CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

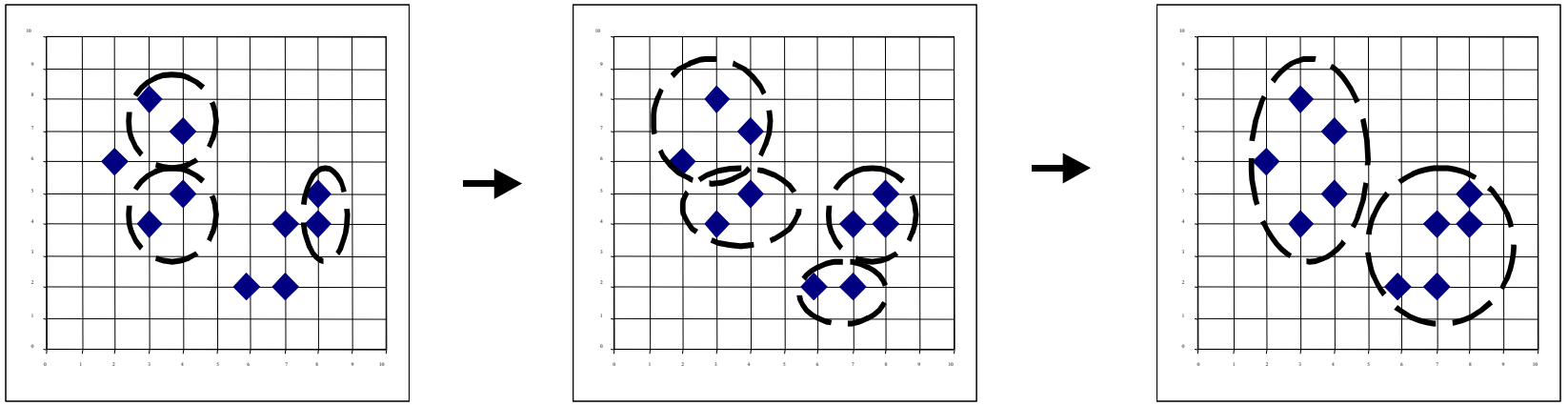
giri@cis.fiu.edu

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

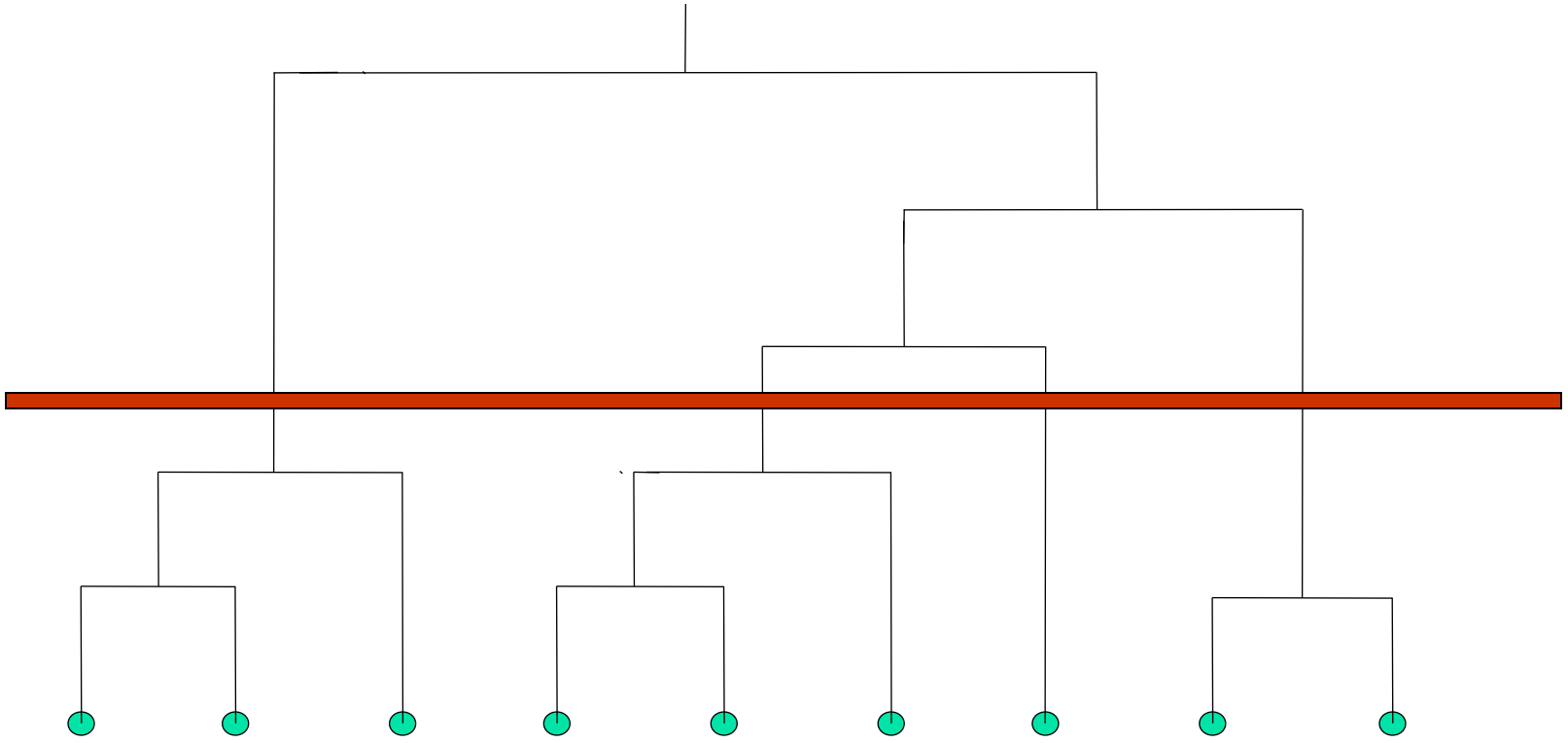
Clustering

- 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 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

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

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

EXHIBIT 3.4 *Joint Probability Model for the Ratings of Two People*

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

x	y = 1	y = 2	y = 3	Total
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 = 1	y = 2	y = 3	Total
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 = 1	y = 2	y = 3	Total
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) $\rho_{XY} = \frac{1}{2}$

x	y = 1	y = 2	y = 3	Total
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) $\rho_{XY} = -\frac{1}{2}$

x	y = 1	y = 2	y = 3	Total
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) $\rho_{XY} = \frac{2}{3}$

