

CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfF18.html

Gene Expression

- ❑ Process of transcription and/or translation of a gene is called **gene expression**.
- ❑ Every cell of an organism has the same genetic material, but different genes are **expressed** at different times.
- ❑ Patterns of gene expression in a cell is indicative of its state.

Hybridization

- If two complementary strands of DNA or mRNA are brought together under the right experimental conditions they will hybridize.
- A hybridizes to $B \Rightarrow$
 - A is reverse complementary to B , or
 - A is reverse complementary to a subsequence of B .
- It is possible to experimentally verify whether A hybridizes to B , by labeling A or B with a radioactive or fluorescent tag, followed by excitation by laser.

Measuring gene expression

- Gene expression for a single gene can be measured by extracting mRNA from the cell and doing a simple **hybridization** experiment.
- Given a sample of cells, gene expression for every gene can be measured using a single microarray experiment.

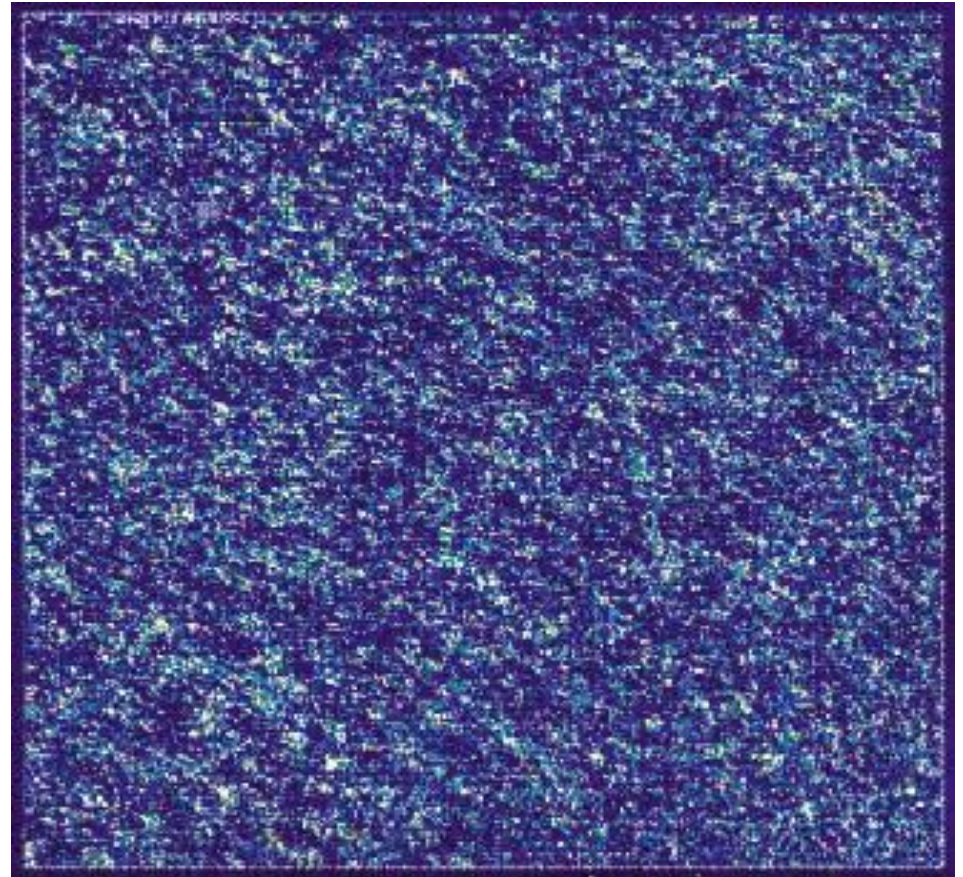
Microarray/DNA chip technology

- High-throughput method to study gene expression of thousands of genes simultaneously.
- Many applications:
 - Genetic disorders & Mutation/polymorphism detection
 - Study of disease subtypes
 - Drug discovery & toxicology studies
 - Pathogen analysis
 - Differing expressions over time, between tissues, between drugs, across disease states

