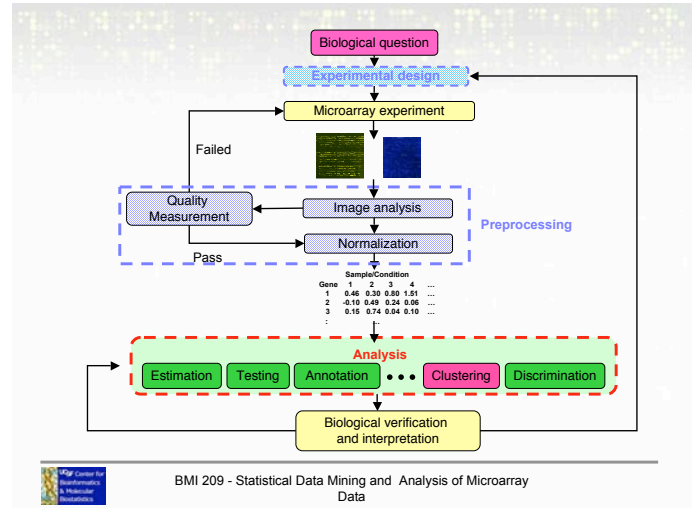


## Lecture 7b: Clustering

Jane Fridlyand

Oct 27, 2005

Fall 2005, BMI 209  
Statistical Data Mining and Analysis of Microarray Data



## Tumor Classification Using Gene Expression Data

Three main types of statistical problems associated with tumor classification:

- Identification of new/unknown tumor classes using gene expression profiles (unsupervised learning – clustering)
- Classification of malignancies into known classes (supervised learning – discrimination)
- Identification of “marker” genes that characterize the different tumor classes (feature or variable selection).

## Clustering

- “Clustering” is an exploratory tool for looking at associations within gene expression data
- These methods allow us to hypothesize about relationships between genes and classes.
- We should use these methods for visualization, hypothesis generation, selection of genes for further consideration
- We should not use these methods inferentially.
- Generally, there is no convincing measure of “strength of evidence” or “strength of clustering structure” provided.
- Hierarchical clustering specifically: we are provided with a picture from which we can make many/any conclusions.

## More specifically....

- Cluster analysis arranges samples and genes into groups based on their expression levels.
- Arrangements are sensitive to choices made with regards to cluster components
- In hierarchical clustering, the VISUALIZATION of the arrangement (the dendrogram) is not unique!

*Just because two samples are situated next to each other does not mean that they are similar.*

## Generic Clustering Tasks

- Assigning objects to the groups
- Estimating number of clusters
- Assessing strength/confidence of cluster assignments for individual objects

## Basic principles of clustering

**Aim:** to group observations that are “similar” based on predefined criteria.

Clustering can be applied to rows (genes) and / or columns (arrays) of an expression data matrix.

Clustering allows for reordering of the rows/columns of an expression data matrix which is appropriate for visualization.



## Basic principles of clustering

### Issues:

- Which genes / arrays to use?
- Which similarity or dissimilarity measure?
- Which method to use to join clusters/observations?
- Which clustering algorithm?
- How to validate the resulting clusters?

It is advisable to **reduce** the number of genes from the full set to some more manageable number, before clustering. The basis for this reduction is usually quite context specific and varies depending on what is being clustered, genes or arrays.



## Clustering microarray data

- Clustering leads to readily interpretable figures and can be helpful for identifying patterns in time or space.

### Examples:

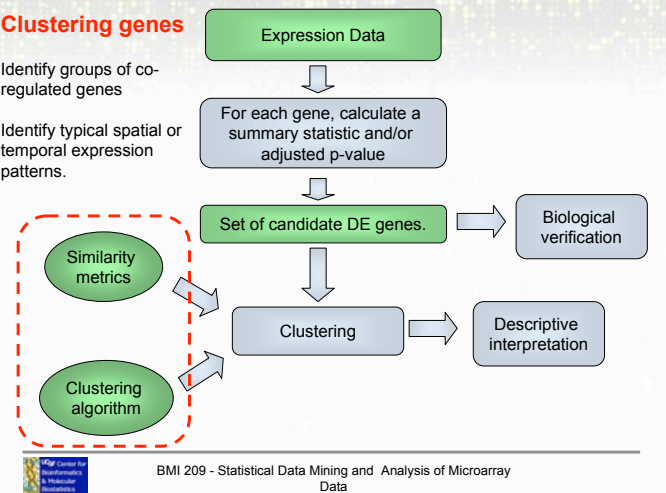
- We can **cluster cell samples** (cols), e.g. 1) for identification (profiles). Here, we might want to estimate the number of different neuron cell types in a set of samples, based on gene expression.  
2) the identification of new / unknown tumor classes using gene expression profiles.
- We can **cluster genes** (rows), e.g. using large numbers of yeast experiments, to identify groups of co-regulated genes.
- We can **cluster genes** (rows) to reduce redundancy (cf. variable selection) in predictive models.



### Clustering genes

Identify groups of co-regulated genes

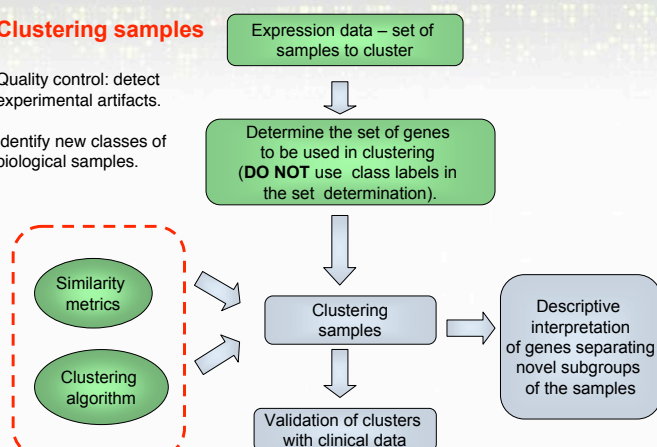
Identify typical spatial or temporal expression patterns.



