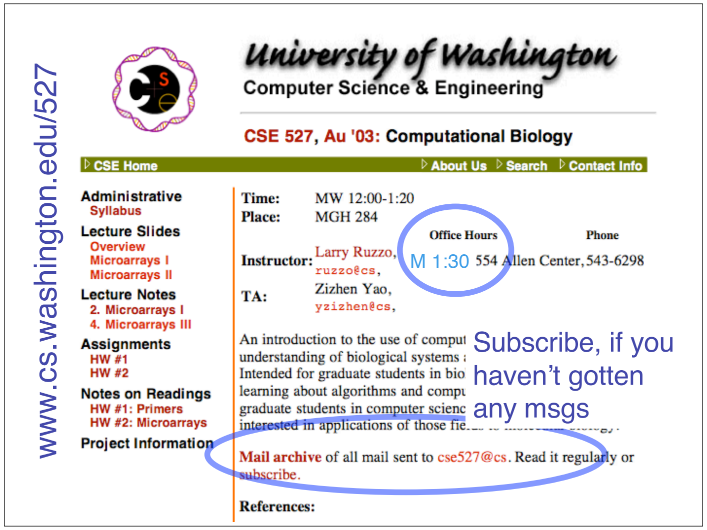


The following slides partly from
http://staff.washington.edu/kayee/research.html
Errors are mine.

Clustering 101

Ka Yee Yeung
Center for Expression Arrays
University of Washington

Overview

- What is clustering?
- Similarity/distance metrics
- Hierarchical clustering algorithms
  - E.g. [Eisen et al. 1998]
- K-means
  - E.g. [MacQueen, 1965] [Tavazoie et al. 1999]
- Self-organizing map (SOM)
  - E.g. [Tamayo et al. 1999]

What is clustering?

- Group *similar* objects together
- Objects in the same cluster (group) are more similar to each other than objects in different clusters
- Data exploratory tool

Clustering Expression Data

- Why cluster gene expression data?
  - Tissue classification } columns
  - Find biologically related genes
  - First step in inferring regulatory networks
  - Look for common promoter elements } rows

  - Hypothesis generation
  - One of the tools of choice for expression analysis

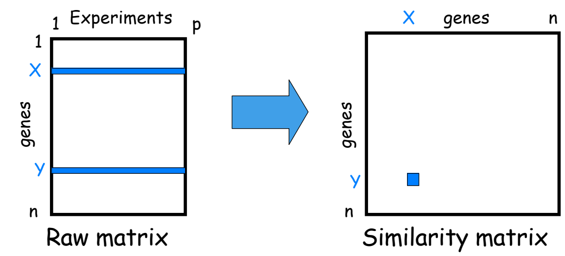
Clustering Expression Data

- What has been done?
  - Partitional
    - CAST (Ben-Dor et al. 1999)
    - k-means, variously initialized (Hartigan 1975)
  - Hierarchical
    - single-, average-, complete-, centroid-link [Eisen et al. 98]
  - Self Organizing Maps (SOM) [Tamayo et al. 99]
  - Support Vector Machines (SVM) [Grundy et al. 00]
  - etc., etc., etc.

Clustering Expression Data

- Why so many methods?
  - Clustering is NP-hard, even with simple objectives, data
  - Hard problem: high dimensionality, noise, …
  - ∴ many heuristic, local search, & approximation algorithms
  - No clear winner

How to define similarity?



- Similarity metric:
  - A measure of pairwise similarity or dissimilarity
  - Examples:
    - Correlation coefficient
    - Euclidean distance

Similarity metrics

- Euclidean distance

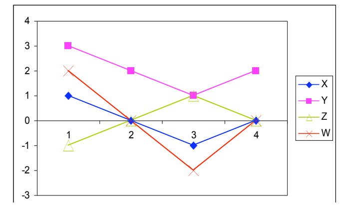
$$\sqrt{\sum_{j=1}^{p}(X[j]-Y[j])^2}$$

- Correlation coefficient

$$\frac{\sum_{j=1}^{p}(X[j]-\overline{X})(Y[j]-\overline{Y})}{\sqrt{\sum_{j=1}^{p}(X[j]-\overline{X})^2\sum_{j=1}^{p}(Y[j]-\overline{Y})^2}}, \quad where\ \overline{X}=\frac{\sum_{j=1}^{p}X[j]}{p}$$

