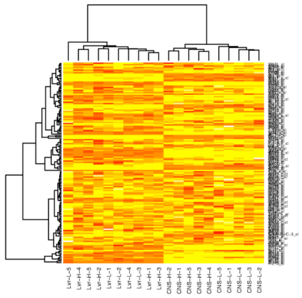
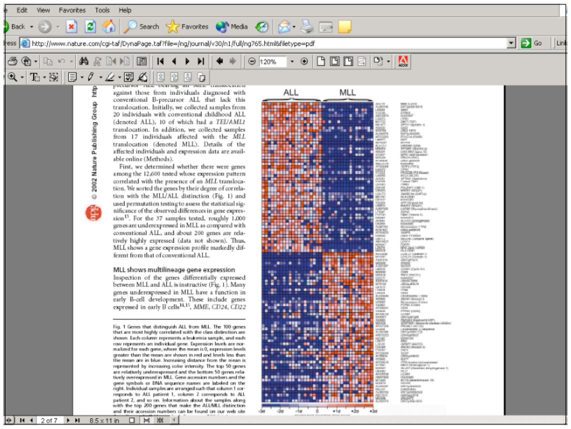
Distances, Clustering, and Classification

Heatmaps




Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Media
http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v30/n1/full/ng765.html&filetype=pdf
Go
Links
120%
© 2002 Nature Publishing Group http
ALL
MLL
2 of 7
8.5 x 11 in

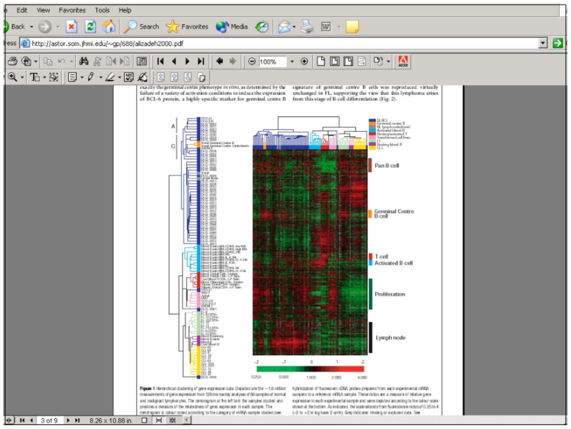
against those from individuals diagnosed with conventional B-precursor ALL that lack this translocation. Initially, we collected samples from 20 individuals with conventional childhood ALL (denoted ALL), 10 of which had a *TEL/AML1* translocation. In addition, we collected samples from 17 individuals affected with the *MLL* translocation (denoted MLL). Details of the affected individuals and expression data are available online (Methods).

First, we determined whether there were genes among the 12,600 tested whose expression pattern correlated with the presence of an *MLL* translocation. We sorted the genes by their degree of correlation with the MLL/ALL distinction (Fig. 1) and used permutation testing to assess the statistical significance of the observed differences in gene expression[13]. For the 37 samples tested, roughly 1,000 genes are underexpressed in MLL as compared with conventional ALL, and about 200 genes are relatively highly expressed (data not shown). Thus, MLL shows a gene expression profile markedly different from that of conventional ALL.

MLL shows multilineage gene expression

Inspection of the genes differentially expressed between MLL and ALL is instructive (Fig. 1). Many genes underexpressed in MLL have a function in early B-cell development. These include genes expressed in early B cells[14,15], *MME*, *CD24*, *CD22*

Fig. 1 Genes that distinguish ALL from MLL. The 100 genes that are most highly correlated with the class distinction are shown. Each column represents a leukemia sample, and each row represents an individual gene. Expression levels are normalized for each gene, where the mean is 0, expression levels greater than the mean are shown in red and levels less than the mean are in blue. Increasing distance from the mean is represented by increasing color intensity. The top 50 genes are relatively underexpressed and the bottom 50 genes relatively overexpressed in MLL. Gene accession numbers and the gene symbols or DNA sequence names are labeled on the right. Individual samples are arranged such that column 1 corresponds to ALL patient 1, column 2 corresponds to ALL patient 2, and so on. Information about the samples along with the top 200 genes that make the ALL/MLL distinction and their accession numbers can be found on our web site