x	y = 1	y = 2	y = 3	Total
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

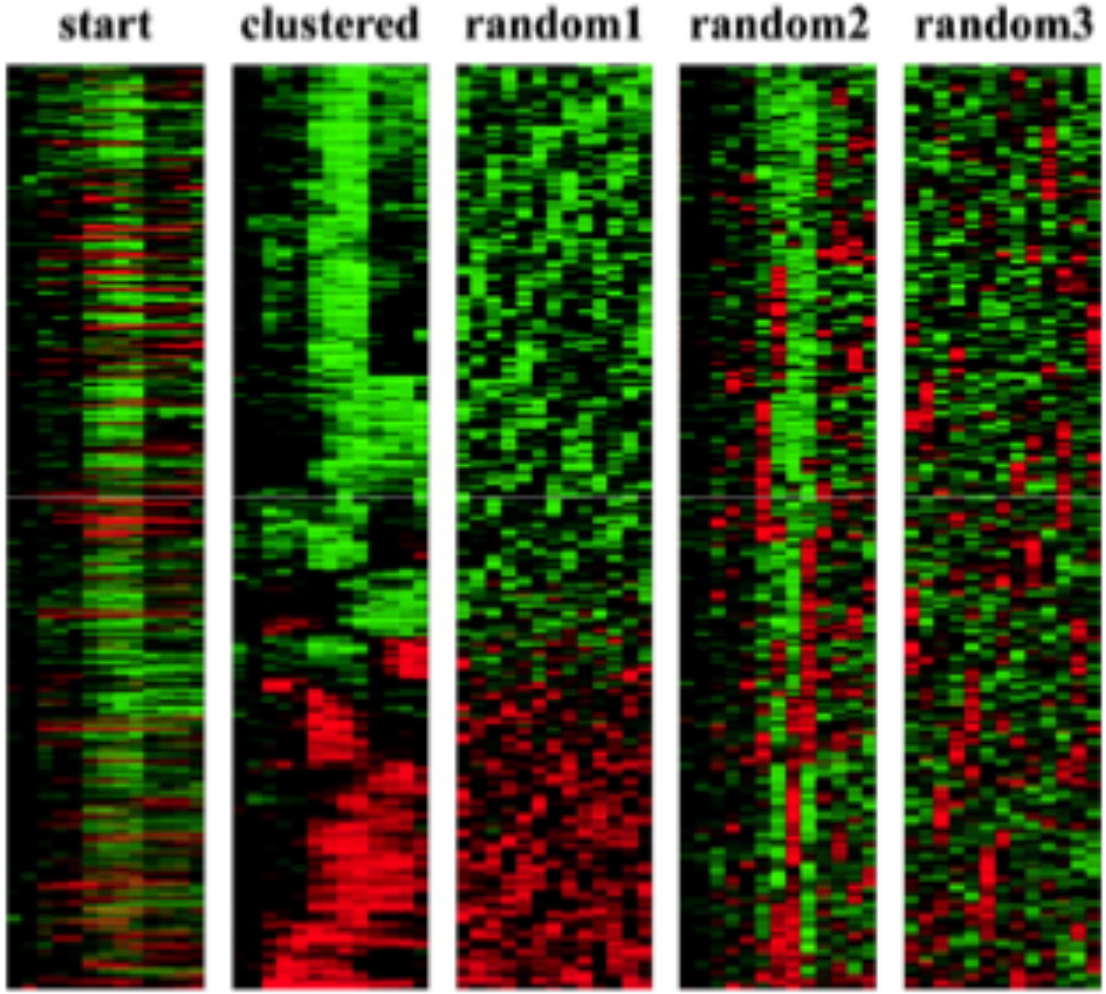
(g) $\rho_{XY} = -\frac{2}{3}$

x	y = 1	y = 2	y = 3	Total
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

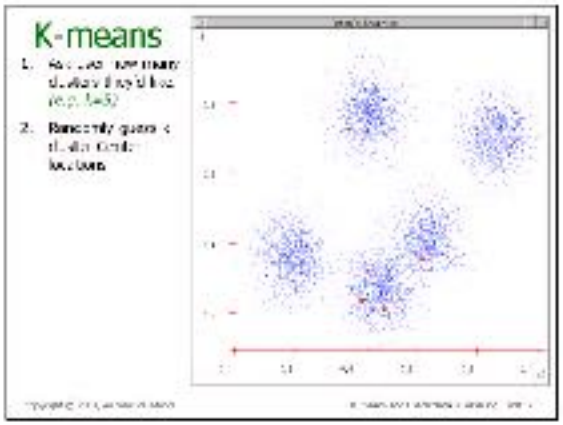


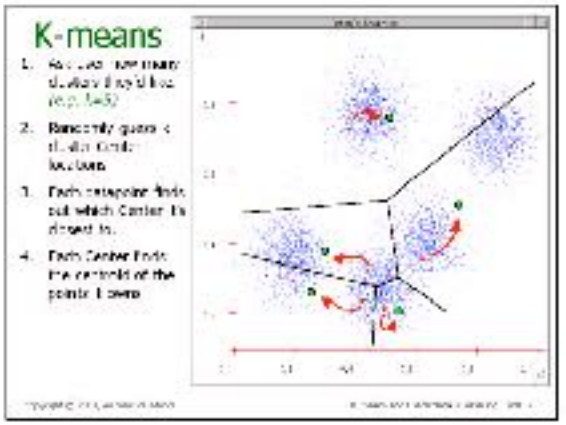
From *Eisen MB, et al, PNAS 1998 95(25):14863 -8*