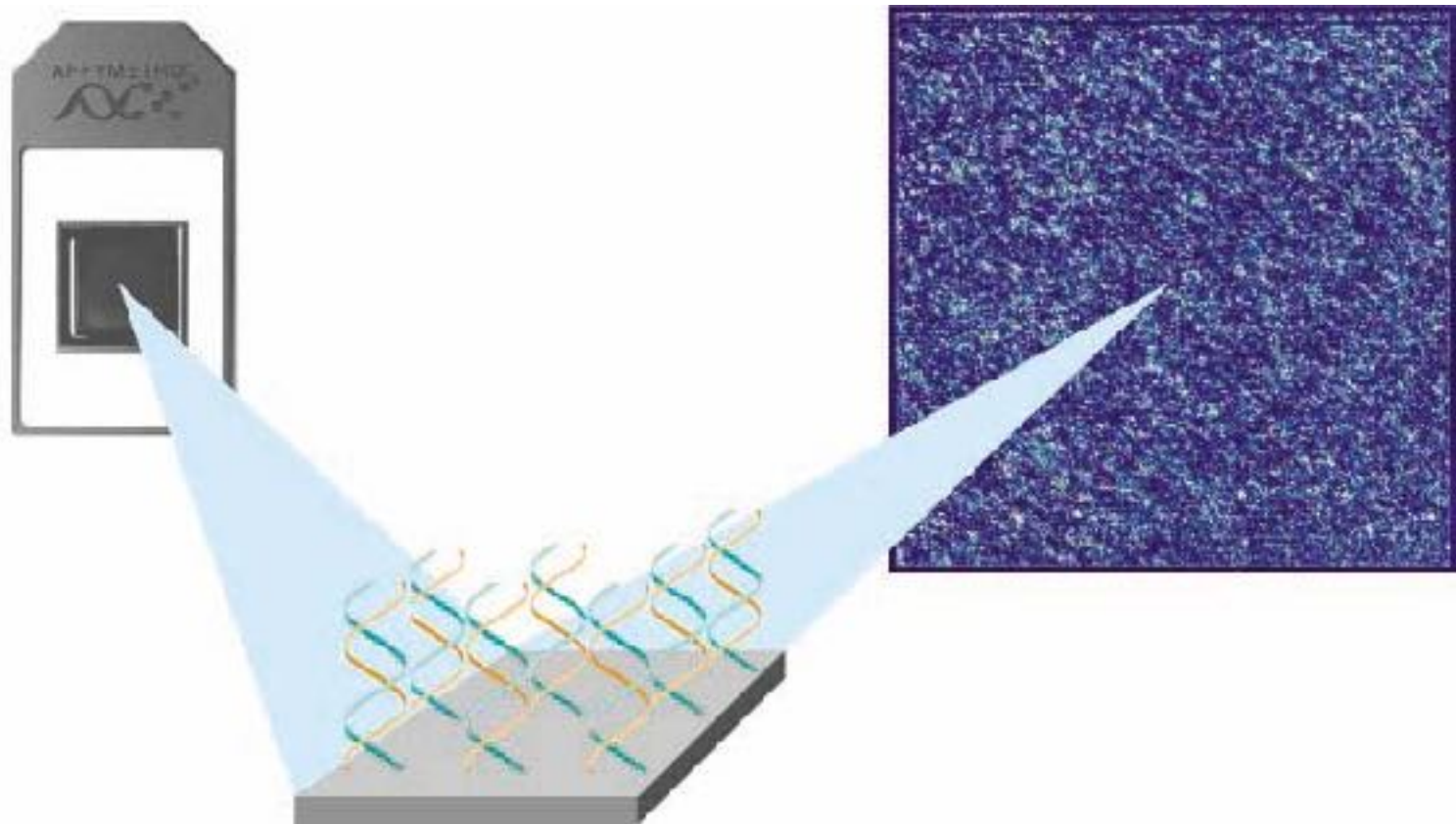
Microarray Data

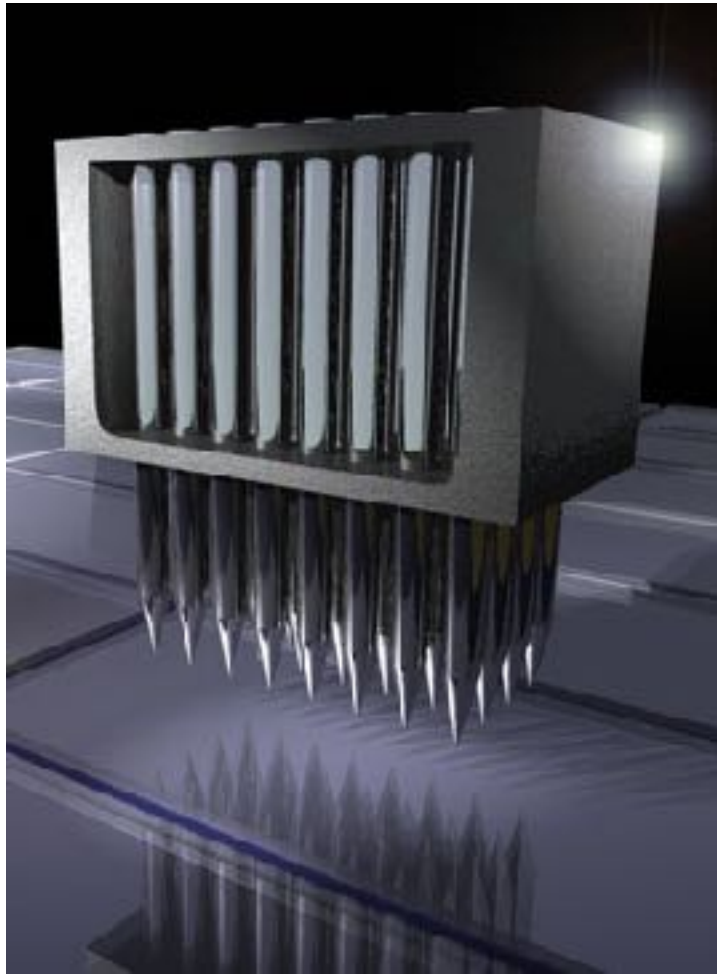
Gene	Expression Level
Gene1	
Gene2	
Gene3	
...	