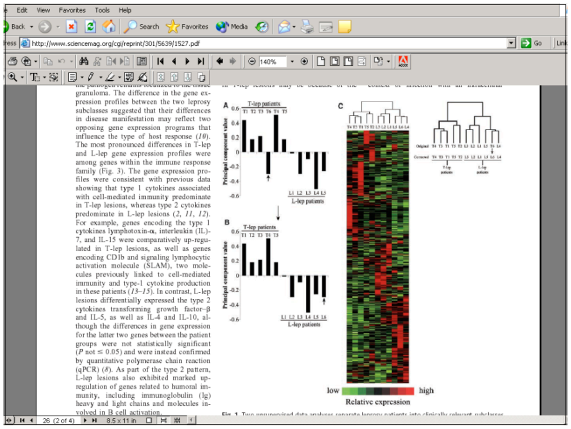
http://astor.som.jhmi.edu/~gp/688/alizadeh2000.pdf

exactly the germinal centre phenotype *in vitro*, as determined by the failure of a variety of activation conditions to induce the expression of BCL-6 protein, a highly specific marker for germinal centre B

signature of germinal centre B cells was reproduced virtually unchanged in FL, supporting the view that this lymphoma arises from this stage of B-cell differentiation (Fig. 2).

Figure 1 Hierarchical clustering of gene expression data. Depicted are the ~1.8 million measurements of gene expression from 128 microarray analyses of 96 samples of normal and malignant lymphocytes. The dendrogram at the left lists the samples studied and provides a measure of the relatedness of gene expression in each sample. The dendrogram is colour coded according to the category of mRNA sample studied (see hybridization of fluorescent cDNA probes prepared from each experimental mRNA samples to a reference mRNA sample. These ratios are a measure of relative gene expression in each experimental sample and were depicted according to the colour scale shown at the bottom. As indicated, the scale extends from fluorescence ratios of 0.25 to 4 (–2 to +2 in log base 2 units). Grey indicates missing or excluded data. See

3 of 9 8.26 x 10.88 in



http://www.sciencemag.org/cgi/reprint/301/5639/1527.pdf

the pathogen remains localized to the tissue granuloma. The difference in the gene expression profiles between the two leprosy subclasses suggested that their differences in disease manifestation may reflect two opposing gene expression programs that influence the type of host response (*10*). The most pronounced differences in T-lep and L-lep gene expression profiles were among genes within the immune response family (Fig. 3). The gene expression profiles were consistent with previous data showing that type 1 cytokines associated with cell-mediated immunity predominate in T-lep lesions, whereas type 2 cytokines predominate in L-lep lesions (*2*, *11*, *12*). For example, genes encoding the type 1 cytokines lymphotoxin-α, interleukin (IL)-7, and IL-15 were comparatively up-regulated in T-lep lesions, as well as genes encoding CD1b and signaling lymphocytic activation molecule (SLAM), two molecules previously linked to cell-mediated immunity and type-1 cytokine production in these patients (*13*–*15*). In contrast, L-lep lesions differentially expressed the type 2 cytokines transforming growth factor–β and IL-5, as well as IL-4 and IL-10, although the differences in gene expression for the latter two genes between the patient groups were not statistically significant (*P* not ≤ 0.05) and were instead confirmed by quantitative polymerase chain reaction (qPCR) (*8*). As part of the type 2 pattern, L-lep lesions also exhibited marked up-regulation of genes related to humoral immunity, including immunoglobulin (Ig) heavy and light chains and molecules involved in B cell activation.

in T-lep lesions may be because of the context of infection with an intracellular

Fig. 1. Two unsupervised data analyses separate leprosy patients into clinically relevant subclasses

26 (2 of 4) 8.5 x 11 in

Distance

- Clustering organizes things that are *close* into groups
- What does it mean for two genes to be close?
- What does it mean for two samples to be close?
- Once we know this, how do we define groups?

Distance

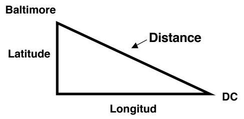
- We need a mathematical definition of distance between two points
- What are points?
- If each gene is a point, what is the mathematical definition of a point?

Points

- Gene1= $(E_{11}, E_{12}, \ldots, E_{1N})'$
- Gene2= $(E_{21}, E_{22}, \ldots, E_{2N})'$
- Sample1= $(E_{11}, E_{21}, \ldots, E_{G1})'$
- Sample2= $(E_{12}, E_{22}, \ldots, E_{G2})'$
- E_{gi}=expression gene g, sample i

Most Famous Distance

- Euclidean distance
  - Example distance between gene 1 and 2:
  - Sqrt of Sum of $(E_{1i} - E_{2i})^2$, $i=1,\ldots,N$
- When N is 2, this is distance as we know it:

