Microarray Data Analysis

Clustering

Analysis of data in many dimensions

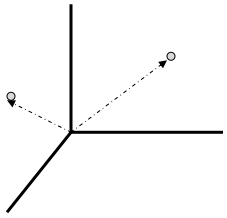
- Example: Time Course (could also use set of mutants, or set of cellular conditions)
- [Background covered previously in class: Using Scanalyze, Cluster, Treeview, and Excel, went through an example of clustering one microarray experiment (one plate)]
- Experimental Procedure Review:
  - **"genomic" "cDNA" spotted onto plate**
  - **obtain labelled, expressed "cDNA" from time 0,1,2…**
  - **wash 1 spotted plate simultaneously with "cDNA" from time 0 and 1; another plate with "cDNA" from time 0 and 2; etc. (what sample will be green? what sample will be red?)**

- Open Table 4.2 (Campbell) from Donna's homepage
- note the data are in two dimensions; only ratio data is reported for each time point (ratio of what to what?)
- you could put a color (green, black, or red) into each cell on the table to color code the expression level of each gene relative to reference for each time point; we're going to do this, but first we're going to log transform the data
- $\log_2$ transform the data in each column:
  - **select cell H2**
  - **type =log(**
  - **then select cell B2**
  - **then type ,2)**
  - **the entry in the fx window will look something like =log(B2,2) ; B2 is the cell you selected, and 2 is the base for the logarithm**
  - **hit enter**
  - **drag down column H to fill in the values**
  - **now do the same for the other columns; put headers on your log transformed columns**

- Now color each log transformed number according to the following color scheme:
- For all values from 3x to > 20 x induction, color the log transformed numbers red (what value of $\log_2$ does 3x induction correspond to?)
- For all values from 3x to > 20 x repression, color the log transformed numbers green (what value of $\log_2$ does 3x repression correspond to?)
- For all values in between 3x induction and 3x repression, color the log transformed numbers black
- make a prediction about which genes have similar expression patterns

Pearson Correlation Coefficient

- Let's say you have two points, and you want to know the distance between them
- (x_1, y_1, z_1), (x_2, y_2, z_2)



- This is just finding the third side of the triangle:
- $\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2 + (z_2-z_1)^2}$

- You can extend this idea of distance to n dimensions
- $d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- You can normalize this idea of distance by using the average and standard deviation, so that the actual distance metric is
- $[1/(\mathbf{E-1})] \sum_{e=1}^{E}(x_{ie}-x_{iav}/s_i)(x_{je}-x_{jav}/s_j)$

where x_{ie} is the normalized expression level for gene i in experiment e, s_i is the standard deviation of the expression levels for gene i across the experiments, and E is the number of experiments

- For each pair of genes, you can calculate the correlation coefficient
  - first find the average value for each gene across all the experiments x_{ie}
    - **select cell N2**
    - **type =average(**
    - **then select cell H2, drag through cell M2**
    - **then type) and enter**
    - **now select cell H2 and drag down column**
  - then find standard deviation for each gene across all the experiments s_i
    - **similar procedure to average, but function is stdeva**
  - then calculate the individual normalized value
    - **e.g., for the gene C at time 0, the normalized value is (H2-N2)/O2**
    - **you can drag down the columns again after you do the first cell**

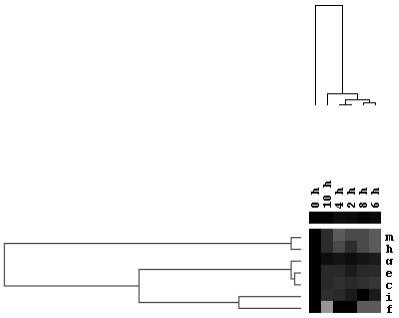
- Now calculate the correlation coefficients
  - first make a correlation table with gene letters for row and column identifiers
  - select the c-e cell in the correlation table; in that cell, type =(sumproduct(
  - then select just the normalized values for c that we just calculated
  - then type ,
  - then select the normalized values for e that we just calculated
  - then type))/5
  - it will look something like this: =(SUMPRODUCT(P2:U2,P3:U3))/5
  - note there are two parentheses before the division by 5
  - this takes each member of the c array and multiplies it with the corresponding member of the e array, and sums all the products; this corresponds to $[1/(\mathbf{E-1})] \Sigma (x_{ie}-x_{iav}/s_i)(x_{je}-x_{jav}/s_j)$
  - do this same operation for all the other cells
- The correlation coefficient closest to 1 is the set of experiments that have the most similar pattern

- Group the closest two, then calculate the correlation coefficients again in a new table
  - the correlation coefficient of a pair to another value is the average of the individual correlation coefficients
- Normally, you would do this until there are no more to group
- We're going to do this for the first 4 genes in the log-transformed table you generated
- From this, generate a distance tree for these 4 genes

Follow-up to this exercise

- Cluster/TreeView used to analyze sample data
- Homework assignment: posted at Donna's homepage
  - homework asks for a complete clustering of the data used in the original table from Campbell (14 genes) using the Excel methods illustrated, including moving color-coded rows of the spreadsheet around according to how they group
  - homework also asks for getting the Cluster/Treeview programs to cluster the same data
  - and to compare their spreadsheet results to the Cluster/Treeview results

Cluster/TreeView vs. Excel results – modified Campbell table



	0 hr	2 hr	4hr	6hr	8hr	10 hr
c	0	3	3.585	4	3.585	3
e	0	2	3	3	3	3
g	0	1	1.585	2	1.585	1
i	0	2	3	2	0	-1
f	0	0	0	-2	-2	-3.32193
m	0	-1.6	-2	-2	-1.6	-1
h	0	-1	-1.6	-2	-1.6	-1

wheaton June 2003, copyright Susan M. E. Smith 11