Gene Chips

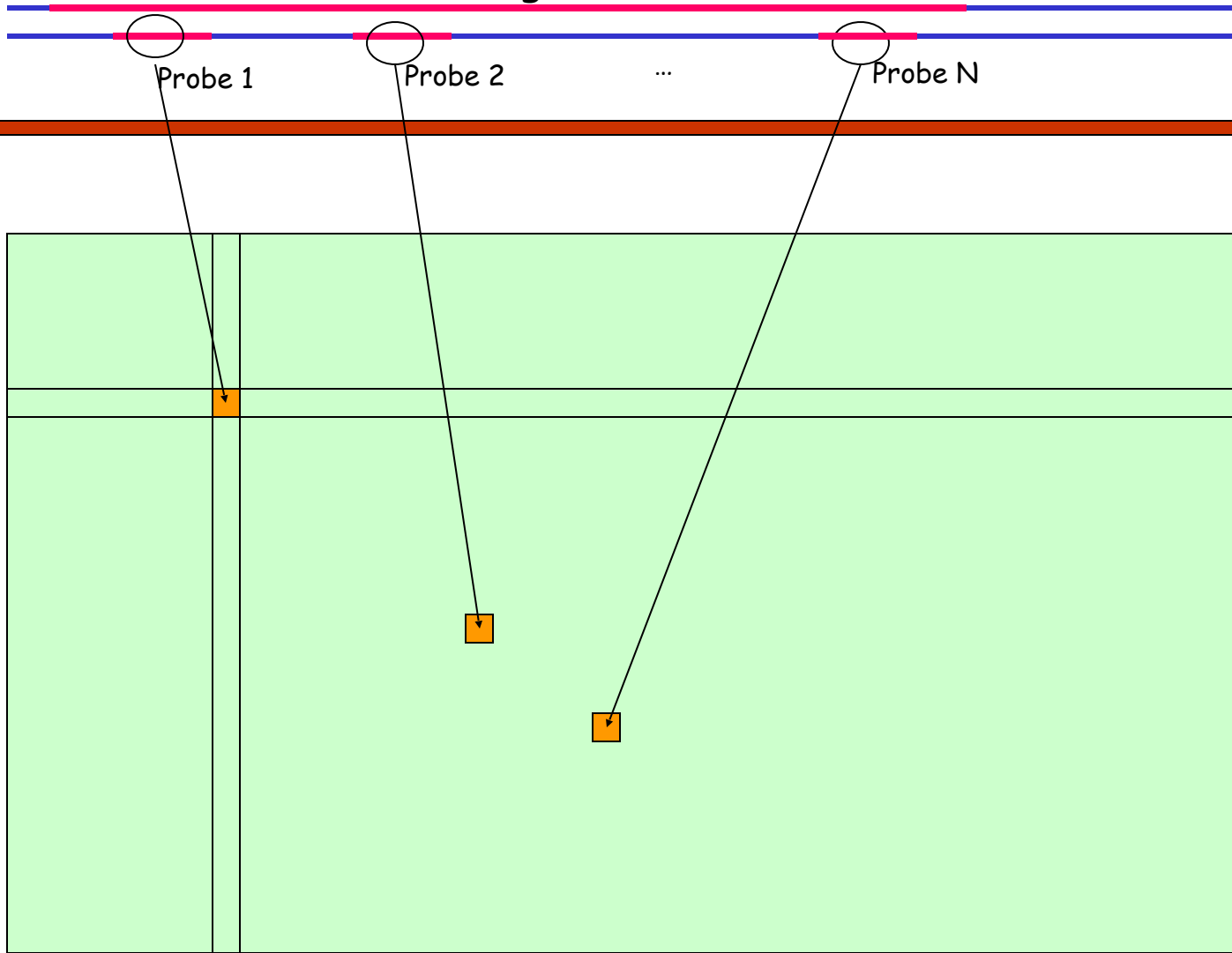


DNA Chips & Images





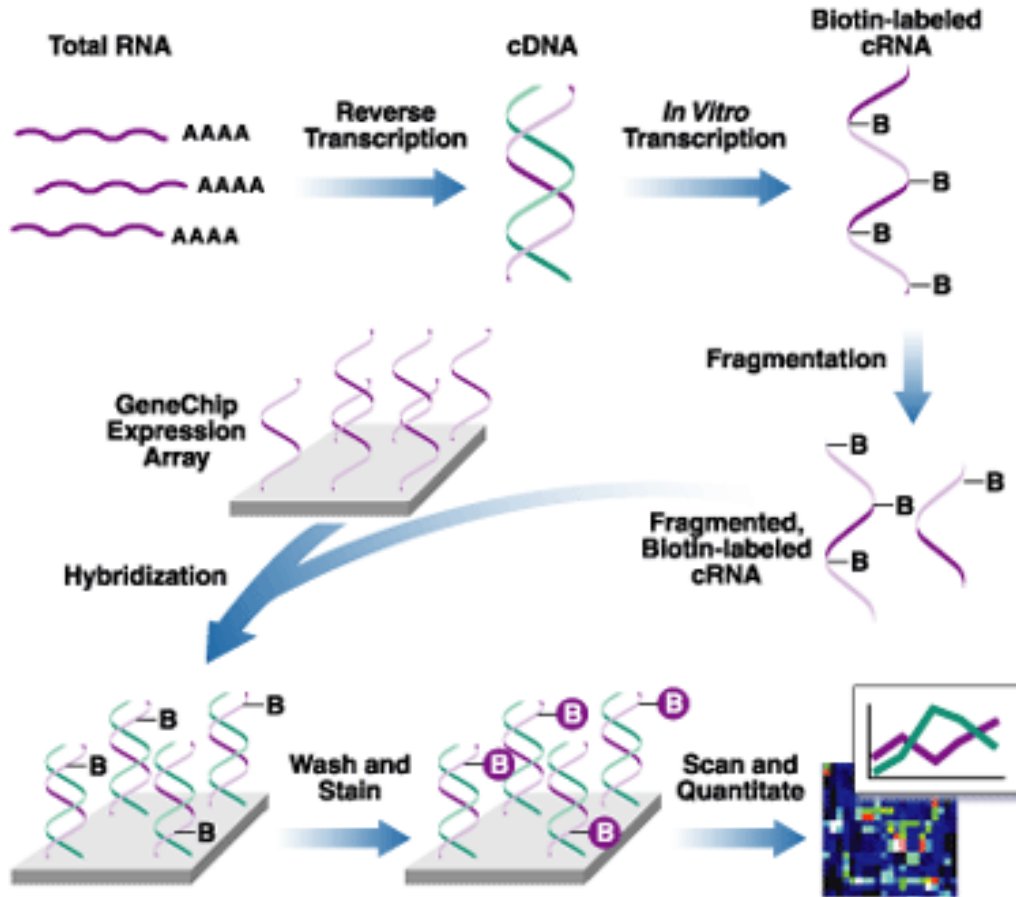
Gene g



Microarray/DNA chips (Simplified)

- ❑ Construct **probes** corresponding to reverse complements of genes of interest.
- ❑ Microscopic quantities of probes placed on solid surfaces at defined spots on the chip.
- ❑ Extract mRNA from sample cells and **label** them.
- ❑ Apply labeled sample (mRNA extracted from cells) to every spot, and allow hybridization.
- ❑ Wash off unhybridized material.
- ❑ Use optical detector to measure amount of fluorescence from each spot.

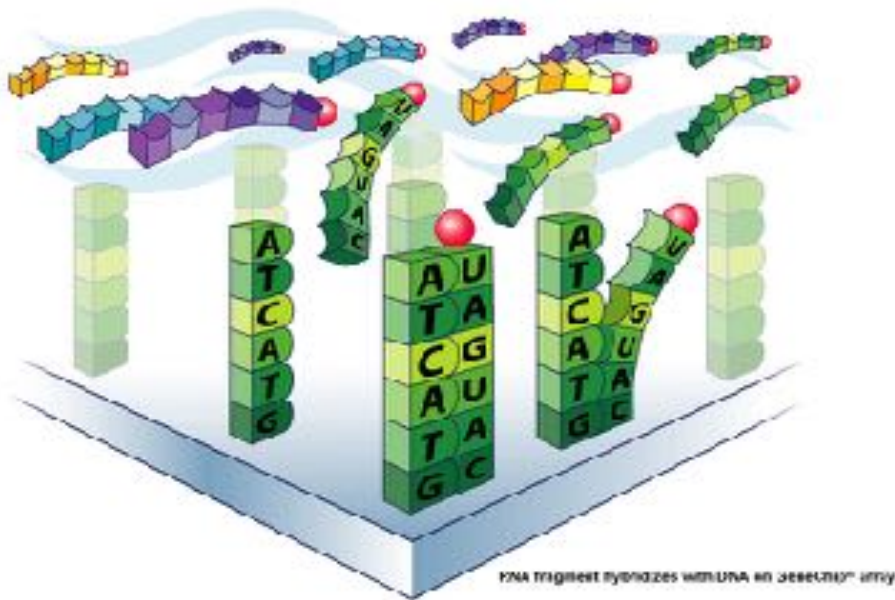
Affymetrix DNA chip schematic



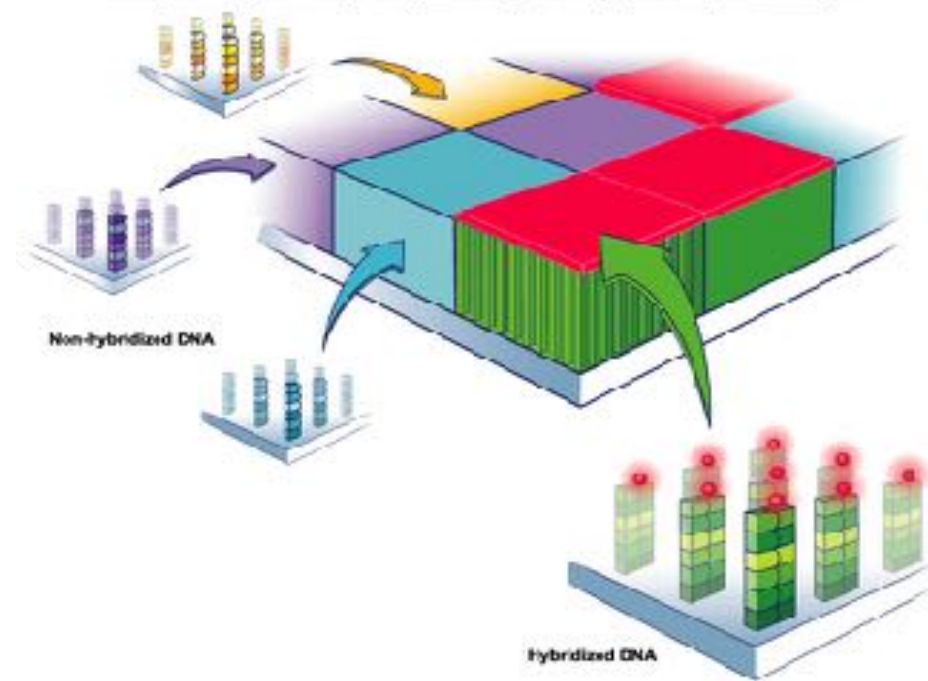
www.affymetrix.com

What's on the slide?