### Clustering samples

Quality control: detect experimental artifacts.

Identify new classes of biological samples.



## Commonly used measure?

- A metric is a measure of the **similarity** or **dissimilarity** between two data objects and it's used to form data points into clusters
- Two main classes of distance:
  - **1- Correlation coefficients** (scale-invariant)
  - **Distance metric** (scale-dependent)



## Some correlations to choose from

- Pearson Correlation:

$$s(x_1, x_2) = \frac{\sum_{k=1}^K (x_{1k} - \bar{x}_1)(x_{2k} - \bar{x}_2)}{\sqrt{\sum_{k=1}^K (x_{1k} - \bar{x}_1)^2 \sum_{k=1}^K (x_{2k} - \bar{x}_2)^2}}$$

- Uncentered Correlation:

$$s(x_1, x_2) = \frac{\sum_{k=1}^K x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^K x_{1k}^2 \sum_{k=1}^K x_{2k}^2}}$$

- Absolute Value of Correlation:

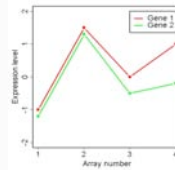
$$s(x_1, x_2) = \frac{\left| \sum_{k=1}^K (x_{1k} - \bar{x}_1)(x_{2k} - \bar{x}_2) \right|}{\sqrt{\sum_{k=1}^K (x_{1k} - \bar{x}_1)^2 \sum_{k=1}^K (x_{2k} - \bar{x}_2)^2}}$$



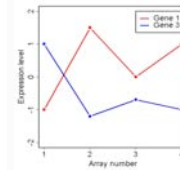
BMI 209 - Statistical Data Mining and Analysis of Microarray Data Adapted from Elizabeth Garrett-Mayer

## Correlation (a measure between -1 and 1)

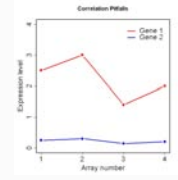
- Others include Spearman's  $\rho$  and Kendall's  $\tau$
- You can use **absolute correlation** to capture both positive and negative correlation



Positive correlation



Negative correlation



Potential pitfalls  
Correlation = 1

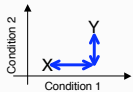


BMI 209 - Statistical Data Mining and Analysis of Microarray Data

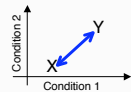
## Distance metrics

- City Block (Manhattan) distance:
  - Sum of differences across dimensions
  - Less sensitive to outliers
  - Diamond shaped clusters
- Euclidean distance:
  - Most commonly used distance
  - Sphere shaped cluster
  - Corresponds to the geometric distance into the multidimensional space

$$d(X, Y) = \sum_i |x_i - y_i|$$



$$d(X, Y) = \sqrt{\sum_i (x_i - y_i)^2}$$



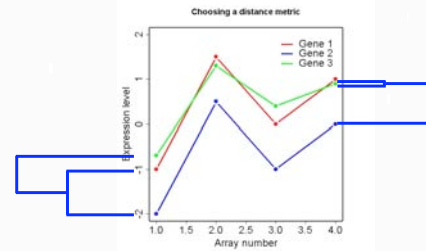
where gene  $X = (x_1, \dots, x_n)$  and gene  $Y = (y_1, \dots, y_n)$



BMI 209 - Statistical Data Mining and Analysis of Microarray Data

## Euclidean vs Correlation

- Euclidean distance
- Correlation



BMI 209 - Statistical Data Mining and Analysis of Microarray Data

## How to Compute Group Similarity?

### Four Popular Methods:

Given two groups  $g_1$  and  $g_2$ ,

- Single-link algorithm:  $s(g_1, g_2)$  = similarity of the **closest** pair
- Complete-link algorithm:  $s(g_1, g_2)$  = similarity of the **furthest** pair
- Average-link algorithm:  $s(g_1, g_2)$  = **average** of similarity of all pairs
- Centroid algorithm:  $s(g_1, g_2)$  = distance between **centroids** of the two clusters



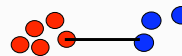
BMI 209 - Statistical Data Mining and Analysis of Microarray Data

Supplementary slide

Adapted from internet

## Distance between clusters

### Between-cluster dissimilarity measures



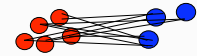
Single (minimum)



Complete (maximum)



Distance between centroids



Average (Mean) linkage



BMI 209 - Statistical Data Mining and Analysis of Microarray Data

## Comparison of the Three Methods

- Single-link
  - Elongated clusters
  - Individual decision, sensitive to outliers
- Complete-link
  - Compact clusters
  - Individual decision, sensitive to outliers
- Average-link or centroid
  - "In between"
  - Group decision, insensitive to outliers

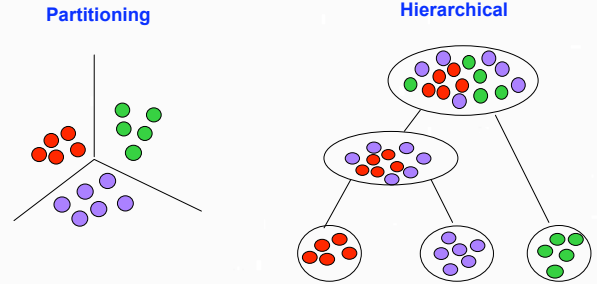
*Which one is the best? Depends on what you need!*

Adapted from internet.



