



# Cluster Analysis

...of genes with similar expression pattern



## Clustering

### Introduction

- Correlated genes are likely to be involved in the same biological pathway.
- Genes transcribed (ie. Activated/Inhibited) by the same transcription factor(s) have a correlation higher than average.
- The above two statements form the basis for the identification of new genes, gene functions and the reconstruction (ie. Reversed engineering) of gene regulatory networks.
- Many clustering algorithms exists.
- Here we shall review four:
  - Hierarchical Clustering
  - Agglomerative Linkage Methods
  - K-Means/Medians
  - Self-Organizing Maps (SOM)



## Clustering

### The Guru

<http://www.hsph.harvard.edu/faculty/JohnQuackenbush.html>

HARVARD SCHOOL OF PUBLIC HEALTH

[Home] [Calendar] [Directory] [Search]

#### John Quackenbush

Professor of Computational Biology and Bioinformatics  
Department of Biostatistics



#### Contact Information

[Department of Biostatistics](#)  
Dana-Farber Cancer Institute  
Mayer 232  
44 Binney Street  
Boston, MA 02115  
Phone: 617-582-8163  
Fax: 617-632-2444  
Email: [johnq@jimmy.harvard.edu](mailto:johnq@jimmy.harvard.edu)

Armidale Animal Breeding Summer Course, UNE, Feb. 2006



## Clustering

### Introduction

- Goal is identify genes (or experiments) which have “similar” patterns of expression
- This is a problem in data mining
- “Clustering Algorithms” are most widely used although many others exist
- Types
  - Agglomerative clustering: Hierarchical
  - Divisive clustering: *k*-means, SOMs
  - Others: Principal Component Analysis (PCA)
- All depend on how one measures **DISTANCE**
- Although a set defined rules exists, clustering is an art.

Armidale Animal Breeding Summer Course, UNE, Feb. 2006



### Clustering

#### Distance Metrics

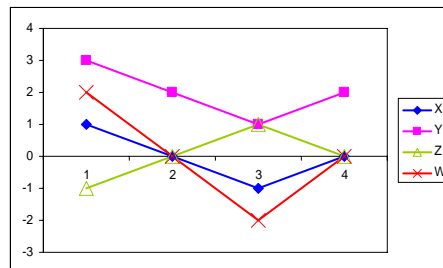
- Distances are measures of “between” expression vectors
- Distance metrics define the way we measure distances
- Many different ways to measure distance:
  - Euclidean distance  $\rightarrow \sqrt{\sum_{j=1}^p (X[j]-Y[j])^2}$
  - Pearson correlation coefficient (r)  $\rightarrow \frac{\sum_{j=1}^p (X[j]-\bar{X})(Y[j]-\bar{Y})}{\sqrt{\sum_{j=1}^p (X[j]-\bar{X})^2 \sum_{j=1}^p (Y[j]-\bar{Y})^2}}$ , where  $\bar{X} = \frac{\sum_{j=1}^p X[j]}{p}$
  - $r^2$
  - Manhattan distance
  - Mutual information
  - Kendall's Tau
  - etc.
- Each has different properties and can reveal different features of the data



### Clustering

#### Correlation vs Euclidean Distance

X	1	0	-1	0
Y	3	2	1	2
Z	-1	0	1	0
W	2	0	-2	0



Correlation (X,Y) = 1      Eucl. Distance (X,Y) = 4  
 Correlation (X,Z) = -1      Eucl. Distance (X,Z) = 2.83  
 Correlation (X,W) = 1      Eucl. Distance (X,W) = 1.41



### Clustering

#### Distance Matrix

- Once a distance metric has been selected, the starting point for all clustering methods is a “distance matrix”