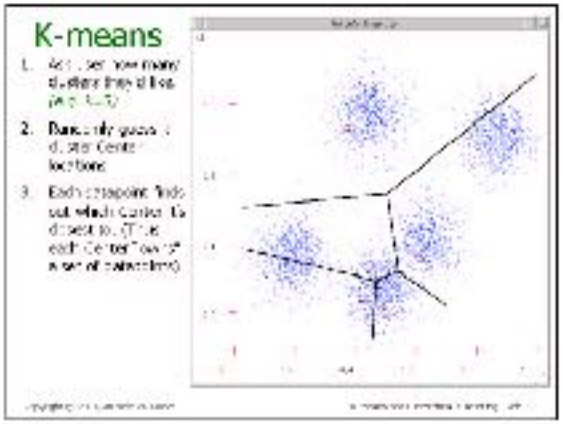
K-Means Clustering: Example

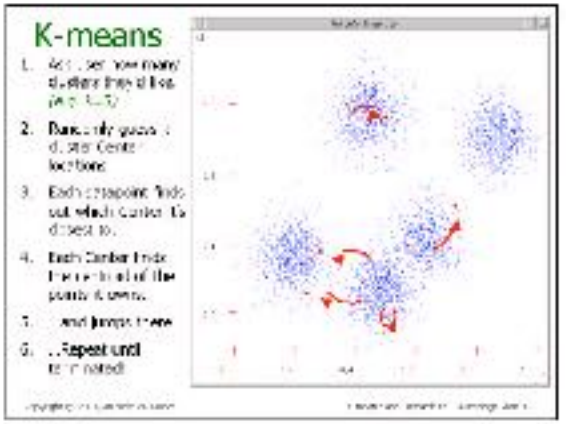
Example from Andrew Moore's tutorial on Clustering.

Start



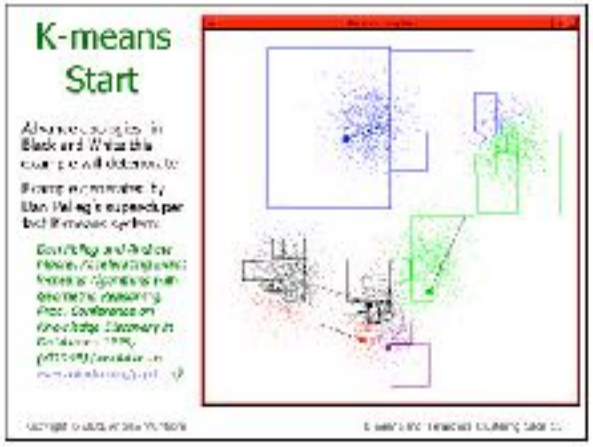


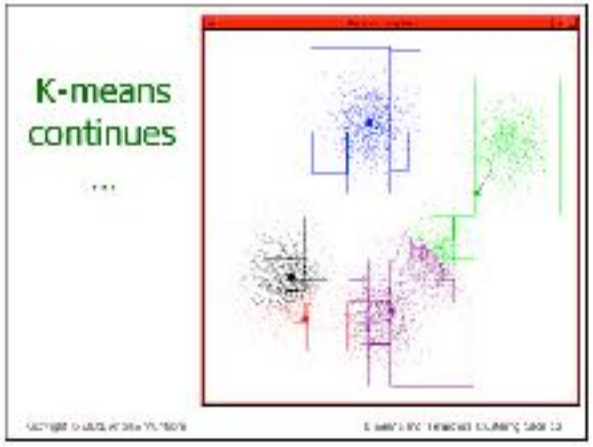


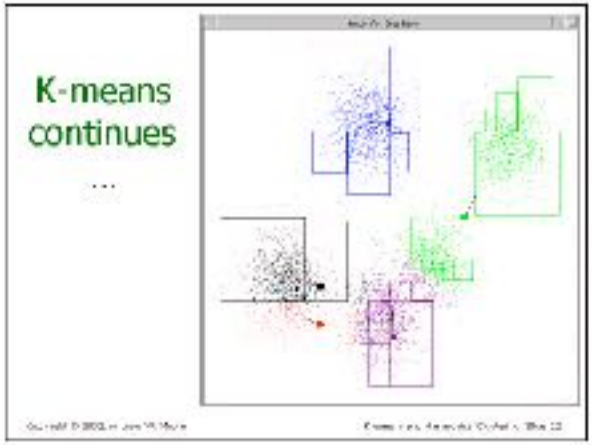


4

5


K-means
Start

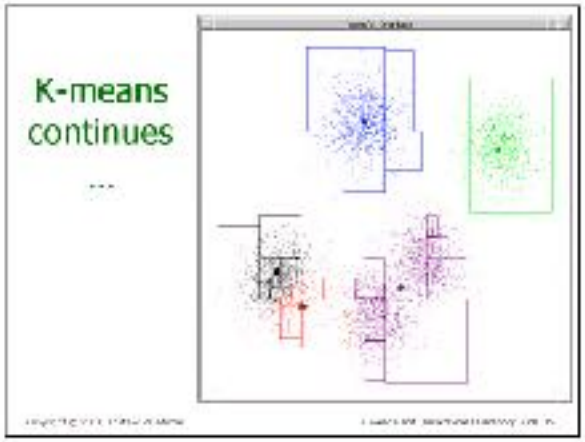

K-means
continues
...