## Clustering algorithms

- Clustering algorithm comes in 2 basic flavors



## Hierarchical Clustering

- The most overused statistical method in gene expression analysis
- Gives us pretty red-green picture with patterns
- But, pretty picture tends to be pretty unstable.
- Many different ways to perform hierarchical clustering
- Tend to be sensitive to small changes in the data
- Provided with clusters of every size: where to "cut" the dendrogram is user-determined

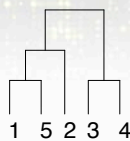


## Agglomerative Methods

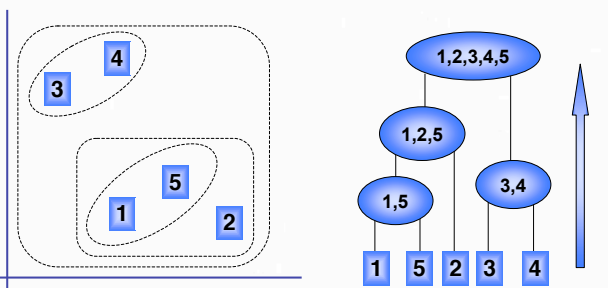
- Start with  $n$  mRNA sample (or  $g$  gene) clusters
- At each step, **merge** the two closest clusters using a measure of between-cluster dissimilarity which reflects the shape of the clusters
- The distance between clusters is defined by the method used (e.g., if complete linkage, the distance is defined as the distance between furthest pair of points in the two clusters)



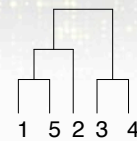
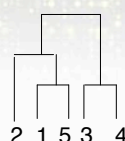
Illustration of points  
In two dimensional  
space



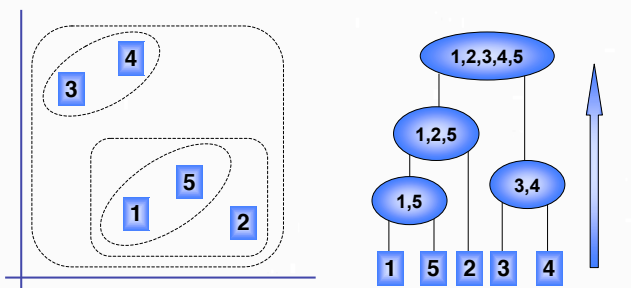
Agglomerative

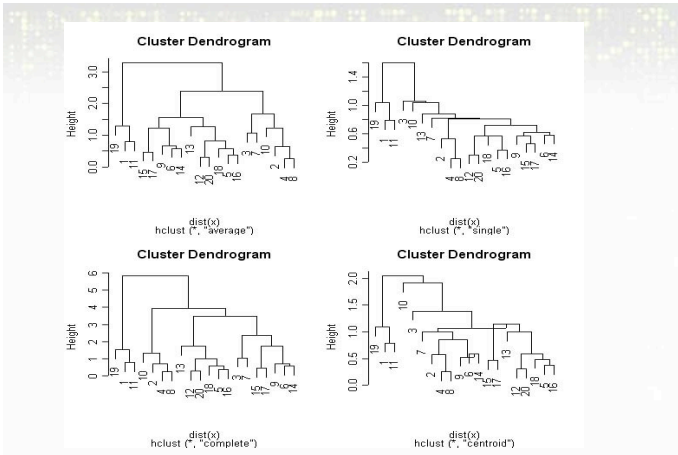


Tree re-ordering?



Agglomerative



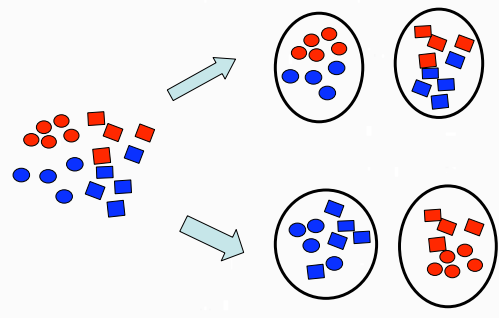


## Partitioning methods

- Partition the data into a **pre-specified** number  $k$  of mutually exclusive and exhaustive groups.
- Iteratively reallocate the observations to clusters until some criterion is met, e.g. minimize within cluster sums of squares. Ideally, dissimilarity *between* clusters will be maximized while it is minimized *within* clusters.
- Examples:
  - k-means, self-organizing maps (SOM), PAM, etc.;
  - Fuzzy (each object is assigned probability of being in a cluster): needs stochastic model, e.g. Gaussian mixtures.

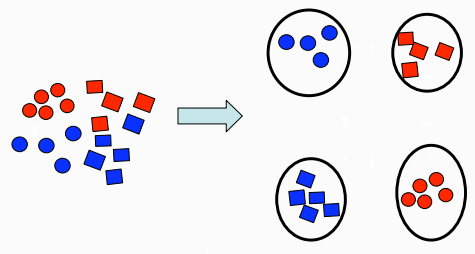
## Partitioning methods

**K = 2**



## Partitioning methods

**K = 4**



## K-means and K-medoids

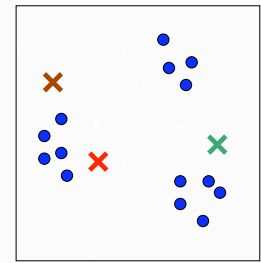
- Partitioning Method
- Don't get pretty picture
- MUST choose number of clusters  $K$  a priori
- More of a "black box" because output is most commonly looked at purely as assignments
- Each object (gene or sample) gets assigned to a cluster
- Begin with initial partition
- Iterate so that objects within clusters are most similar

## How to make a K-means clustering

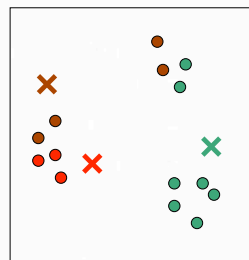
- Choose samples and genes to include in cluster analysis
- Choose similarity/distance metric (generally Euclidean or correlation)
- Choose number of clusters  $K$ .
- Perform cluster analysis.
- Assess cluster fit and stability
- Interpret resulting cluster structure

