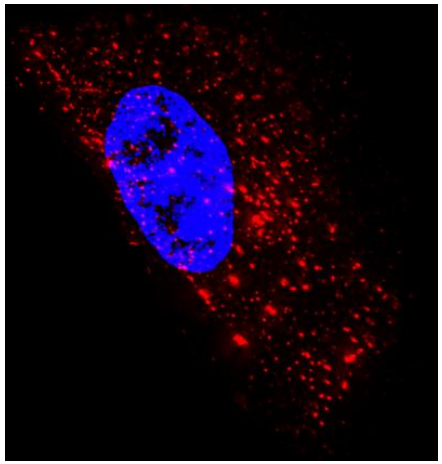
Gene Expression Assays

RNA-seq: What Is It Good For?

High-throughput RNA sequencing experiments (RNA-seq) offer the ability to measure simultaneously the expression level of thousands of genes in a single experiment!

RNA



What Are We "Comparing"?

- Different cell populations
- Pre-treatment vs. post-treatment
- Cell cycle variations
- Environmental response of cells
- Genetically heterogeneous diseases (cancers, heart disease, multiple sclerosis, diabetes, etc.)
- DNA and RNA interactions

Expression Assay Limitations

- Gene expression may not be indicative of protein expression
- Error and variability in results
  - Not all mRNA is reverse transcribed to cDNA with the same efficiency
  - Sequencing errors
  - mRNAs degrade at different rates
  - Post-transcriptional regulation

Data… And Lots of It!

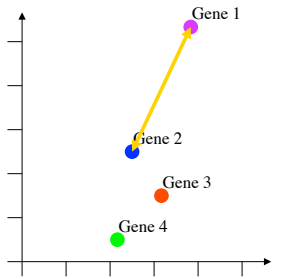
	Experiment 1	Experiment 2	Experiment 3	Experiment 4	…	Experiment *n*-1	Experiment *n*
Gene 1	0.6	4.4	1.3	1.0	…	3.1	2.2
Gene 2	1.5	2.6	5.2	0.8	…	2.8	2.9
Gene 3	0.7	3.7	2.4	1.9	…	1.5	1.6
Gene 4	0.3	0.7	0.2	1.3	…	4.9	3.0
Gene 5	3.1	3.0	2.1	1.4	…	4.2	0.9
…	…	…	…	…	…	…	…
Gene *n*-1	1.8	2.5	1.8	0.7	…	2.7	3.1
Gene *n*	0.5	3.4	3.0	0.5	…	1.8	2.5

Finding Similarly Expressed Genes

- It may be useful to partition the *n* genes into groups of similarly expressed genes
- Clustering is the art of finding groups of genes, such that genes in the same group are as similar to each other as possible and as dissimilar to genes in other groups as possible

Distance Measure

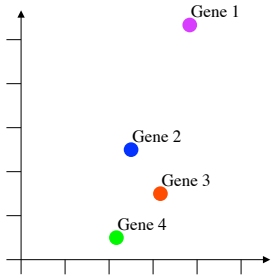
	Experiment 1	Experiment 2
Gene 1	3.8	5.4
Gene 2	2.6	2.6
Gene 3	3.1	1.5
Gene 4	2.1	0.5



$$distance(\text{Gene 1}, \text{Gene 2}) = \sqrt[2]{(3.8 - 2.6)^2 + (5.4 - 2.6)^2}$$

Distance Measure - L^2 Norm

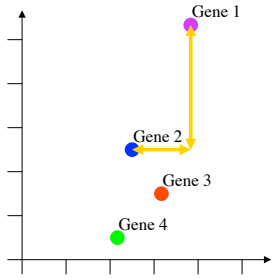
	Experiment 1	Experiment 2
Gene 1	3.8	5.4
Gene 2	2.6	2.6
Gene 3	3.1	1.5
Gene 4	2.1	0.5



$$distance(\text{Gene } a, \text{Gene } b) = \sqrt[2]{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

