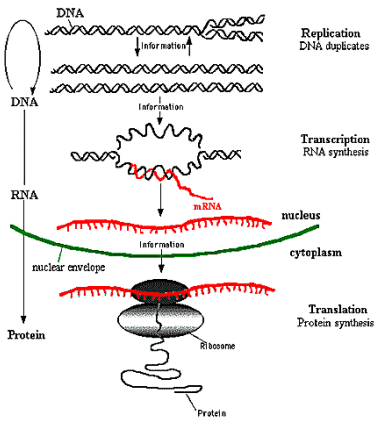
Microarray analysis

CS426, week 12
Golan Yona



The Central Dogma of Molecular Biology

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.

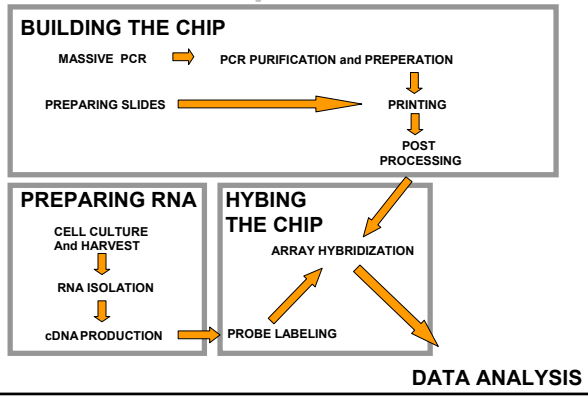
Based on old technique
Many flavors- majority are of two essential varieties

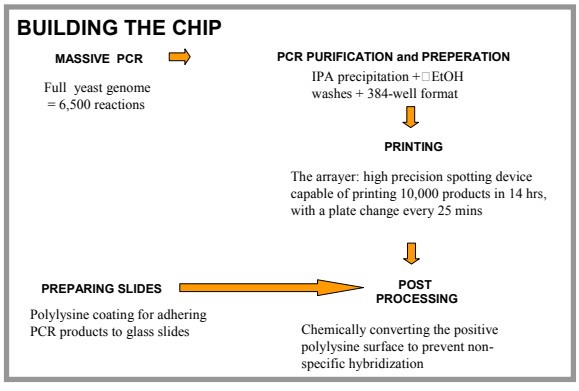
1. cDNA Arrays (discussed today)
   printing on glass slides, miniaturization, throughput fluorescence based detection

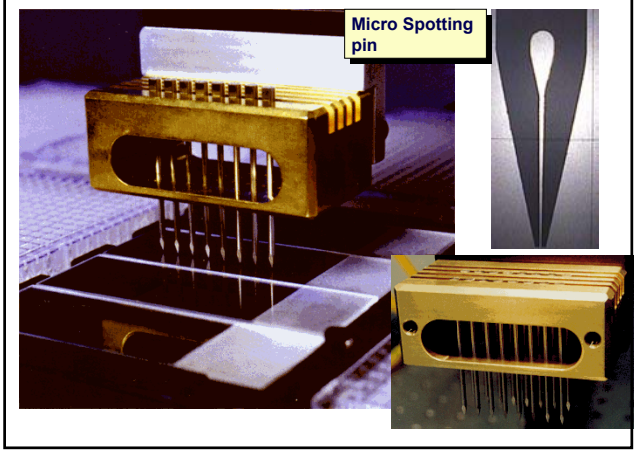
2. Affymetrix Arrays
   in situ synthesis of oligonucleotides.

