A Microarray Analysis Teaching Module

for
Hamilton College

July 2008

Megan Cole
Post-doctoral Associate
Whitehead Institute, MIT

Lecture Topics

I. Uses of microarrays – developed in 1987
    a. To measure gene expression
        i. Different expression patterns in different cell types
        ii. Different expression over cell cycle
        iii. Cells treated with drug vs untreated
        iv. Cells with genetic mutation vs wild type
        v. Cancer vs normal tissue
    b. SNP analysis
    c. To measure binding of proteins along DNA (ChIP-chip)
    d. Measure alternative splicing of mRNAs
    e. Tiling arrays – measure expression of entire genome – finding that much more of it is transcribed than we previously thought (non-coding RNAs)

II. Experimental procedure
    a. collect RNA or DNA sample
    b. label it using labeled nucleotides
    c. hybe it to array
    d. measure the amount of hybridization to each spot
    e. computationally normalize the data
    f. analyze the data for biological insights

III. Types of arrays
    a. Spotted arrays vs synthesized arrays
        i. Spotted arrays 'spot' premade cDNA or oligos onto glass slides
        ii. Synthesized arrays synthesize oligos directly onto array
            1. affymetrix arrays light activate certain spots and then wash slide with nucleotide
            2. agilent arrays light activate whole slide but use inkjet printer to print nucleotides on subset of spots
    b. 2-channel vs 1-channel
        i. 2-channel arrays hybridize 2 differently labeled samples to the same array so a direct comparison is done
        ii. 1-channel arrays hybridize 1 sample to an array and measure absolute expression levels (but these need to be compared to a reference sample that must be hybed to a separate array)

IV. Sources of noise
    a. Biological (differences between cell batches etc.)
    b. Dye biases
    c. Bad spots on array
    d. Different amounts of starting material
    e. Different labeling efficiencies
    f. Crosshybridization

Useful references

V. Microarray reviews
   a. Gary A. Churchill (2002) Fundamentals of experimental design for cDNA microarrays. Nature Genetics.
   b. John Quackenbush (2002) Microarray data normalization and transformation. Nature Genetics.
   c. Jin Hwan Do and Dong-Kug Choi (2006) Normalization of microarray data: single-labeled and dual-labeled arrays. Molecules Genes and Cells.

VI. Textbooks
   a. Mount, David W. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2001.
   b. Glantz, S. A. Primer of Biostatistics. 4th ed. New York, NY: McGraw-Hill, Health Professions Division, 1997. ISBN: 9780070242685.
   c. Beginning Perl for Bioinformatics by James Tisdall
   d. R Programming for Bioinformatics (Chapman & Hall/Crc Computer Science & Data Analysis) by Robert Gentleman
   e. An Introduction to Bioinformatics Algorithms (Computational Molecular Biology) by Neil C. Jones and Pavel A. Pevzner

VII. MIT computational biology course notes
   a. http://ocw.mit.edu/OcwWeb/Biology/7-91JSpring2004/CourseHome/index.htm

VIII. Software for more sophisticated microarray analyses
   a. www.bioconductor.org/
   b. www.broad.mit.edu/cancer/software/genepattern/
   c. http://www.affymetrix.com/products/software/index.affx
   d. http://www.chem.agilent.com/scripts/pcol.asp?lpage=7038

Background Information

The purpose of this lab is to work with real microarray data and to perform common analyses such as data normalization, identification of up and down regulated genes, clustering and identification of enriched gene ontologies. For this lab we will use data from Su et al. 2002 where microarray expression data was gathered from over 90 human and mouse samples from a diverse array of tissues. Although this lab will analyze expression data the techniques are applicable to other uses of microarrays – such as SNP analysis, ChIP-chip and protein arrays. If students are interested in exploring additional datasets they can go to www.ncbi.nlm.nih.gov/geo/.

Data:

Data used is from
Andrew I. Su, Michael P. Cooke, Keith A. Ching, Yaron Hakak, John R. Walker, Tim Wiltshire, Anthony P. Orth, Raquel G. Vega, Lisa M. Sapinoso, Aziz Moqrich, Ardem patapoutian, Garret H. Hampton, Peter G. Schultz and John B. Hogenesch. Large-scale analysis of the human and mouse transcriptomes PNAS 2002 99(7): 4465-70.

Normalization:

Microarray data must be normalized in order to account for many potential sources of noise including differences in labeling efficiency, different amounts of starting RNA material, and biological noise. There are may ways to normalize microarray data but in this lab we will use the straightforward technique of 'median normalization'. The basic assumption behind this method is that the vast majority of genes on the array will not change expression. For cases where this assumption is not true scientists often perform 'control normalization', which is basically median normalization but using a subset of genes that are thought to not change expression between samples.