RNA fragments with fluorescent tags from sample to be tested



Shining a laser light at GeneChip array causes tagged DNA fragments that hybridized to glow



Microarrays: competing technologies

□ Affymetrix & Agilent

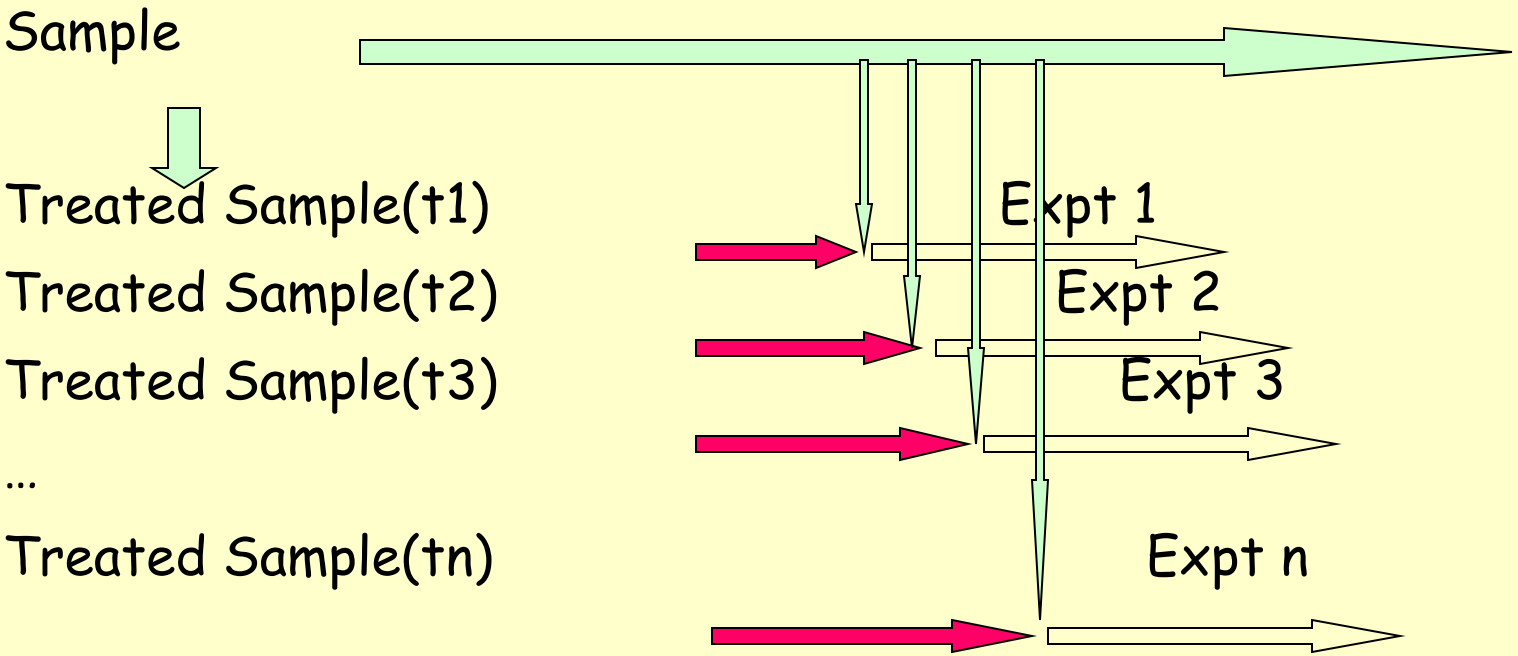
□ Differ in:

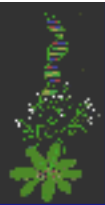
● method to place DNA: Spotting vs. photolithography

● Length of probe

● Complete sequence vs. series of fragments

Study effect of treatment over time





AFGC

2-color DNA microarray



Treated

mRNA

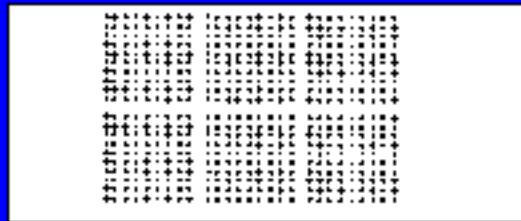
Cy5 Probe



Control

mRNA

Cy3 Probe

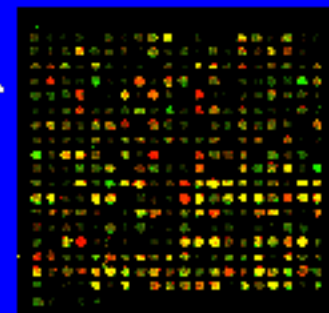


Simultaneous hybridization

Normalization

Data extraction

Scanning



How to compare 2 cell samples with Two-Color Microarrays?

- ❑ mRNA from sample 1 is extracted and labeled with a **red fluorescent** dye.
- ❑ mRNA from sample 2 is extracted and labeled with a **green fluorescent** dye.
- ❑ Mix the samples and apply it to every spot on the microarray. Hybridize sample mixture to probes.
- ❑ Use optical detector to measure the amount of **green** and **red** fluorescence at each spot.

Sources of Variations & Experimental Errors

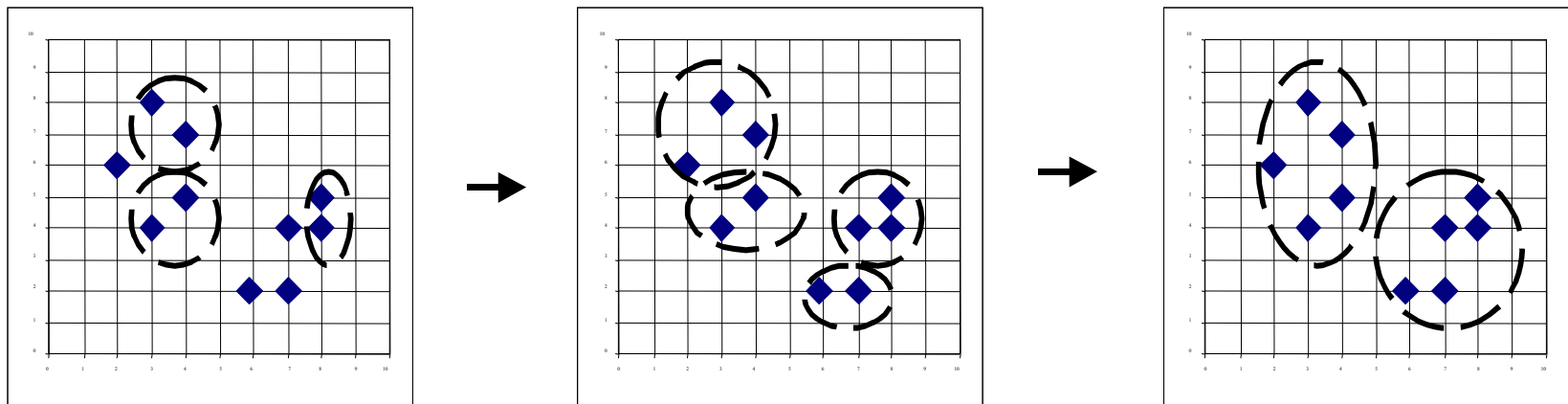
- ❑ Variations in cells/individuals
- ❑ Variations in mRNA extraction, isolation, introduction of dye, variation in dye incorporation, dye interference
- ❑ Variations in probe concentration, probe amounts, substrate surface characteristics
- ❑ Variations in hybridization conditions and kinetics
- ❑ Variations in optical measurements, spot misalignments, discretization effects, noise due to scanner lens and laser irregularities
- ❑ Cross-hybridization of sequences with high sequence identity
- ❑ Limit of factor 2 in precision of results
- ❑ Variation changes with intensity: larger variation at low or high expression levels