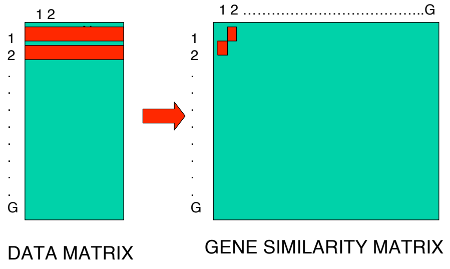


When N is 20,000 you have to think abstractly

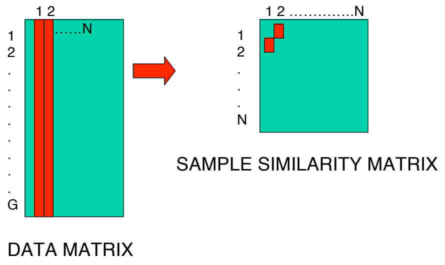
Similarity

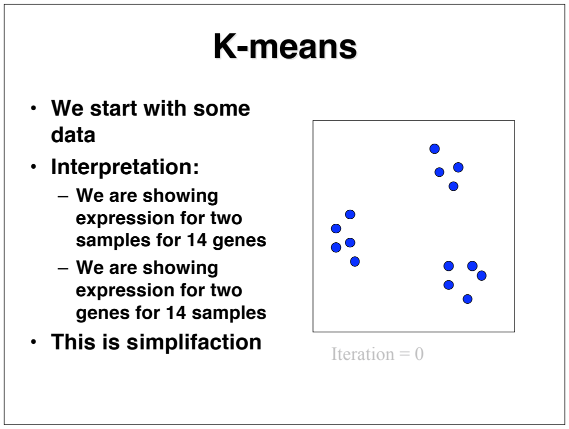
- Instead of distance, clustering can use *similarity*
- If we standardize points then Euclidean distance is equivalent to using absolute value of correlation as a similarity index
- Other examples:
  - Spearman correlation
  - Categorical measures

The similarity/distance matrices



The similarity/distance matrices

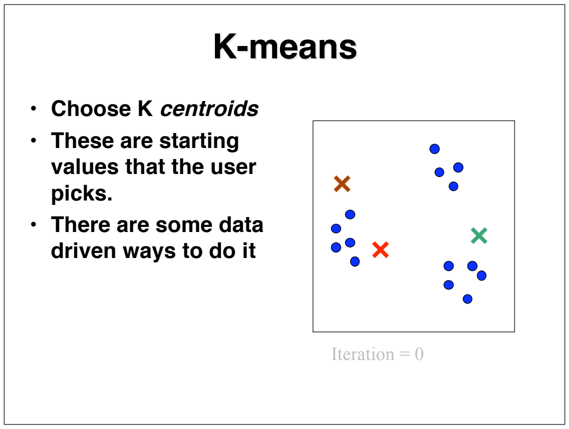




K-means

- We start with some data
- Interpretation:
  - We are showing expression for two samples for 14 genes
  - We are showing expression for two genes for 14 samples
- This is simplifaction

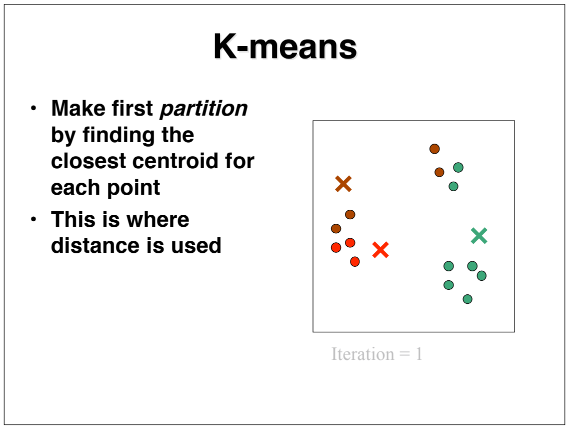
Iteration = 0



K-means

- Choose K *centroids*
- These are starting values that the user picks.
- There are some data driven ways to do it

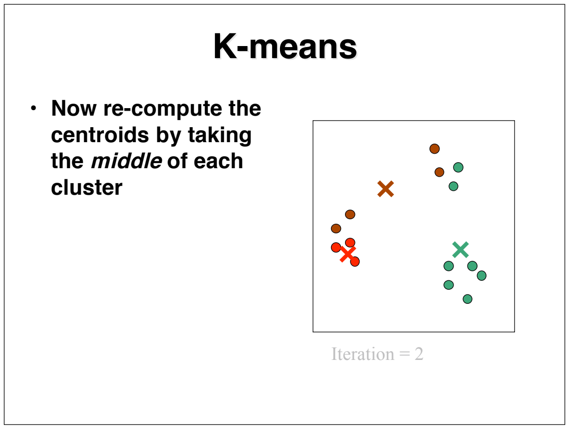
Iteration = 0



K-means

- Make first *partition* by finding the closest centroid for each point
- This is where distance is used

