Differentially Expressed Genes
Oligo arrays: Negative values

- In many cases oligo array can return values that are less than 0 (Why?)
- There are a number of ways to handle these values
- The most common is to threshold at a certain positive value
- A more sophisticated way is to use the negative values to learn something about the variance of the specific gene
What you should know

• The different noise factors that influence microarray results
• The assumptions and corresponding methods for normalizing oligo (single channel) arrays.
• Methods for normalizing two channel (cDNA) arrays.
Data analysis

- Normalization
- Combining results from replicates
- Identifying differentially expressed genes
- Dealing with missing values
- Static vs. time series
Motivation

• In many cases, this is the goal of the experiment.

• Such genes can be key to understanding what goes wrong / or get fixed under certain condition (cancer, stress etc.).

• In other cases, these genes can be used as ‘features’ for a classifier.

• These genes can also serve as a starting point for a model for the system being studied (e.g. cell cycle, phermone response etc.).
Problems

• As mentioned in the previous lecture, differences in expression values can result from many different noise sources.

• Our goal is to identify the ‘real’ differences, that is, differences that can be explained by the various errors introduced during the experimental phase.

• Need to understand both the experimental protocol and take into account the underlying biology / chemistry
The ‘wrong’ way

• During the early days (though some continue to do this today) the common method was to select genes based on their fold change between experiments.

• The common value was 2 (or absolute log of 1).

• Obviously this method is not perfect …
Significance bands for Affy arrays
Value dependent variance

Value Specific Variance

Variance

Log fold change
Typical experiment: replicates

Technical replicates: same sample using multiple arrays
Dye swap: reverse the color code between arrays
Clinical replicates: samples from different individuals

Many experiments have all three kinds of replicates
Hypothesis testing

• A general way of identifying differentially expressed genes is by testing two hypothesis
• Let $g_A$ denote the mean expression of gene $g$ under condition $A$ (say healthy) and $g_B$ be the mean expression under condition $B$ (cancer).
• In this case we can test the following hypotheses:

\[
\begin{align*}
\mathbf{H}_0 \text{ (or the null hypothesis)}: & \quad g_A = g_B \\
\mathbf{H}_1 \text{ (or the alternative hypothesis)}: & \quad g_A \neq g_B
\end{align*}
\]

• If we reject $\mathbf{H}_0$ then gene $g$ has a different mean under the two conditions, and so is \textit{differentially expressed}
P-value

• Using hypothesis testing we need determine our confidence in the resulting decision
• This is done using a test statistics which indicates how strongly the data we observe supports our decision
• A p-value (or probability value) measures how likely it is to see the data we observed under the null hypothesis
• Small p-values indicate that it is very unlikely that the data was generated according to the null hypothesis
Example: Measurements for one gene in 40 (20+20) experiments of two conditions
Hypothesis testing: Log likelihood ratio test

- If we have a probabilistic model for gene expression we can compute the likelihood of the data given the model.
- In our case, let's assume that gene expression is normally distributed with different mean under the different conditions and the same variance.
- Thus for the alternative hypothesis we have:
  \[ y_A \sim N(\mu_A, \sigma^2) \quad y_B \sim N(\mu_B, \sigma^2) \]
  
  and for the null hypothesis we have:

- We can compute the estimated means and variance from the data (and thus we will be using the **sample mean** and **sample variance**).
Example mean

Blue mean: -0.81
Red mean: 0.84
Combined mean: 0.02
Data likelihood

• Given our model, the likelihood of the data under the two hypothesis is:

\[ L(0) = \prod_{i \in A} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y^i - \mu)^2}{2\sigma^2}} \prod_{i \in B} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y^i - \mu)^2}{2\sigma^2}} \]

\[ L(1) = \prod_{i \in A} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y^i - \mu_A)^2}{2\sigma^2}} \prod_{i \in B} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y^i - \mu_B)^2}{2\sigma^2}} \]

• We can also compute the ratio of the likelihoods (L(1)/L(0))

• Intuitively, the higher this ratio the more likely it is that the data was indeed generated according to the alternative hypothesis (and thus the genes are differentially expressed).
Log likelihood ratio test

- We use the log of the likelihood ratio, and after simplifying arrive at:

\[
T = 2 \sum_{i \in A} (y^i - \mu_A)^2 + \sum_{i \in B} (y^i - \mu_B)^2 \\
- \sum_{i \in A} (y^i - \mu)^2 + \sum_{i \in B} (y^i - \mu)^2
\]
Log likelihood ratio test

- We use the *log of the likelihood ratio*, and after simplifying arrive it:

\[
T = 2 \sum_{i \in A} (y_i - \mu_A)^2 + \sum_{i \in B} (y_i - \mu_B)^2 \\
+ \sum_{i \in A} (y_i - \mu)^2 + \sum_{i \in B} (y_i - \mu)^2
\]

- \(T\) is our test statistics, and in this case can be shown to be distributed as \(\chi^2\)
Log likelihood ratio test

- We use the *log of the likelihood ratio*, and after simplifying arrive it:

\[
T = 2 \sum_{i \in A} (y^i - \mu_A)^2 + \sum_{i \in B} (y^i - \mu_B)^2 \\
= 2 \left( \sum_{i \in A} (y^i - \mu)^2 + \sum_{i \in B} (y^i - \mu)^2 \right)
\]

Log likelihood ratio is an important test statistics which is very commonly used in hypothesis testing.
Degrees of freedom