## K-means Algorithm

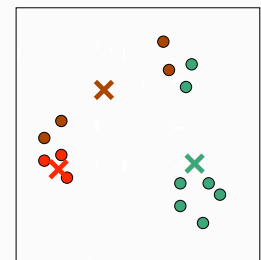
1. Choose K centroids at random or using hierarchical clustering
2. Make initial partition of objects into k clusters by assigning objects to closest centroid
3. Calculate the centroid (mean) of each of the k clusters.
  - a. For object i, calculate its distance to each of the centroids.
  - b. Allocate object i to cluster with closest centroid.
  - c. If object was reallocated, recalculate centroids based on new clusters.
5. Repeat 3 for object  $i = 1, \dots, N$ .
6. Repeat 3 and 4 until no reallocations occur.
7. Assess cluster structure for fit and stability



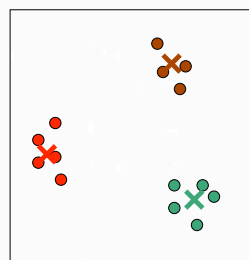
Iteration = 0



Iteration = 1



Iteration = 2

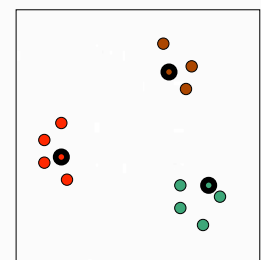


Iteration = 3

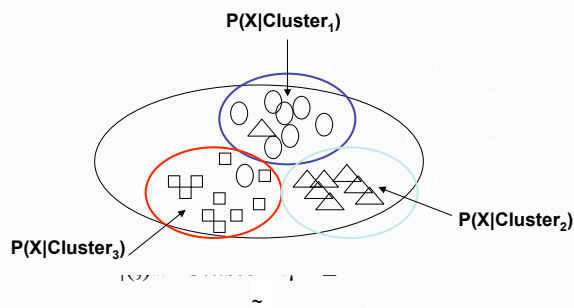


## PAM: Partitioning Around Medoids or K-medoids

- A little different
- Centroid: The average of the samples within a cluster
- Medoid: The "representative object" within a cluster.
- Initializing requires choosing medoids at random.



## Mixture Model for Clustering



## Mixture Model Estimation

- Likelihood function (generally Gaussian)
- Parameters: e.g.,  $\lambda_1, \sigma_1, \mu_1$
- Using EM algorithm
  - Similar to "soft" K-mean
- Number of clusters can be determined using a model-selection criterion, e.g. AIC or BIC (Raftery and Fraley, 1998)



## Some digression into model selection

- Principle of Parsimony: use the smallest number of parameters necessary to represent the data adequately
  - with increasing K (number of parameters), trade-off
  - low K: underfit, miss important effects
  - high K: overfit, include spurious effects and "noise"
  - parsimony – "proper" balance between these 2 effects so that you can repeat results across replications

*AIC/BIC approach – seek a balance between overfit and underfit*

$$AIC = -2 \ln(\text{likelihood}) + 2K; K = \text{number of parameters.}$$



## Partitioning vs. hierarchical

### Partitioning:

#### Advantages

- Optimal for certain criteria.
- Genes automatically assigned to clusters

#### Disadvantages

- Need initial  $k$ ;
- Often require long computation times.
- All genes are forced into a cluster.

### Hierarchical

#### Advantages

- Faster computation.
- Visual

#### Disadvantages

- Unrelated genes are eventually joined
- Rigid, cannot correct later for erroneous decisions made earlier.
- Hard to define clusters.



## How many clusters?

### Global Criteria:

1. Statistics based on within- and between-clusters matrices of sums-of-squares and cross-products (30 methods reviewed by Milligan & Cooper, 1985).
2. Average silhouette (Kaufman & Rousseeuw, 1990).
3. Graph theory (e.g.: cliques in CAST) (Ben-Dor et al., 1999).
4. Model-based methods: EM algorithm for Gaussian mixtures, Fraley & Raftery (1998, 2000) and McLachlan et al. (2001).

### Resampling methods:

1. Gap statistic (Tibshirani et al., 2000).
2. WADP (Bittner et al., 2000).
3. Clest (Dudoit & Fridlyand, 2001).
4. Bootstrap (van der Laan & Pollard, 2001).



## Estimating number of clusters using silhouette (see PAM)

Define silhouette width of the observation is :

$$S = (b-a)/\max(a,b)$$

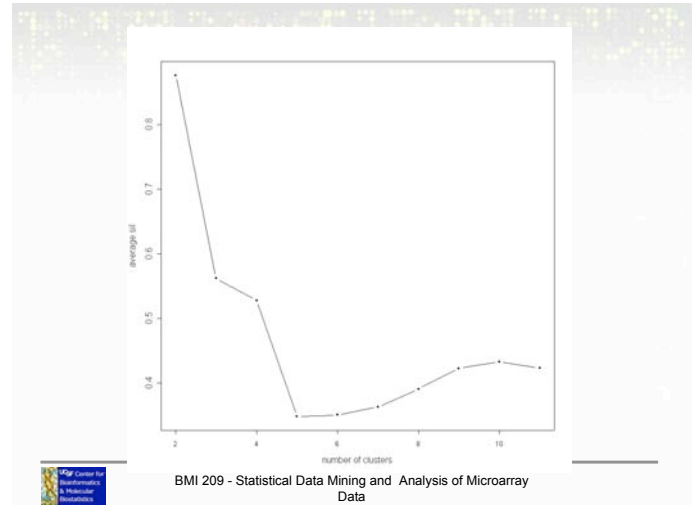
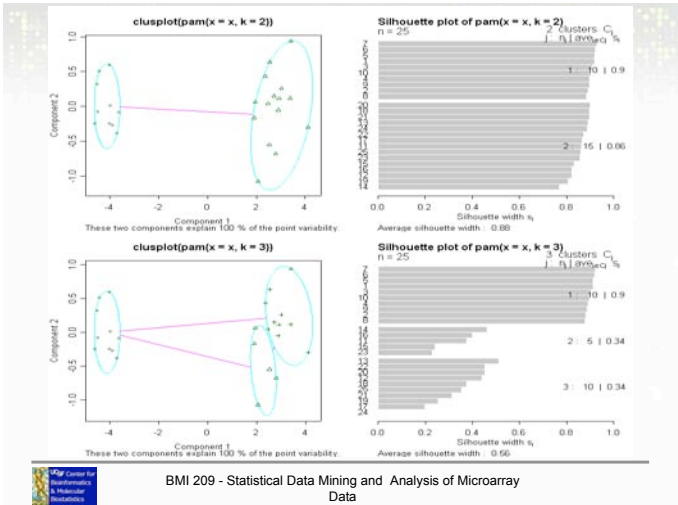
Where  $a$  is the average dissimilarity to all the points in the cluster and  $b$  is the minimum distance to any of the objects in the other clusters.

