. \ colnames(RG)\ 
14. \ MA\ <-\ normalizeWithinArrays(RG)\ 
   This function normalizes the samples as indicated above in the normalization section.
15. \ design\ <-\ cbind(“WT-Ref”=1, “KO-WT”=rep(0:1,c(8,8)))\ 
   This command creates the linear model we will use to run our experiment.
16. \ design\ 
17. \ fit\ <-\ lmFit(MA,design=design)\ 
18. \ colnames(fit)\ 
19. \ names(fit)\ 
20. \ fit\ <-\ eBayes(fit)\ 
   This function creates a best fit using an Empirical Bayes algorithm.
21. \ plotMA(fit)\ 
22. \ names(fit)\ 
23. \ summary(fit)\ 
24. \ topTable(fit,coef=“KO-WT”,adjust=“fdr”)\ 
   This is an important table since it shows you the results of the analysis. Important to note here are the p-values (indicate which results are significant) and the M/A values. Large M values indicate large differentials and hence differential gene expression between populations.
25. \ modelMatrix(RG/targets,ref=“Pool”)\ 
   These seven commands perform the same functions as above, but we first remove the control data points. While this does not alter the overall outcome of the experiment, it does make for cleaner (easier-to-read) results.
26. \ table(RG/genes TYPE)\ 
27. \ isGene<-RG/genes$TYPE==“cDNA”\ 
28. \ MA2<-MA[isGene,]\ 
29. \ fit<-lmFit(MA,design=design)\ 
30. \ fit<-eBayes(fit)\ 
31. \ plotMA(fit,2)\ 
   This command creates our M vs. A plot. M represents differential ratio between the samples as determined via the formula $M = (\log R - \log G)$. A represents the fluorescence intensity between samples as determined via the formula $A = (\log R + \log G)/2$.
   This plot is of particular importance for a variety of reasons. First, since M represents the differential ratio of gene expression between samples, those data points that are further from the M axis demonstrate a variation in expression level as defined by our linear model. Basically, those that are further from the axis are the genes we are
looking for. Secondly, this plot gives us an understanding of how good our data is with respect to intensity vs. differential.

An example of poor intensity to differential data transformation quality is given in Figure 3. Notice that at low fluorescence intensity, there is a wide variation of gene expression, while at high intensity, there is very little fluorescence variation. This indicates that either our transformation or optical data was poor. A desirable transformation would show uniform distribution of differential across intensities.

32. top10<-order(fit$lods[,“KO-WT”],decreasing=TRUE)[1:10]
   This and the next five commands will assign labels to the data points in the graph that correspond to the ten most differentially expressed genes in our experiment.

33. A<-fit$Amean
34. M<-fit$coef[,2]
35. shortlabels<-substring(fit$genes[,“NAME”],1,5)
36. text(A[top10],M[top10],labels=shortlabels[top10],cex=0.8,col=“blue”)

Figure 3. Example of a poor MAplot

4 Conclusion

By observing the resulting data, it is possible to derive a couple of conclusions. First, it is important to note that ApoAI is very distant from the M axis (large differential expression
between the two sample groups). This (as Dr. Snell put it) serves as a sanity check since it was ApoAI that we knocked out originally. Additionally, we note that ESTH is also expressed in a largely differential manner. After evaluating the p-values, we verify that this is a significant data point. If I were a researcher, this is the first gene I would start looking at in a possible signaling pathway.

5 Lab Description

Initially I had suggested that I reproduce the experiment presented by Tang in [9]. However, after investing many hours, I came to the realization that this was not a feasible project to accomplish within the time frame presented for this class. Some of the obstacles I encountered are described below. I suggested that I might instead apply some of what I had learned to creating a tutorial that other students could use to become somewhat oriented with microarrays. This proposal was approved by Dr. Clement with the stipulation that I make my tutorial available and expound somewhat on my experience. This section is dedicated to the later stipulation.

5.1 Hurdles

When I initially tried to recreate the experiment, I honestly had no idea what a microarray analysis entailed. It took me many hours to simply decipher what the article was getting at what it meant. Essentially, they are presenting an extension to Granular Support Vector Machine gene classification [9]. By applying this technique to a collection of prostate cancer microarray data, they were able to report perfect classification. While I do have my reservations about one-hundred percent classification results (often an indication of over fitting the data), I was intrigued by their methods and wanted to see what I might come up with. However, it soon became apparent that I would not be able to accomplish this easily.

After downloading the data set they used to run their experiment, I turned my focus on reproducing their experiment. Unfortunately, the links they provided for the MATLAB Ohio State University Support Vector Machine Classifier Matlab Toolbox they had used for the experiment was dead. I began looking around for other Support Vector Machine (SVM) packages to extend in the same way they had done [2, 3]. It soon became apparent, however, that without the package they had used, or an extensive knowledge of the problem (attacking it from another angle), I would not be able to reproduce the experiment. During the course of my searchings, I did come across many SVM packages that either did basic microarray analysis (Bioconductor), or were strict SVM implementations (that did not correlate with the variables they claimed they were setting). I became intrigued with the whole microarray analysis problem and began to check out books and look up articles on the subject. Overall, it was a beneficial experience for me. This area of bioinformatics greatly appeals to me as a science that will eventually penetrate all areas of biological and medical research. I am
fortunate to now have a baseline understanding of this analysis method, and hope you (the reader) were able to walk away with a similar impression.

Bibliography


