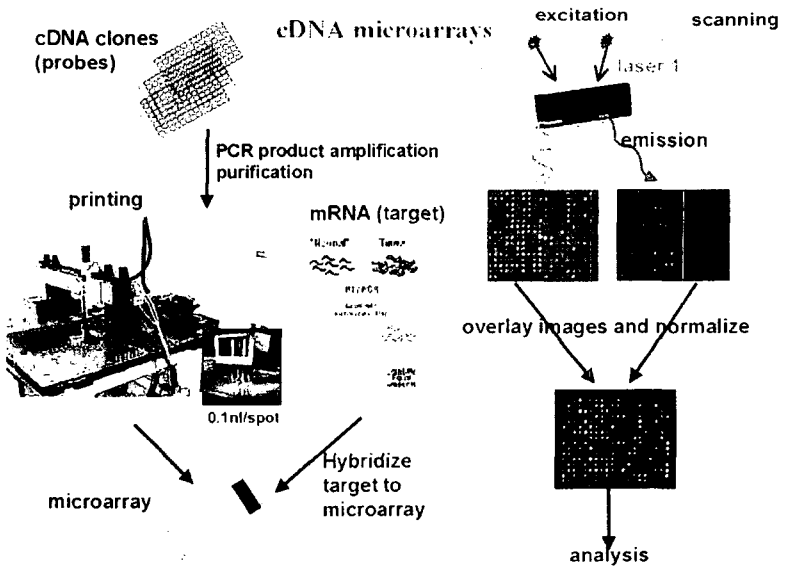
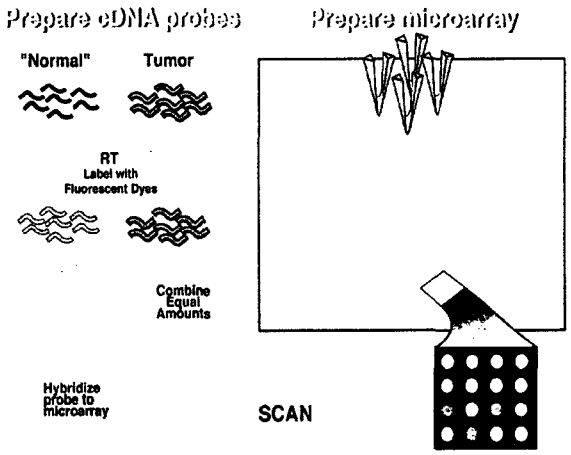
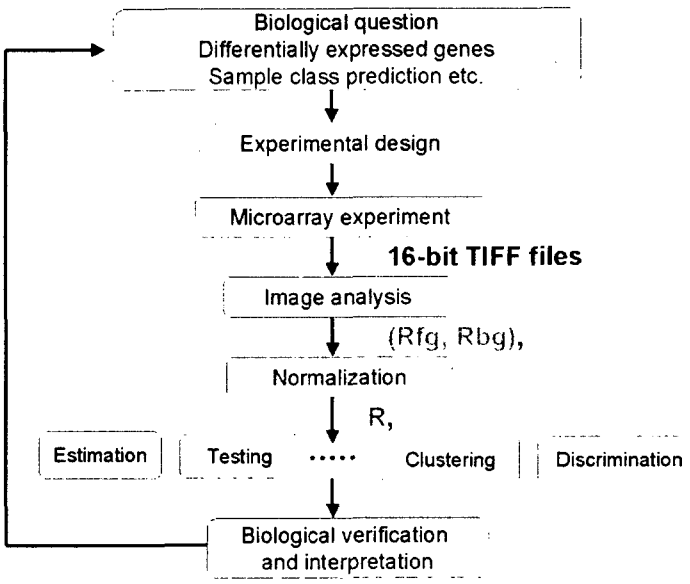
A Brief Introduction to DNA microarrays Data Analysis

Seyeon Weon
sywon@bioinformatics.pe.kr
Bioinformatics Research
Laboratories, Co.





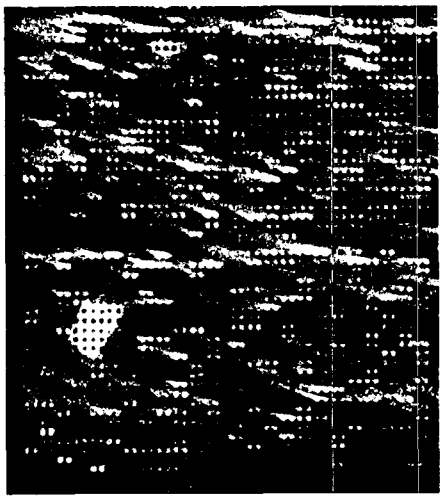
- Many of the slides are from Terry Speed, Department of Statistics, UC Berkeley
- And some slides adopted from Prabhakar Raghavan, Department of Computer Science, Stanford
- Also, some from Russ Altman, Biomedical Informatics, Stanford.