Distance Measure

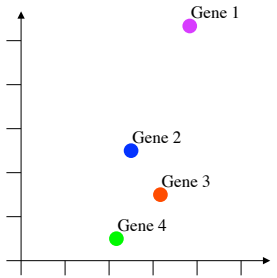
	Experiment 1	Experiment 2
Gene 1	3.8	5.4
Gene 2	2.6	2.6
Gene 3	3.1	1.5
Gene 4	2.1	0.5



$$distance(\text{Gene 1}, \text{Gene 2}) = \sqrt[1]{(3.8 - 2.6)^1 + (5.4 - 2.6)^1}$$

Distance Measure - L^1 Norm

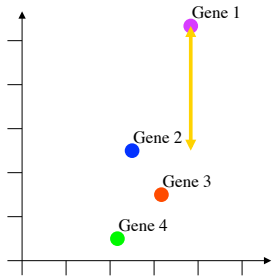
	Experiment 1	Experiment 2
Gene 1	3.8	5.4
Gene 2	2.6	2.6
Gene 3	3.1	1.5
Gene 4	2.1	0.5



$$distance(\text{Gene } a, \text{Gene } b) = \sqrt[1]{(a_1 - b_1)^1 + (a_2 - b_2)^1}$$
$$= (a_1 - b_1) + (a_2 - b_2)$$

Distance Measure

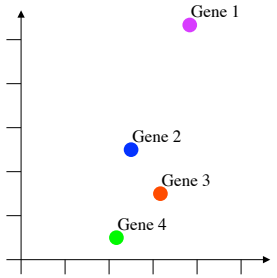
	Experiment 1	Experiment 2
Gene 1	3.8	5.4
Gene 2	2.6	2.6
Gene 3	3.1	1.5
Gene 4	2.1	0.5



$$distance(\text{Gene 1}, \text{Gene 2}) = \sqrt[\infty]{(3.8 - 2.6)^\infty + (5.4 - 2.6)^\infty}$$

Distance Measure - L^∞ Norm

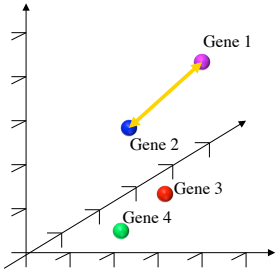
	Experiment 1	Experiment 2
Gene 1	3.8	5.4
Gene 2	2.6	2.6
Gene 3	3.1	1.5
Gene 4	2.1	0.5



$$distance(\text{Gene } a, \text{Gene } b) = \sqrt[\infty]{(a_1 - b_1)^\infty + (a_2 - b_2)^\infty}$$
$$= \max\{(a_1 - b_1), (a_2 - b_2)\}$$

Distance Measure

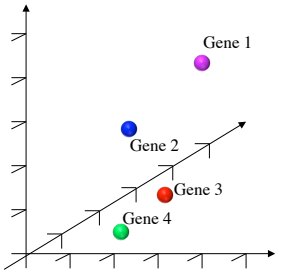
	Experiment 1	Experiment 2	Experiment 3
Gene 1	3.8	5.4	4.7
Gene 2	2.6	2.6	2.6
Gene 3	3.1	1.5	2.2
Gene 4	2.1	0.5	1.2



$$distance(\text{Gene } 1, \text{Gene } 2) = \sqrt[2]{(3.8 - 2.6)^2 + (5.4 - 2.6)^2 + (4.7 - 2.6)^2}$$

Distance Measure - L^2 Norm

	Experiment 1	Experiment 2	Experiment 3
Gene 1	3.8	5.4	4.7
Gene 2	2.6	2.6	2.6
Gene 3	3.1	1.5	2.2
Gene 4	2.1	0.5	1.2



$$distance(\text{Gene } a, \text{Gene } b) = \sqrt[2]{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$