Intuitively, objects with large  $S$  are well-clustered while the ones with small  $S$  tend to lie between clusters.

**How many clusters:** Perform clustering for a sequence of the number of clusters  $k$  and choose the number of components corresponding to the largest average silhouette.

*Issue of the number of clusters in the data is most relevant for novel class discovery, i.e. for clustering samples.*





### Estimating number of clusters

There are other resampling (e.g. Dudoit and Fridlyand, 2002) and non-resampling based rules for estimating the number of clusters (for review see Milligan and Cooper (1978) and Dudoit and Fridlyand (2002) ).

The bottom line is that none work very well in complicated situation and, to a large extent, clustering lies outside a usual statistical framework.

It is always reassuring when you are able to characterize a newly discovered clusters using information that was not used for clustering.

### Estimating number of clusters using reference distribution

Idea: Define a *goodness of clustering score* to minimize, e.g. pooled Within clusters Sum of Squares (WSS) around the cluster means, reflecting compactness of clusters.

$$W_k = \sum_{i=1}^k \frac{1}{2n_i} D_i$$

where  $n_i$  and  $D_i$  are the number of points in the cluster and sum of all pairwise distances, respectively.

Then gap statistic for  $k$  clusters is defined as:

$$Gap_n(k) = E_n^*(\log(W_k)) - \log(W_k)$$

Where  $E^*n$  is the average under a sample of the same size from the reference distribution. Reference distribution can be generated either parametrically (e.g. from a multivariate) or non-parametrically (e.g. by sampling from marginal distributions of the variables). The first local maximum is chosen to be the number of clusters (slightly more complicated rule) (Tibshirani et al, 2001)

### Clest

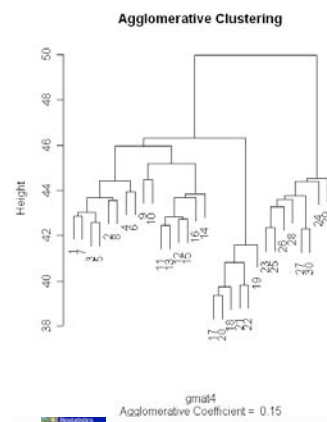
Combines supervised and unsupervised approaches:

For each  $K$  in  $2 \dots K_{max}$

- Repeatedly split the observations into training and test set
- Cluster training and test sets into  $K$  clusters
- Use training set to build a predictor using the resulting cluster labels
- Assess how well predicted labels match the cluster results on the training set

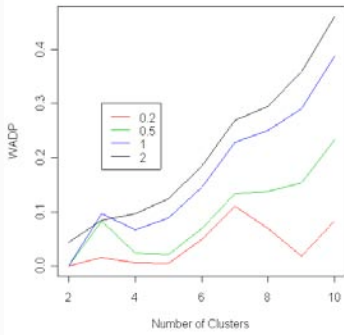
Assessment is done by considering null distribution of the "agreement" statistic.

### WADP: Weighted Average Discrepancy Pairs



- Add perturbations to original data
- Calculate the number of paired samples that cluster together in the original cluster that didn't in the perturbed
- Repeat for every cutoff (i.e. for each  $k$ )
- Do iteratively
- Estimate for each  $k$  the proportion of discrepant pairs.

## WADP



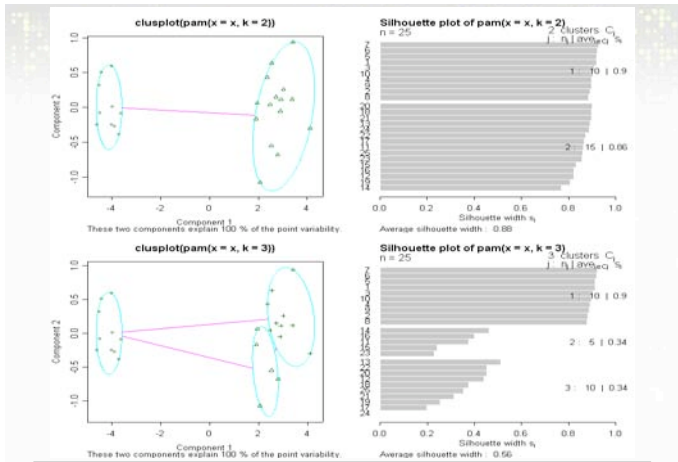
- Different levels of noise have been added
- We look for largest  $k$  before WADP gets big.
- Note that different levels of noise provide different suggested cut-off



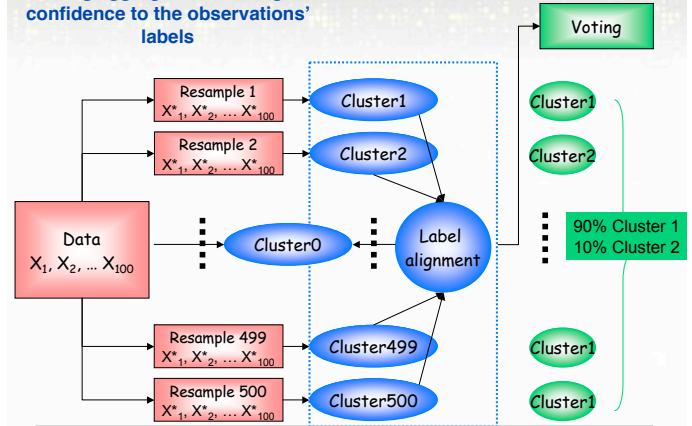
## Confidence in of the individual cluster assignments

Want to assign confidence to individual observations of being in their assigned clusters.