Brief introduction to image analysis issues

Comet Tails

Possibly caused by insufficiently rapid immersion of the slides in the succinic anhydride blocking solution.

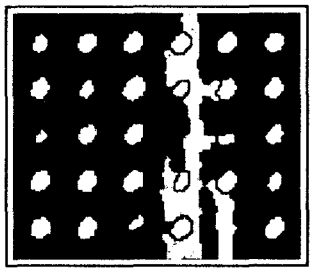


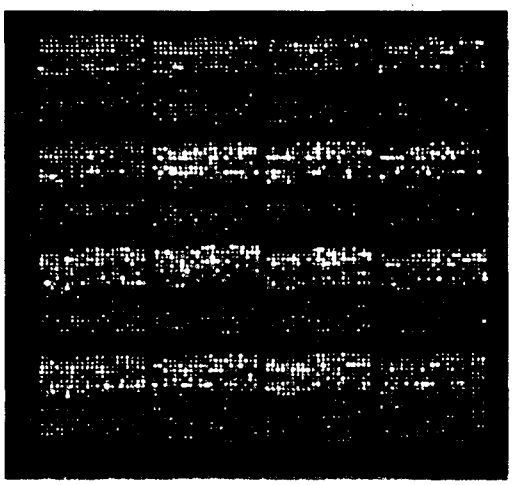
Steps in Images Processing

1. Addressing: **locate centers**

2. Segmentation: **classification of pixels either as signal or background (using seeded region growing).**

3. Information extraction: **for each spot of the array, calculates signal intensity pairs, background and quality measures.**





KO #8

Probes ~6,000 cDNAs, including 200 related to lipid metabolism.
Arranged in a 4x4 array of 19x21 sub-arrays called pin-groups.

Draw the lowess curve within print-tip group

- M=log(R/G)
- A=(logR + logG)/2

Which genes are (relatively) up/down regulated?

Samples: liver tissue from each of two kinds of mice, e.g. KO *vs.* WT, or mutant *vs.* WT

- For each gene form the *t* statistic:

$$\frac{\text{average of } n \text{ trt } Ms}{\text{sqrt}(1/n \text{ (SD of n trt Ms)}^2)}$$

Which genes are (relatively) up/down regulated?

: as before, but also pooled control (reference) liver tissue

T × n C*

C × n C*

- For each gene form the t statistic:
  $$\frac{\text{average of n trt Ms - average of n ctl Ms}}{\text{sqrt}(1/n\ (\text{SD of n trt Ms})^2 + (\text{SD of n ctl Ms})^2)}$$

Many comparisons of interest

T1 T2 T3 T4

x 2 x 2 x 2 x 2

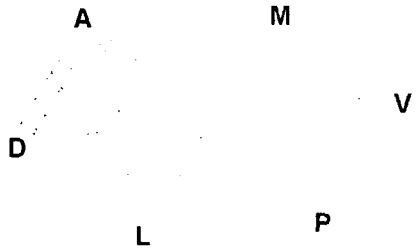
C

Liver tissue from mice treated by cholesterol modifying drugs.

Question 1: **Find genes that respond differently between the treatment and the control.**

Question 2: **Find genes that respond similarly across two or more treatments relative to control.**

The olfactory bulb experiments



tissues from different regions of the olfactory bulb.

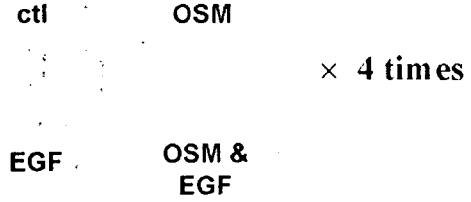
Question 1 **differences between different regions.**

Question 2 **identify genes with pre-specified patterns across regions.**

Interaction?

treated cell lines at 4 time points (30 minutes, 1 hour, 4 hours, 24 hours)

Question **Which genes contribute to the enhanced inhibitory effect of OSM when it is combined with EGF? Role of time?**



Gene expression data from cDNA microarrays

Data on *p* genes (typically 1000s) for *n* samples; always ratios

Genes	sample1	sample2	sample3	sample4	sample5	...
1	0.46	0.30	0.80	1.51	0.90	...
2	-0.10	0.49	0.24	0.06	0.46	...
3	0.15	0.74	0.04	0.10	0.20	...
4	-0.45	-1.03	-0.79	-0.56	-0.32	...
5	-0.06	1.06	1.35	1.09	-1.09	...