Distance Measure - L^2 Norm

	Experiment 1	Experiment 2	Experiment 3	Experiment 4	⋱	Experiment m
Gene 1	3.8	5.4	4.7	5.0	…	4.2
Gene 2	2.6	2.6	2.6	2.6	…	2.6
Gene 3	3.1	1.5	2.2	1.9	…	2.7
Gene 4	2.1	0.5	1.2	0.9	…	1.7

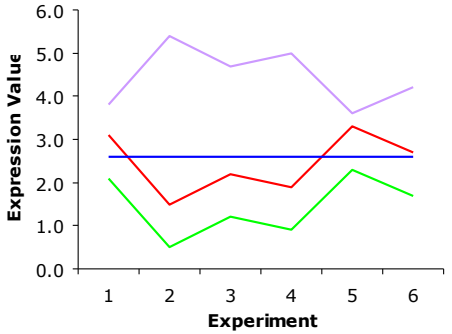
$$distance(\text{Gene } a, \text{Gene } b) = \sqrt[2]{\sum_{i=1}^{m} (a_i - b_i)^2}$$

Distance vs. Similarity Measure

	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6
Gene 1	3.8	5.4	4.7	5.0	3.6	4.2
Gene 2	2.6	2.6	2.6	2.6	2.6	2.6
Gene 3	3.1	1.5	2.2	1.9	3.3	2.7
Gene 4	2.1	0.5	1.2	0.9	2.3	1.7

Similarity Measure

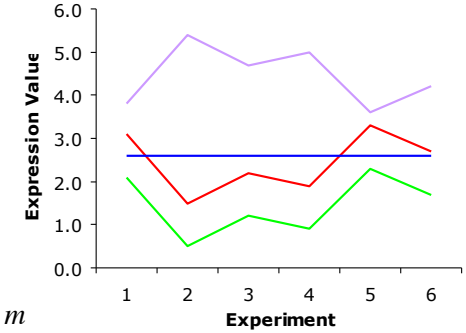
	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6
Gene 1	3.8	5.4	4.7	5.0	3.6	4.2
Gene 2	2.6	2.6	2.6	2.6	2.6	2.6
Gene 3	3.1	1.5	2.2	1.9	3.3	2.7
Gene 4	2.1	0.5	1.2	0.9	2.3	1.7



Correlation Measure

$\rho_{1,4} = -1.0$ $\rho_{2,4} = 0.0$ $\rho_{3,4} = 1.0$ $\rho_{4,4} = 1.0$

	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6
Gene 1	3.8	5.4	4.7	5.0	3.6	4.2
Gene 2	2.6	2.6	2.6	2.6	2.6	2.6
Gene 3	3.1	1.5	2.2	1.9	3.3	2.7
Gene 4	2.1	0.5	1.2	0.9	2.3	1.7



$$\rho_{a,b} = \frac{\sum_{i=1}^{m} a_i b_i - (\sum_{i=1}^{m} a_i)(\sum_{i=1}^{m} b_i) / m}{\sqrt[2]{(\sum_{i=1}^{m} a_i^2 - (\sum_{i=1}^{m} a_i)^2/m)\,(\sum_{i=1}^{m} b_i^2 - (\sum_{i=1}^{m} b_i)^2/m)}}$$

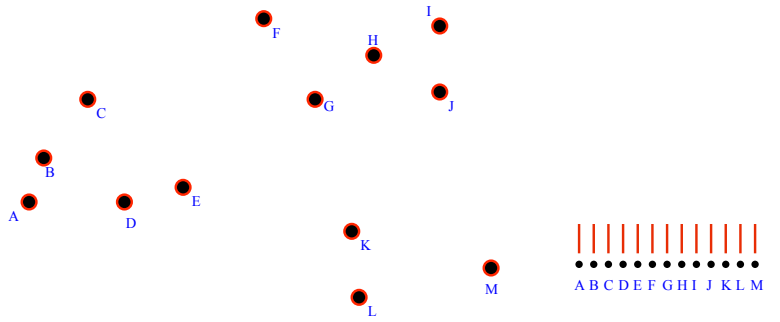
Example with 2 Experiments

	Experiment 1	Experiment 2
Gene 1	0.6	4.4
Gene 2	1.5	2.6
Gene 3	0.7	3.7
Gene 4	0.3	0.7
Gene 5	3.1	3.0
…	…	…
Gene *n*-1	1.8	2.5
Gene *n*	0.5	3.4