Need to Normalize data

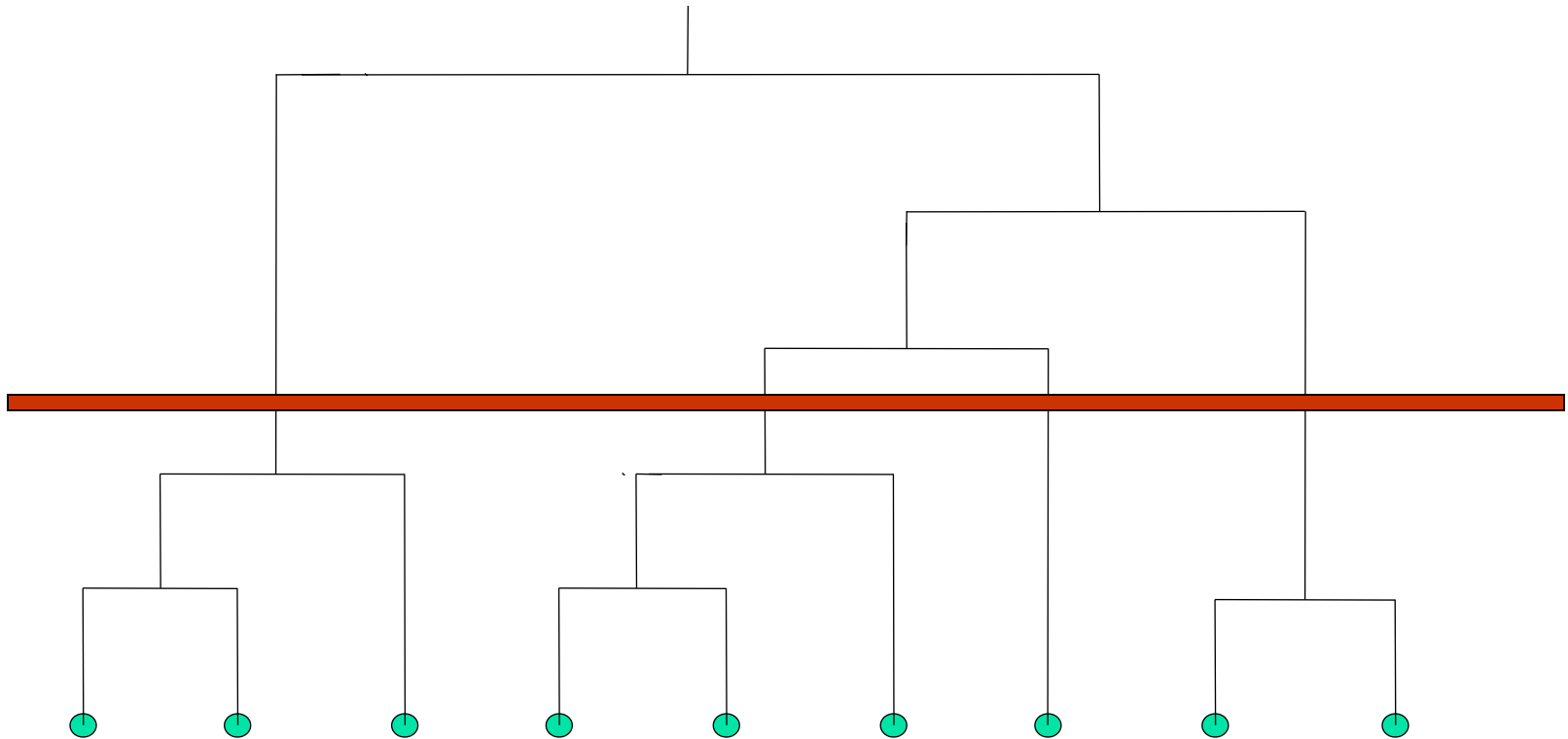
Clustering

- Clustering is a general method to study patterns in gene expressions.
- Several known methods:
 - *Hierarchical Clustering* (Bottom-Up Approach)
 - *K-means Clustering* (Top-Down Approach)
 - *Self-Organizing Maps* (SOM)

Hierarchical Clustering: Example



A Dendrogram



Hierarchical Clustering [Johnson, SC, 1967]

- Given n points in \mathbb{R}^d , compute the distance between every pair of points
- While (not done)
 - Pick closest pair of points s_i and s_j and make them part of the same cluster.
 - Replace the pair by an average of the two s_{ij}

Try the applet at: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

Distance Metrics

□ For clustering, define a distance function:

● Euclidean distance metrics

$$D_k(X, Y) = \left[\sum_{i=1}^d (X_i - Y_i)^k \right]^{1/k}$$

k=2: Euclidean Distance

● Pearson correlation coefficient

$$\rho_{xy} = \frac{1}{d} \sum_{i=1}^d \left(\frac{X_i - \bar{X}}{\sigma_x} \right) \left(\frac{Y_i - \bar{Y}}{\sigma_y} \right)$$

$-1 \leq \rho_{xy} \leq 1$

EXHIBIT 3.4 Joint Probability Model for the Ratings of Two People

(a) $\rho_{XY} = 0$

x	y			Total
	1	2	3	
3	1/9	1/9	1/9	1/3
2	1/9	1/9	1/9	1/3
1	1/9	1/9	1/9	1/3
Total	1/3	1/3	1/3	1

(b) $\rho_{XY} = \frac{1}{3}$

x	y			Total
	1	2	3	
3	1/18	1/18	4/18	1/3
2	1/18	4/18	1/18	1/3
1	4/18	1/18	1/18	1/3
Total	1/3	1/3	1/3	1

(c) $\rho_{XY} = -\frac{1}{3}$

x	y			Total
	1	2	3	
3	4/18	1/18	1/18	1/3
2	1/18	4/18	1/18	1/3
1	1/18	1/18	4/18	1/3
Total	1/3	1/3	1/3	1

(d) $\rho_{XY} = \frac{1}{3}$

x	y			Total
	1	2	3	
3	1/27	2/27	6/27	1/3
2	2/27	5/27	2/27	1/3
1	6/27	2/27	1/27	1/3
Total	1/3	1/3	1/3	1

(e) $\rho_{XY} = -\frac{2}{3}$

x	y			Total
	1	2	3	
3	6/27	2/27	1/27	1/3
2	2/27	5/27	2/27	1/3
1	1/27	2/27	6/27	1/3
Total	1/3	1/3	1/3	1