- Model-based clustering: natural probability interpretation
- Partitioning methods: silhouette
- Dudoit and Fridlyand (2003) have presented a resampling-based approach that assigns confidence by computing the proportion of resampling times that an observation ends up in the assigned cluster.



## Using aggregation to assign confidence to the observations' labels



- Number of clusters  $K$  needs to be fixed a-priori
- Has been shown on simulated data to improve quality of cluster assignment
- Interesting alternative by-product:
  - For each pair of samples, compute proportion of bootstrap iterations where they were co-clustered
  - Use 1-proportion as a new distance metric
  - Re-cluster observations using this new distance metric



## Hybrid methods: HOPACH

- Hierarchical Ordered Partitioning and Collapsing Hybrid (between divisive and agglomerative methods)
- Reference: van der Laan & Pollard (2001).
  - Apply a partitioning algorithm iteratively to produce a hierarchical tree of clusters.
  - At each node, a cluster is partitioned into **two or more** smaller clusters. Splits are not restricted to be binary. E.g., choose  $K$  based on average silhouette.



## Tight clustering (genes)

Identifies small stable gene clusters by not attempting to cluster all the genes. Thus, it does not necessitate estimation of the number of clusters and assignment of all points into the clusters. Aids interpretability and validity of the results. (Tseng et al, 2003)

### Algorithm:

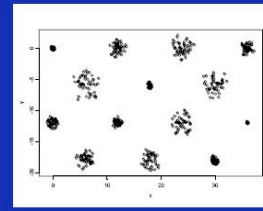
For sequence of  $k > k_0$ :

1. Identify the set of genes that are consistently grouped together when genes are repeatedly sub-sampled. Order those sets by size. Consider the top largest  $q$  sets for each  $k$ .
2. Stop when for  $(k, (k+1))$ , the two sets are nearly identical. Take the set corresponding to  $(k+1)$ . Remove that set from the dataset.
3. Set  $k_0 = k_0 - 1$  and repeat the procedure.



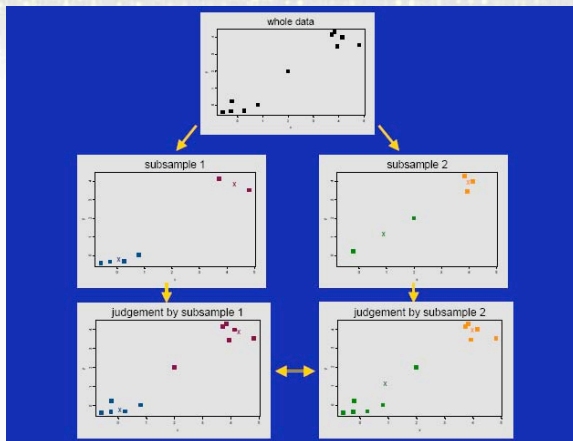
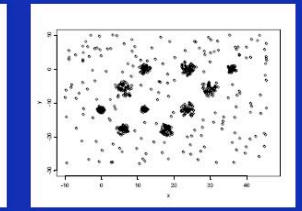
### Traditional:

- Estimate the number of clusters,  $k$  (except for hierarchical clustering)
- Perform clustering through assigning all genes into clusters.



### Tight Clustering:

- Directly identify informative, tight and stable clusters with reasonable size, say, 20-60 genes.
- Need not estimate  $k$  !!
- Need not assign all genes into clusters.

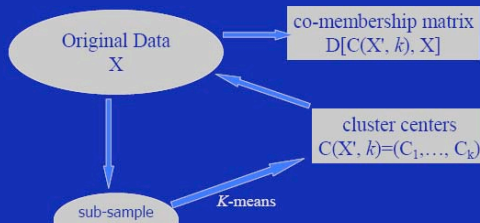


## Algorithm (Tight Clustering)

- $X = \{x_{ij}\}_{n \times d}$ : data to be clustered.
- $X' = \{x'_{ij}\}_{n/2 \times d}$ : random sub-sample
- $C(X', k) = (C_1, C_2, \dots, C_k)$ : the cluster centers obtained from clustering  $X'$  into  $k$  clusters.
- $D[C(X', k), X]$ : an  $n \times n$  matrix denoting co-membership relations of  $X$  classified by  $C(X', k)$ . (Tibshirani 2001)
 
$$D[C(X', k), X]_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ in the same cluster.} \\ 0 & \text{o.w.} \end{cases}$$
- $s(V_i, V_j) = \frac{|V_i \cap V_j|}{|V_i \cup V_j|}$ : a measure of similarity of two sets of genes



## Algorithm (Tight Clustering)



## Algorithm (Tight Clustering)

Algorithm 1 (when fixing  $k$ ):

1. Fix  $k$ . Random sub-sampling  $X^{(1)}, \dots, X^{(B)}$ . Define the average co-membership matrix to be

$$\bar{D} = \text{mean}(D[C(X^{(1)}, k), X], \dots, D[C(X^{(B)}, k), X])$$

Note:

- a.  $\bar{D}_{ij} = 1 \Rightarrow i$  and  $j$  always clustered together in each sub-sampling judgment.
- b.  $\bar{D}_{ij} = 0 \Rightarrow i$  and  $j$  never clustered together in each sub-sampling judgment.
- c.  $\bar{D}_{ii} = 1 \quad \forall i$



## Algorithm (Tight Clustering)

Algorithm 1 (when fixing  $k$ ): (cont'd)

2. Search for a large set of points

$$V = \{v_1, \dots, v_m\} \subseteq \{1, \dots, n\} \text{ such that } \bar{D}_{v_i, v_j} \geq 1 - \alpha \quad \forall i, j$$

$\alpha$  close to 0. Sets with this property are candidates of tight clusters. Order sets with this property by their size to obtain  $V_{k1}, V_{k2}, \dots$

...