Example

X	1	0	-1	0
Y	3	2	1	2
Z	-1	0	1	0
W	2	0	-2	0



Correlation (X,Y) = 1 Distance (X,Y) = 4

Correlation (X,Z) = -1 Distance (X,Z) = 2.83

Correlation (X,W) = 1 Distance (X,W) = 1.41

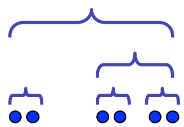
Lessons from the example

- Correlation – direction only
- Euclidean distance – magnitude & direction
- Min # attributes (experiments) to compute pairwise similarity
  - >= 2 attributes for Euclidean distance
  - >= 3 attributes for correlation
- Array data is noisy ➔ need many experiments to robustly estimate pairwise similarity

Clustering algorithms

- Inputs:
  - Raw data matrix or similarity matrix
  - Number of clusters or some other parameters
- Many different classifications of clustering algorithms:
  - Hierarchical vs partitional
  - Heuristic-based vs model-based
  - Soft vs hard

Hierarchical Clustering
[Hartigan 1975]

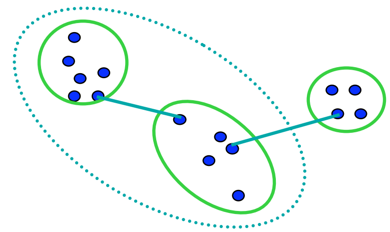


dendrogram

- Agglomerative (bottom-up)
- Algorithm:
  - Initialize: each item a cluster
  - Iterate:
    - select two most similar clusters
    - merge them
  - Halt: when required number of clusters is reached

Hierarchical: Single Link

- cluster similarity = similarity of two most similar members

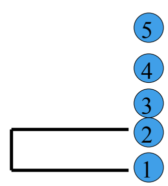


- Potentially long and skinny clusters
+ Fast

Example: single link

$$
\begin{array}{c|c}
 & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix}
\end{array}
\Rightarrow
\begin{array}{c|c}
 & \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\
\begin{array}{c} (1,2) \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 9 & 7 & 0 & \\ 8 & 5 & 4 & 0 \end{bmatrix}
\end{array}
$$

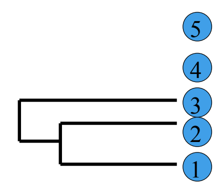
$$d_{(1,2),3} = \min\{ d_{1,3}, d_{2,3}\} = \min\{ 6,3\} = 3$$
$$d_{(1,2),4} = \min\{ d_{1,4}, d_{2,4}\} = \min\{ 10,9\} = 9$$
$$d_{(1,2),5} = \min\{ d_{1,5}, d_{2,5}\} = \min\{ 9,8\} = 8$$



Example: single link

$$
\begin{array}{c|c}
 & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix}
\end{array}
\Rightarrow
\begin{array}{c|c}
 & \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\
\begin{array}{c} (1,2) \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 9 & 7 & 0 & \\ 8 & 5 & 4 & 0 \end{bmatrix}
\end{array}
\Rightarrow
\begin{array}{c|c}
 & \begin{array}{ccc} (1,2,3) & 4 & 5 \end{array} \\
\begin{array}{c} (1,2,3) \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 5 & 4 & 0 \end{bmatrix}
\end{array}
$$

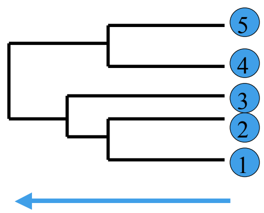
$$d_{(1,2,3),4} = \min\{ d_{(1,2),4}, d_{3,4}\} = \min\{ 9,7\} = 7$$
$$d_{(1,2,3),5} = \min\{ d_{(1,2),5}, d_{3,5}\} = \min\{ 8,5\} = 5$$



Example: single link

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 9 & 7 & 0 & \\ 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2,3) \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccc} (1,2,3) & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 5 & 4 & 0 \end{bmatrix} \end{array}$$

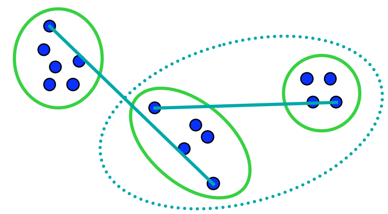
$$d_{(1,2,3),(4,5)} = \min\{ d_{(1,2,3),4}, d_{(1,2,3),5} \} = 5$$



Sometimes drawn to a scale

Hierarchical: Complete Link

- cluster similarity = similarity of two least similar members

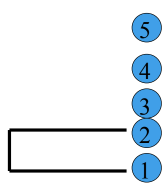


+ tight clusters
- slow

Example: complete link

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 6 & 0 & & \\ 10 & 7 & 0 & \\ 9 & 5 & 4 & 0 \end{bmatrix} \end{array}$$