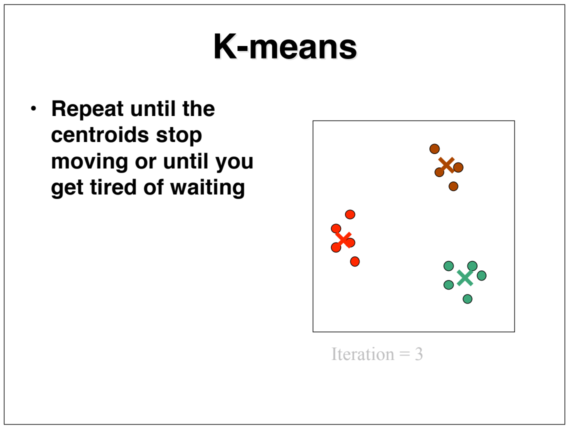
Iteration = 1



K-means

- Now re-compute the centroids by taking the *middle* of each cluster

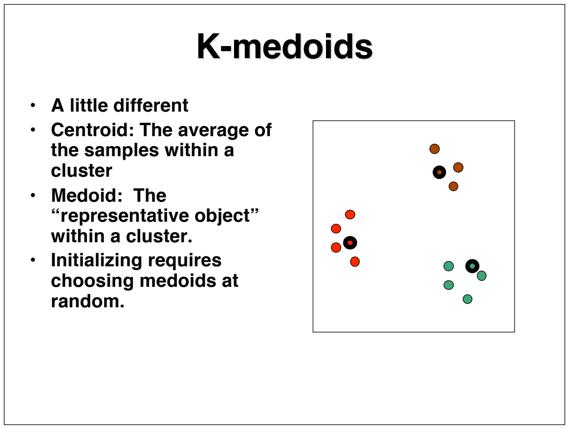
Iteration = 2



K-means

- Repeat until the centroids stop moving or until you get tired of waiting

Iteration = 3



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.

K-means Limitations

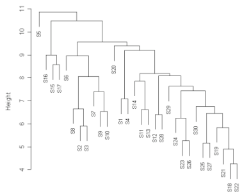
- Final results depend on starting values
- How do we chose K? There are methods but not much theory saying what is best.
- Where are the pretty pictures?

Hierarchical

- Divide all points into 2. Then divide each group into 2. Keep going until you have groups of 1 and can not divide further.
- This is divisive or top-down hierarchical clustering. There is also agglomerative clustering or bottom-up

Dendrograms

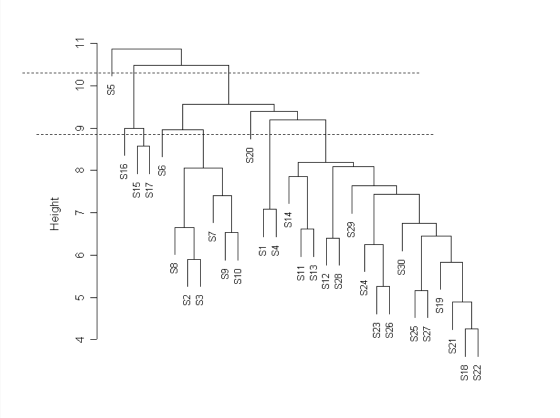
- We can then make dendrograms showing divisions
- The y-axis represents the distance between the groups divided at that point



Note: Left and right is assigned arbitrarily.
Look at the height of division to find out distance.
For example, S5 and S16 are very far.

But how do we form actual clusters?

We need to pick a height



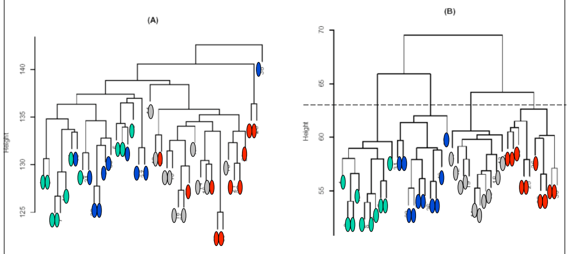
How to make a hierarchical clustering

1. Choose samples and genes to include in cluster analysis
2. Choose similarity/distance metric
3. Choose clustering direction (top-down or bottom-up)
4. Choose linkage method (if bottom-up)
5. Calculate dendrogram
6. Choose height/number of clusters for interpretation
7. Assess cluster fit and stability
8. Interpret resulting cluster structure

1. Choose samples and genes to include

- Important step!
- Do you want housekeeping genes included?
- What to do about replicates from the same individual/tumor?
- Genes that contribute noise will affect your results.
- Including all genes: dendrogram can't all be seen at the same time.
- Perhaps screen the genes?