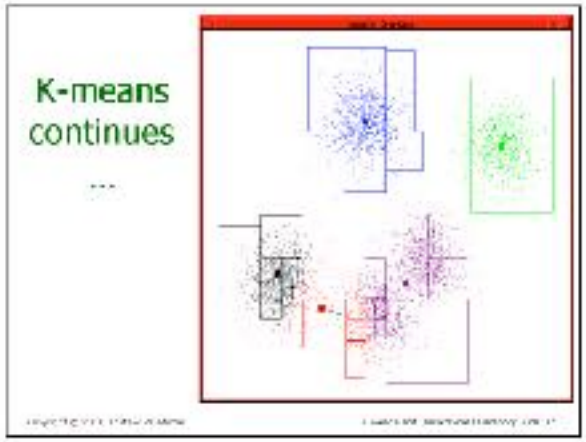

K-means
continues
...

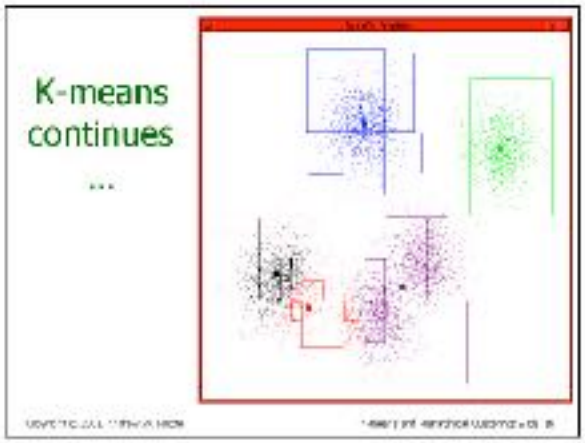

K-means
continues
...

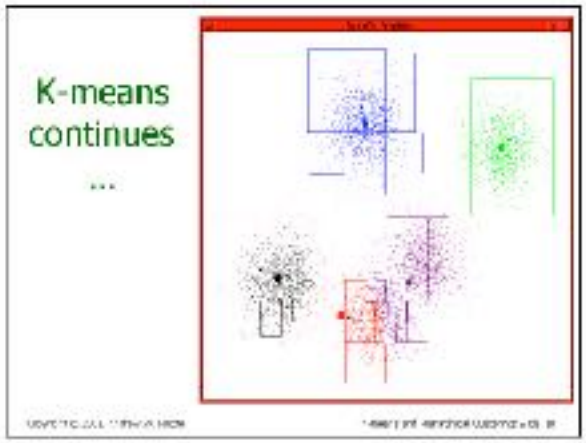
6

7


K-means
continues
...


K-means
continues
...

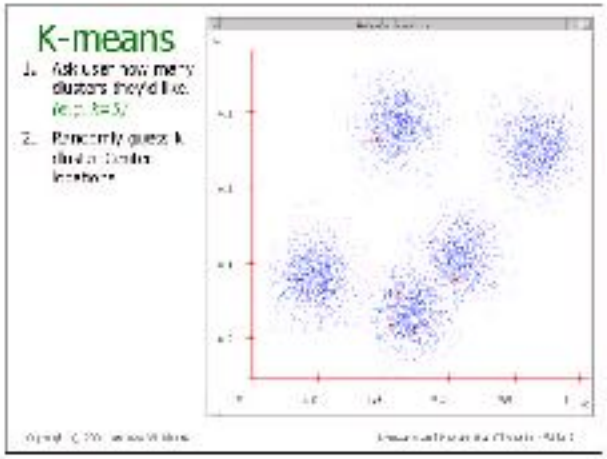

K-means
continues
...

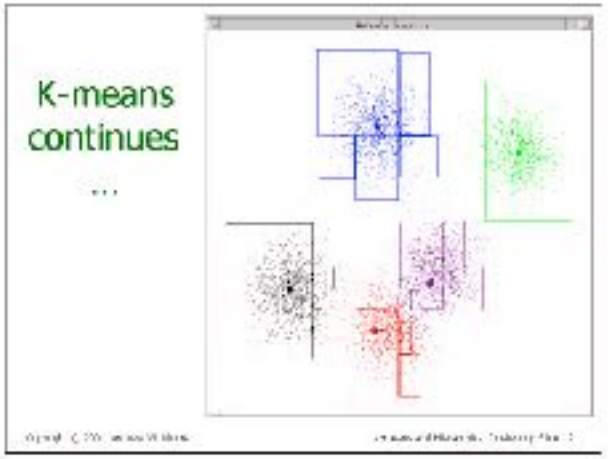

K-means
continues
...