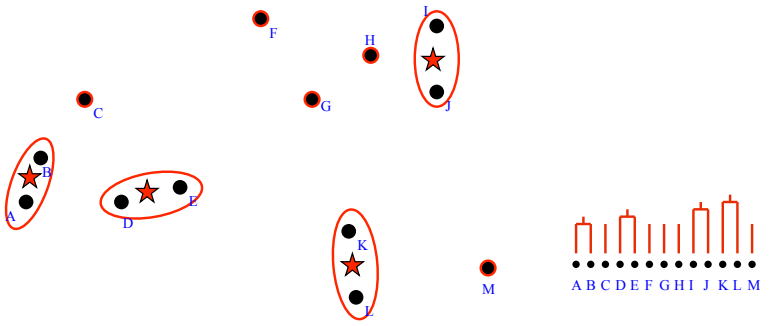
Hierarchical Clustering Algorithm

- Assign each point to its own cluster
- Repeat the following step until the desired number of clusters is reached
  - Merge together the two closest clusters

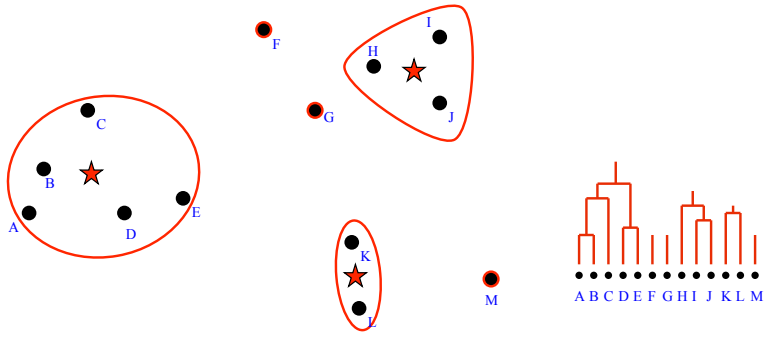
Hierarchical Clustering



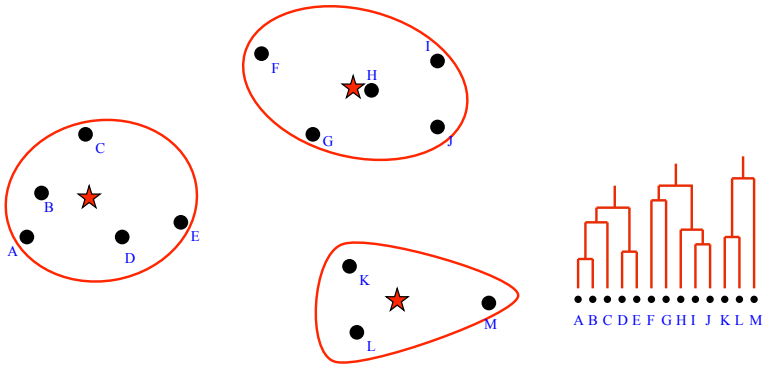
Hierarchical Clustering



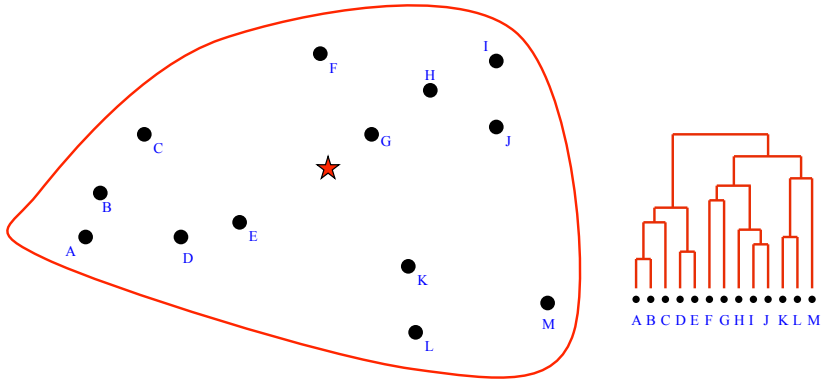
Hierarchical Clustering



Hierarchical Clustering



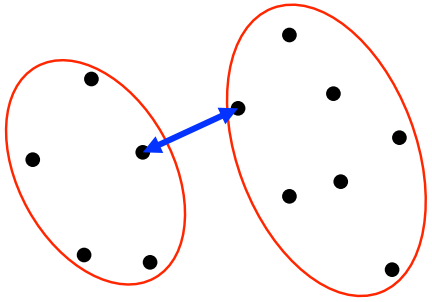
Hierarchical Clustering



Variations on Measure of "Distance" Between Two Clusters

- Single-linkage

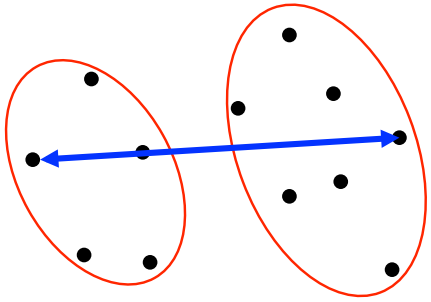
  The distance between two clusters is the distance between the closest pair of points (one from each cluster) in the clusters



Variations on Measure of "Distance" Between Two Clusters

- Complete-linkage

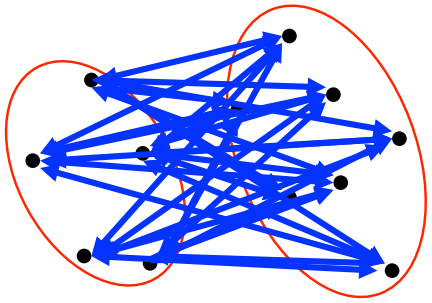
  The distance between two clusters is the distance between the farthest pair of points (one from each cluster) in the clusters



Variations on Measure of "Distance" Between Two Clusters