(f) $\rho_{XY} = \frac{1}{3}$

x	y			Total
	1	2	3	
3	1/36	2/36	9/36	1/3
2	2/36	8/36	2/36	1/3
1	9/36	2/36	1/36	1/3
Total	1/3	1/3	1/3	1

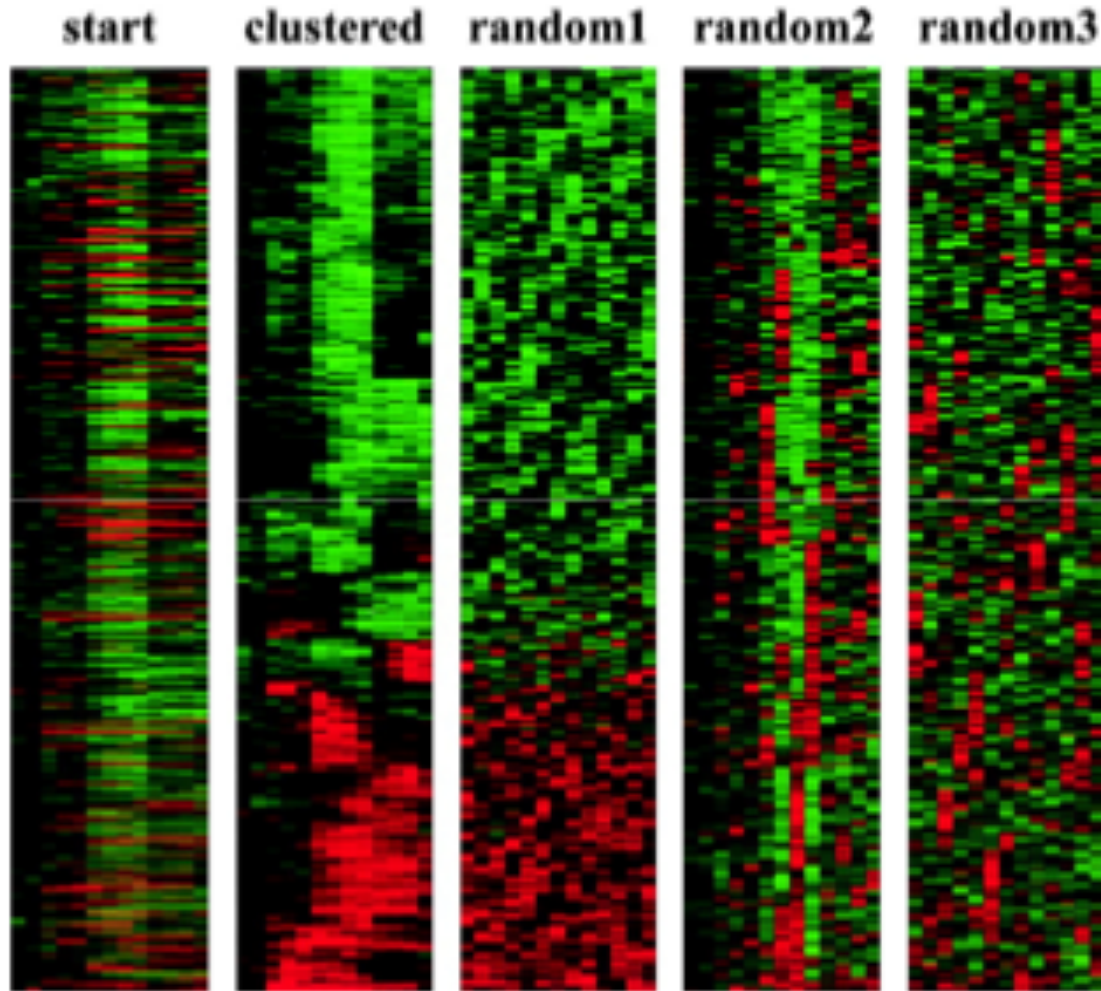
(g) $\rho_{XY} = -\frac{1}{3}$

x	y			Total
	1	2	3	
3	9/36	2/36	1/36	1/3
2	2/36	8/36	2/36	1/3
1	1/36	2/36	9/36	1/3
Total	1/3	1/3	1/3	1

Clustering of gene expressions

- Represent each gene as a vector or a point in d -space where d is the number of arrays or experiments being analyzed.

Clustering Random vs. Biological Data

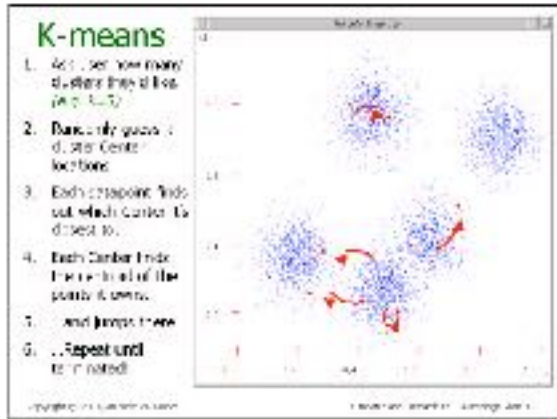
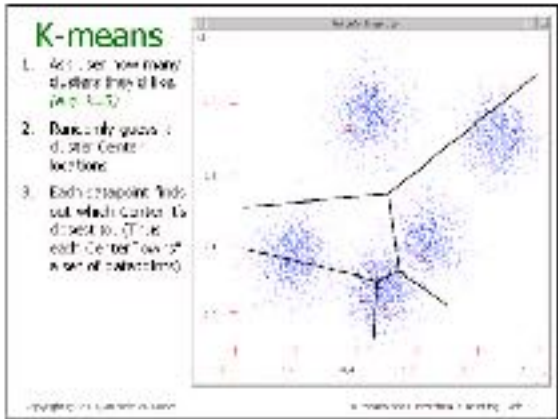
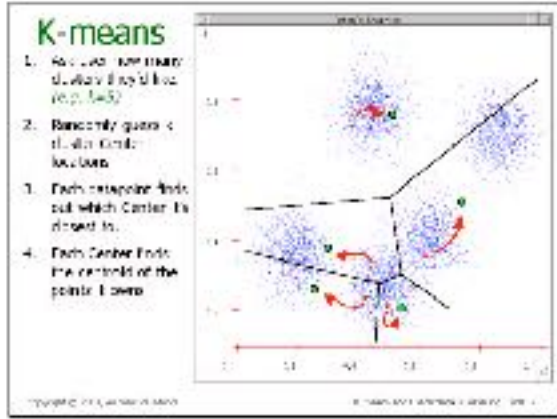
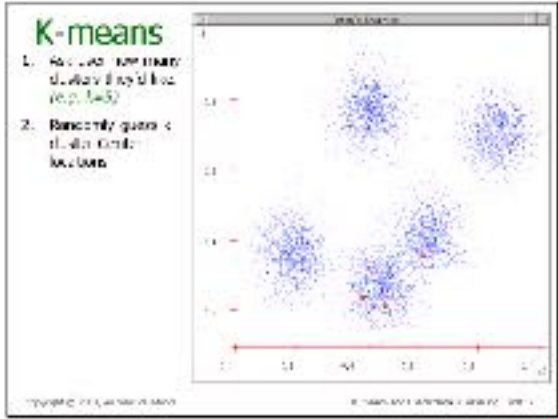


