Clustering and Cluster Evaluation

Josh Stuart

Tuesday, Feb 24, 2004

Read chap 4 in Causton

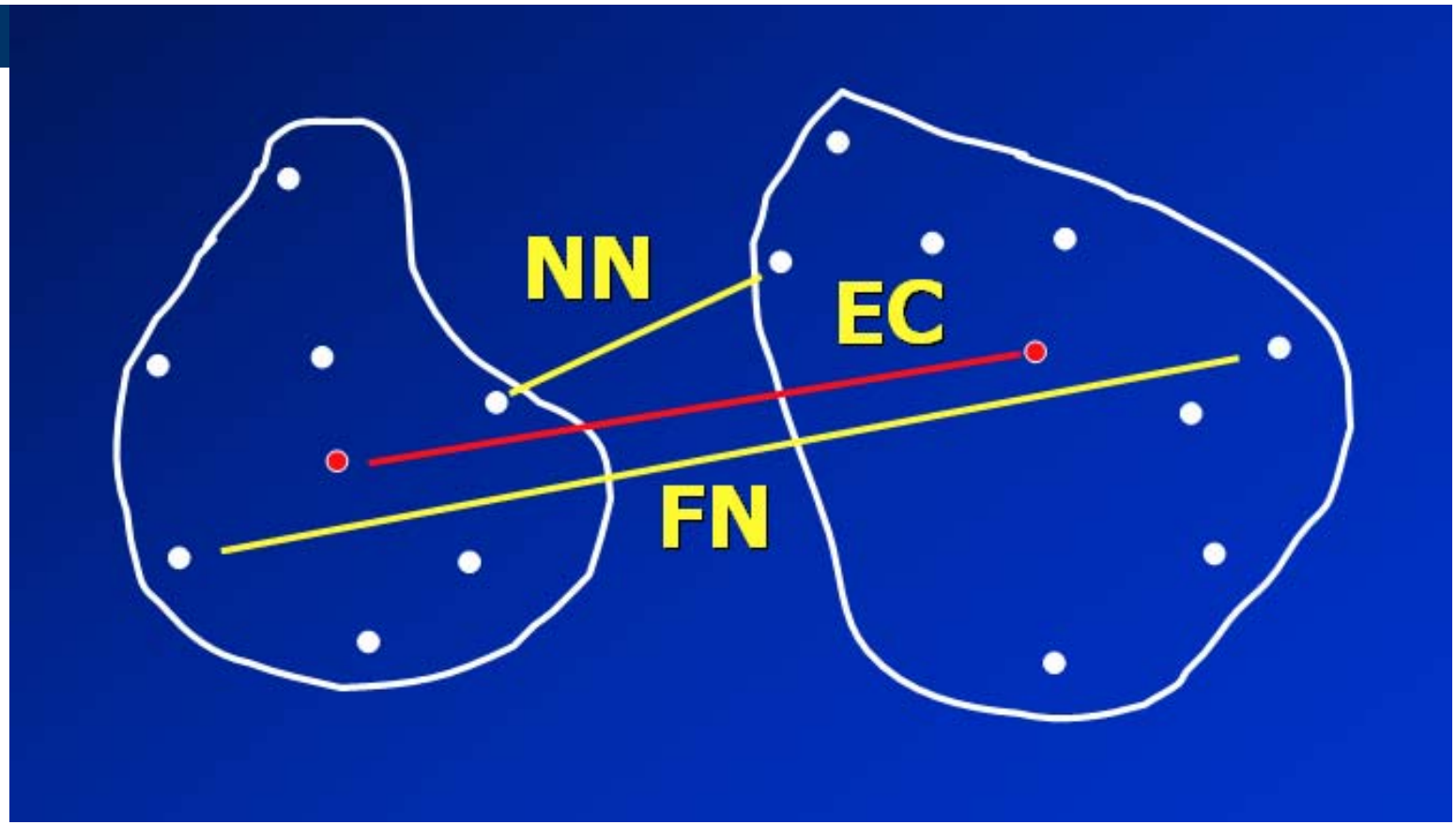
Clustering Methods

- Agglomerative
  - Start with all separate, end with some connected
- Partitioning / Divisive
  - Start with all connected, end with some separate
- Dimensional Reduction
  - Find dominant information in the data

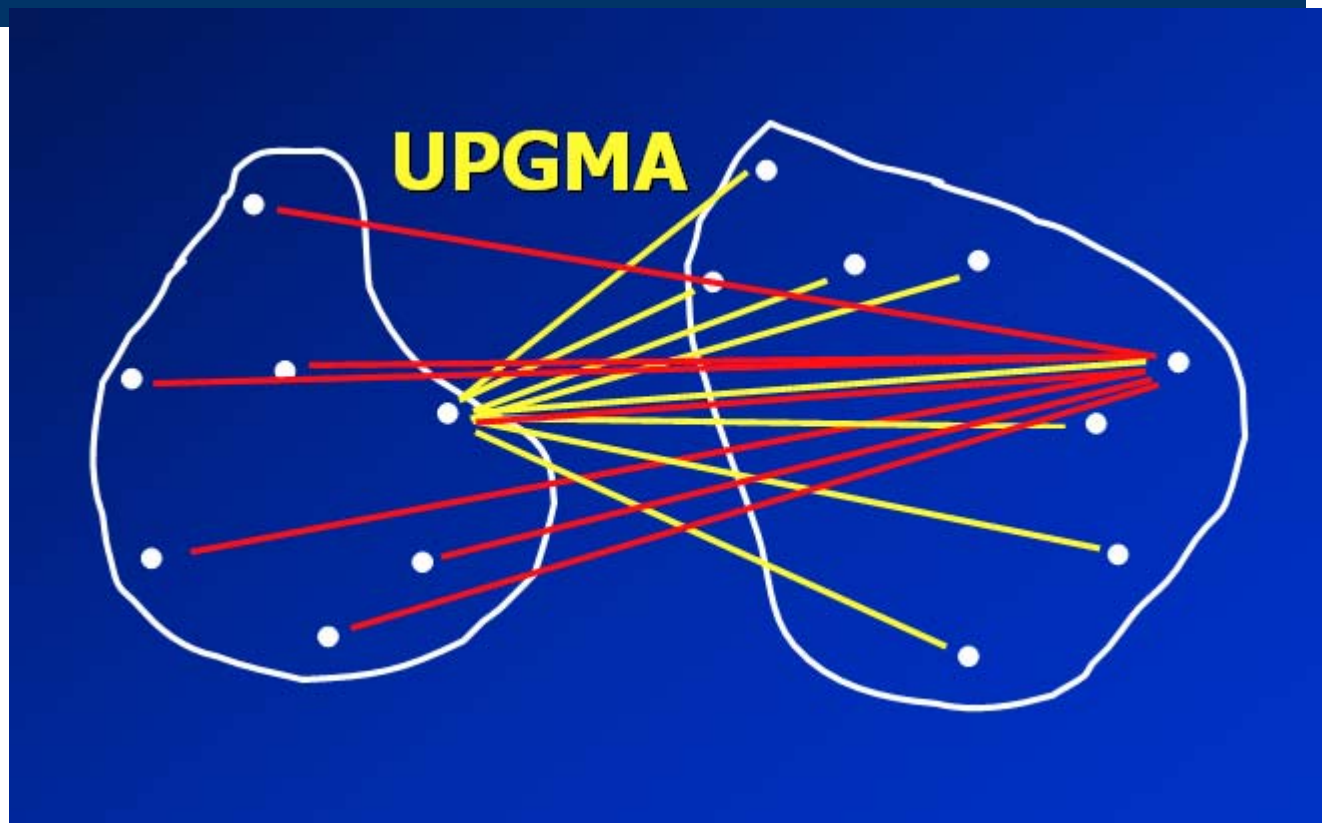
Hierachical Clustering (HCA)

- Algorithm
  - Join 2 most similar genes.
  - Merge genes into a new super gene
  - Repeat until all merged
- Group Proximity
  - Single Linkage (Nearest Neighbor)
  - Complete Linkage (Furthest Neighbor)
  - Average Linkage (UPGMA)
  - Euclidean distance between centroids