• We are almost done …
• We still need to determine one more value in order to use the test
• Degrees of freedom for likelihood ratio tests depends on the difference in the number of free parameters
• In this case, our free parameters are the mean and variance
• Thus the difference is …

• In this case, the difference is 1 (two means vs. one)
Example: Log likelihood ratio

\[ T = 2 \times \frac{64.3}{37.1} \]
\[ = 3.46 \]

D.O.F = 1

P-value = 0.06
Limitations

- We assumed a specific probabilistic model (Gaussian noise) which may not actually capture the true noise factors
- We may need many replicates to derive significant results
- Multiple hypothesis testing
Multiple hypothesis testing

• A p-value is meaningful when one test is carried out
• However, when thousands of tests are being carried out, it is hard to determine the real significance of the results based on the p-value alone.
• Consider the following two cases:

<table>
<thead>
<tr>
<th>we test 100 genes</th>
<th>we test 1000 genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>we find 10 to be differentially expressed with a p-value &lt; .01</td>
<td>we find 10 to be differentially expressed with a p-value &lt; .01</td>
</tr>
</tbody>
</table>

• We need to correct for the multiple tests we are carrying out!
Bonferroni Correction

• Bonferroni Correction is a simple and widely used method to correct for multiple hypothesis testing
• Using this approach, the significance value obtained is divided by the number of tests carried out.
• For example, if we are testing 1000 genes and are interested in a (gene specific) \( p \)-value of 0.05 we will only select genes with a \( p \)-value of \( \frac{0.05}{1000} = 0.00005 = 5 \times 10^{-5} \)
• Motivation: If
  
\[
p(specific \ T_i \ passes | H_0) < \frac{\alpha}{n}
\]
• Then
  
\[
p(some \ T_i \ passes | H_0) < \alpha
\]
Bonferroni Correction

- The Bonferroni Correction is very conservative
- Using it may lead to missing important genes
- An alternative is to use randomization methods
- Other methods rely on the false discovery rate (FDR) as we discuss for SAM
SAM – Significance Analysis of Microarray

• Relies on repeats.
• Avoid using fold change alone.
• Use permutations to determine the false discovery rate.
Data

- Many genes were assigned negative values
- Many were expressed at low levels
- Noise is larger for genes expressed at low levels.
Relative difference

\[ d(i) = \frac{\hat{x}_1(i) - \hat{x}_2(i)}{s(i) + s_0} \]

- Where \( x_1 \) and \( x_2 \) are the observed means and \( s(i) \) is the observed standard deviation.
- \( S_0 \) is chosen so that \( d(i) \) is consistent across the different expression levels.
Different comparisons of repeated experiments.
Identifying differentially expressed genes

• Using the normalized $d(i)$ we can detect differentially expressed genes by selecting a cutoff above (or below for negative values) which we will declare this gene to be differentially expressed.

• However selecting the cutoff is still a hard problem.

• Solution: use the False Discovery Rate (FDR) to choose the best cutoff.
False Discovery Rate

- Percentage of genes wrongly identifies / total gene identified.
- What is the difference between this and a p-value?

P-value: probability under the null hypothesis for observing this value
False Discovery Rate

- Percentage of genes wrongly identifies / total gene identified.
- What is the difference between this and a p-value?

P-value: probability under the null hypothesis for observing this value

- Given a p-value, the total number of genes tested and the number of genes identified as differentially expression, the FDR can be computed directly.
Determining the FDR

- A permutation based method.
- Use all 36 permutations (why 36?).
- For each one compute the $d_p(i)$ for all genes.
- Scatter plot observed $d(i)$ vs. expected $d(i)$. 
Selecting differentially expressed genes
Extensions

• Can be extended to multiple labels.

• Compute average for each label.

• Compute difference between specific class average and global average and corresponding variance.

• As before, adjust variance to correct for low / high level of expression.
Mixture populations

- We may be measuring the transcript levels in a heterogeneous (mixture) cell population
- There are a few surprises:
  - genes co-expressed (correlated) in each cell type may appear uncorrelated in the mixture
  - genes uncorrelated in each cell type may appear perfectly correlated in the mixture
Mixture of cancer and normal cells

Cancer cells

Healthy cells

Exp 1  Exp 2

Exp 1  Exp 2
Example

$$y_A \sim N(\mu, \sigma^2) \quad y_B \sim N(\mu, \sigma^2)$$
Example
What about time series?

- Comparing time points is not always possible (different sampling rates).
- Even if sampling rates are the same, there are differences in the *timing* of the system under different conditions.
- Another problem is lack of repeats.
Time series comparison

knockout = deletion of gene(s) from the sequence

Zhu et al, Nature 2000
Results for the Fkh1/2 Knockout
Clustering expression data
Goal

- Data organization (for further study)
- Functional assignment
- Determine different patterns

- Classification
- Relations between experimental conditions

- Subsets of genes related to subset of experiments

Genes
Experiments
Both
Example: co-expression

For example: grouping together genes active in the same phase of the cell cycle
Clustering experiments

- For example: clustering genes on the basis of how similar their effects are if they are knocked out.
- The "profiles" associated with the genes in this case are the knock-out responses.
Bi-clustering

- Find subsets of genes and experiments such that the genes in the subset behave similarly across the subset of the experiments
What you should know

• Statistical hypothesis testing
• Log likelihood ratio test
• Why SAM is successful:
  - No need to model expression distribution
  - Handles Excel data well