## Algorithm (Tight Clustering)

Tight Clustering Algorithm:

1. Start with a suitable  $k_0$ . Search for consecutive  $k$ 's and choose the top 3 clusters for each  $k$ .

$$\{V_{k_0,1}, V_{k_0,2}, V_{k_0,3}\}, \{V_{(k_0+1),1}, V_{(k_0+1),2}, V_{(k_0+1),3}\}, \dots$$

2. Stop when

$$s(V_{k'l}, V_{(k'+1)m}) \geq \beta, \quad s(V_{(k'+1)m}, V_{(k'+2)n}) \geq \beta$$

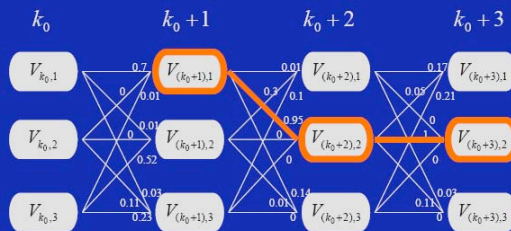
$$k' \geq k_0, \quad l, m, n \in \{1, 2, 3\}, \quad \beta \text{ close to } 1$$

Select  $V_{(k'+1)m}$  to be the tightest cluster.



## Algorithm (Tight Clustering)

Tight Clustering Algorithm:



## Algorithm (Tight Clustering)

Tight Clustering Algorithm: (cont'd)

3. Identify the tightest cluster and remove it from the whole data.

4. Decrease  $k_0$  by 1. Repeat 1.-3. to identify the next tight cluster.

Remark:  $\alpha$ ,  $\beta$  and  $k_0$  determines the tightness and size of resulting clusters.

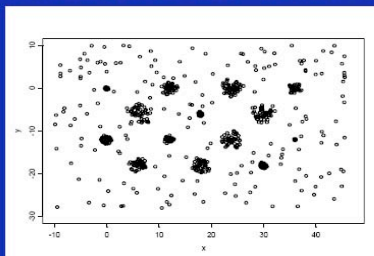


## Simulation

A simple simulation on 2-D:

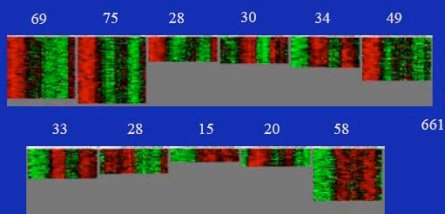
14 clusters normally distributed (50 points each)

175 sporadic points. Stdev=0.1, 0.2, ..., 1.4



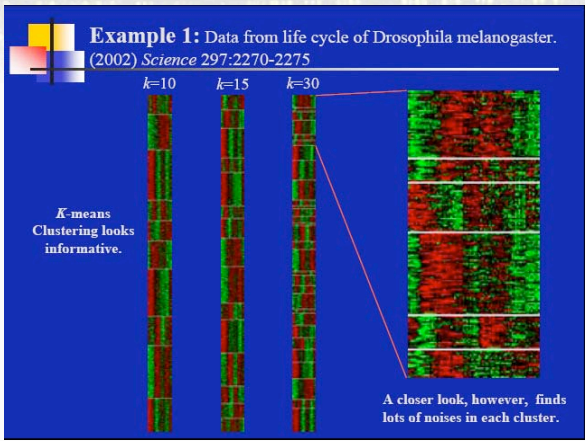
## Example 1: Data from life cycle of *Drosophila melanogaster*. (2002) *Science* 297:2270-2275

Tight Clustering  $\alpha = 0.1, \beta = 0.6, k_0 = 15$



11 clusters and 661 remaining scattered genes





## Two-way clustering of genes and samples.

Refer to the methods that use samples and genes simultaneously to extract information. These methods are not yet well developed.

Some examples of the approaches include *Block Clustering* (Hartigan, 1972) which repeatedly rearranges rows and columns to obtain the largest reduction of total within block variance.

Another method is based on *Plaid Models* (Lazzeroni and Owen, 2002)

Friedman and Meulmann (2002) present an algorithm allowing to cluster samples based on the subsets of attributes, i.e. each group of samples could have been characterized by different gene sets.

## Applications of clustering to the microarray data

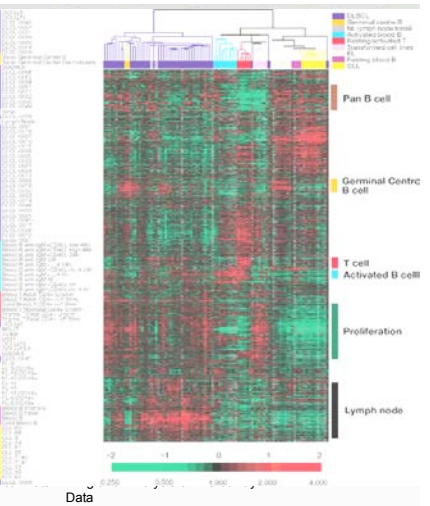
Alizadeh et al (2000) *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.*

- Three subtypes of lymphoma (FL, CLL and DLBCL) have different genetic signatures. (81 cases total)