- Average-linkage

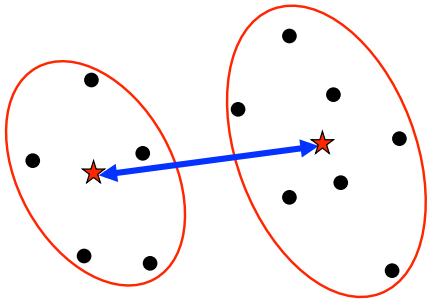
  The distance between two clusters is the average distance between all pairs of points (one from each cluster) in the clusters



Variations on Measure of "Distance" Between Two Clusters

- Centroid-linkage

  The distance between two clusters is the distance between the centroids (centers) of each cluster



Assessing Clustering

- Figure of Merit (FOM)

  Apply clustering algorithm to all but one experimental condition and use the left-out condition to assess the predictive power of the clustering algorithm

Assessing Clustering

- Rand Index

  Measures similarity of two clusterings, *X* and *Y*

$$R = \frac{a + b}{a + b + c + d}$$

a = number of pairs of points in the same cluster in *X* and in the same cluster in *Y*

b = number of pairs of points in different clusters in *X* and in different clusters in *Y*

c = number of pairs of points in the same cluster in *X* and in different clusters in *Y*

d = number of pairs of points in different clusters in *X* and in the same cluster in *Y*