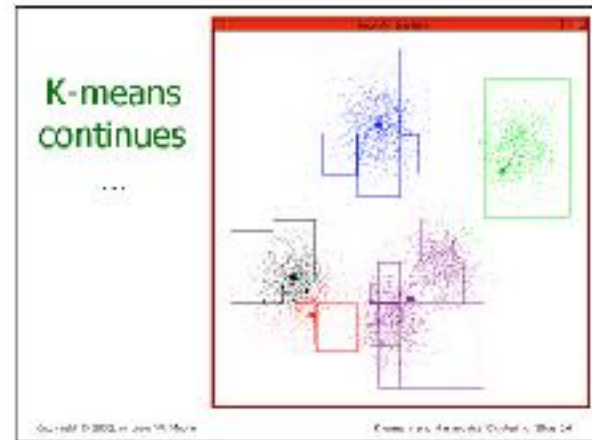
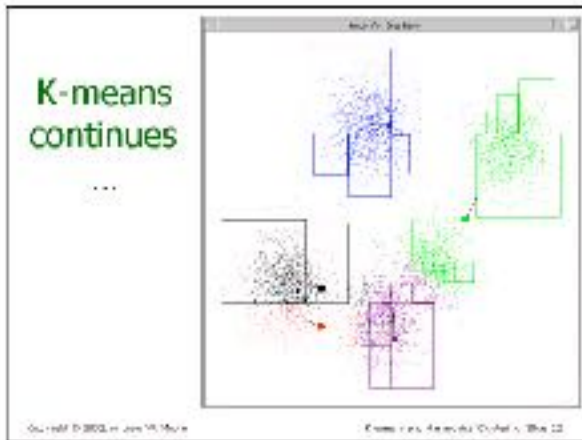
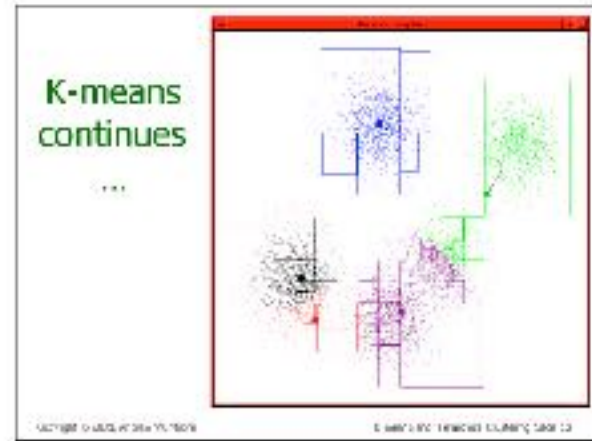
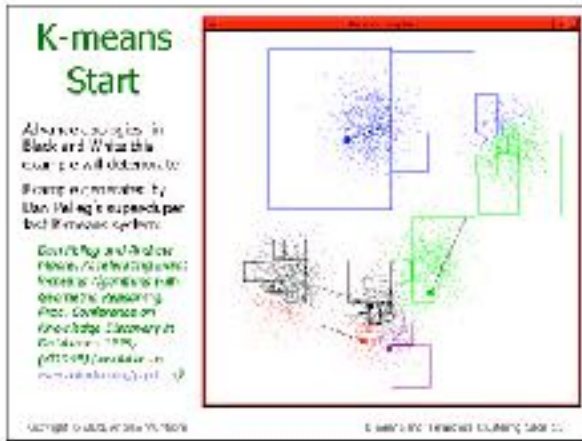
From Eisen MB, et al, PNAS 1998 95(25):14863 -8

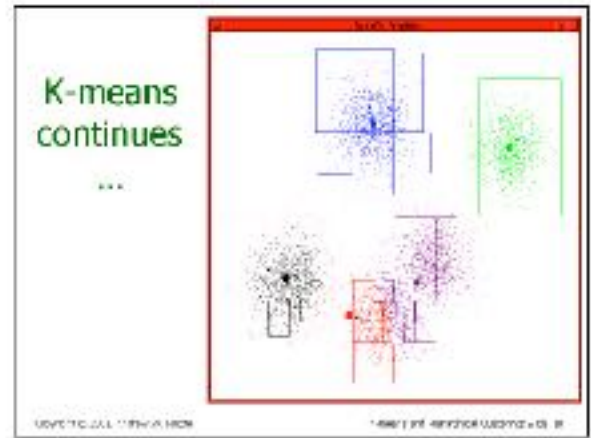
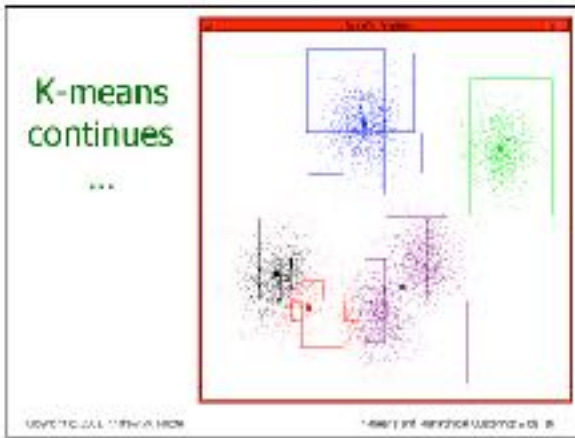
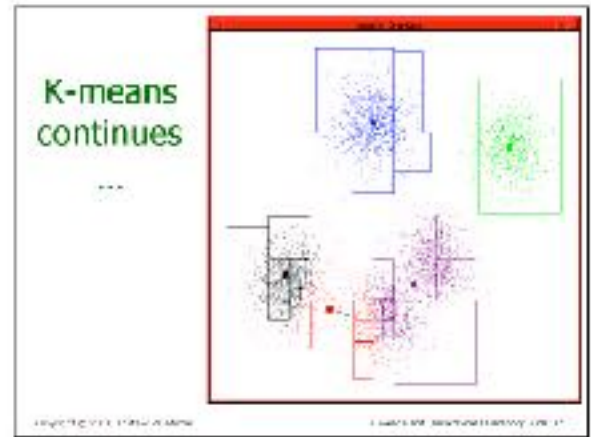
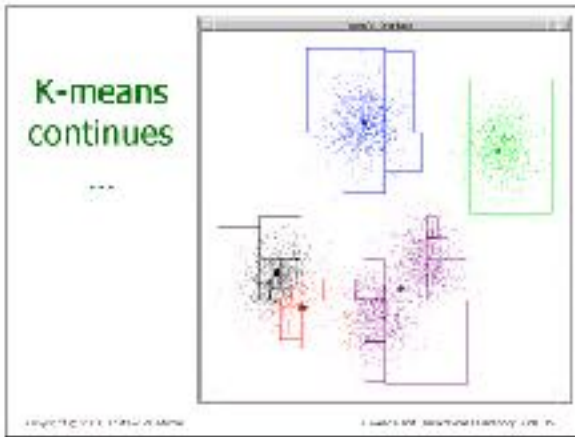
K-Means Clustering: Example

Example from Andrew Moore's tutorial on Clustering.