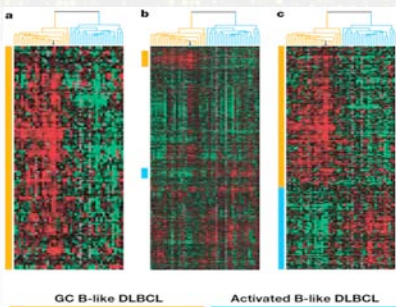
- DLBCL group can be partitioned into two subgroups with significantly different survival. (39 DLBCL cases)

## Clustering both cell samples and genes

Taken from Nature February, 2000 Paper by A Alizadeh et al *Distinct types of diffuse large B-cell lymphoma identified by Gene expression profiling.*

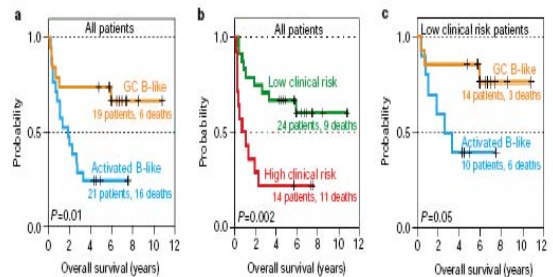


## Clustering cell samples Discovering sub-groups



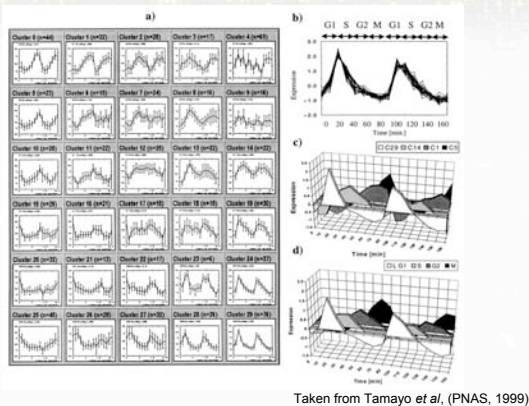
Taken from Alizadeh et al (*Nature*, 2000)

## Attempt at validation of DLBCL subgroups



Taken from Alizadeh et al (*Nature*, 2000)

## Clustering genes Finding different patterns in the data



Yeast Cell Cycle  
(Cho *et al.*, 1998)  
6\_5 SOM with  
828 genes

Taken from Tamayo *et al.*, (PNAS, 1999)



## Summary

### Which clustering method should I use?

- What is the biological question?
- Do I have a preconceived notion of how many clusters there should be?
- Hard or soft boundaries between clusters

### Keep in mind:

- **Clustering cannot NOT work.** That is, every clustering methods will return clusters.
- Clustering helps to group / order information and is a visualization tool for learning about the data. However, clustering results do not provide biological "proof".
- Clustering is generally used as an **exploratory and hypotheses generation** tool.



## Some clustering pitfalls

*The procedure should not bias results towards desired conclusions.*

**Question:** Do expression data cluster according to the survival status.

**Design:** Identify genes with high t-statistic for comparison short and long survivors. Use these genes to cluster samples. Get excited that samples cluster according to survival status.

**Issues:** The genes were already selected based on the survival status. Therefore, it would rather be surprising if samples did \*not\* cluster according to their survival.

**Conclusion:** None are possible with respect to clustering as variable selection was driven by class distinction.



*P-values for differential expression are only valid when the class labels are independent of the current dataset.*

**Question:** Identify genes distinguishing among "interesting" subgroups.

**Design:** Cluster samples into K groups. For each gene, compute F-statistic and its associated p-value to test for differential expression among two subgroups.

**Issues:** Same data was used to create groups as to test for DEs – p-values are invalid.

**Conclusion:** None with respect to DEs p-values. Nevertheless, it is possible to select genes with high value of the statistic and test hypotheses about functional enrichment with, e.g., Gene Ontology. Also, can cluster these genes and use the results to generate new hypotheses.



## Acknowledgements

### SFGH

- Agnes Paquet
- David Erle
- Andrea Barczac
- UCSF Sandler Genomics Core Facility.

### UCSF /CBMB

- Ajay Jain
- Mark Segal
- UCSF Cancer Center Array Core
- Jain Lab

### UCB

- Terry Speed
- Jean Yang



### Some references

1. Hastie, Tibshirani, Friedman "The Elements of Statistical Learning", Springer, 2001
2. Speed (editor) "Statistical Analysis of Gene Expression Microarray Data", Chapman & Hall/CRC, 2003
3. Alizadeh et al, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature, 2000
4. Van 't Veer et al, "Gene expression profiling predicts clinical outcome of breast cancer, Nature, 2002
5. Van de Vijver et al, "A gene-expression signature as a predictor of survival in breast cancer, NEJM, 2002
6. Petricoin et al, "Use of proteomics patterns in serum to identify ovarian cancer", Lancet, 2002 (and relevant correspondence)
7. Golub et al, "Molecular Classification of Cancer: Class Discovery and Class prediction by Gene Expression Monitoring ", Science, 1999
8. Cho et al, A genome-wide transcriptional analysis of the mitotic cell cycle, Mol. Cell, 1999
9. Dudoit, et al, :Comparison of discrimination methods for the classification of tumors using gene expression data, JASA, 2002



### Some references

10. Ambrose and McLachlan, "Selection bias in gene extraction on the basis microarray gene expression data", PNAS, 2002
11. Tibshirani et al, "Estimating the number of clusters in the dataset via the GAP statistic", Tech Report, Stanford, 2000
12. Tseng et al, "Tight clustering : a resampling-based approach for identifying stable and tight patterns in data", Tech Report, 2003
13. Dudoit and Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset ", Genome Biology, 2002
14. Dudoit and Fridlyand, "Bagging to improve the accuracy of a clustering procedure", Bioinformatics, 2003
15. Kaufmann and Rousseeuw, "*Clustering by means of medoids.*", Elsevier/North Holland 1987
16. See many articles by Leo Breiman on aggregation