Hierarchical Clustering

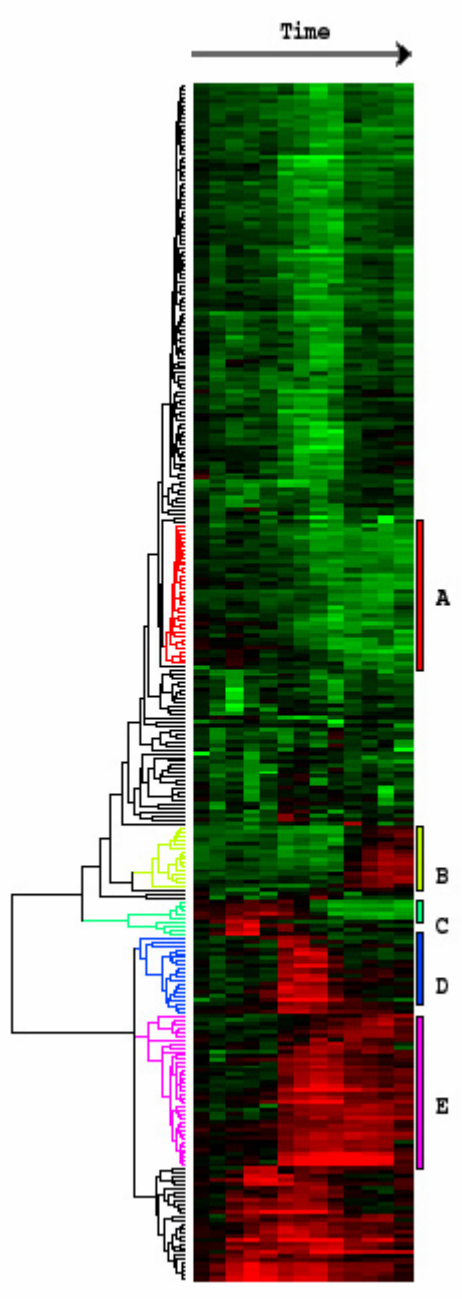


Hierarchical Clustering



HCA: Eisen et al (1998)

height of branch gives an indication of how similar two genes are



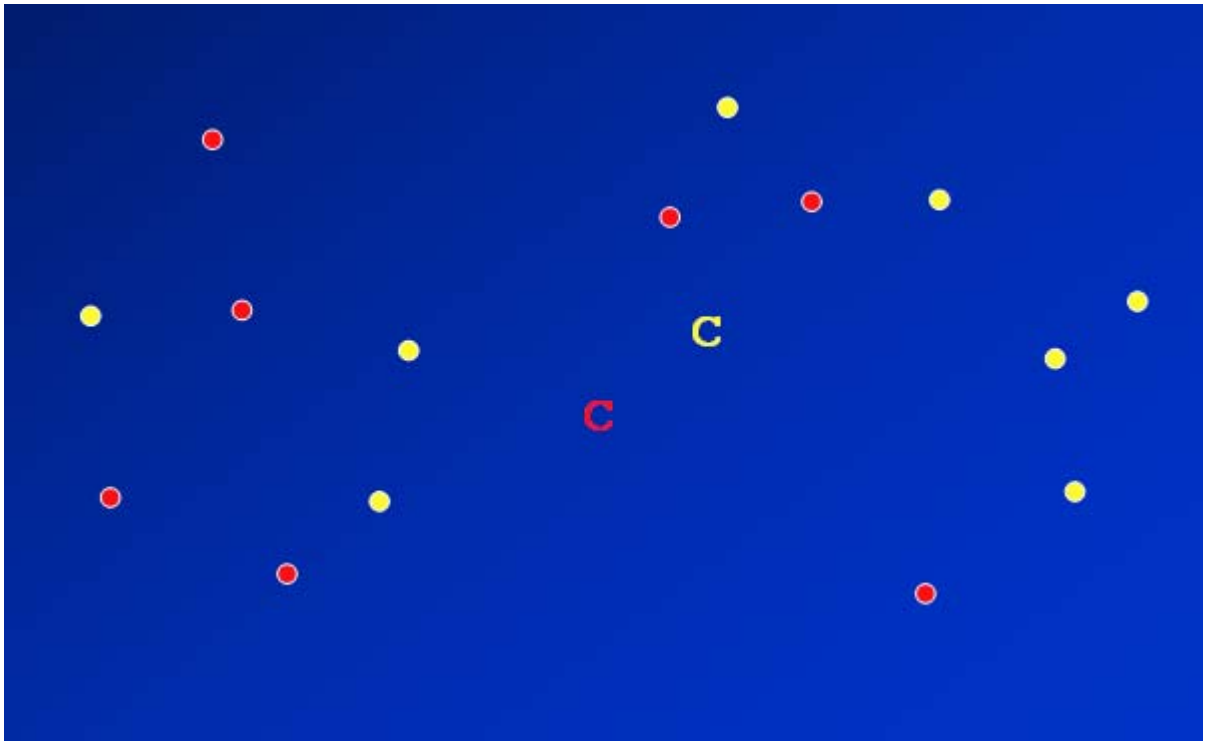
Partitioning: *k*-means Clustering

- Guess what the cluster centers are and then fix.

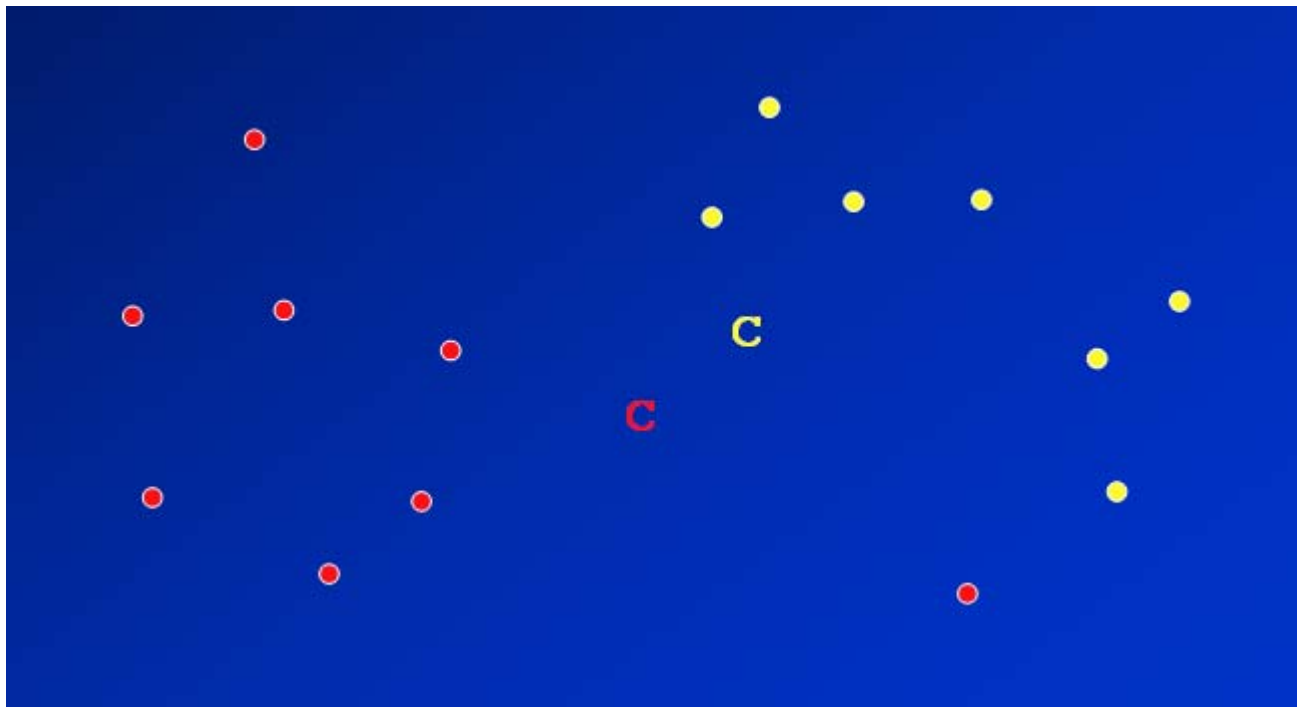
k-means

1. Specify # clusters *k*
2. Randomly create *k* partitions
3. Calculate center of partitions
4. Associate genes to their closest center
5. Repeat until changes very little

