

Tutorial on Statistical Methods and Software for the Analysis of Microarray Experiments

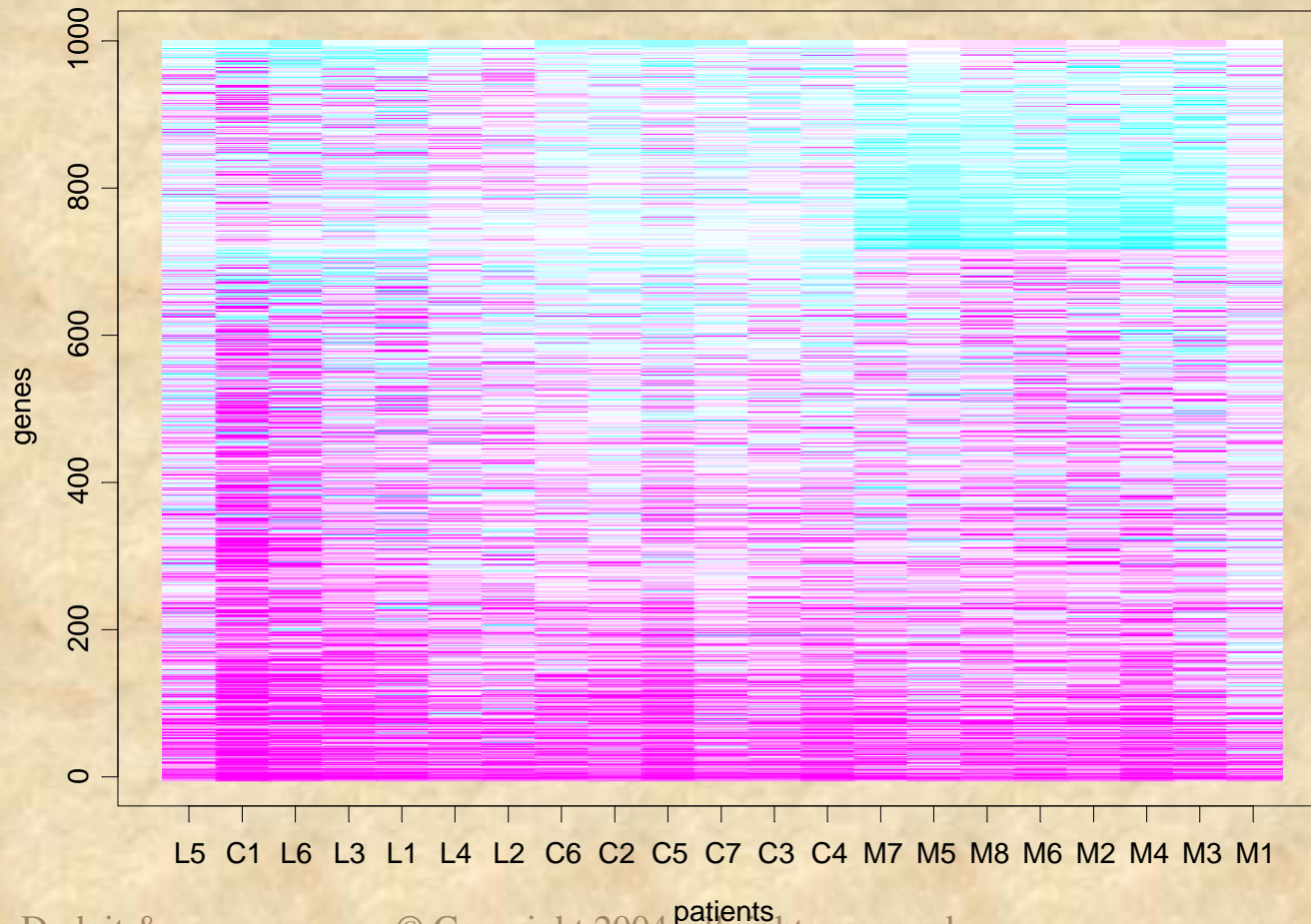


Nicholas P. Jewell
Sandrine Dudoit

September 22, 2004

Class 3
Microarray Experiments:
Clustering

Analysis of Microarray Data



Overview

- ❖ Cells respond to conditions and treatments by regulating gene expression; e.g. activating or repressing the transcription of certain genes
- ❖ DNA microarrays are high-throughput biological assays that can measure DNA or RNA abundance, and hence yield information on gene expression levels
 - In cancer research, microarrays are used to measure transcript levels in tumor samples for tens of thousands of genes at a time

Overview

- ❖ **General Question:** Relate genome-wide microarray measures to biological and clinical covariates and outcomes. That is, make inference about the joint distribution of these random variables
 - **Covariates:** treatment, dose, time, occurrence of DNA sequence motifs, SNP genotypes etc
 - **Outcomes:** affectedness/unaffectedness, quantitative trait, metastasis indicator, response to treatment, etc

Overview

- ❖ Unsupervised learning (today): just consider joint distribution of gene expressions (and covariates)
 - Select expressing genes or differentially expressing genes
 - Which genes tend to express similarly over a variety of patients/conditions?
- ❖ Supervised learning (tomorrow): introduce comparisons w.r.t. covariates and outcomes
 - Which genes differ between groups?
 - Which genes are correlated with or predict an outcome?

Unsupervised Learning

- ❖ Clustering and estimating reproducibility of clusters. (Which genes tend to systematically and similarly over or under express in a correlated fashion? It is possible that different genes may over express but in random patients, that is, not "together"; A different question: which patients tend to have similar gene expressions)

Clustering

Why cluster?