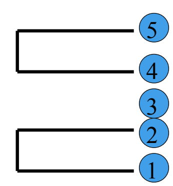
$$d_{(1,2),3} = \max\{ d_{1,3}, d_{2,3} \} = \max\{6,3\} = 6$$
$$d_{(1,2),4} = \max\{ d_{1,4}, d_{2,4} \} = \max\{10,9\} = 10$$
$$d_{(1,2),5} = \max\{ d_{1,5}, d_{2,5} \} = \max\{9,8\} = 9$$



Example: complete link

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 6 & 0 & & \\ 10 & 7 & 0 & \\ 9 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ (4,5) \end{array}\begin{array}{c} \begin{array}{ccc} (1,2) & 3 & (4,5) \end{array} \\ \begin{bmatrix} 0 & & \\ 6 & 0 & \\ 10 & 7 & 0 \end{bmatrix} \end{array}$$

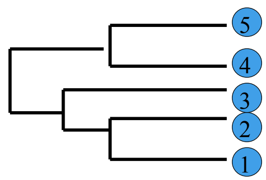
$$d_{(1,2),(4,5)} = \max\{ d_{(1,2),4}, d_{(1,2),5} \} = \max\{10,9\} = 10$$
$$d_{3,(4,5)} = \max\{ d_{3,4}, d_{3,5} \} = \max\{7,5\} = 7$$



Example: complete link

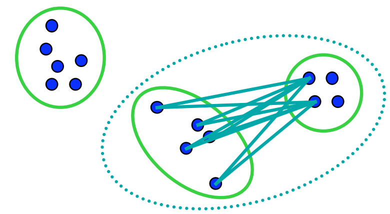
$$
\begin{array}{c|ccccc}
 & 1 & 2 & 3 & 4 & 5 \\
\hline
1 & 0 & & & & \\
2 & 2 & 0 & & & \\
3 & 6 & 3 & 0 & & \\
4 & 10 & 9 & 7 & 0 & \\
5 & 9 & 8 & 5 & 4 & 0
\end{array}
\Rightarrow
\begin{array}{c|cccc}
 & (1,2) & 3 & 4 & 5 \\
\hline
(1,2) & 0 & & & \\
3 & 6 & 0 & & \\
4 & 10 & 7 & 0 & \\
5 & 9 & 5 & 4 & 0
\end{array}
\Rightarrow
\begin{array}{c|ccc}
 & (1,2) & 3 & (4,5) \\
\hline
(1,2) & 0 & & \\
3 & 6 & 0 & \\
(4,5) & 10 & 7 & 0
\end{array}
$$

$$d_{(1,2,3),(4,5)} = \max\{d_{(1,2),(4,5)}, d_{3,(4,5)}\} = 10$$



Hierarchical: Average Link

- cluster similarity = average similarity of all pairs



+ tight clusters
- slow

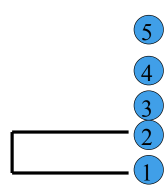
Example: average link

$$
\begin{array}{c|ccccc}
 & 1 & 2 & 3 & 4 & 5 \\
\hline
1 & 0 & & & & \\
2 & 2 & 0 & & & \\
3 & 6 & 3 & 0 & & \\
4 & 10 & 9 & 7 & 0 & \\
5 & 9 & 8 & 5 & 4 & 0
\end{array}
\Rightarrow
\begin{array}{c|cccc}
 & (1,2) & 3 & 4 & 5 \\
\hline
(1,2) & 0 & & & \\
3 & 4.5 & 0 & & \\
4 & 9.5 & 7 & 0 & \\
5 & 8.5 & 5 & 4 & 0
\end{array}
$$

$$d_{(1,2),3} = \frac{1}{2}(d_{1,3} + d_{2,3}) = \frac{6+3}{2} = 4.5$$

$$d_{(1,2),4} = \frac{1}{2}(d_{1,4} + d_{2,4}) = \frac{10+9}{2} = 9.5$$

$$d_{(1,2),5} = \frac{1}{2}(d_{1,5} + d_{2,5}) = \frac{9+8}{2} = 8.5$$

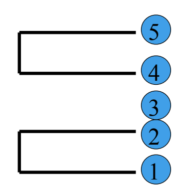


Example: average link

$$
\begin{array}{c|ccccc}
 & 1 & 2 & 3 & 4 & 5 \\
\hline
1 & 0 & & & & \\
2 & 2 & 0 & & & \\
3 & 6 & 3 & 0 & & \\
4 & 10 & 9 & 7 & 0 & \\
5 & 9 & 8 & 5 & 4 & 0
\end{array}
\Rightarrow
\begin{array}{c|cccc}
 & (1,2) & 3 & 4 & 5 \\
\hline
(1,2) & 0 & & & \\
3 & 4.5 & 0 & & \\
4 & 9.5 & 7 & 0 & \\
5 & 8.5 & 5 & 4 & 0
\end{array}
\Rightarrow
\begin{array}{c|ccc}
 & (1,2) & 3 & (4,5) \\
\hline
(1,2) & 0 & & \\
3 & 4.5 & 0 & \\
(4,5) & 9 & 6 & 0
\end{array}
$$