Other arrays – protein arrays, combinatorial chemistry

The process

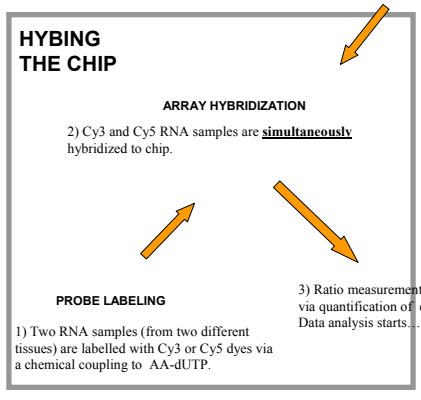




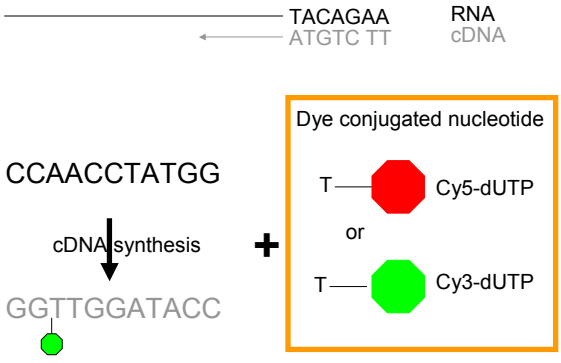


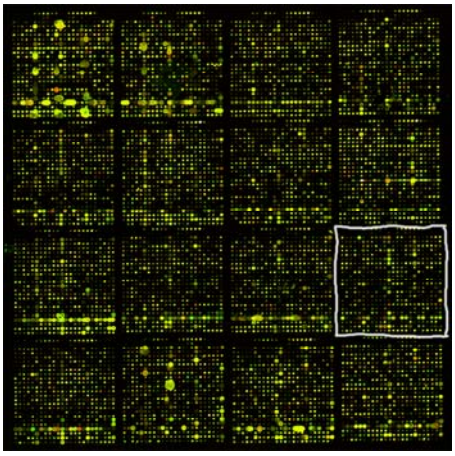
Practical problems

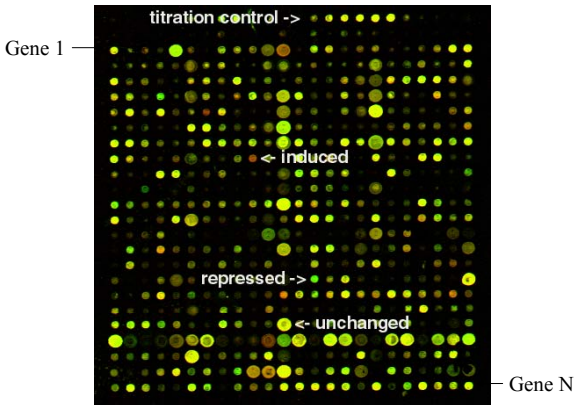
- **Surface chemistry: uneven surface may lead to high background.**
- **Dipping the pin into large volume -> pre-printing to drain off excess sample.**
- **Spot variation can be due to mechanical difference between pins. Pins could be clogged during the printing process.**
- **Spot size and density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



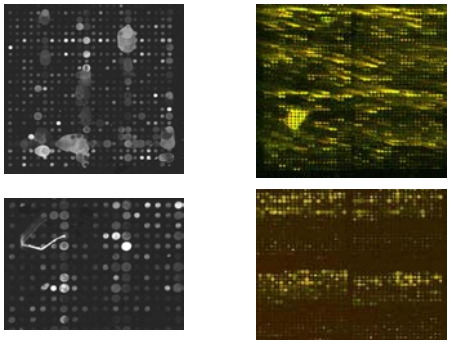
Direct labeling of RNA





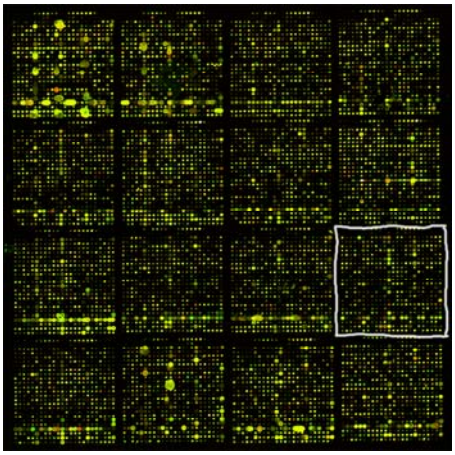


Practical problems



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 E1 is the expression of the gene in experiment 1 and E2 is the expression of the gene in experiment 2 then the spot is converted to a number representing the ratio E1/E2 or log(E1/E2)

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

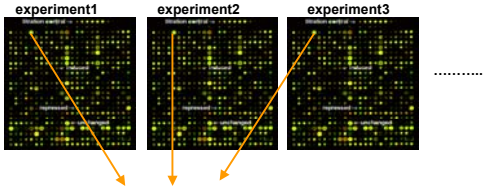
Absolute values are misleading

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: (x1,x2,x3,…)

This is an expression profile of a gene

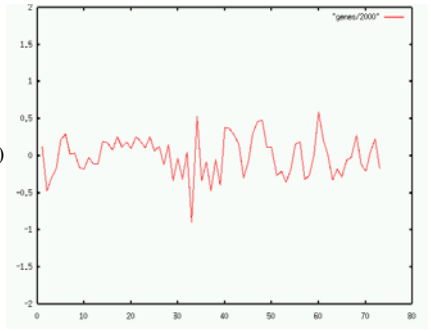
x1,x2,x3,.. are usually in log scale

Example: Yeast cell-cycle data

Expression at time t

$\log(\frac{Et}{Ec})$

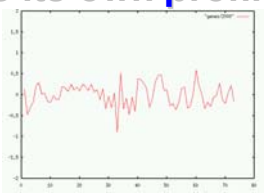
Background Expression (control)



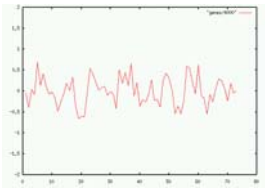
Time

Each gene has its own profile

Gene X



Gene Y



How to measure similarity?

Given two expression vectors U,V of dimension d

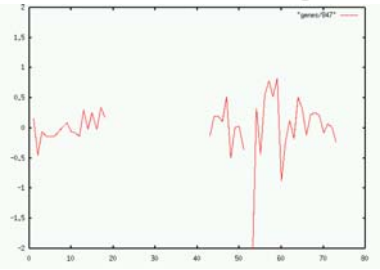
- The normalized Euclidean metric

$$Dist_{euc}(\mathbf{V},\mathbf{U}) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(V_i - U_i)}$$

- Pearson correlation

$$Corr(\mathbf{V},\mathbf{U}) = \sum_{i=1}^{d}\frac{(V_i - < V >)(U_i - < U >)}{\sigma_V \sigma_U}$$

Problem with missing values



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