Microarray analysis

CS426, week 12
Golan Yona

What makes one cell different from another?
- liver vs. brain
- Cancerous vs. non-cancerous
- Treatment vs. control

There are about 100,000 genes in mammalian genome
- each cell expresses only ~15,000 of these genes
- genes can be expressed at a different level

Gene expression can be measured by #copies of mRNA/cell
- 1-5 copies/cell - “rare” (~30% of all genes)
- 10-200 copies/cell - “moderate”
- 200 copies/cell and up - “abundant”

What makes one cell different from another?
Which genes are expressed
How much of each gene is expressed

Traditional biology:
Try and find genes that are differentially expressed
Study the function of these genes
Find which genes interact with your favorite gene

Extremely time consuming!
Impractical for gene mining!
Microarrays
Massively parallel analysis of gene expression
- screen an entire genome at once
- find not only individual genes that differ, but groups of genes that differ.
- find relative expression level differences
Shifting the interest from analysis of single molecules to large complexes and networks
Effective for
- Functional analysis
- Identify regulatory networks and cellular procedures
- Tune medical diagnosis and treatment

The technology
Measure interactions between mRNA-derived target molecules and genome-derived probes.
Based on old technique
Many flavors - majority are of two essential varieties
1. cDNA Arrays (discussed today)
   printing on glass slides, miniaturization, throughput fluorescence based detection
2. Affymetrix Arrays
   in situ synthesis of oligonucleotides.
Other arrays – protein arrays, combinatorial chemistry

The process
BUILDING THE CHIP
MASSIVE PCR
PCR PURIFICATION and PREPARATION
PREPARING SLIDES
PRINTING
POST PROCESSING
PREPARING RNA
CELL CULTURE
And HARVEST
RNA ISOLATION
cDNA PRODUCTION
HYBING THE CHIP
ARRAY HYBRIDIZATION
PROBE LABELING
DATA ANALYSIS
BUILDING THE CHIP
MASSIVE PCR
PCR PURIFICATION and PREPARATION
IPA precipitation + EtOH washes + 384-well format
PRINTING
The arrayer: high precision spotting device capable of printing 10,000 products in 14 hrs, with a plate change every 25 mins
PREPARING SLIDES
Polylysine coating for adhering PCR products to glass slides
POST PROCESSING
Chemically converting the positive polylysine surface to prevent non-specific hybridization
**Practical problems**

- Surface chemistry: uneven surface may lead to high background.
- Dipping the pin into large volume -> pre-printing to drain off excess sample.
- Spot variation can be due to mechanical difference between pins. Pins could be clogged during the printing process.
- Spot size and density depends on surface and solution properties.
- Pins need good washing between samples to prevent sample carryover.

**Direct labeling of RNA**

1. Two RNA samples (from two different tissues) are labelled with Cy3 or Cy5 dyes via a chemical coupling to AA-dUTP.
2. Cy3 and Cy5 RNA samples are simultaneously hybridized to chip.
3. Ratio measurements are determined via quantification of emission values. Data analysis starts...

**HYBING THE CHIP**

- ARRAY HYBRIDIZATION
- PROBE LABELING

**HYBRIDIZATION**

- TACAGAA
- ATGTC TT
- CCAACCTATGG
- CCAACCTATGG

**Direct labeling of RNA**

- TACAGAA
- ATGTC TT
- RNA
- cDNA

**Probe Labeling**

- CCAACCTATGG
- cDNA synthesis
- GGTGGGATACC

**Dye conjugated nucleotide**

- Cy5-dUTP
- Cy3-dUTP
Practical problems

- Definition of what a real signal is: what is a spot, and how to determine what should be included in the analysis?
- How to determine background local (surrounding spot) vs. global (across slide)
- How to correct for dye effect
- How to correct for spatial effect e.g. print-tip, others
- How to correct for differences between slides e.g. scale normalization

Pre-processing issues

- Definition of what a real signal is: what is a spot, and how to determine what should be included in the analysis?
- How to determine background local (surrounding spot) vs. global (across slide)
- How to correct for dye effect
- How to correct for spatial effect e.g. print-tip, others
- How to correct for differences between slides e.g. scale normalization
Data representation

Each spot on the array corresponds to one gene.

If $E_1$ is the expression of the gene in experiment 1 and $E_2$ is the expression of the gene in experiment 2 then the spot is converted to a number representing the ratio $E_1/E_2$ or $\log(E_1/E_2)$

Data analysis

1) Identify individual genes that are over or under expressed

2) Identify groups of genes that are co-expressed

3) Reconstruct the gene networks that underlie the observed expression levels

1) Analysis of individual genes

Absolute values are misleading

Need to establish the baseline in order to derive a measure of statistical significance for individual genes

Define distributions over the whole array or a control group. Use mean/variance to determine the significance – normalization, t-tests...statisticians like this stuff
2) Analysis of co-expression

Search for similarly expressed genes

Gene i: \((x_1, x_2, x_3, \ldots)\)

This is an expression profile of a gene

\(x_1, x_2, x_3, \ldots\) are usually in log scale

Each gene has its own profile

How to measure similarity?

Given two expression vectors \(U, V\) of
dimension \(d\)

- The normalized Euclidean metric

\[
\text{Dist}_{\text{eucl}}(V, U) = \sqrt{\frac{1}{d} \sum_{i=1}^{d} (V_i - U_i)^2}
\]

- Pearson correlation

\[
\text{corr}(V, U) = \frac{\sum_{i=1}^{d} (V_i - \bar{V})(U_i - \bar{U})}{\sigma_V \sigma_U}
\]

Example: Yeast cell-cycle data
Problem with missing values

Possible solutions: averaging, nearest-neighbor approach, EM approach, shuffling