Clustering

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!

Data… And Lots of It!

	Experiment 1	Experiment 2	Experiment 3	Experiment 4	…	Experiment *m*-1	Experiment *m*
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

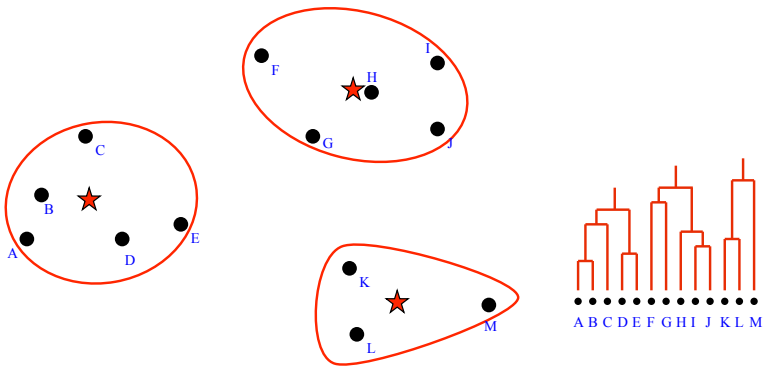
Clustering Algorithms

- Hierarchical clustering
- CAST
- *k*-means clustering
- Model-based clustering

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

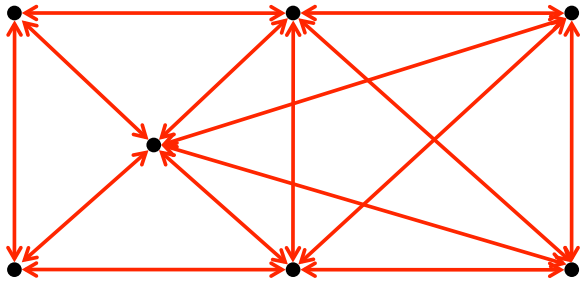
Hierarchical Clustering



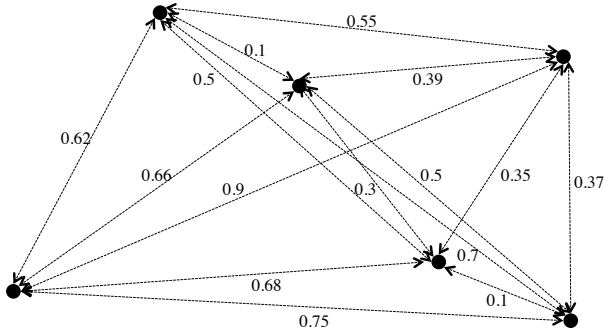
Clustering Algorithms

- Hierarchical clustering
- CAST
- *k*-means clustering
- Model-based clustering

Cliques in Graphs

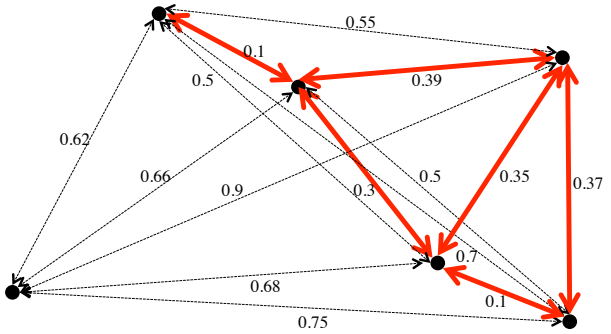


Cluster Affinity Search Technique (CAST)



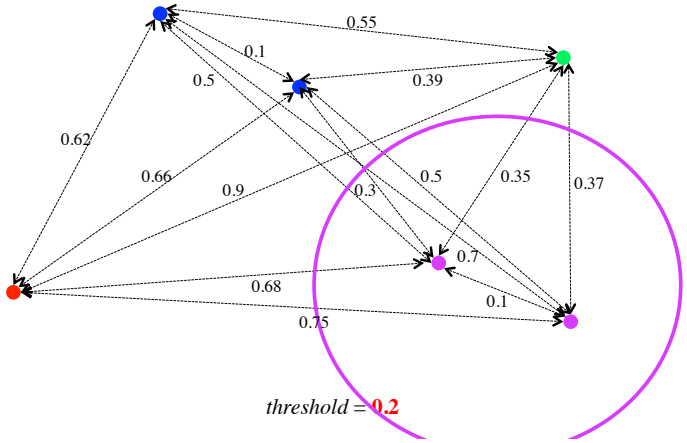
threshold = **0.0**

Cluster Affinity Search Technique (CAST)