1. Download data from Su et al. by going to http://expression.gnf.org/data_public_U95.gz and open the unzipped file in excel. The first column is the affymetrix ID associated with that spot/gene and the first row describes the samples. Each column represents the data from a single microarray.
2. This is the primary/raw data that has only been run through Affymetrix software to combine probes into probesets (using perfect match and mismatch probe pairs). Negative values mean that there was more cross-hybridization (hybridization to the mismatch probe) than real hybridization (hybridization to the perfect match probe) and reflect the fact that there is very little (or no) mRNA for that gene. Since negative numbers create difficulties we should 'floor' all small numbers to 20.
    a. Open a new workbook by clicking file/new workbook.
    b. Now copy the labels to the new workbook. In the old worksheet highlight the entire first column by clicking the 'A' at the top of the column and then copy (apple-c). Now go to the new workbook, highlight column A and paste (apple-v).
    c. Now copy the sample labels to the new worksheet. In the old worksheet highlight the entire first row by clicking the '1' to the left of the row and then copy (apple-c). Now go to the new workbook, highlight row 1 and paste (apple-v).

   d. Now copy the 'floored' data values to the new worksheet. In cell B2 type '=if(' then select the cell in the original worksheet in B2 then type '<20,20,' and again click the cell in the original worksheet in B2 and then type ')' and enter. The formula should look something like this: '=IF(data_032502.idx.sort.rep.txt!B2<20,20,data_032502.idx.sort.rep.txt!B2)' and basically says if the value in B2 is less than 20 use 20 otherwise use the value in B2.
   e. Now let's copy the rest of the data. Highlight the cell you just copied in Sheet1 and copy it (apple-c). Now highlight the rest of the cells that you need to copy data into (from B2 to CX12601) and paste (apple-v).
   f. While all of the cells are still highlighted copy them (apple-c) and then click edit/paste special/values/ok. Then save this workbook as FlooredData.
3. Now let's median normalize the data.
   a. At the bottom of column B in the FlooredData workbook find the median value for that column by typing '=median(B2:B12601)'. Now find the median for the rest of the columns by copying this formula to the other columns. Note: we use median here but mean could also be used, or a combination of the two called 'trimmed mean'. What are the pros and cons of median, mean and trimmed mean normalization?
   b. Now let's normalize each array to a median of 80. Calculate the normalization factor by dividing 80 by each median (in B12603 type '=80/B12602 and then copy this formula for the other columns.
   c. Now let's open a new workbook where we will put the mean-normalized data. Copy the data labels into the new workbook.
   d. In cell B2 of the new workbook type '=' and then click on B2 in the FlooredData sheet and then type '*' and click on the normalization factor in B12603. Edit this formula in the new workbook by inserting a '$' before the 12603 (the dollar sign means that when you copy this formula 12603 will always be 12603). Then copy this formula to the other data cells to get the median-normalized data.
   e. While all of the cells are still highlighted copy them (apple-c) and then click edit/paste special/values/ok. Then save this workbook as MedianNormalizedData.
4. Now let's take a look at some of the data. Highlight columns B and C in the MedianNormalizedData and graph them as a scatterplot. These should be replicates so they should graph along the 1:1 line. Do they? Are there any outliers?
5. Now change the axes on the graph so that they are log scaled – does this make it easier to see all the data? Why is the bottom left part of the graph noisier than the top right?
6. Now let's plot two different samples in the same way. Try plotting columns C and D – samples from fetal brain vs adult cerebellum. How does this look different and why?

Identification of Up and Down Regulated Genes:

Scientists are often interested in identifying the set of genes whose expression is changed in their sample. In order to identify genes with expression changes we must

compare expression levels for the sample to a reference expression level. For this dataset we will use the median expression level of a gene across all of the samples for the reference. This will allow us to identify which genes are up or down regulated in a particular sample compared to its typical expression level.

1. Now we want to see which genes are high/low in each sample. Let's start by calculating the median expression value of each gene.
    a. Label column CY in the MedianNormalizedData worksheet 'Median Expression'. Now in CY2 type '=median(B2:CX2)' then copy this formula to the rest of the rows. Then copy (apple-c) the entire column and click edit/paste special/values/ok. Save your data.
    b. Open a new worksheet and copy the data labels into it.
    c. Now let's find the expression ratios for each sample. In cell B2 of the new worksheet type '=log(' and then click on B2 of the MedianNormalizedData then type '/' and click on CY2 then finish the formula with ',2)'. This gives us the log base 2 expression ratio for this gene and sample relative to its median expression for all the samples. Now edit the formula to add a '$' in front of CY and copy the formula to the rest of the data cells.
    d. While all of the cells are still highlighted copy them (apple-c) and then click edit/paste special/values/ok. Then save this workbook as RatioData.
    e. What relative expression (ratio) would a gene that has twice as much mRNA in fetal brain (compared to its average expression) have? What about a gene that has half as much mRNA? What would these be if we hadn't taken the log ratio?
7. Now let's visualize the data using an MA-plot. We need to plot the log(ratio) vs the average log(intensity). We already have the log(ratios) calculated for all the samples but will need to calculate the average log(intensity) for a sample in order to make a plot.
    a. In a new workbook copy the log(ratios) for your favorite sample into column A.
    b. In column B you'll need to use the formula '=0.5*log(MedianNormalizedData for you favorite sample * MedianNormalizedData for the median value for that gene,2).
    c. Now make a scatterplot of this data. Assuming that most genes have no change in expression where should most spots plot? For spots with expression changes where do they plot?