$$d_{(1,2),(4,5)} = \frac{1}{4}(d_{1,4} + d_{1,5} + d_{2,4} + d_{2,5}) = 9$$

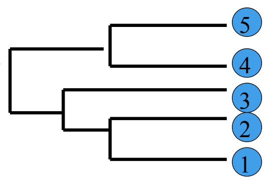
$$d_{3,(4,5)} = \frac{1}{2}(d_{3,4} + d_{3,5}) = 6$$



Example: average link

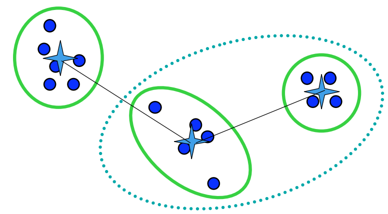
$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 4.5 & 0 & & \\ 4 & 9.5 & 7 & 0 & \\ 5 & 8.5 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2) & 3 & (4,5) \\ \hline (1,2) & 0 & & \\ 3 & 4.5 & 0 & \\ (4,5) & 9 & 6 & 0 \end{array}$$

$$d_{(1,2,3),(4,5)} = \frac{1}{6}(d_{1,4}+d_{1,5}+d_{2,4}+d_{2,5}+d_{3,4}+d_{3,5}) = 8$$



Hierarchical: Centroid Link

- cluster centroid = average of all points
- cluster similarity = distance between centroids



In Expression literature, often called "Average link"

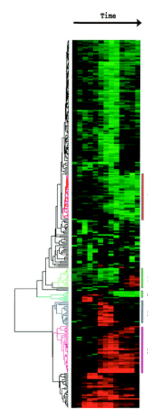
+ faster

- discards shape

Algorithm Analysis

(see class notes)

Software: TreeView [Eisen et al. 1998]



- Fig 1 in Eisen's PNAS 99 paper
- Time course of serum stimulation of primary human fibrolasts
- cDNA arrays with approx 8600 spots
- centroid-link
- Free download at: http://rana.lbl.gov/EisenSoftware.htm

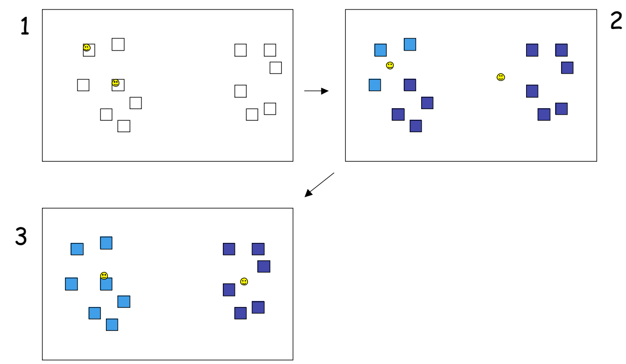
- **Another Good Package: TMEV**
  - http://www.tigr.org/software/tm4/

Hierarchical divisive clustering algorithms

- Top down
  - Start with all the objects in one cluster
  - Successively split into smaller clusters
- Tend to be less efficient than agglomerative
- Resolver implemented a deterministic annealing approach from [Alon et al. 1999]

Partitional: K-Means

[MacQueen 1965]



Details of k-means

- Iterate until converge:
  - Assign each data point to the closest centroid
  - Compute new centroid

Objective function:

Minimize

$$\sum_{x} \left(x - Centroid(Cluster(x))\right)^2$$

Properties of k-means

- Fast
- Proved to converge to local optimum
- In practice, converge quickly
- Tend to produce spherical, equal-sized clusters
- Related to the model-based approach (next lecture)

Summary

- Definition of clustering
- Pairwise similarity:
  - Correlation
  - Euclidean distance
- Clustering algorithms:
  - Hierarchical (single-, complete-, average-, centroid-ink)
  - K-means
  - SOM
- Different clustering algorithms ➔ different clusters

Misc Notes

- Greedy algorithms. Can get trapped in local minima. Can be sensitive to addition of new points, order of points,…

+ simple, intuitive algorithms, reasonably fast, ok on simple data, no obvious preconception about structure

\- no model of structure; biases unclear