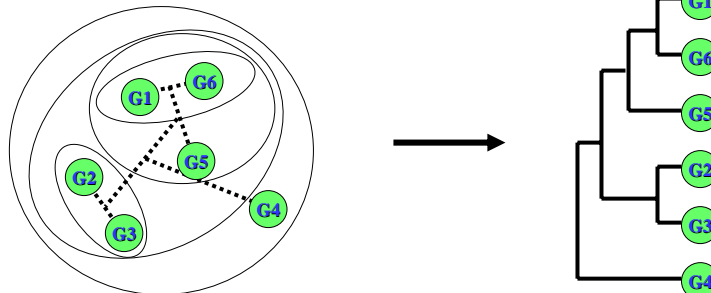
	Gene <sub>1</sub>	Gene <sub>2</sub>	Gene <sub>3</sub>	Gene <sub>4</sub>	Gene <sub>5</sub>	Gene <sub>6</sub>
Gene <sub>1</sub>	0	1.5	1.2	0.25	0.75	1.4
Gene <sub>2</sub>	1.5	0	1.3	0.55	2.0	1.5
Gene <sub>3</sub>	1.2	1.3	0	1.3	0.75	0.3
Gene <sub>4</sub>	0.25	0.55	1.3	0	0.25	0.4
Gene <sub>5</sub>	0.75	2.0	0.75	0.25	0	1.2
Gene <sub>6</sub>	1.4	1.5	0.3	0.4	1.2	0

- The elements of this matrix are the pair-wise distances. Note that the matrix is symmetric about the diagonal.



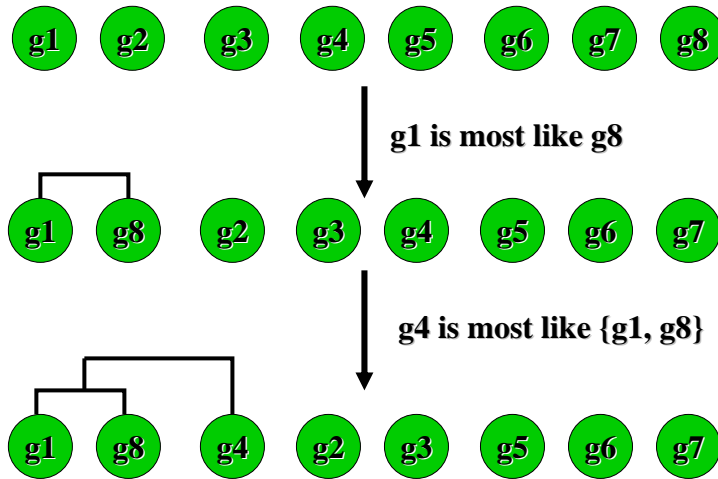
### Hierarchical Clustering

1. Calculate the distance between all genes. Find the smallest distance. If several pairs share the same similarity, use a predetermined rule to decide between alternatives.
2. Fuse the two selected clusters to produce a new cluster that now contains at least two objects. Calculate the distance between the new cluster and all other clusters.
3. Repeat steps 1 and 2 until only a single cluster remains.
4. Draw a tree representing the results.