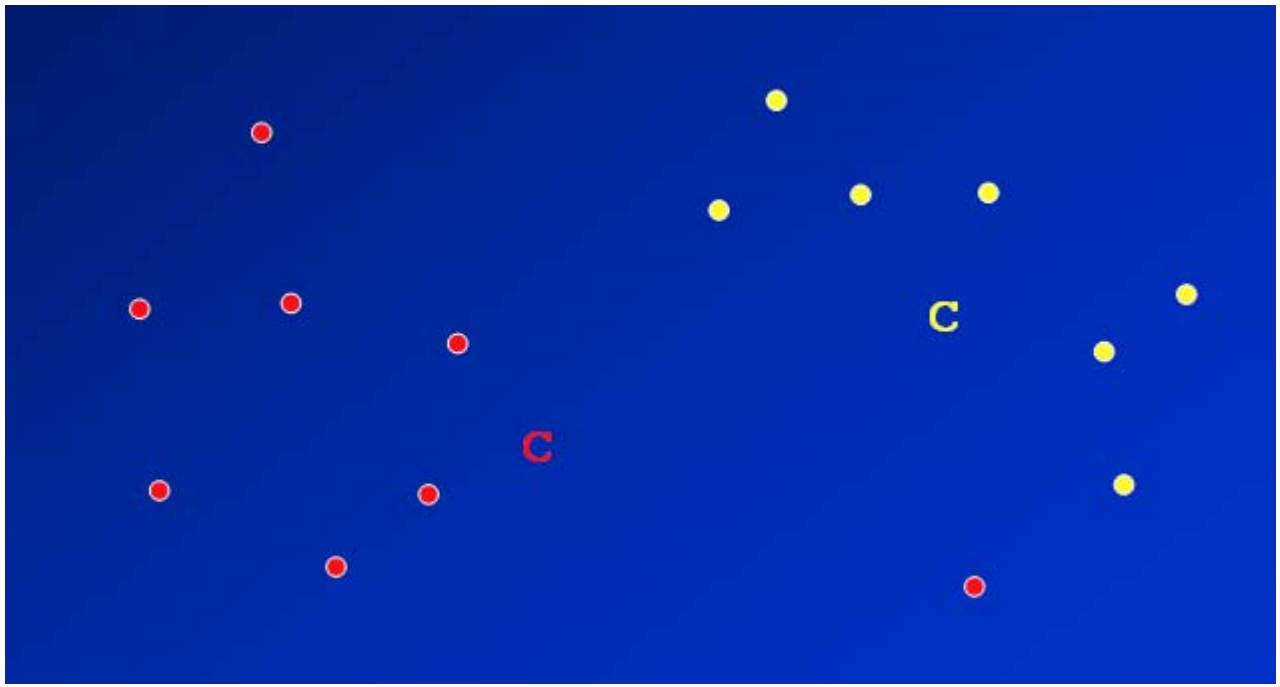
k-means



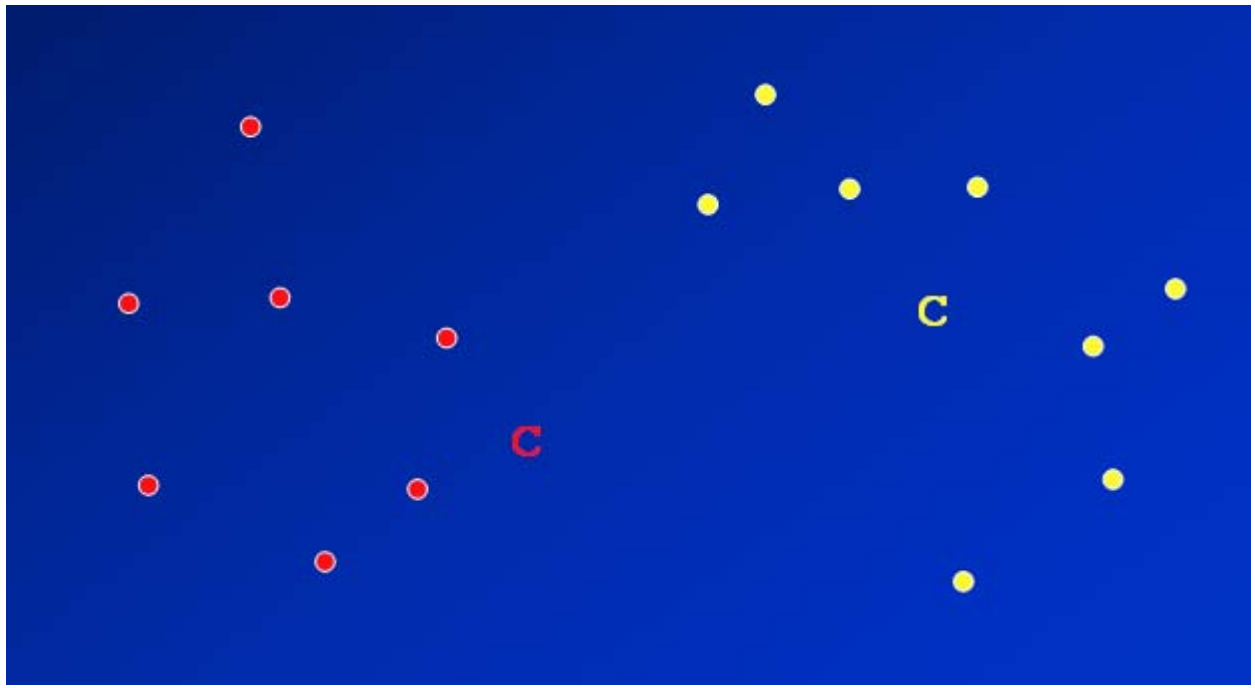
k-means



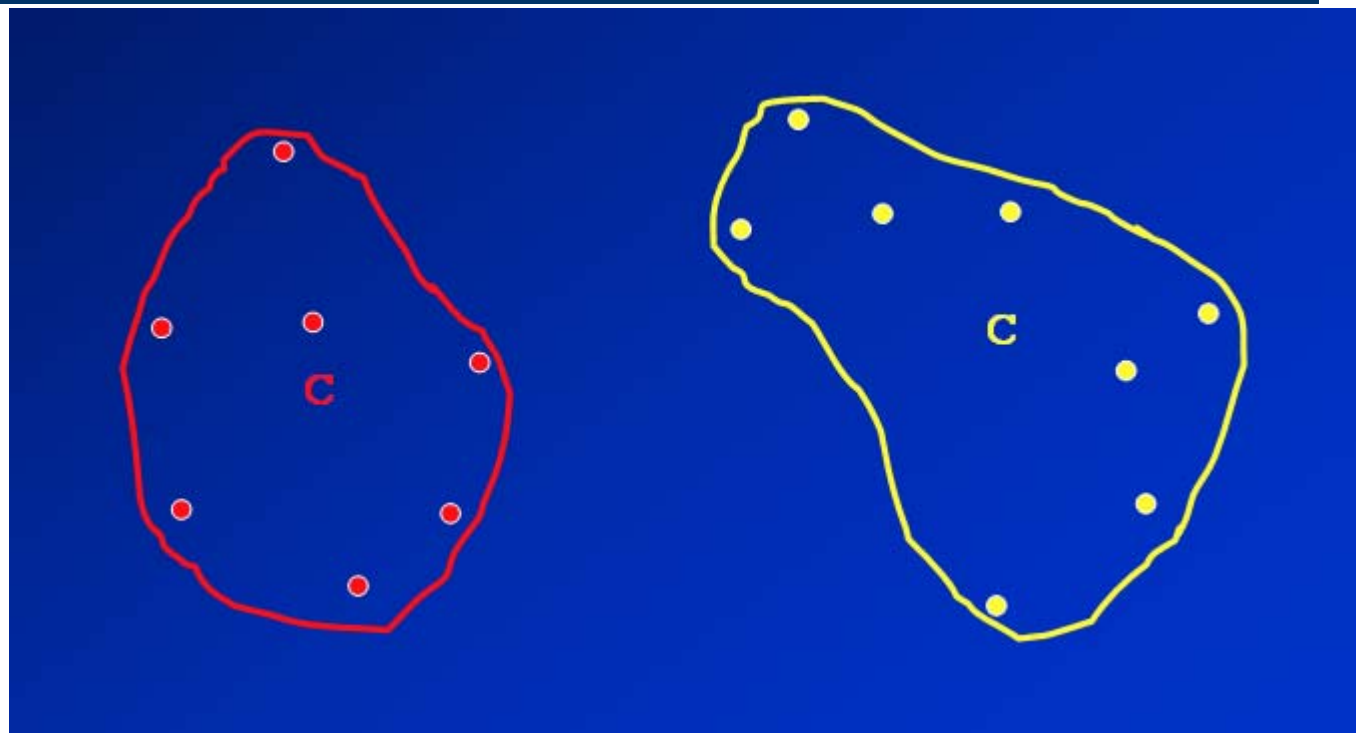
k-means



k-means



k-means



Choice of *k*

- Informed by type of data (bi-state, time series)
- Initialize from HCA
- Try different values of *k* and pick best one

Fuzzy *k*-means

Don’t hard assign genes to clusters.

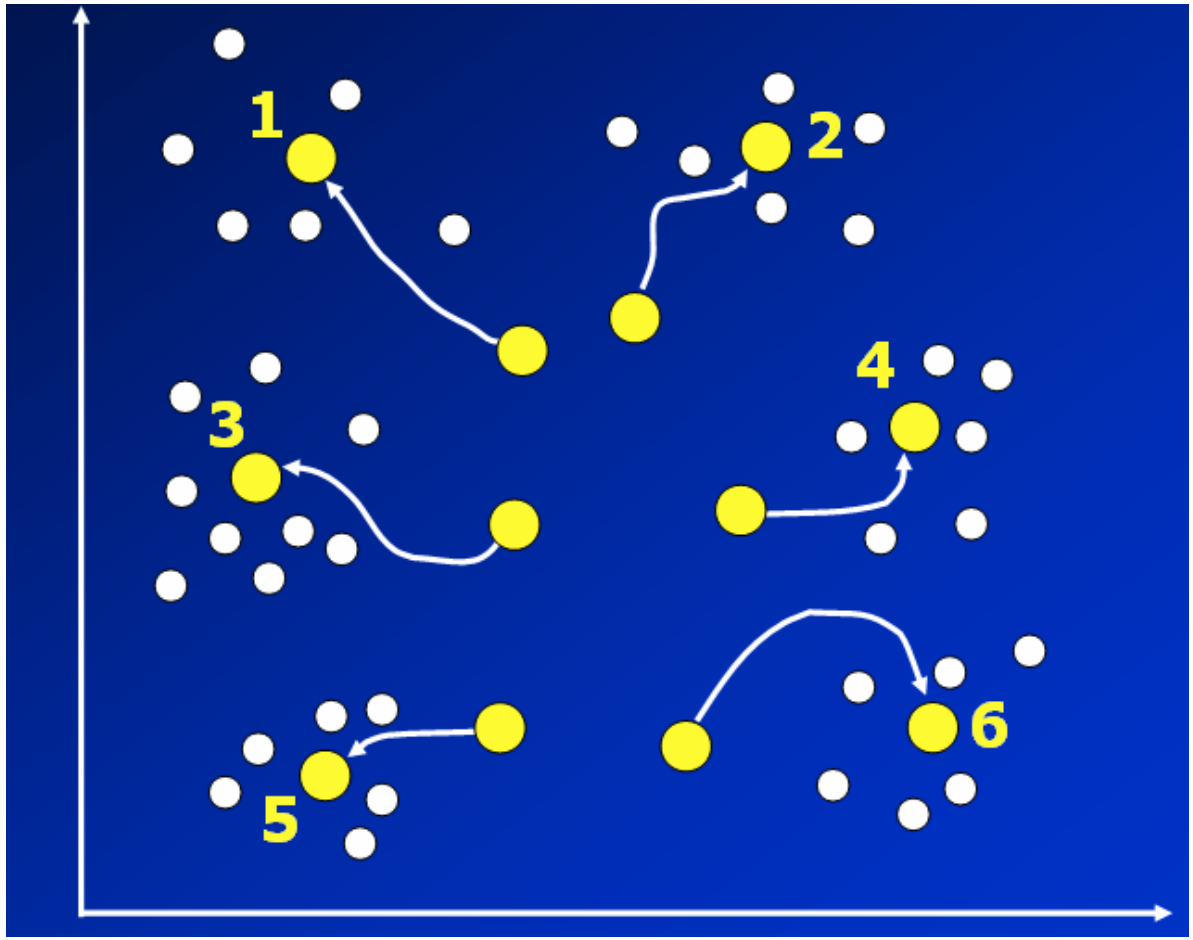
Record distance to final clusters.