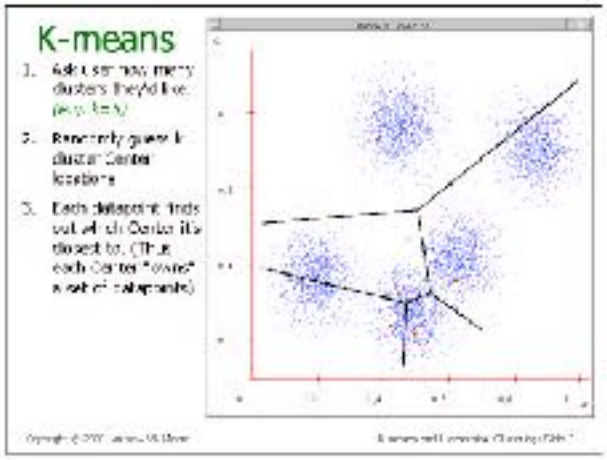
8

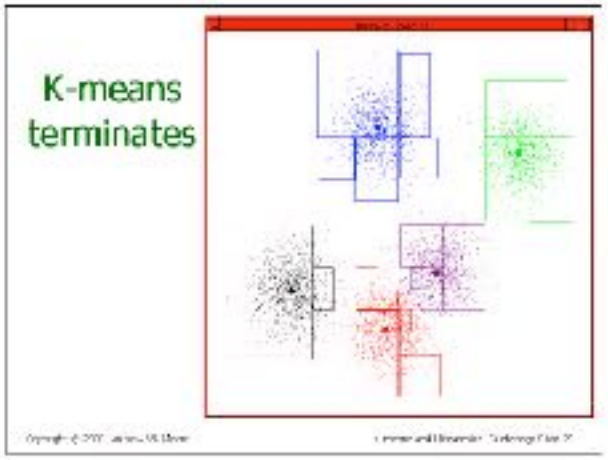
9

Start









End

4

10

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

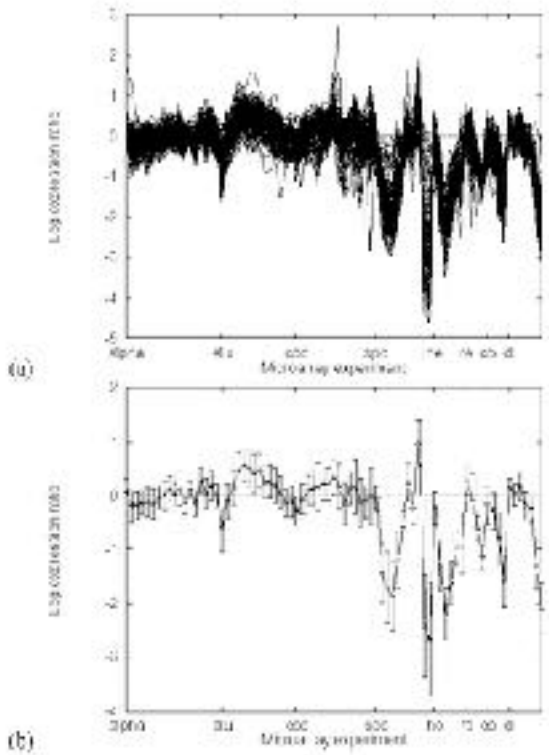


Figure 1 **Expression profiles of the cytoplasmic ribosomal proteins.** Figure (a) shows the expression profiles from the data in [Eisen et al., 1998] of 121 cytoplasmic ribosomal proteins, as classified by MYGD [MYGD, 1999]. The logarithm of the expression ratio is plotted as a function of DNA microarray experiment. Ticks along the X-axis represent the beginnings of experimental series. They are, from left to right, cell division cycle after synchronization with α factor arrest (alpha), cell division cycle after synchronization by centrifugal elutriation (elu), cell division cycle measured using a temperature sensitive *cdc15* mutant (cdc), sporulation (spo), heat shock (ht), reducing shock (re), cold shock (co), and diauxic shift (dif). 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