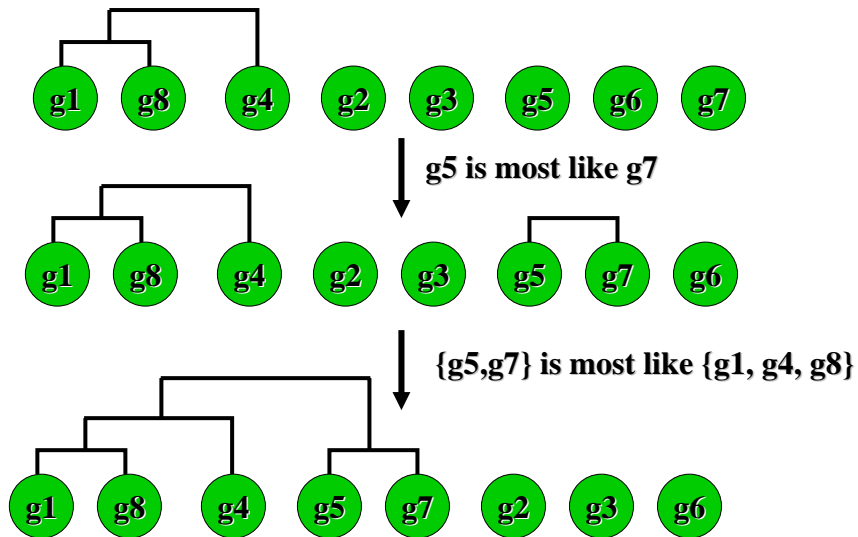




### Hierarchical Clustering

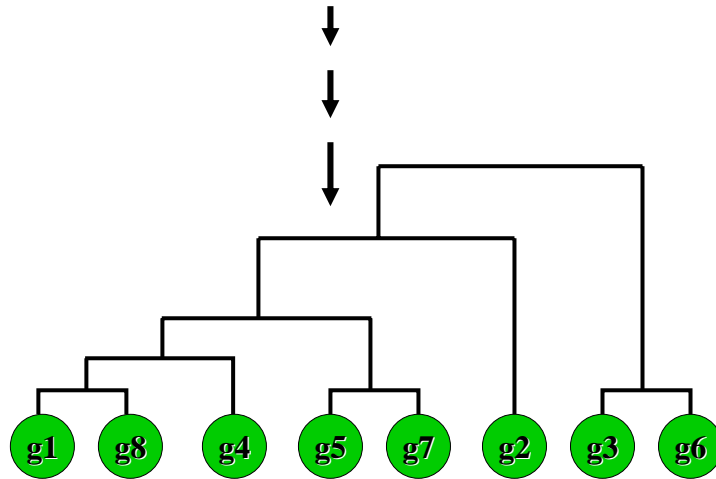


### Hierarchical Clustering





Hierarchical Clustering

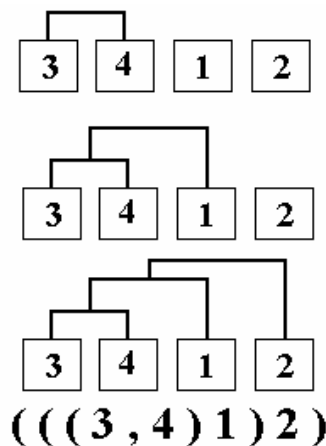


Armidale Animal Breeding Summer Course, UNE, Feb. 2006



Hierarchical Clustering

	exp1	exp2	exp3	exp4
exp1	0			
exp2	6	0		
exp3	3	5	0	
exp4	4	4	2	0
	exp1	exp2	exp(3,4)	
exp1	0			
exp2	6	0		
exp(3,4)	3.5	4.5	0	
	exp2	exp((3,4)1)		
exp2				
exp((3,4)1)	5.75			



Armidale Animal Breeding Summer Course, UNE, Feb. 2006



## Hierarchical Clustering

### Advantages

- Computationally efficient
- Produces tree-like structure

### Disadvantage

- Clusters are not optimal. Once branches split, it's permanent. There is no way to reevaluate whether it was the best division based on whole data set.



## Agglomerative Linkage Methods

Linkage methods are rules or metrics that return a value that can be used to determine which elements (clusters) should be linked.

Three linkage methods that are commonly used are:

- Single Linkage
- Average Linkage
- Complete Linkage



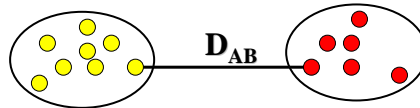
Agglomerative Linkage Methods

## Single Linkage

Cluster-to-cluster distance is defined as the *minimum distance* between members of one cluster and members of the another cluster. Single linkage tends to create 'elongated' clusters with individual genes chained onto clusters.

$$D_{AB} = \min ( d(u_i, v_j) )$$

where  $u$  belongs to  $A$  and  $v$  belongs to  $B$   
for all  $i = 1$  to  $N_A$  and  $j = 1$  to  $N_B$