threshold = **0.4**

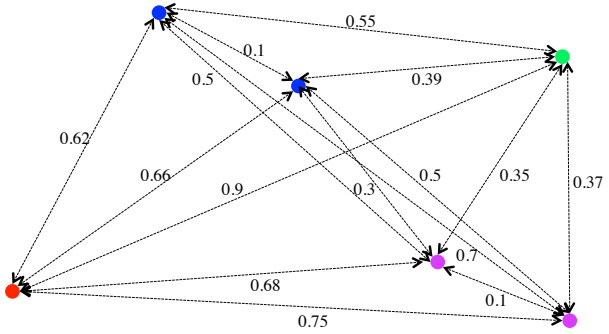
Cluster Affinity Search Technique (CAST)



threshold = **0.2**

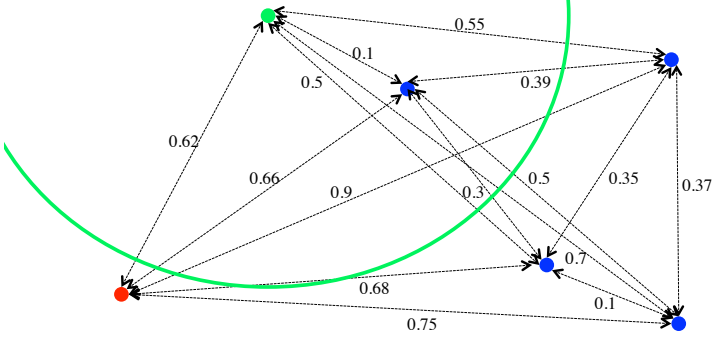
Cluster Affinity Search Technique (CAST)

4 Clusters



threshold = **0.2**

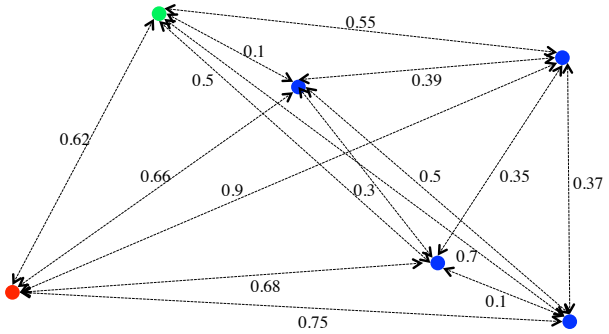
Cluster Affinity Search Technique (CAST)



threshold = **0.45**

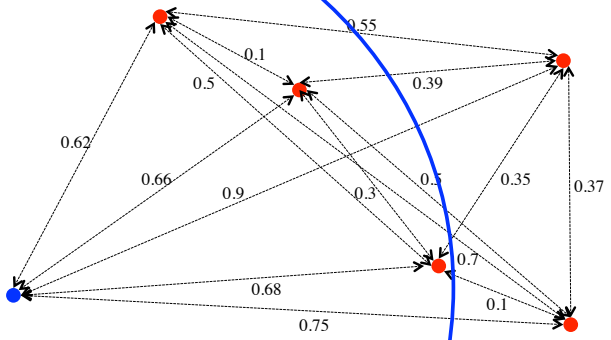
Cluster Affinity Search Technique (CAST)

3 Clusters



threshold = **0.45**

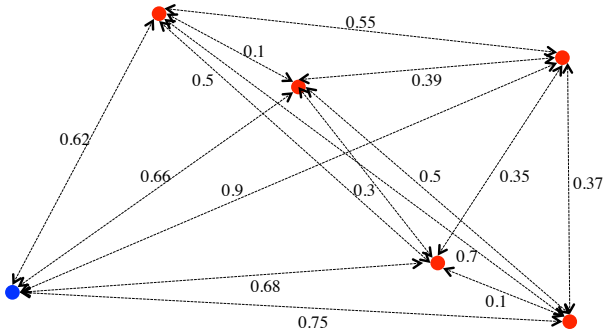
Cluster Affinity Search Technique (CAST)



threshold = **0.7**

Cluster Affinity Search Technique (CAST)

2 Clusters



threshold = **0.7**

CAST Clustering Algorithm

- Repeat until all points (genes) are assigned to a cluster
  - Choose a point (gene) not already assigned to a cluster and assign it to a new cluster *C*
  - Repeat until the cluster *C* converges
    - Add to *C* any unassigned points (genes) that are closer to the points (genes) in *C*, on average, than some *threshold*
    - Remove from *C* any points (genes) that are farther from the other points (genes) in *C*, on average, than some *threshold*