Gasch & Eisen (2002)

Self Organizing (Kohonen) Map

- Specify # of *nodes* (*m* X *n*)
- Map nodes to random expression profiles
- Choose a gene, assign it to its nearest node
- Adjust the position of all of the nodes
- Do for all of the genes

SOM – et al (1999)



New approaches

CAST – graph-theoretic based derisive approach based on finding minimum cut sets. (Ben Dor 2000)

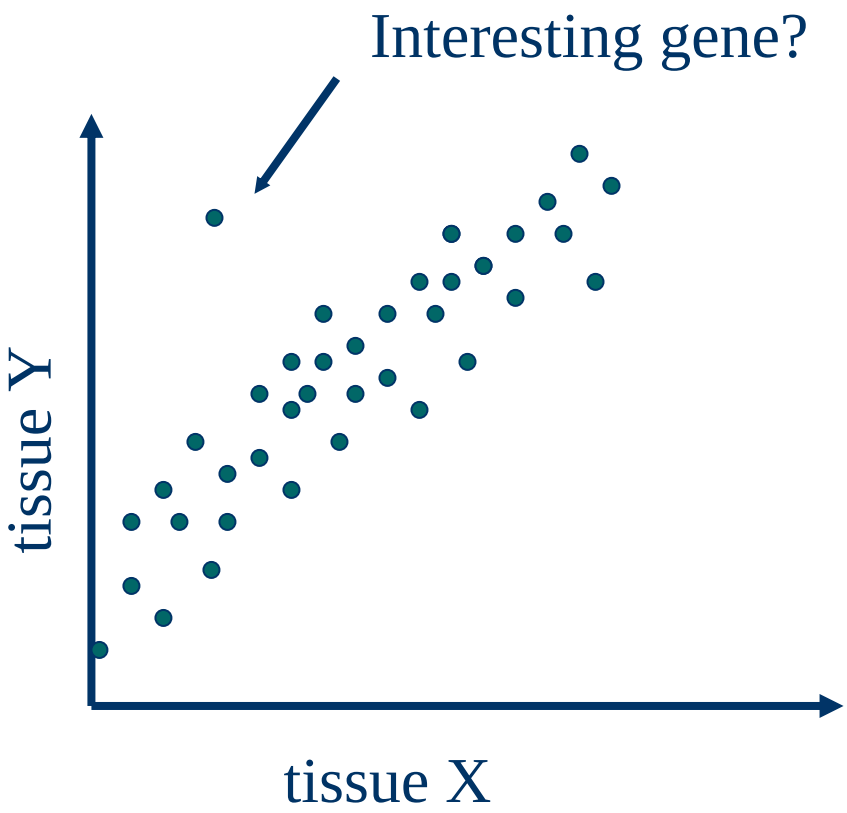
Bi-clustering – cluster both the genes *and* the experiments simultaneously to find appropriate context for clustering (Lazzeroni & Owen 2000)

Bayesian Network – model observations as deriving from modules (gene clusters) and processes (experiment clusters). Search for most likely model. (Segal and Koller 2002).

Dimension Reduction: Principal Components Analysis (PCA)

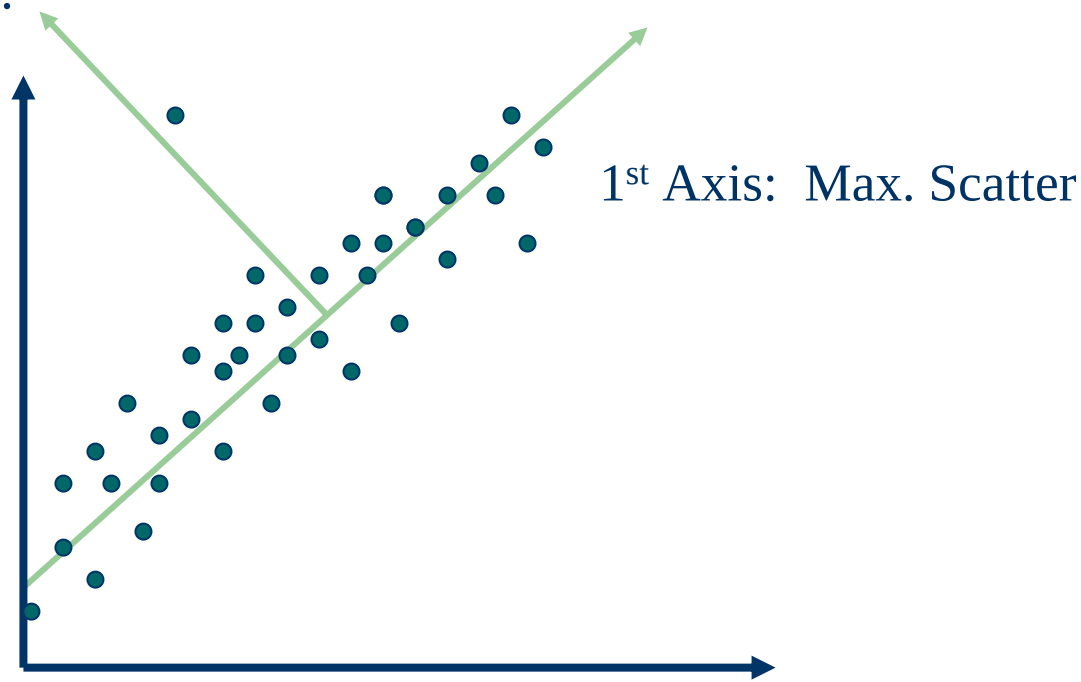
- Construct a new data set with fewer conditions
- New conditions are *linear combinations* of original conditions