Simulated Data with 4 clusters: 1-10, 11-20, 21-30, 31-40



A: 450 relevant genes plus
450 "noise" genes.

B: 450 relevant genes.

2. Choose similarity/distance matrix

- Think hard about this step!
- Remember: garbage in ➔ garbage out
- The metric that you pick should be a valid measure of the distance/similarity of genes.
- Examples:
  - Applying correlation to highly skewed data will provide misleading results.
  - Applying Euclidean distance to data measured on categorical scale will be invalid.
- Not just "wrong", but which makes most sense

Some correlations to choose from

- **Pearson Correlation:**

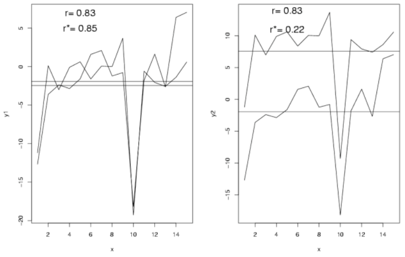
$$s(x_1, x_2) = \frac{\sum_{k=1}^{K} (x_{1k} - \bar{x}_1)(x_{2k} - \bar{x}_2)}{\sqrt{\sum_{k=1}^{K} (x_{1k} - \bar{x}_1)^2 \sum_{k=1}^{K} (x_{2k} - \bar{x}_2)^2}}$$

- **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:**

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



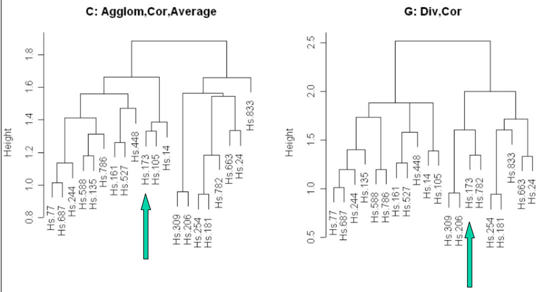
The difference is that, if you have two vectors X and Y with identical shape, but which are offset relative to each other by a fixed value, they will have a standard Pearson correlation (centered correlation) of 1 but will not have an uncentered correlation of 1.

3. Choose clustering direction (top-down or bottom-up)

- **Agglomerative clustering (bottom-up)**
  - **Starts with as each gene in its own cluster**
  - **Joins the two most similar clusters**
  - **Then, joins next two most similar clusters**
  - **Continues until all genes are in one cluster**
- **Divisive clustering (top-down)**
  - **Starts with all genes in one cluster**
  - **Choose split so that genes in the two clusters are most similar (maximize “distance” between clusters)**
  - **Find next split in same manner**
  - **Continue until all genes are in single gene clusters**

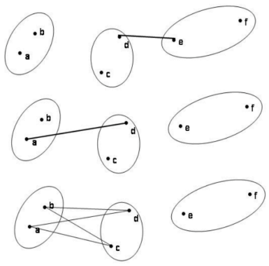
Which to use?

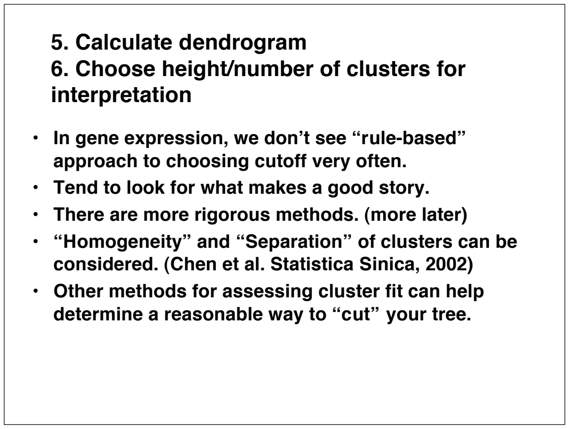
- **Both are** only **'step-wise' optimal: at each step the optimal split or merge is performed**
- **This does not imply that the final cluster structure is optimal!**
- **Agglomerative/Bottom-Up**
  - **Computationally simpler, and more available.**
  - **More "precision" at bottom of tree**
  - **When looking for small clusters and/or many clusters, use agglomerative**
- **Divisive/Top-Down**
  - **More "precision" at top of tree.**
  - **When looking for large and/or few clusters, use divisive**
- In gene expression applications, divisive makes more sense**.**
- **Results ARE sensitive to choice!**