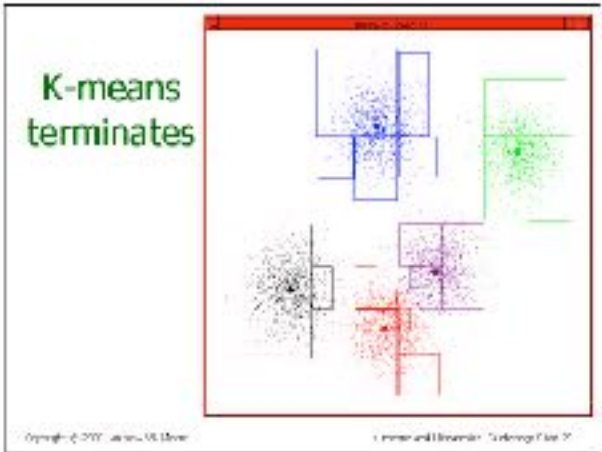
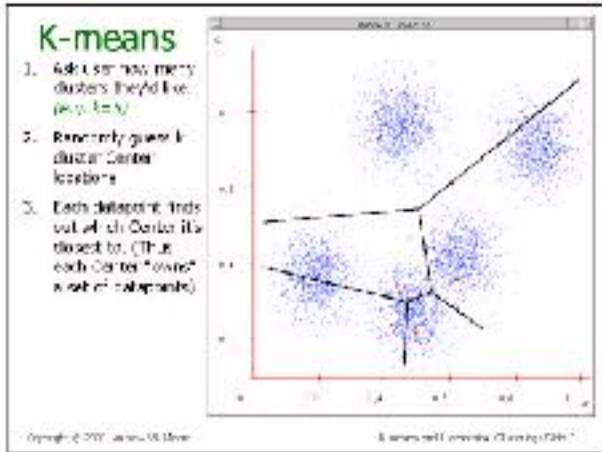
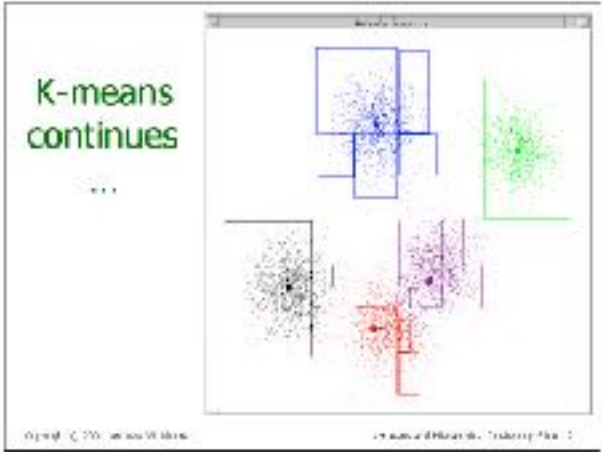
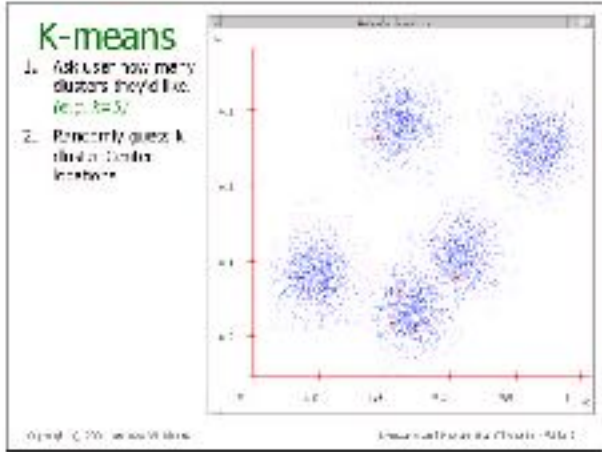
Start







Start



End

K-Means Clustering [McQueen '67]

Repeat

- Start with randomly chosen cluster centers
- Assign points to give greatest increase in score
- Recompute cluster centers
- Reassign points

until (no changes)

Try the applet at: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

Comparisons

□ Hierarchical clustering

- Number of clusters not preset.
- Complete hierarchy of clusters
- Not very robust, not very efficient.

□ K-Means

- Need definition of a **mean**. Categorical data?
- More efficient and often finds optimum clustering.

Functionally related genes behave similarly across experiments

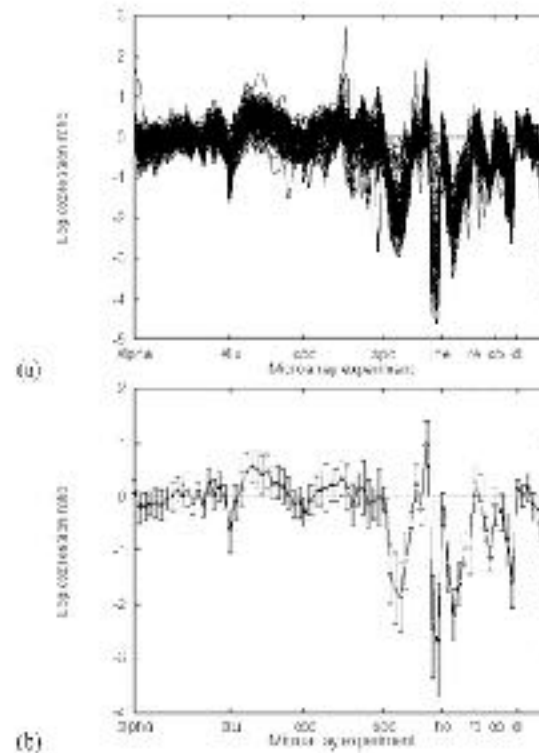


Figure 1. Expression profiles of the cytoplasmic ribosomal proteins. Figure (a) shows the expression profiles from the data in [Eisen et al., 1998] of 121 cytoplasmic ribosomal proteins, as classified by MYOD [MYOD, 1999]. The logarithm of the expression ratio is plotted as a function of DNA microarray experiment. Ticks along the X-axis represent the beginnings of experimental periods. They are, from left to right, cell division cycle after synchronization with α factor arrest (alpha), cell division cycle after synchronization by centrifugal elutriation (sta), cell division cycle measured using a temperature sensitive *rad5* mutant (cdc), sporulation (spo), heat shock (hs), reducing shock (rc), cold shock (cs), and diauxic shift (di). Sporulation is the generation of a yeast spore by meiosis. Diauxic shift is the shift from anaerobic (fermentation) to aerobic (respiration) metabolism. The medium starts rich in glucose, and yeast cells ferment, producing ethanol. When the glucose is used up, they switch to ethanol as a source for carbon. Heat, cold, and reducing shock are various ways to stress the yeast cell. Figure (b) shows the average, plus or minus one standard deviation, of the data in Figure (a).

Self-Organizing Maps [Kohonen]

- ❑ Kind of neural network.
- ❑ Clusters data and find complex relationships between clusters.
- ❑ Helps reduce the dimensionality of the data.
- ❑ Map of 1 or 2 dimensions produced.
- ❑ Unsupervised Clustering
- ❑ Like K-Means, except for visualization

SOM Architectures

- ❑ 2-D Grid
- ❑ 3-D Grid
- ❑ Hexagonal Grid