PCA



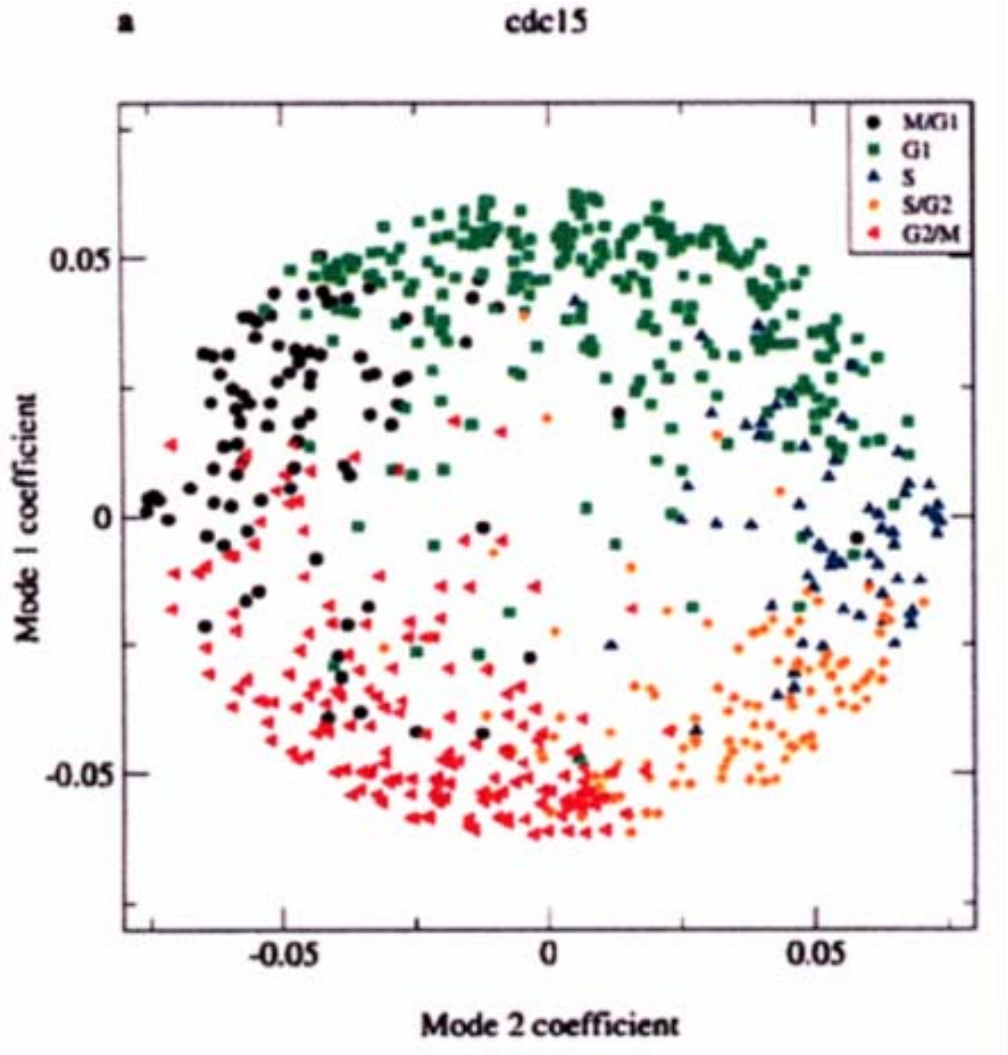
PCA

2nd Axis: 90o to 1st & maximize the rest of the Scatter.



1st Axis: Max. Scatter

PCA Example: Cell cycle



Holter et al. (2000)

PCA Example: Sporulation

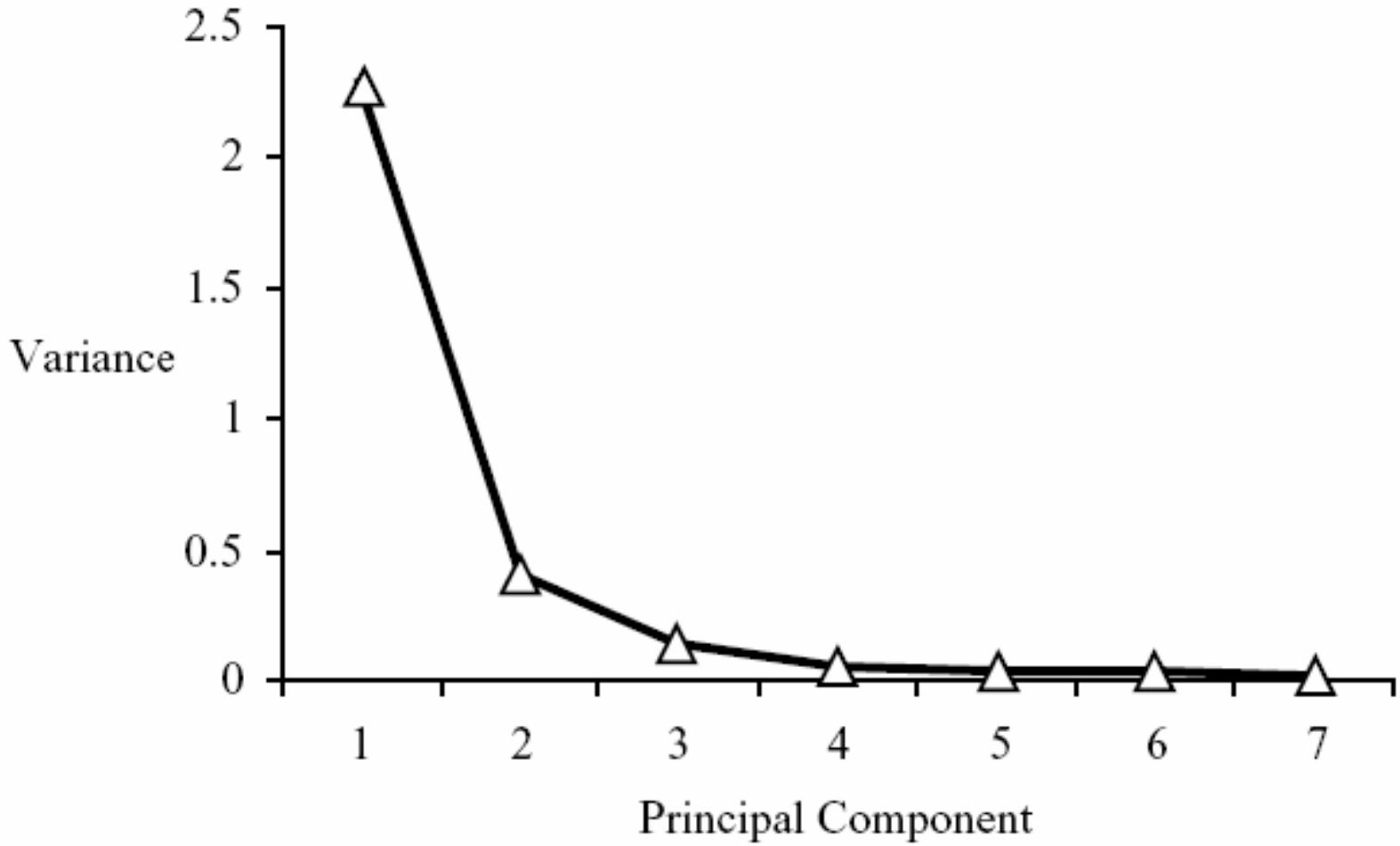
- 7 timepoints during yeast sporulation (Chu et al 1998): 0h, 30min, 2h, 5h, 7h, 9h, 11h

PCA Example: Sporulation

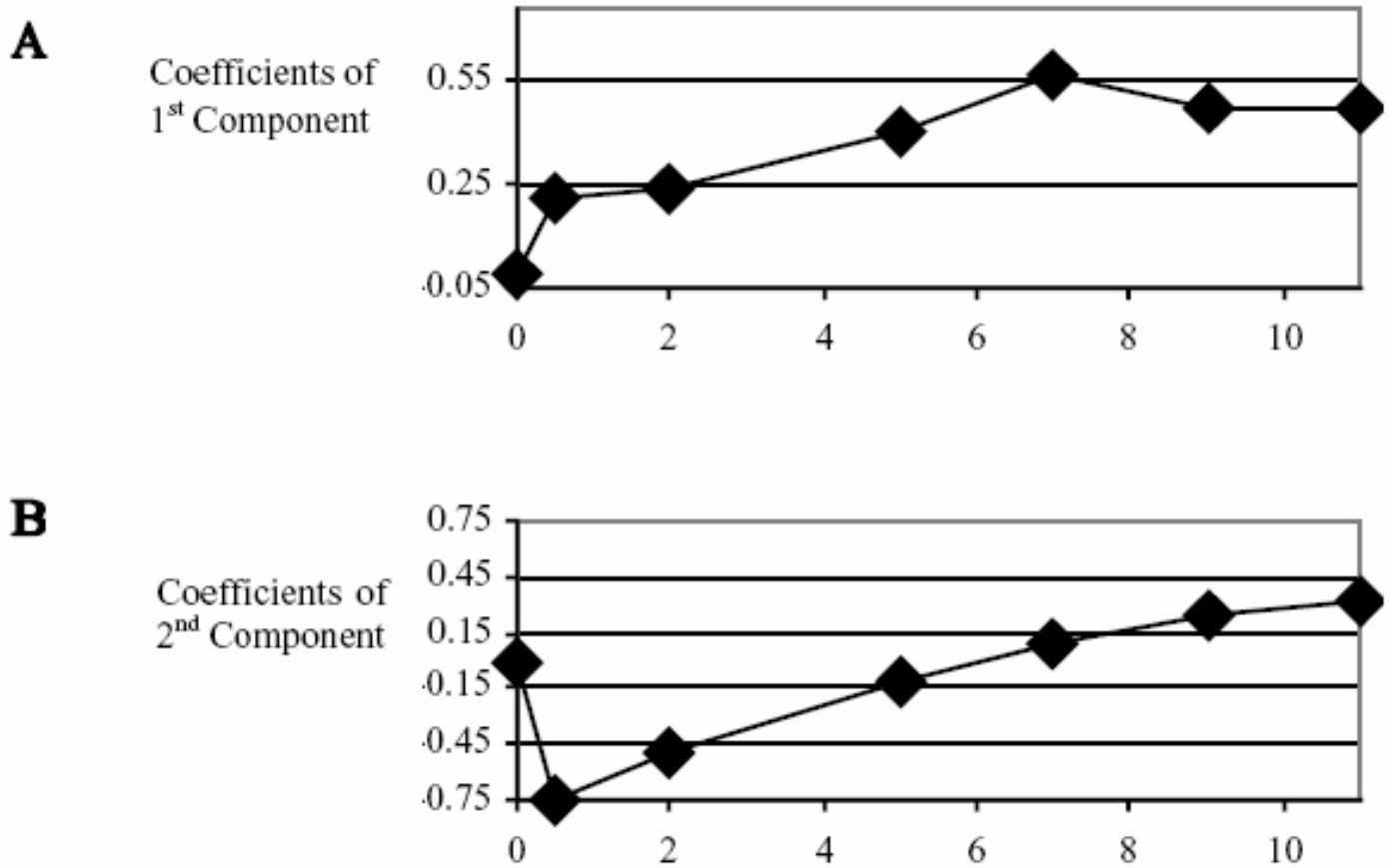
Projection On condition	Principal Components						
	1	2	3	4	5	6	7
T = 0	-0.0072	-0.0116	-0.0631	-0.2166	0.0764	-0.7433	0.625
T = .5	0.2076	-0.7524	-0.5373	0.2606	0.1545	-0.0683	-0.0756
T = 2	0.2358	-0.4925	0.3296	-0.5935	-0.453	0.1713	0.0803
T = 5	0.3975	-0.1156	0.5612	-0.002	0.5919	-0.2532	-0.3151
T = 7	0.554	0.0862	0.1869	0.4959	-0.1112	0.2889	0.5559
T = 9	0.4671	0.2517	-0.153	0.1169	-0.5413	-0.4488	-0.4324
T = 11	0.4671	0.3273	-0.4748	-0.5229	0.3307	0.254	0.044
Eigenvalue	2.2928	0.401	0.1322	0.0594	0.0406	0.0288	0.025
% variance	76.9 %	13.5 %	4.4 %	2.0 %	1.4 %	1.0 %	0.8 %