- ❖ Once differentially expressed genes have been selected, the researcher is interested in usually interested in discovering groups of similarly expressed (related?) genes.
- ❖ Clustering is a natural way to explore potentially related genes.
- ❖ Want to find groups of patients whose genes similarly express

Types of Clustering

❖ Partitioning Methods

- Partitioning around medoids
- K-means
- Etc.

❖ Hierarchical Methods

- Agglomerative
- Divisive
- → Phylogenetic trees

Example of Hierarchical Clustering

(b) Clustering tree of agnes(agriculture)

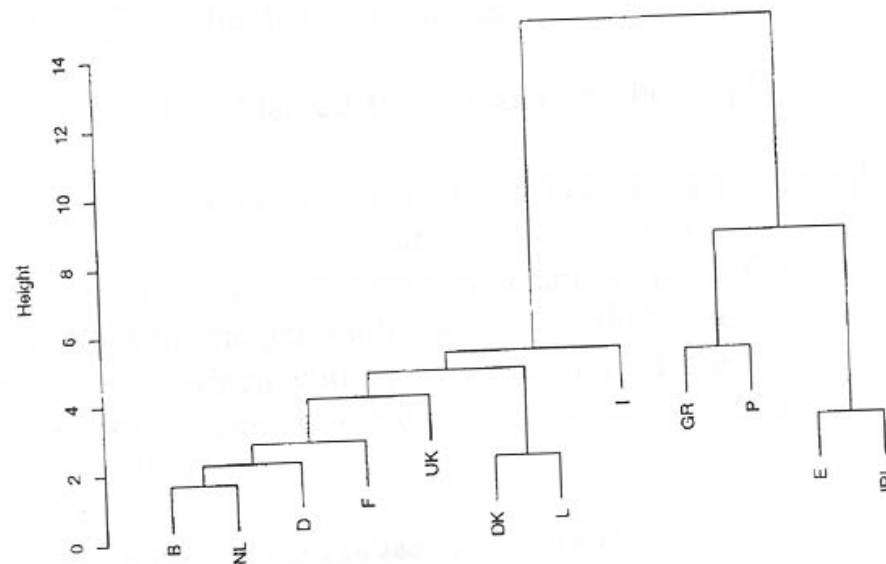


Fig. 6. Result of applying agnes to the agriculture data, represented in the form of (a) a banner; and (b) a clustering tree.

Distances

- ❖ Clustering procedures rely on defining a distance between units (in our case, either p genes or n units)
- ❖ One can cluster on the units or the genes
 - e.g., groups of people with similar gene expression profiles (distance matrix will be $n \times n$)
 - e.g., groups of genes with similar expressions across people (distance matrix will be $p \times p$).
- ❖ Many choices for distance, none objectively better than the other (nb clustering only depends on the distance matrix).

Distances

❖ $d_{jk} = 1 - \rho_{jk}$ Correlation

❖ $d_{jk} = 1 - |\rho_{jk}|$ Absolute correlation

❖ $d_{jk} = 1 - \rho^0_{jk}$ Cosine-angle

❖ $d_{jk} = \sqrt{\sum_{i=1}^n (Y_{ij} - Y_{ik})^2}$ Euclidean

❖ ρ_{jk} = the empirical correlation between j th, k th gene (unit)

$$\rho^0_{jk} = \frac{\sum_{i=1}^n Y_{ij} Y_{ik}}{\sqrt{\sum_{i=1}^n Y_{ij}^2} \sqrt{\sum_{i=1}^n Y_{ik}^2}}$$

Partitioning Methods: Partitioning Around Medoids

- ❖ Abbreviated as PAM
- ❖ Discuss both clustering genes and clustering units (pay attention to context).
- ❖ In PAM, one sets the number of clusters, say k .
- ❖ The procedure then chooses the "optimal" set of k medoids (genes if clustering genes).

PAM continued

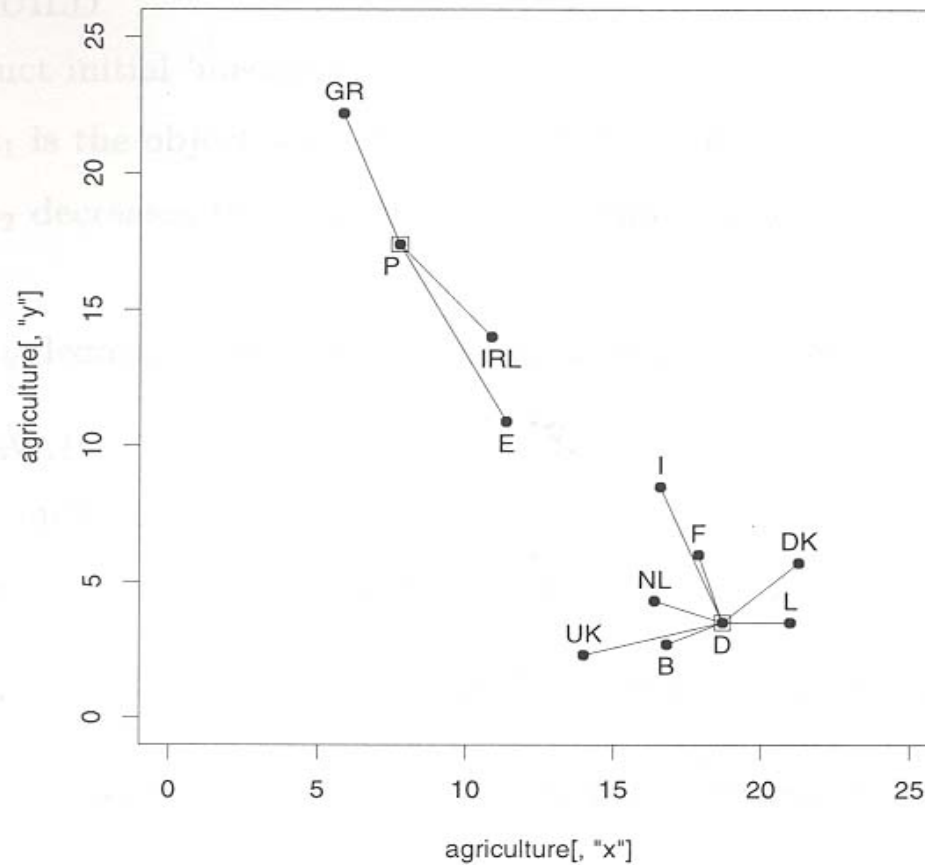
- ❖ Clustering techniques generally try to find the set of clusters (eg groups of genes) that minimize some criterion related to the chosen distance.
- ❖ Let m_1, m_2, \dots, m_k be a potential vector of medoids.
- ❖ PAM finds the set of medoids that minimizes: $\sum_{j=1}^q \min d(j, m_t), \quad t = 1, \dots, k$