2-Color Arrays:

This exercise uses data from affymetrix (1-color) arrays but it is applicable to 2-color array analysis as well. When analyzing 2-color arrays you simply normalize each color as if it were a separate array. To calculate ratios you simply find the ratio of 1-channel to the other channel, making sure to use channels 1 and 2 from the same array. 2-channel array data is harder to compare between different arrays unless the exact same reference sample was used for each array.

Clustering:

When analyzing microarray data it can be useful to identify which samples and/or genes behave similarly. Examining patterns or 'clusters' within the data can be a more powerful technique than individual examination of genes or samples. Samples can be clustered to determine which tissues/cell types are similar in transcriptome while genes can be clustered to identify sets of genes that behave in a certain way across different samples.

Hierarchical Clustering

Hierarchical clustering orders samples or genes into a hierarchical tree. To cluster samples the Euclidean distance between samples can be calculated in n-dimensional space where each dimension represents the expression level of a gene. At each step in this algorithm the two clusters with the shortest distance between them are joined.

K-means Clustering

K-means clustering will cluster samples or genes into 'k' number of clusters/sets. The user must specify the number, k, of clusters desired. The clustering of genes is done by first randomly assigning k genes to be in different clusters and then assigning all other genes to their closest cluster. The center of each cluster is then calculated and all genes are reassigned to their closest cluster. This step is repeated until no genes change their assigned cluster. As this algorithm begins with a random assignment of genes to clusters it is stochastic and will not always produce the same result. For this reason it is often repeated many times to ensure that a 'good' result has been found.

1. For the RatioData worksheet you need to save-as a tab delimited txt file so that the clustering software can read it.
2. Open Cluster 3.0
3. Click file/open and select the tab delimited file.
4. Let's first do hierarchical clustering for the samples.
    d. Click on the hierarchical tab and check the cluster box for arrays.
    e. Change the pull-down menu to Euclidean distance
    f. Click average linkage to perform the clustering.
    g. To view your clustered data open the Java Treeview program. Then click file/open and open the clustered file (it will have the same name as the tab delimited file but have a '.cdt' extension.
    h. To view all of the data at once click settings/pixel settings and click 'fill' for global X and Y. Then click close.
    i. To more easily see the green (downregulated) and red (upregulated) genes click settings/pixel settings and change the contrast value to 2.0.
    j. From the picture (called a heatmap) you can see which samples were clustered together first – what samples are clustered together first and what does this say about them?
5. Now let's cluster the genes using K-means clustering.
    k. Go back to the cluster program and click on the K-means tab.
    l. Click organize genes and then choose the number of clusters/sets you want.
    m. When it's done visualize the data using Java Treeview again. You can change the pixel settings as before.
    n. Try clustering again but change the number of clusters and see what happens. How would you decide how many clusters is best?

o. If you want to see which genes are in each cluster then open the '.atr' file in excel.

Identification of Enriched Gene Ontologies:

Gene ontologies describe categories of genes, for example genes involved in neurogenesis or genes that regulate transcription. When scientists deal with large sets of genes it is often useful to examine whether any particular gene ontologies are over-represented in the data. For example, if one examines the set of genes up-regulated in neurons we might expect to see many neurotransmitter genes.

1. Now let's see what types of genes are up regulated in various tissues.
    a. Go back to the RatioData excel workbook.
    b. Pick your favorite sample and sort the array based on that sample.
    c. How many genes are more than 2-fold up-regulated for this sample?
    d. Now go to the David Gene Ontology site at http://david.abcc.ncifcrf.gov/
    e. Click on 'start analysis'.
    f. Now click 'upload'.
    g. Paste in the list of affymetrix Ids that are >2-fold upregulated in your tissue, check the 'genelist' box and then click 'submit list'.
    h. If it asks you to choose a species select homo sapiens.
    i. Now go back to upload and paste in all the affymetrix Ids from the array. This time check the 'background' box and then submit. Again choose homo sapiens.
    j. Now choose functional annotation chart and then click on functional annotation chart.
    k. Which ontologies are enriched in your dataset?
    l. To see which genes from you dataset are in a particul ontology click on the blue bar for that ontology.
    m. If you want to learn about a particular gene click on its ID.
    n. Try this for a different tissue sample – do you get a different set of enriched ontologies? What do the ontologies tell you about the samples?
2. If you want to see the gene names for the affymetrix Ids upregulated in your samples use the 'gene ID conversion tool' and convert to 'gene symbol'.