Raychaudhuri et al (2000)

PCA Example: Sporulation



PCA Example: Sporulation



Clustering Method Comparision

- Hierarchical slow (K-means fast)
- K-means and SOMs have to choose # clusters

Clustering Software

- Cluster & TreeView
- http://rana.lbl.gov/EisenSoftware.htm

YORF	NAME	GWEIGHT	spo0	spo30
EWEIGHT			1	1
YAL003W	EFB1	1	0.23	-1.79
YAL004W	YAL004W	1	0.41	-0.38
YAL005C	SSA1	1	0.61	-0.07
YAL010C	MDM10	1	0.16	-0.15
YAL012W	CYS3	1	0.03	1.39

Clustering Software

- JavaTreeView
- MapleView
- GeneXPress
  - http://genexpress.stanford.edu

Stretch Break (10 minutes)

Cluster Evaluation

- Intrinsic
- Extrinsic

Intrinsic Evaluation

- Measure cluster quality based on how “tight” the clusters are.
- Do genes in a cluster appear more similar to each other than genes in other clusters?

Intrinsic Evaluation Methods

- Sum of squares
- Silhouette
- Rand Index
- Gap statistic
- Cross-validation

Sum of squares

A good clustering yields clusters where genes have small within-cluster sum-of-squares (and high between-cluster sum-of-squares).

Within-cluster variance

B(k) = between cluster sum of squares

W(k) = within cluster sum of squares

Maximize CH(k) over the clusters:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}$$

Calinski & Harabasz (1974)

Silhouette

- Good clusters are those where the genes are close to each other compared to their next closest cluster.

Silhouette

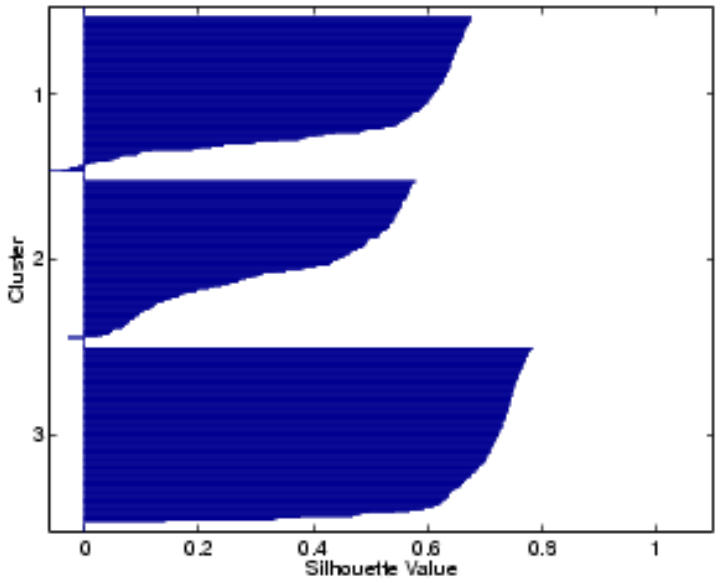
b(i) = min(AVGD_BETWEEN(i,k))

a(i) = AVGD_WITHIN(i)

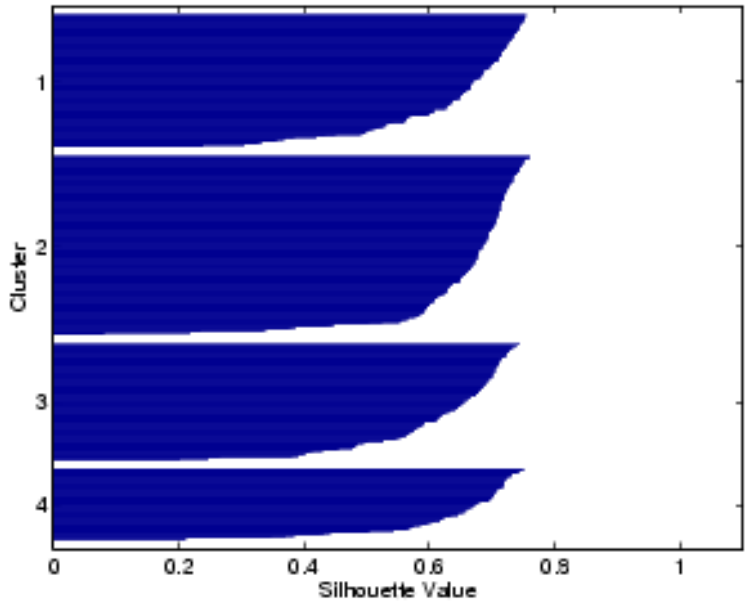
$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Kaufman & Rousseeuw (1990)

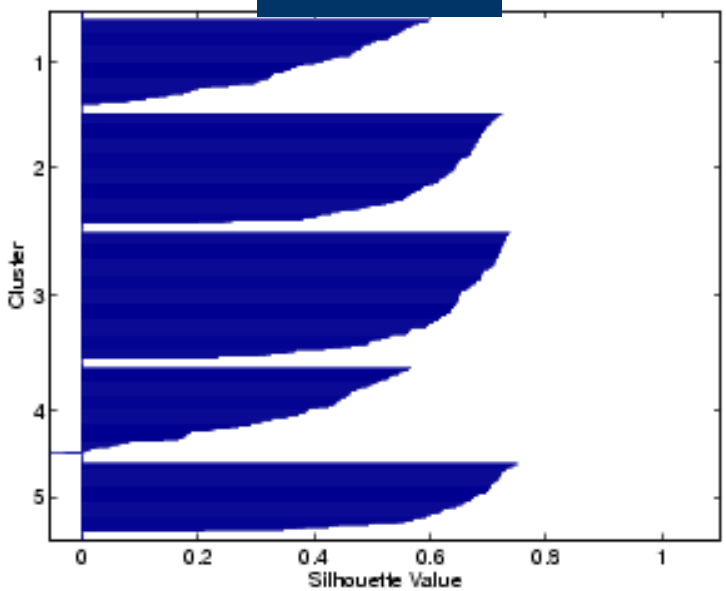
3 clusters



4 clusters



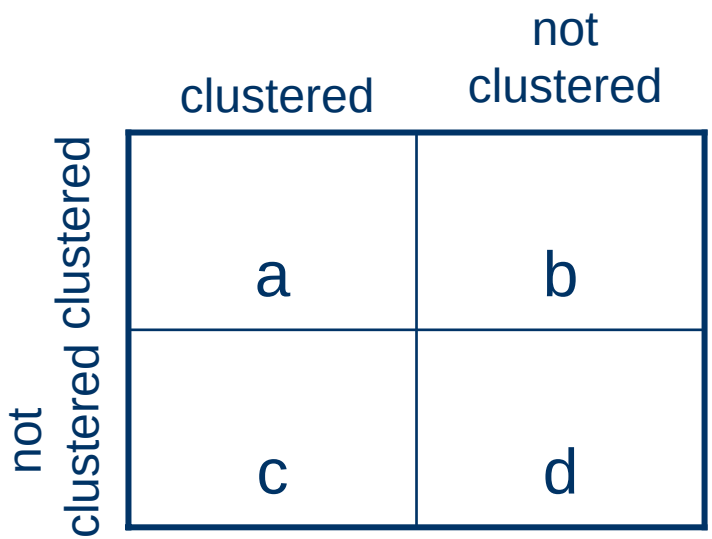
5 clusters



Kaufman & Rousseeuw (1990)

Rand Index

Rand = (a + d) / (a + b + c + d)