PAM continued

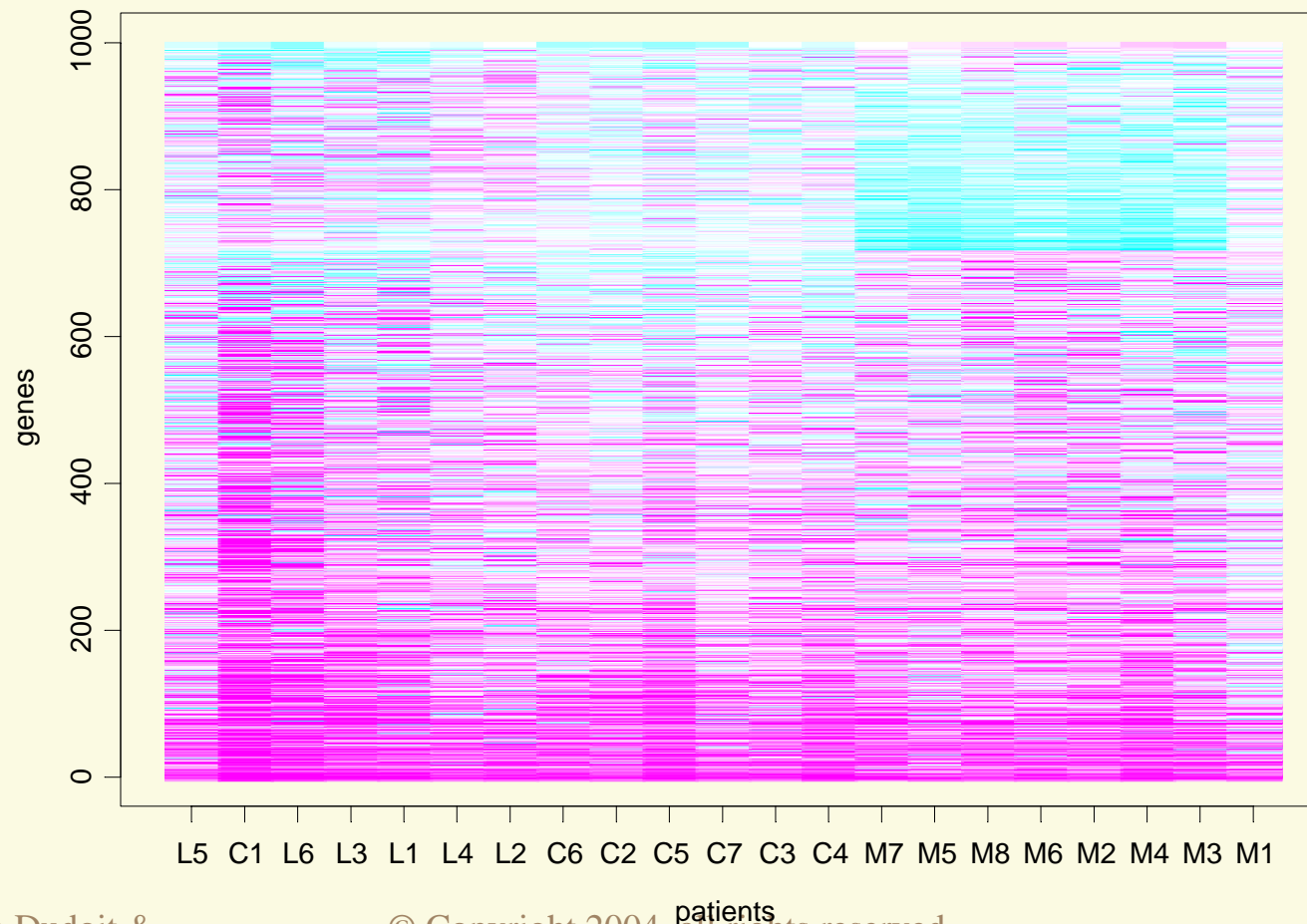
- ❖ After the medoids are chosen, then the assignment of an object to a cluster is to its nearest medoid (gene).
- ❖ One can represent this algorithmically as, object j (eg gene j) is assigned to cluster v_j when medoid m_{v_j} is nearer to j than any other medoid, i.e.,

$$d(j, m_{v_j}) \leq d(j, m_t) \text{ for } t = 1, \dots, k$$

Example 1 - GNP (x) vs. % farming (y).



Example 2 - log relative expression data



Choosing the Number of Clusters

- ❖ Besides giving the medoids and cluster assignment, PAM also calculates a quantity known as the *silhouette*, which is a measure of how distinct the clusters are.

- ❖ Let
$$a(j) = \frac{1}{|A| - 1} \sum_{k \in A, k \neq j} d(j, k)$$

which is the average distance of j to all other objects of A (the cluster to which j belongs).

