Clustering

Section 6

J Singh

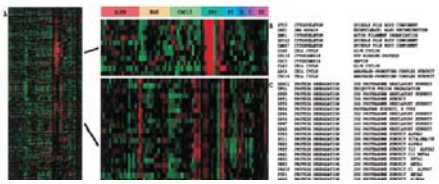
Oct. 28th, 2003

Outline

- Application Area
- Tree Clustering
- K-means Clustering

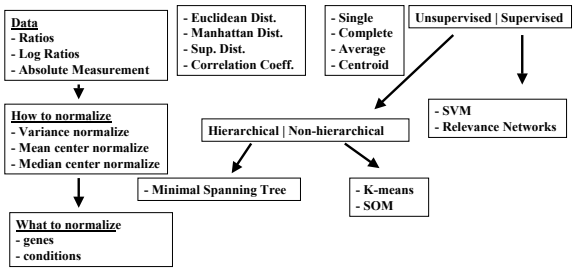
Application Area

Cluster analysis and display of genome-wide expression patterns:
Eisen et al. *Proc. Natl. Acad. Sci.* Vol 95, pp. 14863 – 14868, December 1998. http://www.pnas.org/cgi/reprint/95/25/14863.pdf



Gene Expression Clustering Decision Tree

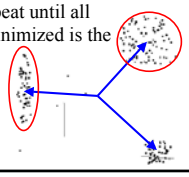
Data Normalization | Distance Metric | Linkage | Clustering Method



Clustering hierarchical & non-

•**Hierarchical:** a series of successive fusions of data until a final number of clusters is obtained; e.g. Minimal Spanning Tree: each component of the population to be a cluster. Next, the two clusters with the minimum distance between them are fused to form a single cluster. Repeated until all components are grouped.

• **Non-:** e.g. K-mean: K clusters chosen such that the points are mutually farthest apart. Each component in the population assigned to one cluster by minimum distance. The centroid's position is recalculated and repeat until all the components are grouped. The criterion minimized is the within-clusters sum of the variance.



Key Terms in Cluster Analysis

- Distance measures
- Similarity measures
- Hierarchical and non-hierarchical
- Single/complete/average linkage
- Dendrogram

Distance Measures: Minkowski Metric

Suppose two objects x and y both have p features :

$$x = (x_1 x_2 \Lambda\ x_p)$$

$$y = (y_1 y_2 \Lambda\ y_p)$$

The Minkowski metric is defined by

$$d(x, y) = \sqrt[r]{\sum_{i=1}^{p} |x_i - y_i|^r}$$

Most Common Minkowski Metrics

1, $r = 2$ (Euclidean distance)

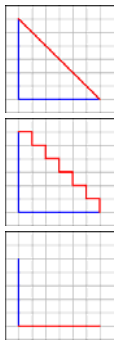
$$d(x, y) = \sqrt[2]{\sum_{i=1}^{p} |x_i - y_i|^2}$$

2, $r = 1$ (Manhattan distance)

$$d(x, y) = \sum_{i=1}^{p} |x_i - y_i|$$

3, $r = +\infty$ ("sup" distance)

$$d(x, y) = \max_{1 \le i \le p} |x_i - y_i|$$



Similarity Measures: Correlation Coefficient

$$s(x, y) = \frac{\sum_{i=1}^{p} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p} (x_i - \bar{x})^2 \times \sum_{i=1}^{p} (y_i - \bar{y})^2}}$$

where $\bar{x} = \frac{1}{p} \sum_{i=1}^{p} x_i$ **and** $\bar{y} = \frac{1}{p} \sum_{i=1}^{p} y_i$.

$$|s(x, y)| \le 1$$

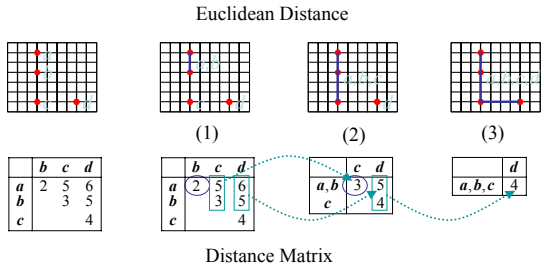
Hierarchical Clustering Techniques

At the beginning, each object (gene) is a cluster. In each of the subsequent steps, two *closest* clusters will merge into one cluster until there is only one cluster left.

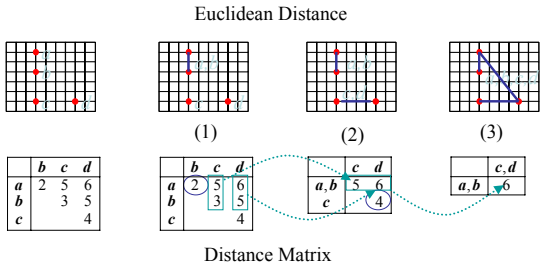
The distance between two clusters is defined as the distance between

- Single-Link Method / Nearest Neighbor: their closest members.
- Complete-Link Method / Furthest Neighbor: their furthest members.
- Centroid: their centroids.
- Average: average of all cross-cluster pairs.