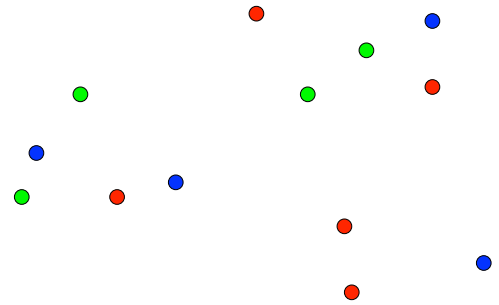
Clustering Algorithms

- Hierarchical clustering
- CAST
- *k*-means clustering
- Model-based clustering

k-means Clustering Algorithm

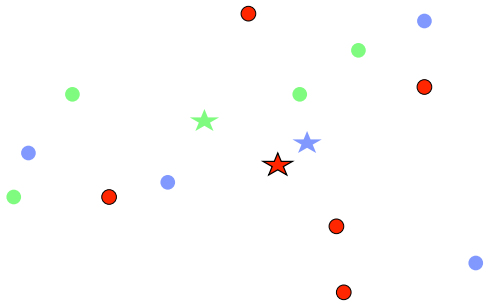
- Randomly assign each point (gene) to one of *k* clusters
- Repeat until convergence
  - Calculate *mean* of each of the *k* clusters
  - Assign each point (gene) to the cluster with the closest *mean*

k-means Clustering Example



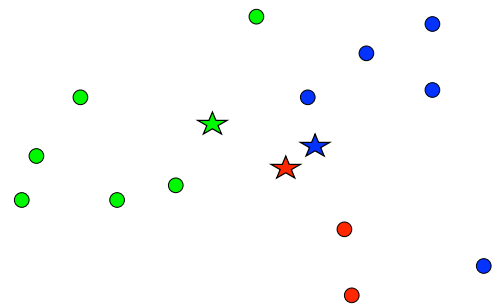
Randomly assign each point to one of *k* clusters

k-means Clustering Example



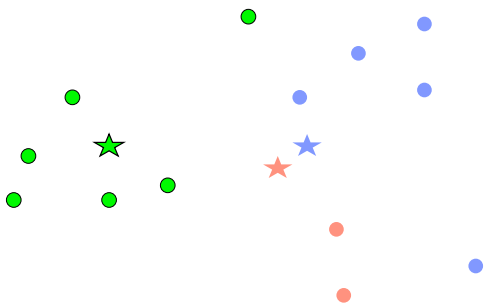
Calculate mean of each cluster

k-means Clustering Example



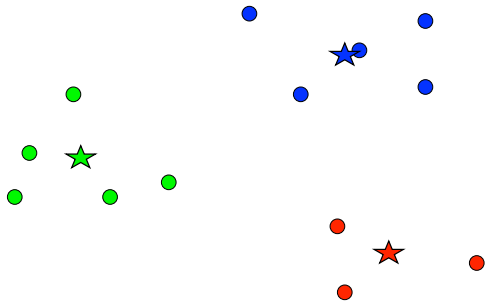
Assign each point to closest cluster mean

k-means Clustering Example



Calculate mean of each cluster

k-means Clustering Example



Convergence

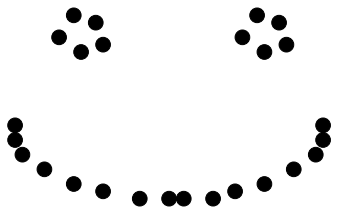
How Good Is Our Clustering?

For a given number of clusters, *k*, one measure of a clustering's quality is the sum of the distances between each point and the mean of the point's cluster

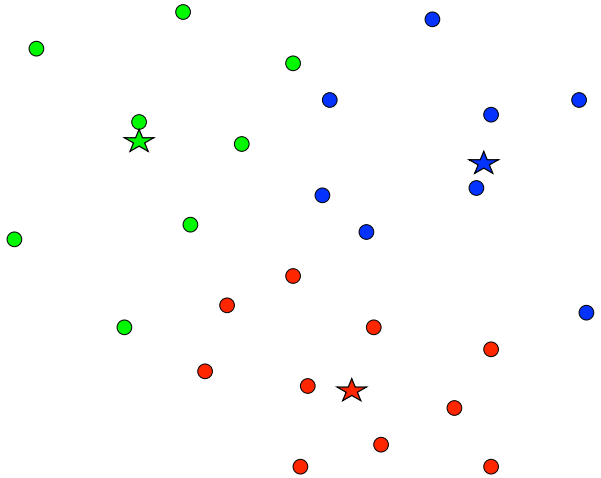
Clustering Problem

- *Clustering Problem*: Partition *n* data points into *k* clusters such that the total distance from each point to its cluster center is minimized.
- Clustering is an NP-complete problem

Does *k*-means Always Work?



Does *k*-means Always Work?