Choosing the Number of Clusters

- ❖ Now, consider any cluster, C , different from A .

$$d(j, C) = \frac{1}{|C|} \sum_{k \in C} d(j, k)$$

which is the average distance of j to the objects in cluster C .

- ❖ After computing $d(j, C)$ for all clusters $C \neq A$, find the smallest of the $d(j, C)$,

$$b(j) = \min_{C \neq A} d(j, C)$$

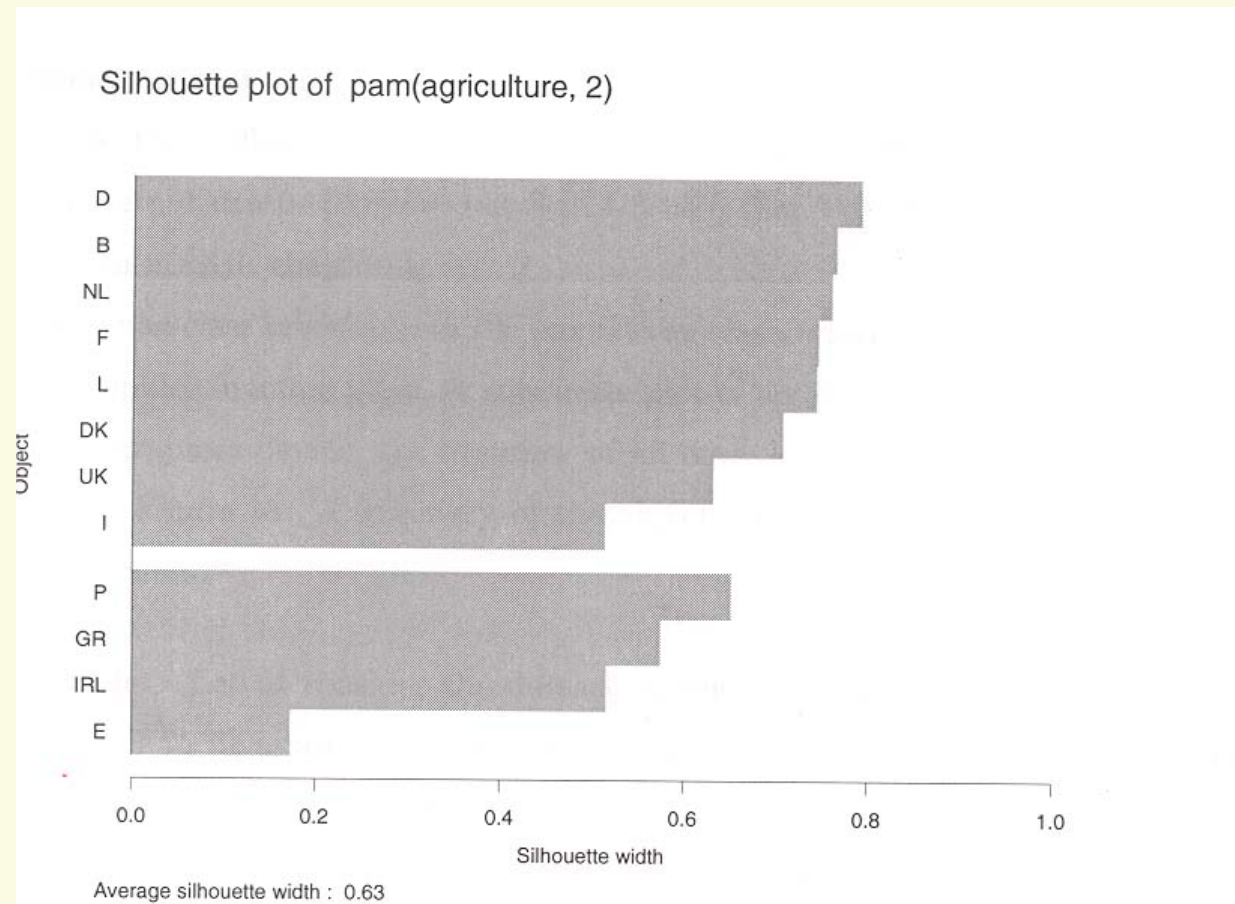
Choosing the number of clusters, cont.

- ❖ The cluster B that minimizes this distance is called the "neighbor".
- ❖ The silhouette is defined as:

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

- ❖ $s(j) \approx 1$ then object is well classified
- $s(j) \approx 0$ then object (gene) is intermediate between 2 clusters
- $s(j) \approx -1$ then object is poorly classified.

Silhouette Plot of Agriculture Example



Sandrine Dudoit
Nicholas P. Jewell

Figure 3: Silhouette plot of the pam clustering of Figure 1.