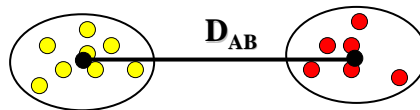
Agglomerative Linkage Methods

## Average Linkage

Cluster-to-cluster distance is defined as the *average distance* between all members of one cluster and all members of another cluster. Average linkage has a slight tendency to produce clusters of similar variance.

$$D_{AB} = 1/(N_A N_B) \sum \sum ( d(u_i, v_j) )$$

where  $u$  belongs to  $A$  and  $v$  belongs to  $B$   
for all  $i = 1$  to  $N_A$  and  $j = 1$  to  $N_B$





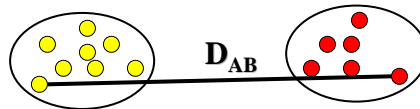
Agglomerative Linkage Methods

# Complete Linkage

Cluster-to-cluster distance is defined as the *maximum distance* between members of one cluster and members of the another cluster. Complete linkage tends to create clusters of similar size and variability.

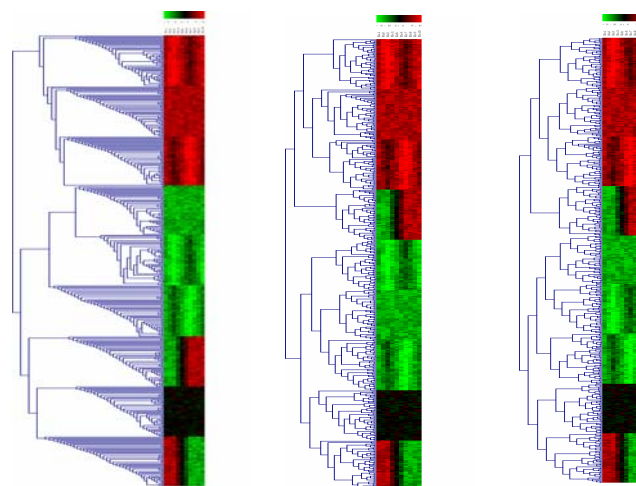
$$D_{AB} = \max ( d(u_i, v_j) )$$

where  $u$  belongs to  $A$  and  $v$  belongs to  $B$   
for all  $i = 1$  to  $N_A$  and  $j = 1$  to  $N_B$



Agglomerative Linkage Methods

# Comparison of Linkage Methods



Single

Average

Complete



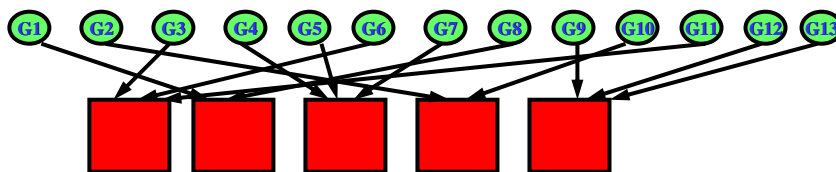
### K-Means/Medians Clustering

#### Process

1. Specify number of clusters, e.g., 5.



2. Randomly assign genes to clusters.



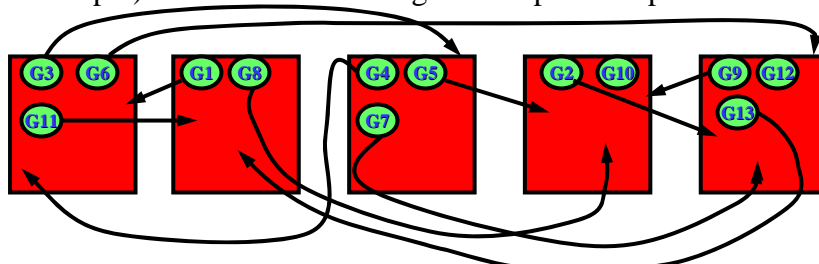
Armidale Animal Breeding Summer Course, UNE, Feb. 2006



### K-Means/Medians Clustering

3. Calculate mean/median expression profile of each cluster.

4. Shuffle genes among clusters such that each gene is now in the cluster whose mean expression profile (calculated in step 3) is the closest to that gene's expression profile.