4. Choose linkage method (if bottom-up)

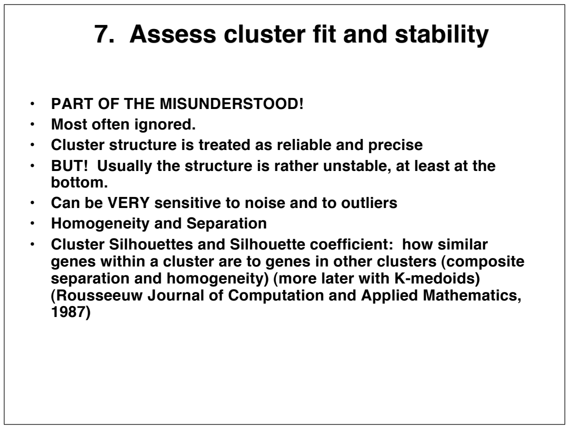
- **Single Linkage: join clusters whose distance between closest genes is smallest (elliptical)**
- **Complete Linkage: join clusters whose distance between furthest genes is smallest (spherical)**
- **Average Linkage: join clusters whose average distance is the smallest.**





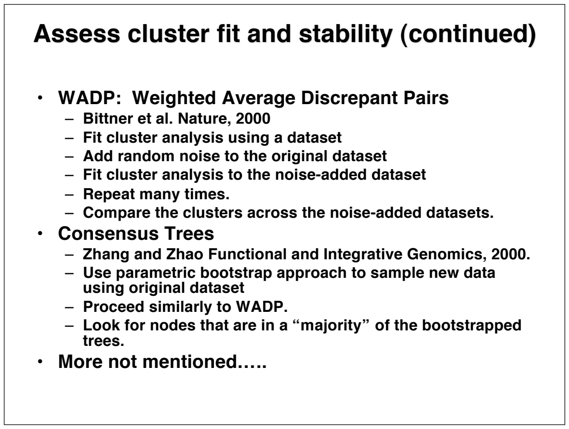
5. Calculate dendrogram
6. Choose height/number of clusters for interpretation

- In gene expression, we don't see "rule-based" approach to choosing cutoff very often.
- Tend to look for what makes a good story.
- There are more rigorous methods. (more later)
- "Homogeneity" and "Separation" of clusters can be considered. (Chen et al. Statistica Sinica, 2002)
- Other methods for assessing cluster fit can help determine a reasonable way to "cut" your tree.



7. Assess cluster fit and stability

- PART OF THE MISUNDERSTOOD!
- Most often ignored.
- Cluster structure is treated as reliable and precise
- BUT! Usually the structure is rather unstable, at least at the bottom.
- Can be VERY sensitive to noise and to outliers
- Homogeneity and Separation
- Cluster Silhouettes and Silhouette coefficient: how similar genes within a cluster are to genes in other clusters (composite separation and homogeneity) (more later with K-medoids) (Rousseeuw Journal of Computation and Applied Mathematics, 1987)



Assess cluster fit and stability (continued)

- WADP: Weighted Average Discrepant Pairs
  - Bittner et al. Nature, 2000
  - Fit cluster analysis using a dataset
  - Add random noise to the original dataset
  - Fit cluster analysis to the noise-added dataset
  - Repeat many times.
  - Compare the clusters across the noise-added datasets.
- Consensus Trees
  - Zhang and Zhao Functional and Integrative Genomics, 2000.
  - Use parametric bootstrap approach to sample new data using original dataset
  - Proceed similarly to WADP.
  - Look for nodes that are in a "majority" of the bootstrapped trees.
- More not mentioned…..

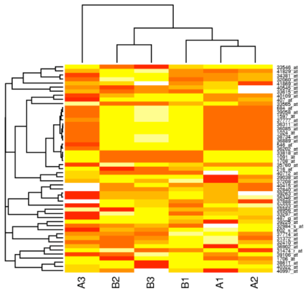
Careful though….

- Some validation approaches are more suited to some clustering approaches than others.
- Most of the methods require us to define number of clusters, even for hierarchical clustering.
  - Requires choosing a cut-point
  - If true structure is hierarchical, a cut tree won't appear as good as it might truly be.

Final Thoughts

- 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

We should not use heatmaps to compare two Populations?



Prediction

Common Types of Objectives

- Class Comparison
  - Identify genes differentially expressed among predefined classes such as diagnostic or prognostic groups.
- Class Prediction
  - Develop multi-gene predictor of class for a sample using its gene expression profile
- Class Discovery
  - Discover clusters among specimens or among genes

What is the task

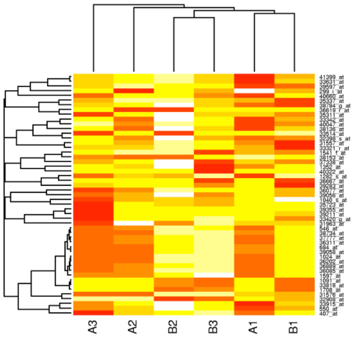
- Given the *gene profile* predict the class
- Mathematical representation: find function *f* that maps *x* to {1,…,K}
- How do we do this?