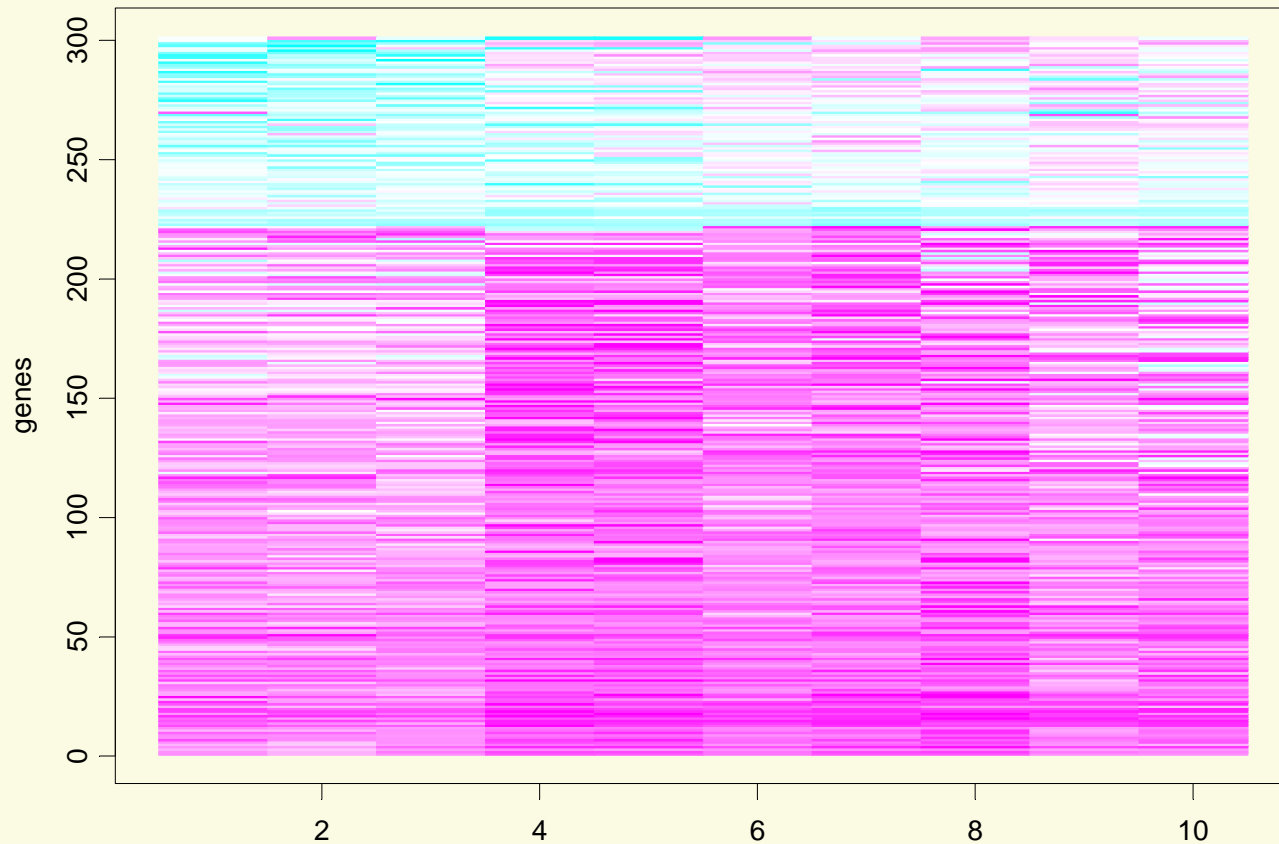
Choosing the number of clusters, cont.

- ❖ Finally, one can use the average silhouette to choose the number of clusters.

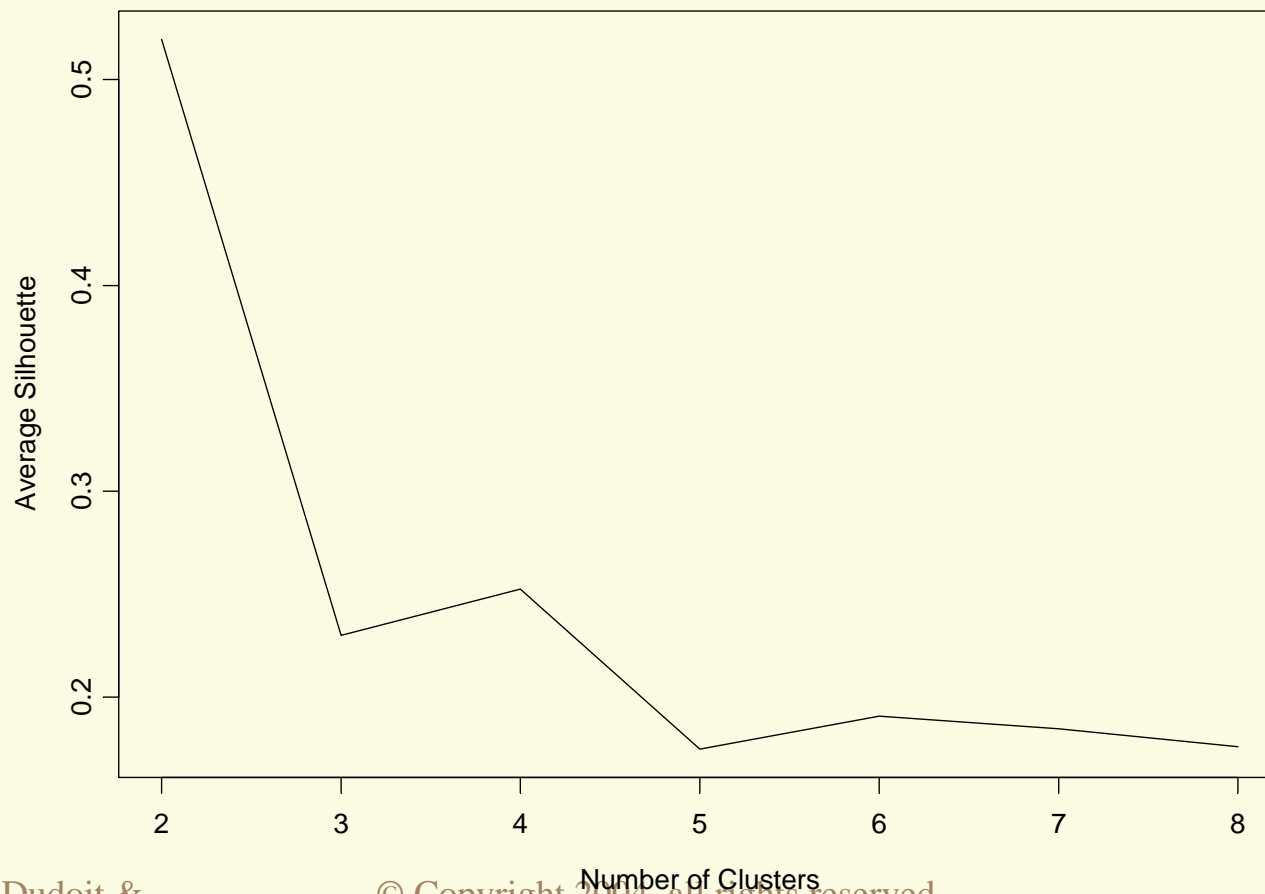
$$\bar{s} = \frac{1}{q} \sum_{j=1}^q s(j)$$