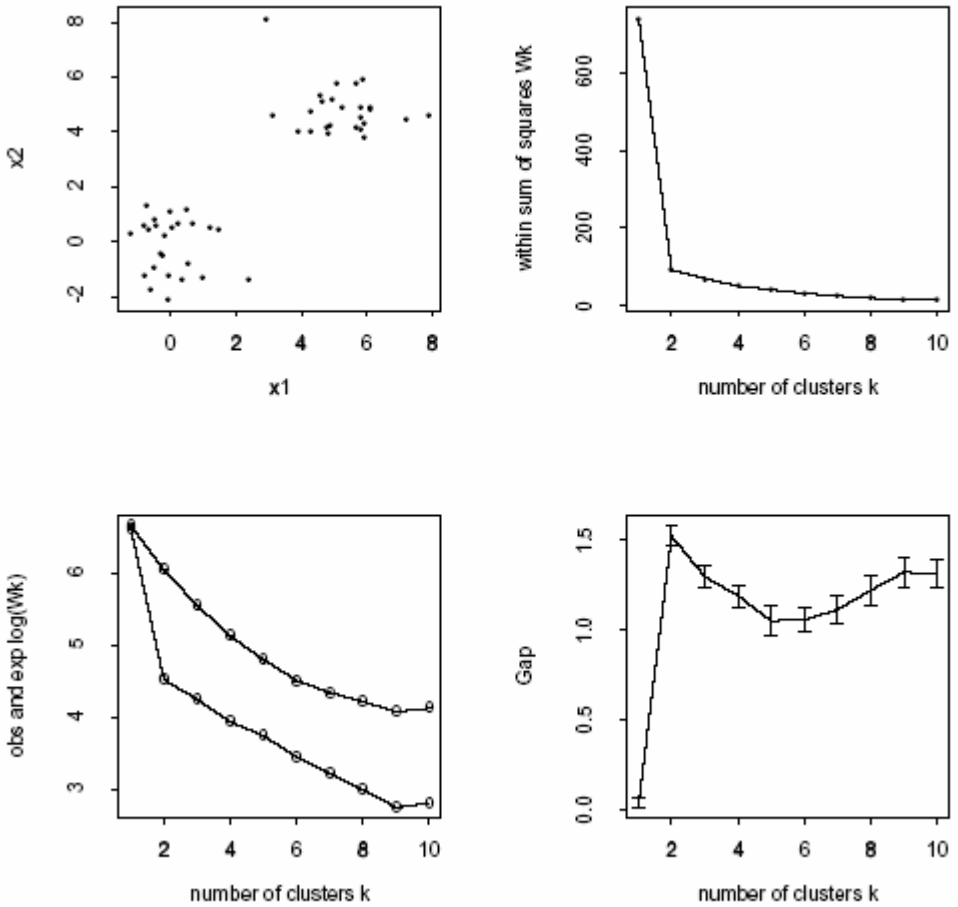
Hubert and Arabie (1985)

Adjusted Rand Index

Adjusts the Rand index to make it vary between 0 and 1 according to expectation:

AdjRand = (Rand – expect) / (max – expect)

Gap statistic



Tibshirani et al. (2000)

Gap statistic

Computation of the Gap statistic

1. Cluster the observed data, varying the total number of clusters from $k = 1, 2, \ldots K$, giving within dispersion measures $W_k, k = 1, 2, \ldots K$.

2. Generate B reference datasets, using the uniform prescription (a) or (b) above, and cluster each one giving within dispersion measures W^*_{kb}, $b = 1, 2, \ldots B$, $k = 1, 2, \ldots K$. Compute the (estimated) Gap statistic:

$$\text{Gap}(k) = (1/B) \sum_b \log(W^*_{kb}) - \log(W_k)$$

3. Let $\bar{l} = (1/B) \sum_b \log(W^*_{kb})$, compute the standard deviation $\text{sd}_k = [(1/B) \sum_b (\log(W^*_{kb}) - \bar{l})^2]^{1/2}$, and define $s_k = \text{sd}_k \sqrt{1 + 1/B}$. Finally choose the number of clusters via

$$\hat{k} = \text{smallest } k \text{ such that } \text{Gap}(k) \geq \text{Gap}(k+1) - s_{k+1}$$

Tibshirani et al. (2000)

Cross-validation approaches

- Leave out *k* experiments (or genes)
- Perform clustering
- Measure how well clusters group in left out experiment(s)
- Or, measure agreement between test and training set.

Figure of Merit

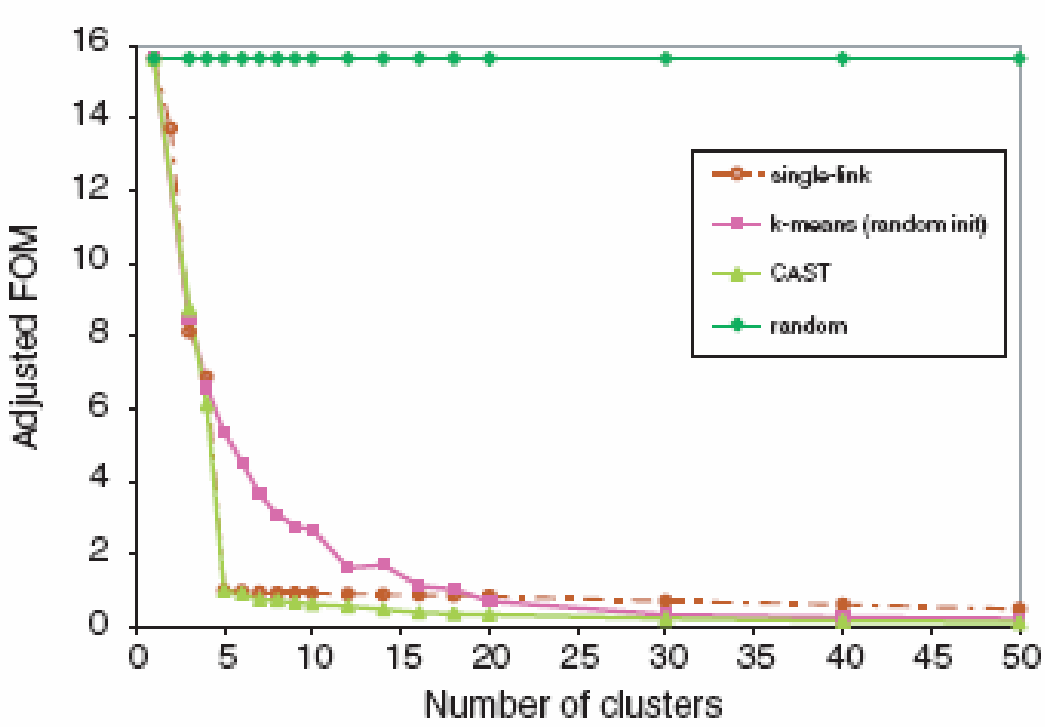
$$FOM(e,k) = \sqrt{\frac{1}{n} * \sum_{i=1}^{k} \sum_{x \in C_i} (R(x,e) - \mu_{C_i}(e))^2}$$

$$FOM(k) = \sum_{e=1}^{m} FOM(e,k)$$

Yeung et al. (2001)

Figure of Merit

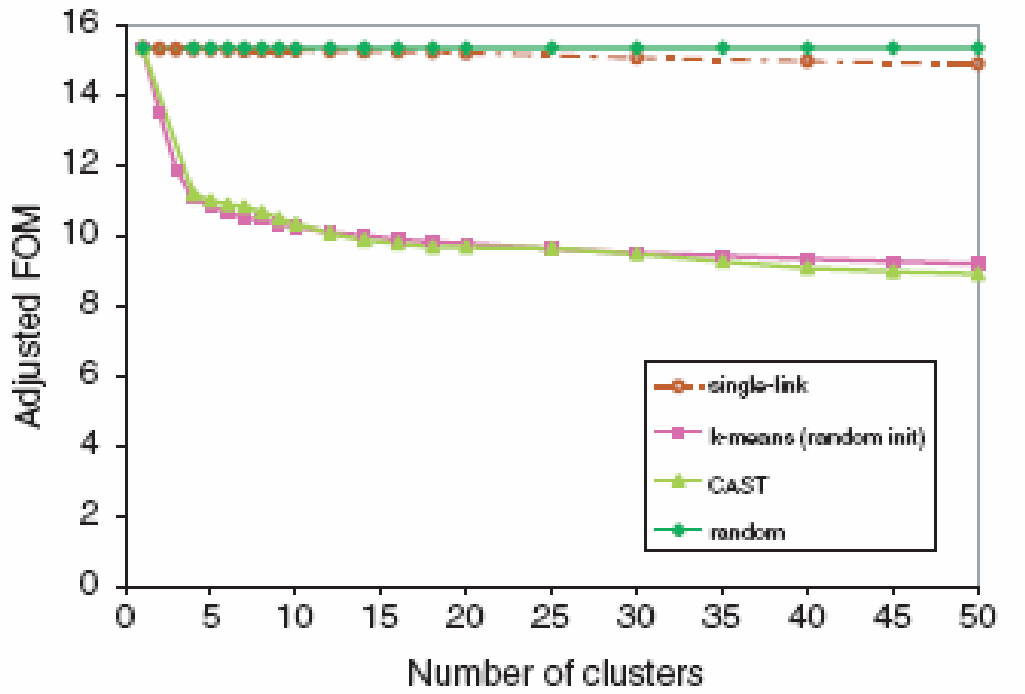
Random Data with 5 clusters



Yeung et al. (2001)