Possibilities

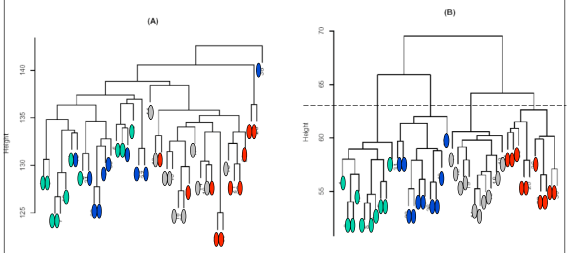
- Have expert tell us what genes to look for being over/under expressed?
- Then we do not really need microarrrays

- Use clustering algorithms?
- Not appropriate for this taks…

Clustering is not a good tool

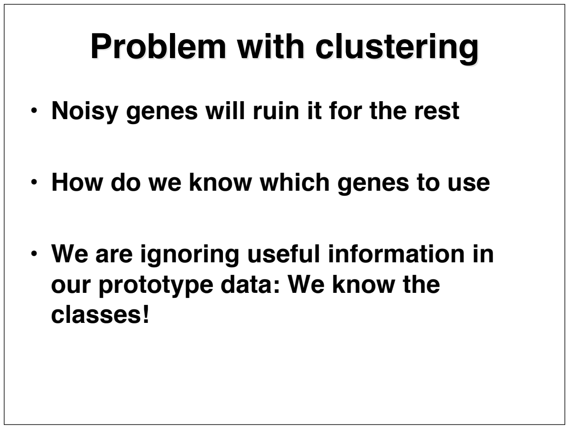


Simulated Data with 4 clusters: 1-10, 11-20, 21-30, 31-40



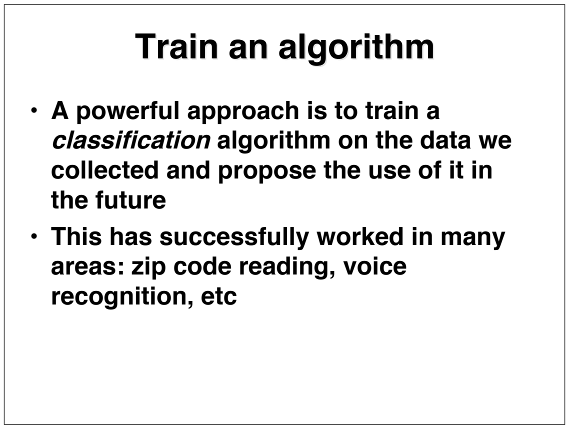
A: 450 relevant genes plus 450 “noise” genes.

B: 450 relevant genes.



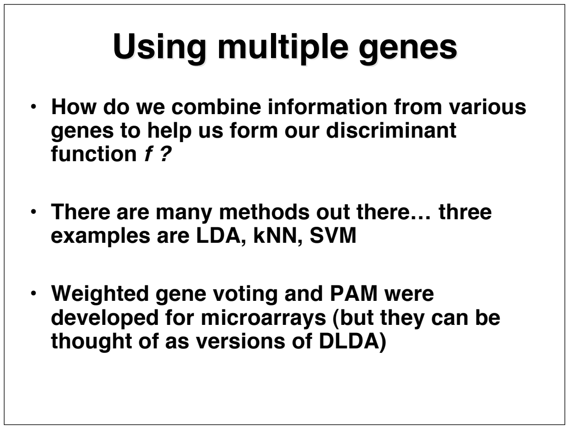
Problem with clustering

- Noisy genes will ruin it for the rest
- How do we know which genes to use
- We are ignoring useful information in our prototype data: We know the classes!



Train an algorithm

- A powerful approach is to train a *classification* algorithm on the data we collected and propose the use of it in the future
- This has successfully worked in many areas: zip code reading, voice recognition, etc



Using multiple genes

- How do we combine information from various genes to help us form our discriminant function *f ?*
- There are many methods out there… three examples are LDA, kNN, SVM
- Weighted gene voting and PAM were developed for microarrays (but they can be thought of as versions of DLDA)

Weighted Gene Voting is DLDA

With equal priors, DLDA is: $\delta_k(x) = \sum_{g=1}^{G} \frac{(x_g - \mu_{kg})^2}{\sigma_g^2}$

With two classes we select class 1 if

$$\sum_{g=1}^{G} \frac{(\bar{x}_{1g} - \bar{x}_{2g})}{\hat{\sigma}_g^2} \left(x_g - \frac{\left(\bar{x}_{1g} + \bar{x}_{2g}\right)}{2} \right) \geq 0$$