- ❖ Method works by finding the average silhouette for $k=2$ (2 clusters), $k=3$ (3 clusters), etc.
- ❖ Then choosing the k that gives the maximum

Breast cancer: log relative expression data



Average silhouette vs. number of clusters



Aside: Bootstrapping -- Estimating the Variance of an Average and of the Median

- ❖ Original Data is: 1.4, 1.7, 2.3, 1.9, 4.1, 7.2

$$\hat{\mu} = \text{average} = 3.1 \quad \text{median is } 2.1$$

- ❖ First random sample: 1.7, 1.7, 2.3, 4.1, 7.2, 4.1

$$\hat{\mu}_b(1) = 3.5 \quad \text{median}_b = 3.2$$

- ❖ Sec. random sample: 1.4, 2.3, 4.1, 2.3, 7.2, 1.4

$$\hat{\mu}_b(2) = 3.1 \quad \text{median}_b = 2.3$$

- ❖ Do this many, say B , times (something like 1000)

Simple Example Using Bootstrapping - Estimating the Variance of an Average and of the Median

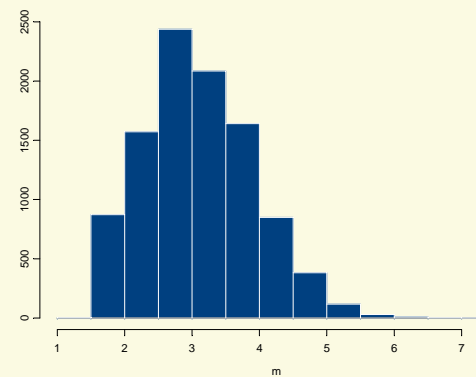
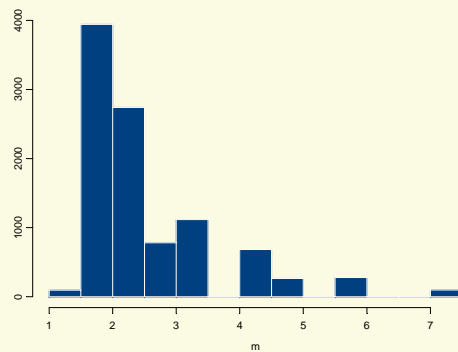
- ❖ Estimate the variance of the sample average as:
median as:

$$\hat{\sigma}^2(\hat{\mu}) = \frac{1}{B} \sum_{l=1}^B (\hat{\mu}_b(l) - \hat{\mu})^2 \quad \hat{\sigma}^2(\text{med}) = \frac{1}{B} \sum_{l=1}^B (\text{med}_b(l) - \text{med})^2$$

- ❖ The trick is to treat the data as the true distribution and bootstrap samples as realizations--then you estimate anything you want that depends on the true distribution.

Variances of Bootstrapped Medians and Means

B	Var(median _b)	Var($\hat{\mu}_b$)
100	0.686	0.785
1000	0.969	0.648
10000	1.099	0.682



Copyright © 2004, by Nicholas P. Jewell
Histogram of bootstrapped medians
Histogram of bootstrapped means

Using bootstrapping to determine the consistency of cluster assignment

- ❖ Assume the number of clusters is fixed at k .
- ❖ Perform PAM on full data set, say PAM^F
- ❖ Then, for each bootstrap sample, perform PAM^b .
- ❖ Medoids (cluster centers) will not be the same in PAM^F and PAM^b .
- ❖ Need to assign the cluster number to those in PAM^b relative to the original in PAM^F .

Assigning cluster number in bootstrap sample relative to original data

- ❖ Use the chosen distance metric and find the pair of medoids in original clustering (PAM^F) and bootstrap clustering (PAM^b) that are the closest.
- ❖ Assign the cluster number from PAM^F to the corresponding one in PAM^b .
- ❖ Remove these medoids from the corresponding sets of medoids in PAM^F and PAM^b .
- ❖ Repeat until all clusters assigned.

Using bootstrapping to determine the consistency of cluster assignment

- ❖ Let $S(j) = m$ be the "true" cluster assignment of gene j .
- ❖ Then $\hat{S}(j)$ is the cluster assignment using the data, i.e., the empirical cluster assignment.
- ❖ Want to estimate the probability that the empirical is the same as the truth, or,