Clustering Algorithms

- Hierarchical clustering
- CAST
- *k*-means clustering
- Model-based clustering

Model-Based Clustering

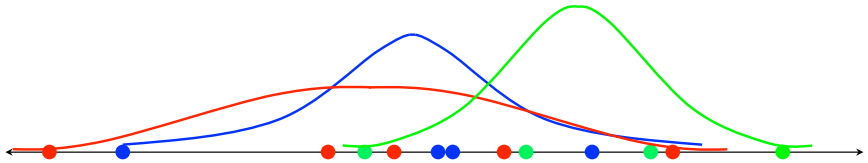
- Randomly assign each point (gene) to one of *k* clusters
- Repeat until convergence
  - Calculate *model* of each of the *k* clusters
  - Assign each point (gene) to the cluster with the closest *model*

Model-Based Clustering

Randomly assign each point to one of *k* clusters (models)

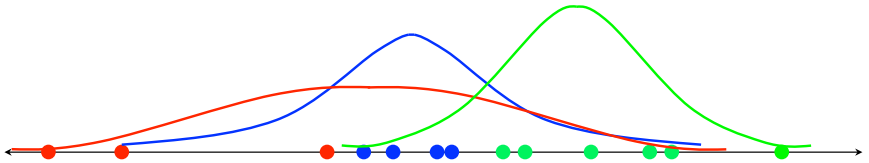
Model-Based Clustering

Calculate model for each of the *k* clusters



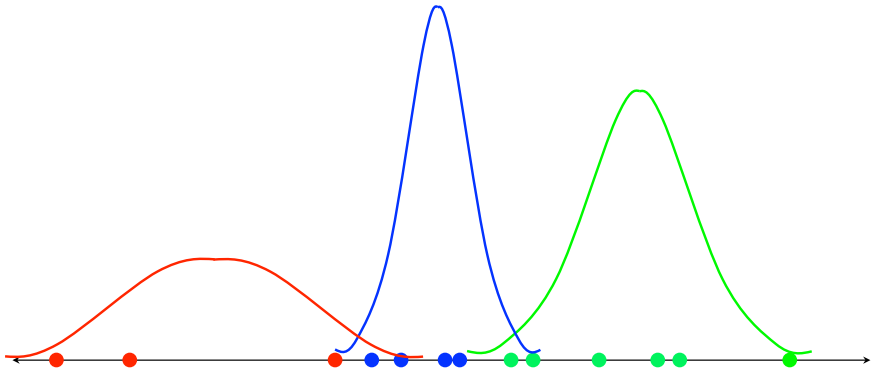
Model-Based Clustering

Assign each point to the most likely model



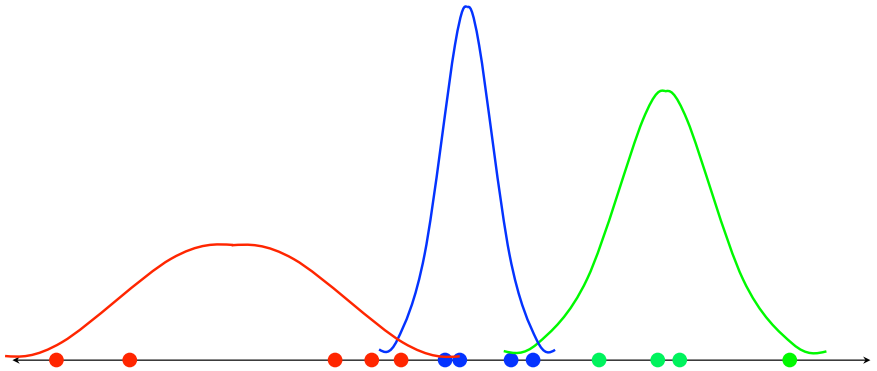
Model-Based Clustering

Calculate model for each of the *k* clusters



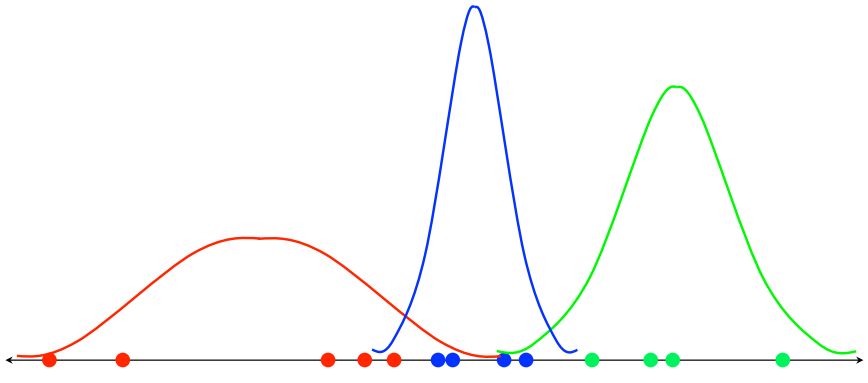
Model-Based Clustering

Assign each point to the most likely model



Model-Based Clustering

Calculate model for each of the *k* clusters



Clustering Genes vs. Clustering Experiments

	Experiment 1	Experiment 2	Experiment 3	Experiment 4	⋱	Experiment *m*-1	Experiment *m*
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

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*