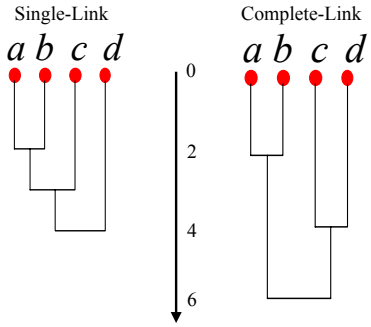
Single-Link Method



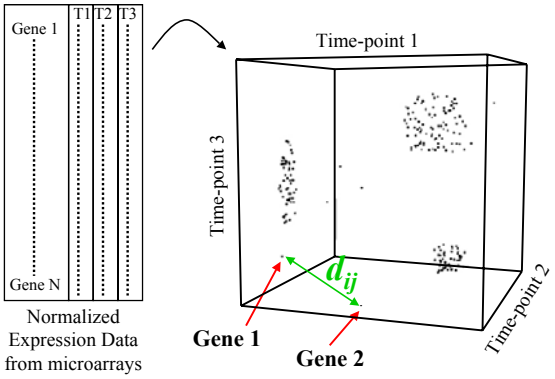
Complete-Link Method



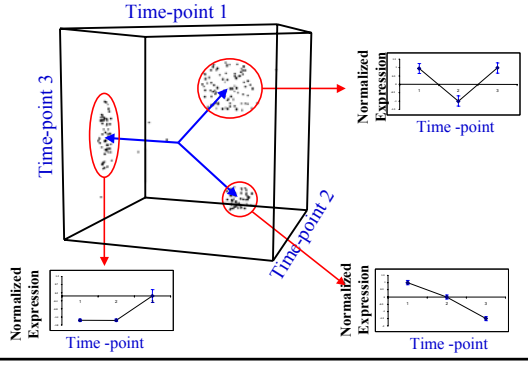
Dendrograms



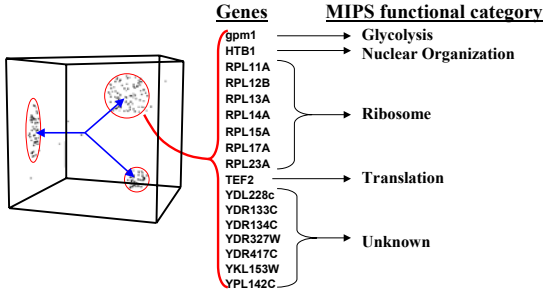
Representation of expression data

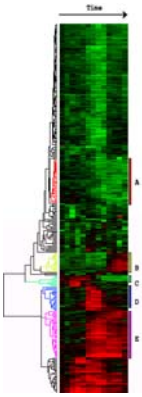


Identifying prevalent expression patterns (gene clusters)



Cluster contents





Eisen *et al.* (1998):

FIG. 1. Cluster display of data from time course of serum stimulation of primary human fibroblasts.

Experiments:
Foreskin fibroblasts were grown in culture and were deprived of serum for 48 hr. Serum was added back and samples taken at time 0, 15 min, 30 min, 1hr, 2 hr, 3 hr, 4 hr, 8 hr, 12 hr, 16 hr, 20 hr, 24 hr.

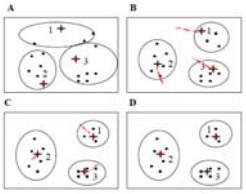
Clustering:
Correlation Coefficient + Centroid Hierarchical Clustering

Clusters:
(A) cholesterol biosynthesis,
(B) the cell cycle,
(C) the immediate-early response,
(D) signaling and angiogenesis,
(E) wound healing and tissue remodeling.

K-means Clustering

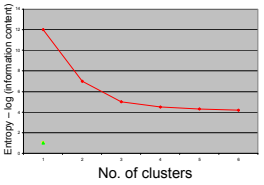
Different approach than Joining

1. Divide the population into *k* clusters
   a. Calculate the mean of each cluster
2. Move objects between the clusters with the goal to minimize variability within clusters and maximize variability between clusters
   a. Calculate distances between means and data points
   b. The data points closest to each mean join its cluster
   c. Recalculate means
3. Repeat until convergence is achieved



What's K?

- Based on analyst's hunch.
- Trial and error
- Graph as a function of the number of clusters: 3 is probably the right number in this case



Next Week

- Hidden Markov Models

Acknowledgement / References

This presentation includes material written by Suzanne Komili.

A number of slides are from the October 21, 2003 Bioinformatics lecture by George Church.

Cluster Analysis, http://www.statsoftinc.com/textbook/stcluan.html