$$P(\hat{S}(j) = m)$$

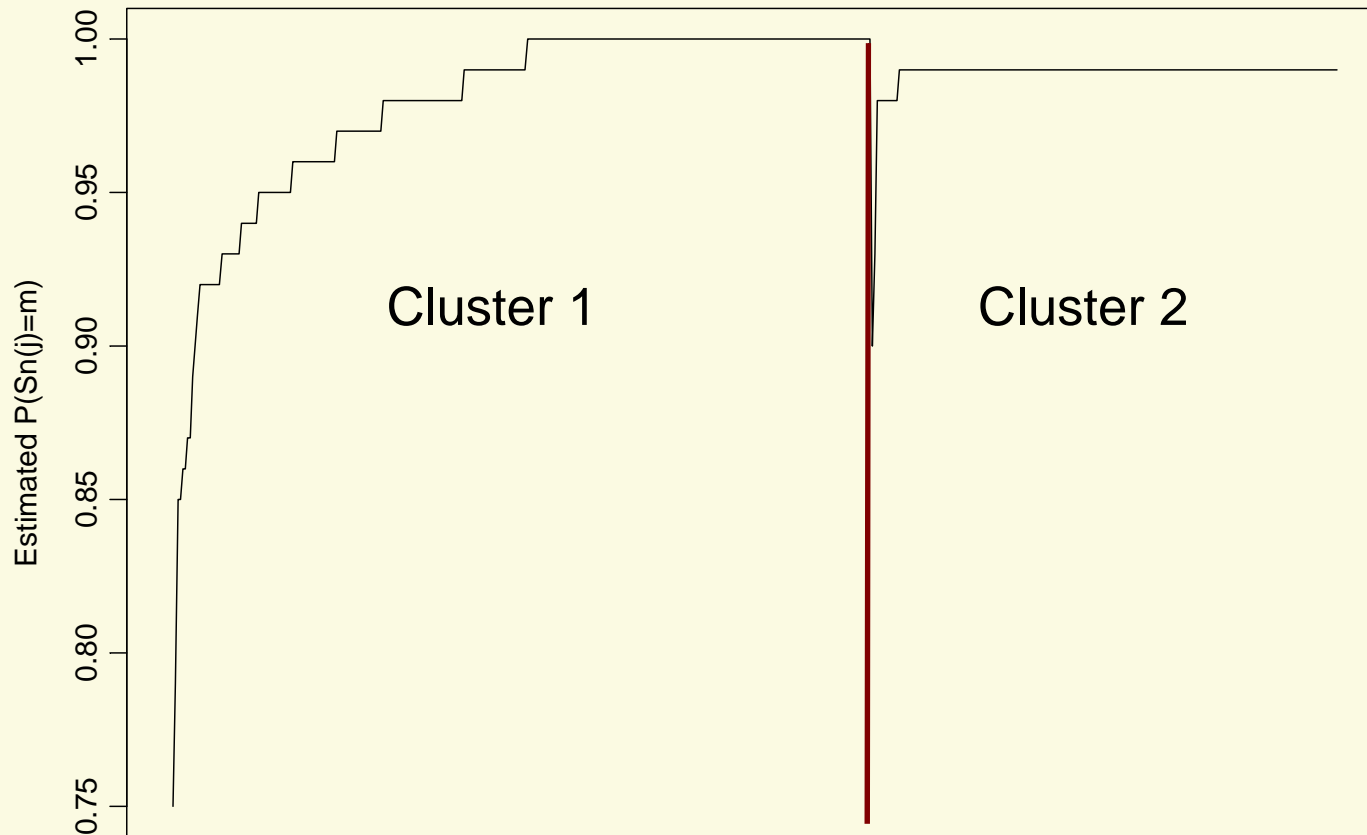
Using bootstrapping to determine the consistency of cluster assignment

- ❖ Estimate this, for each gene separately, as:

$$P(\hat{S}^b(j) = \hat{S}(j)) = \frac{\# \text{ with } \hat{S}^b(j) = \hat{S}(j)}{B}$$

or the proportion of bootstrap samples ($b = 1, 2, \dots, B$) that have the same cluster assignment for gene j in both PAM^F and PAM^b .

Breast Cancer Example



Hierarchical Clustering

- ❖ Two different types of hierarchical clustering methods
 - Partitioning (start with one cluster and split it up successively) (SOM, PAM, K-means)
 - Agglomerative (start with n or p clusters and pool them successively) (DIANA, AGNES, Cluster)
- ❖ Hybrid method : HOPACH (Hierarchical Ordered Partitioning And Collapsing Hybrid)

Example—Alzheimer's Disease

- ❖ cDNA arrays on brain tissue of 20 patients who died of Alzheimer disease.
- ❖ A pooled-reference of normal brain tissues was used for the reference for all samples
- ❖ We only cluster those genes that had a 25% quantile of the ratios of expression (patient vs. normal pool) either $< 1/1.5$ or a 75% quantile > 1.5 , to capture genes that either had at least 25% of the subjects reasonably under or over-expressed relative to the reference (approximately 500 out of the original 20,000 or so genes).

HOPACH

- ❖ General Outline: at each level of the tree
 - Partition: apply PAM, say, to each cluster
 - Order: order the new clusters
 - Collapse: possibly merge some clusters

HOPACH: Partitioning

- ❖ Select the number of "child" clusters for any "parent" cluster by maximizing the average silhouette over a suitable range of cluster sizes

HOPACH: Ordering

- ❖ Always order k child clusters (with medoids M_1, M_2, \dots, M_k) as follows:
 - Distance between clusters = distance between medoids
 - Neighboring cluster of all child clusters is cluster to the "right" of their parent in previous level
 - Order the k child clusters left to right from largest to smallest distance to the neighboring cluster
 - For the right-most group of child clusters, define the neighboring cluster to be the closest one to the left of their parent and proceed similarly
 - For the child clusters at the first level (with no parents), run HOPACH on the medoids themselves, allowing only one split at the first stage

HOPACH: Collapsing

- ❖ Two or more clusters at a given tree level may be very similar whether they share the same parent or not
- ❖ Data driven collapsing: collapse any pair of clusters that improves average silhouette for the whole level
- ❖ A new medoid is picked for the merged cluster
- ❖ When collapsing, choosing which cluster to move over (collapse) to the tree position of the other is arbitrary, or some criterion such as distance of the old medoids to the new one