Figure of Merit

Cho cell cycle data

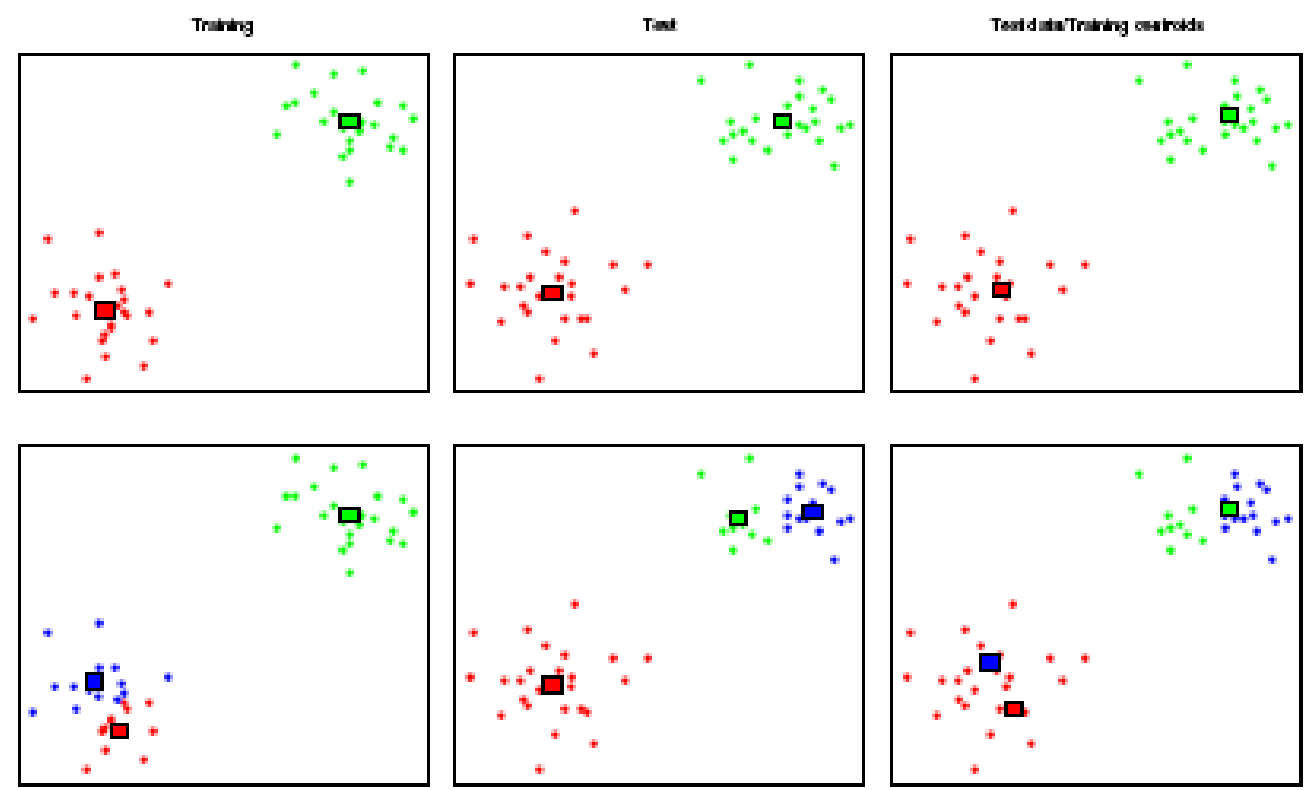


Yeung et al. (2001)

Prediction strength of clustering

- Clustering as classification
- Split data into training and test set
- Apply clustering method to both and measure agreement
- Compute *prediction strength* of clustering

Prediction strength of clustering



Tibshirani et al. (2001)

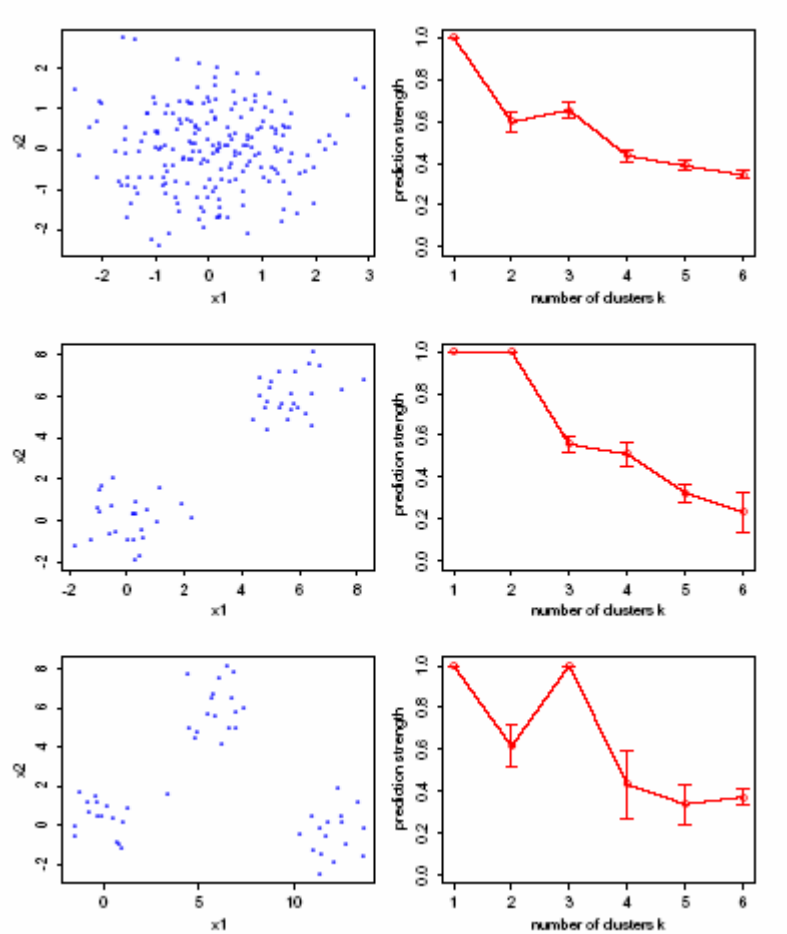
Prediction strength of clustering

Prediction strength:

$$\mathrm{ps}(k) = \min_{1 \le j \le k} \frac{1}{n_{kj}(n_{kj}-1)} \sum_{i \neq i' \in A_{kj}} I(D[C(X_{tr}, k), X_{te}]_{ii'} = 1).$$

Tibshirani et al. (2001)

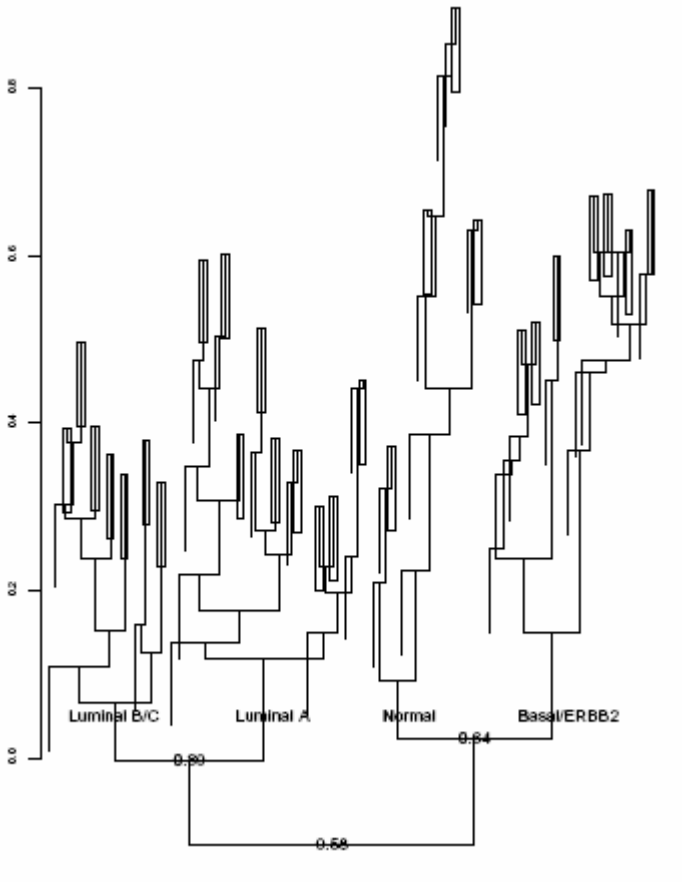
Prediction strength of clustering



Synthetic data.

Tibshirani et al. (2001)

Prediction strength of clustering



- Apply HCA
- On each split, perform *k*-means on test and training set
- Label each split with the prediction strength

Tibshirani et al. (2001)

Further Reading

- Yeung et al. (2001) *Bioinformatics* 17:309.
- Tibshirani et al. (2001). *Stanford Technical Report*. Go to http://www-stat.stanford.edu/~tibs/lab/publications.html
- Calinski & Harabasz (1974) *Communications in Statistics* 3:1-27.
- Kaufman & Rousseeuw (1990) *Finding groups in data.* New York, Wiley.
- Kim et al (2001) *Science* 293:2087
- Gasch & Eisen (2002) *Genome Biol.* 3(11):research0059