10-810 /02-710
Computational Genomics

Microarrays
Why sequence is not enough

Identifying genes and control regions is not enough to decipher the inner workings of the cell:

• We need to determine the function of genes.
• We would like to determine which genes are activated in which cells and under which conditions.
• We would like to know the relationships between genes (protein-DNA, protein-protein interactions etc.).
• We would like to model the various dynamic systems in the cell
Microarrays for molecular biology

Transcription factors

DNA

transcription

mRNA

translation

Proteins

[Diagram showing the process of gene expression from DNA to proteins through transcription and translation]
FDA Approves Gene-Based Breast Cancer Test*

“MammaPrint is a DNA microarray-based test that measures the activity of 70 genes... The test measures each of these genes in a sample of a woman's breast-cancer tumor and then uses a specific formula to determine whether the patient is deemed low risk or high risk for the spread of the cancer to another site.”

*Washington Post, 2/06/2007
What is gene expression?

Expression = level of gene (protein) in this experiment

genes

Experiments (over time)

baseline expression

Higher expression compared to baseline

Lower expression compared to baseline

Genes and Gene Expression
Technology
Display of Expression Information
What is a gene?

Promoter  Protein coding sequence  Terminator

Genomic DNA
How are Genes Regulated?
DNA-binding Activators Are Key To Specific Gene Expression
How are Genes Regulated?
DNA-binding activators are key, but there are additional factors
Genome-wide Gene Expression (mRNA) can be Measured with DNA Microarrays
Yeast Transcriptome (Glucose)

5460 mRNA species
average level: 2.8 copies/cell
median level: 0.8 copies/cell

80% of the transcriptome is expressed at 0.1 - 2 mRNA copies/cell
Genes and Gene Expression
Technology
Display of Expression Information
Microarray Hybridization

- Watson-Crick base pairing of complementary DNA sequences.

- Microarrays have thousands of spots, each representing a piece of one gene, immobilized on a glass slide.

- The intensity (or intensity ratio) of each spot indicates the amount of labeled cDNA hybridized, thus, representing the starting mRNA transcript abundance.
Two major technologies

- **cDNA arrays**
  - probes are placed on the slides
  - allows comparison of different cell types

- **Oligonucleotide arrays**
  - partial sequences are printed on the array
  - measure values in one tissue type
Hybridization and Scanning— cDNA arrays

1. Prepare Cy3, Cy5-labeled ss cDNA
2. Hybridize 600 ng of labeled ss cDNA to glass slide array
3. - Hybridize 600 ng of labeled ss cDNA to glass slide array
4. - Scan
5. - Scan
6. - Scan

Copyright © 1998-9 by Jeremy Buhler
Cartesian PixSys 5500 with quill printing technology

- Complete subsequences are printed on the array
- 10,000 spots/slide
- Spots are 100-200 µm in diameter
- Hybridization volumes: 20-100ul
Array Scanning

Laser based - fluorescent emission
Hybridization and Scanning—oligo arrays
cDNA vs. Oligo: **Pros and Cons**

**cDNA**
- Does not require sequence
- Cheap
- Direct comparisons
- Inaccurate
- Cannot measure individual samples

**Oligo**
- Can be designed to minimize cross hybridization
- Allows for internal control
- Both lead to better accuracy
- Expensive
- Limited to certain species
Errors

Microarrays introduce many errors which should be taken into account when working with measured expression values:
- Scanning errors
- Spotting errors
- Cross hybridization
- Errors related to day / reading device / experimentalist
- Background differences between slides
Error types

Microarrays introduce many types of errors which should be taken into account when working with measured expression values:

• Scanning errors \textit{additive} + \textit{multiplicative}
• Spotting errors \textit{multiplicative}
• Cross hybridization \textit{multiplicative}
• Errors related to day / reading device / experimentalist \textit{additive} + \textit{multiplicative}
• Background differences between slides \textit{additive}
Handling the Different Errors

• Scanning errors
• Spotting errors
• Cross hybridization
• Errors related to day / reading device / experimentalist
• Background differences between slides

Analysis of image data (we assume it was performed)
Handling the Different Errors

- Scanning errors
- Spotting errors
- Cross hybridization
- Errors related to day / reading device / experimentalist
- Background differences between slides

Use ratio instead of individual values:

\[ Y_i = \frac{R_i}{G_i} \]
Handling the Different Errors

• Scanning errors
• Spotting errors
• Cross hybridization
• Errors related to day / reading device / experimentalist
• Background differences between slides

For Oligo arrays, use the match / mismatch spots
• Presence and absent calls can be made using the Match / Mismatch information.

• However, it has been reported that in some cases the mismatch was higher than the match.
Handling the Different Errors

- Scanning errors
- Spotting errors
- Cross hybridization
- Errors related to day / reading device / experimentalist
- Background differences between slides

Normalization (later)
Binding arrays

• Instead of printing the genes on the microarray, we can print the intergenic region (an area upstream of the gene).

• We tag a protein of interest (a transcription factor) and fuse all proteins to DNA.

• Next, we hybridize the extracted portions of DNA onto the array, resulting in areas that are bound by the TF being spotted on the microarray.
Genes and Gene Expression Technology
Display of Expression Information
Yeast cell cycle expression program

Experiments (over time)

Baseline expression

Gene 1

Higher expression compared to baseline

Lower expression compared to baseline

Genomic Reprogramming in Response to Oxidant

6218 genes
Visualization:
Relative vs. absolute expression
Exercising the Genome
Using annotation databases

• Statistical tests to identify the overlap with various functional categories
Genome wide binding

experiments (transcription factors)

genes

probably bound by this TF

no binding by this TF

Lee et al Science 2002
What you should know

• The basic idea behind microarray profiling
• The two different microarray technologies
• Pros and cons for each
• Noise factors in microarray experiments (more next time)
Gene expression analysis
Experiment design

A number of computational issues should be addressed:

• Selecting short subsequences for oligo arrays to minimize cross hybridizations
• Determining the number of replicates for each sample
• Sampling rates for time series experiments
Data analysis

• Normalization
• Combining results from replicates
• Identifying differentially expressed genes
• Dealing with missing values
• Static vs. time series
Data analysis

• Normalization
• Combining results from replicates
• Identifying differentially expressed genes
• Dealing with missing values
• Static vs. time series
Typical experiment: replicates

healthy

cancer

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
Normalizing across arrays

- Consider the following two sets of values:
Lets put them together ...