HOPACH: Visualization

- ❖ At any level, want to visualize the data and distance matrix elements corresponding to an ordering of all elements
 - Clusters are already ordered
 - Elements within a cluster can be ordered (i) at random, (ii) by their distance from the medoid of their cluster (so that badly clustered elements are at the edge of their cluster), or (iii) by their distance to the medoid of a neighboring cluster

❖ Use colors to represent distances

Example—Alzheimer's Disease

- ❖ cDNA arrays on brain tissue of 30 patients who died of Alzheimer disease.
- ❖ A pooled-reference of normal brain tissues was used for the reference for all samples - thus, the null hypothesis is, for example, mean \log_2 relative expression is 0.
- ❖ We only cluster those genes that had a 25% quantile of the ratios of expression (patient vs. normal pool) either $< 1/1.5$ or a 75% quantile > 1.5 , to capture genes that either had at least 25% of the subjects reasonably under or over-expressed relative to the reference (approximately 500 out of the original 20,000 or so genes).

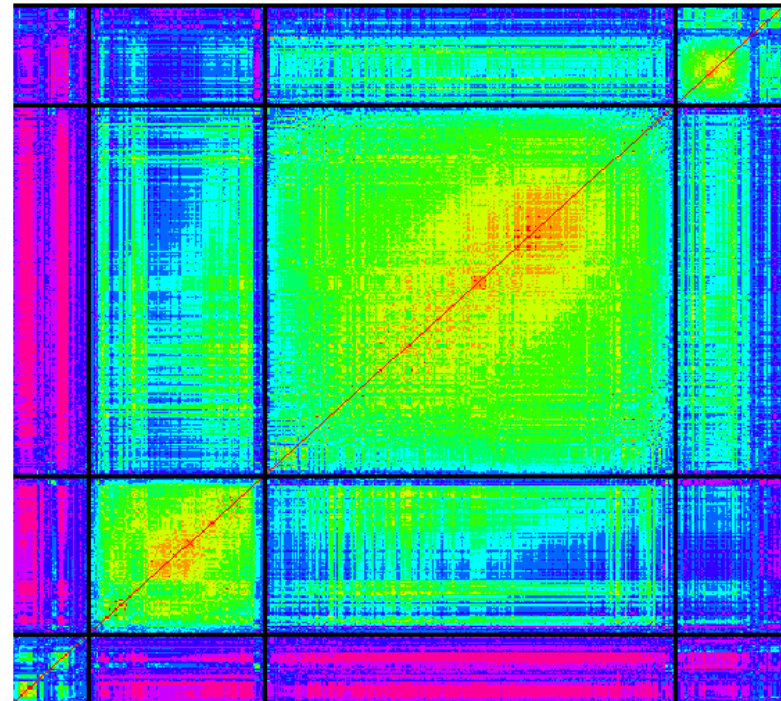
Example—Alzheimer's Disease

- ❖ Relative expressions are logged
- ❖ Distance matrix based on cosine angle

Example—Alzheimer's Disease

Ordered distance matrix of genes with the dark lines surrounding the first level of clustering.

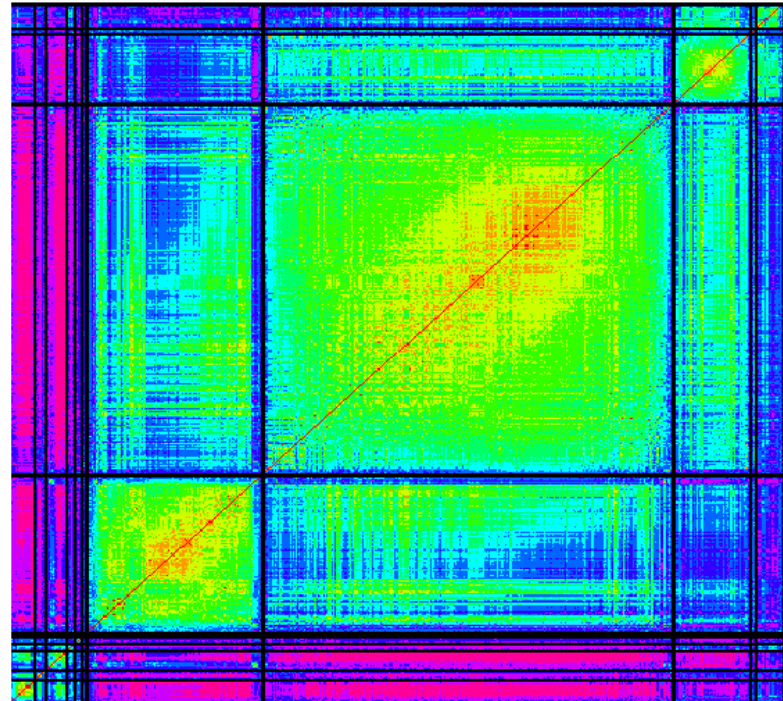
Red to purple is close to far.



Example—Alzheimer's Disease

*Ordered distance matrix of genes
with the second level of clustering
added.*

Red to purple is close to far.



Phylogenetic Trees

- ❖ Construction of phylogenetic trees also depend on hierarchical clustering methods
- ❖ Data are now DNA sequence base information (rather than gene ions expressions)—distances have to be defined for similarity (or dissimilarity) of two sequences
- ❖ Distances can be developed that use stochastic models for base mutations etc

Summary---Lessons Learned

- ❖ Clustering methods are based on distance matrices
- ❖ Bootstrapping procedure very effective in adding statistical estimates of repeatability to clustering analyses.
 - Consistency of cluster assignment.
- ❖ Hierarchical methods provide more information than partitioning methods