This can be written as $\sum_{g=1}^{G} a_g \left(x_g - b_g \right) \geq 0$

with $a_g = \frac{(\bar{x}_{1g} - \bar{x}_{2g})}{\hat{\sigma}_g^2}$ $b_g = \frac{\left(\bar{x}_{1g} + \bar{x}_{2g}\right)}{2}$

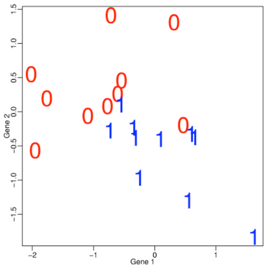
Weighted Gene Voting simply uses $a_g = \frac{(\bar{x}_{1g} - \bar{x}_{2g})}{\hat{\sigma}_{1g} + \hat{\sigma}_{2g}}$

Notice the units and scale fore sum are wrong!

KNN

- Another simple and useful method is K nearest neighbors
- It is very simple

Example



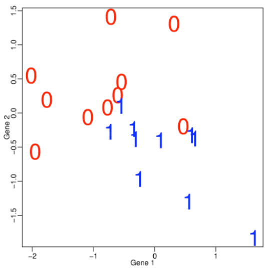
Too many genes

- A problem with most existing approaches: They were not developed for p>>n
- A simple way around this is to filter genes first: Pick genes that, marginally, appear to have good predictive power

Beware of over-fitting

- With p>>n you can always find a prediction algorithm that predicts perfectly on the training set
- Also, many algorithm can be made to me too flexible. An example is KNN with K=1

Example



Split-Sample Evaluation

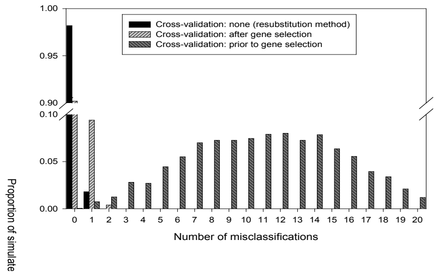
- **Training-set**
  - **Used to select features, select model type, determine parameters and cut-off thresholds**
- **Test-set**
  - **Withheld until a single model is fully specified using the training-set.**
  - **Fully specified model is applied to the expression profiles in the test-set to predict class labels.**
  - **Number of errors is counted**

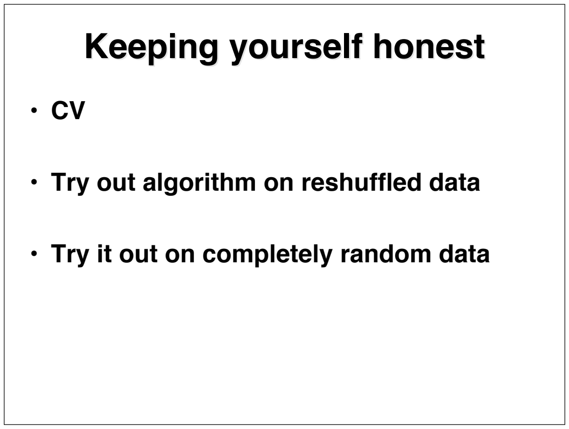
Note: Also called cross-validation

Important

- **You have apply the entire algorithm, from scratch, on the train set**
- **This includes the choice of feature gene, and in some cases normalization!**

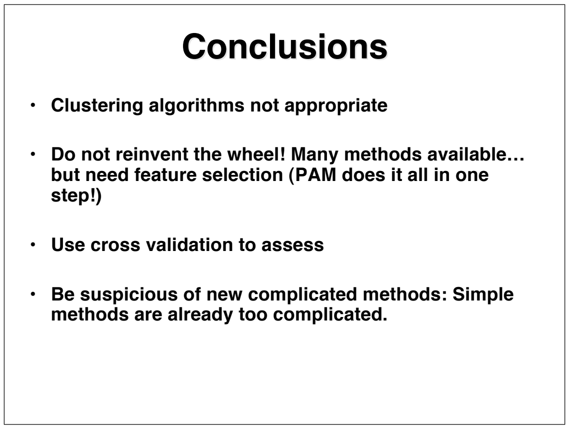
Example





Keeping yourself honest

- CV
- Try out algorithm on reshuffled data
- Try it out on completely random data



Conclusions

- Clustering algorithms not appropriate
- Do not reinvent the wheel! Many methods available… but need feature selection (PAM does it all in one step!)
- Use cross validation to assess
- Be suspicious of new complicated methods: Simple methods are already too complicated.