Gene expression level of gene *i* in mRNA sample *j*

= Log(Red Intensity / [illegible])

Analyzing the Data

- Unsupervised Learning
- Supervised Learning
- Network Modeling

Cosine similarity

Cosine similarity of D_j, D_k:

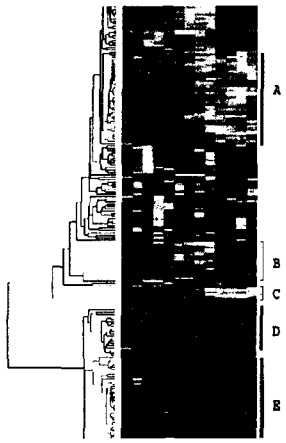
$$sim(D_j, D_k) = \sum_{i=1}^{m} w_{ij} \times w_{ik}$$

Aka normalized inner product.

Supervised vs. unsupervised learning

- Unsupervised learning:
  - Infer structure implicit in the data, without prior training.
- Supervised learning:
  - Train system to recognize classes
  - Decide whether or not new data belong to the class(es) trained on

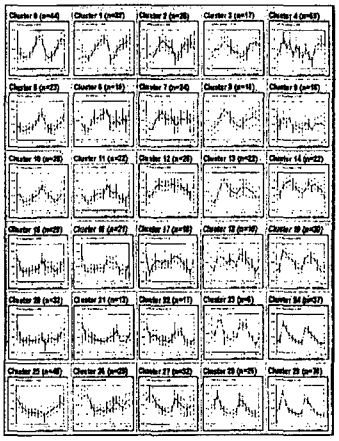
Hierarchical clustering



k-Means Clustering

- At the start of the iteration, we have ***k*** centroids.
- Each data point assigned to the nearest centroid.
- All points assigned to the same centroid are averaged to compute a new centroid;
  - thus have ***k*** new centroids.
- Repeat above

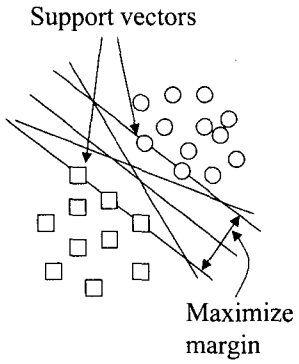
Self Organizing Maps(SOM)



k-Nearest Neighbor Classification

- For a given data point, find out ***k***-closest data points in the training set.
- If more than ***p*** percent of above data points belong to a class, the data point is assigned to the class.

Support Vector Machine(SVM)

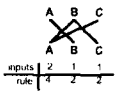


Bayesian Classification

- Assume all expression measurements for a gene are independent.
- Assume p(f) and p(group1) are constant.
- P(f|group 1) = p(f1&f2...fn|group1) = p(f1|group1) * p(f2|group1)...* p(fn|group1)
- Can just multiply these probabilities (or add their logs), which are easy to compute, by counting up frequencies in the set of "known" members of group 1.
- Choose a cutoff probability for saying "Group 1 member."

Boolean Network

Wiring and rules



Basis for rules:

1. A activates B
2. B activates A and C
3. C inhibits A

Trajectory 1 results in a point attractor

Iteration	A	B	C
1	1	1	0
2	1	1	1
3	0	1	1
4	0	0	1
5	0	0	0
6	0	0	0

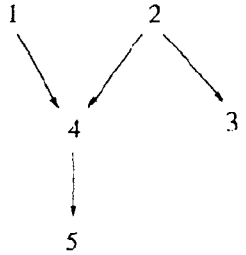
Trajectory 2 results in a 2-state dynamic attractor

Iteration	A	B	C
1	1	0	0
2	0	1	0
3	1	0	1
4	0	1	0

Bayesian Network

- X_i = expression level of gene i
- Arrow = indicate direct regulators of i
- Compute probability that i is expressed based on probability of direct regulators being expressed

$$p(\boldsymbol{X}) = \prod_{i=1}^{n} p(X_i \mid \text{parents}(X_i)).$$



$p(X_1)$, $p(X_2)$, $p(X_4 \mid X_1, X_2)$

$p(X_5 \mid X_4)$, $p(X_3 \mid X_2)$

$$p(\boldsymbol{X}) = p(X_5 \mid X_4)\, p(X_4 \mid X_1, X_2)\, p(X_3 \mid X_2)\, p(X_2)\, p(X_1)$$

$i(X_1; X_2, X_3)$, $i(X_2; X_1)$, $i(X_4; X_3 \mid X_1, X_2)$

$i(X_3; X_1, X_4, X_5 \mid X_2)$, $i(X_5; X_1, X_2, X_3 \mid X_4)$