5. Repeat steps 3 and 4 until genes cannot be shuffled around any more, OR a user-specified number of iterations has been reached.

K-Means is most useful when the user has an *a priori* hypothesis about the number of clusters the genes should group into.

Armidale Animal Breeding Summer Course, UNE, Feb. 2006



### K-Means/Medians Clustering

Example

	<u>Control</u>				<u>Drug</u>			
Expression	9	12	14	17	18	21	23	26
Average	13				22			



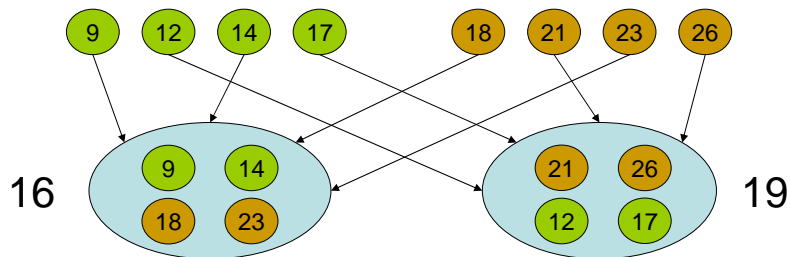
1. Specify the existence of two clusters and assign them at random



### K-Means/Medians Clustering

Example

	<u>Control</u>				<u>Drug</u>			
Expression	9	12	14	17	18	21	23	26
Average	13				22			



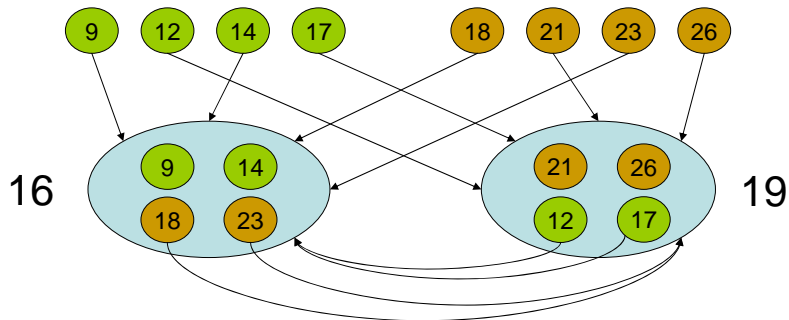
2. Calculate Mean (Median) of each cluster and
3. Shuffle genes around so that each goes to the 'closest' cluster



### K-Means/Medians Clustering

Example

	<u>Control</u>				<u>Drug</u>			
Expression	9	12	14	17	18	21	23	26
Average	13				22			



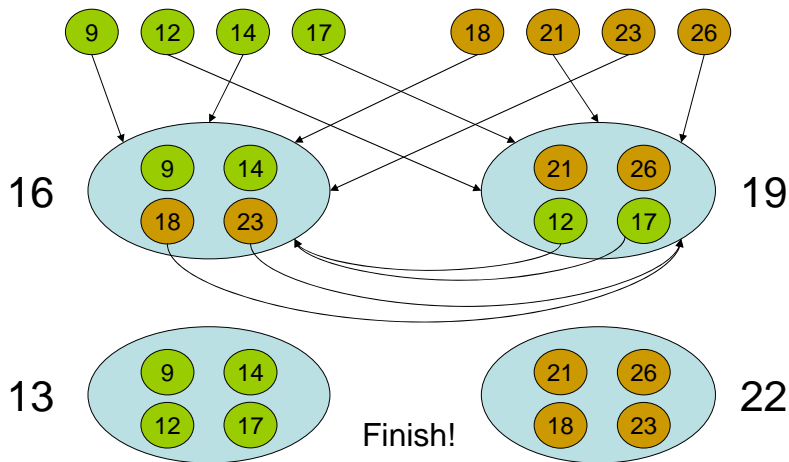
4. Re-compute cluster Means (Medians) and
5. Re-shuffle genes til convergence is reached



### K-Means/Medians Clustering

Example

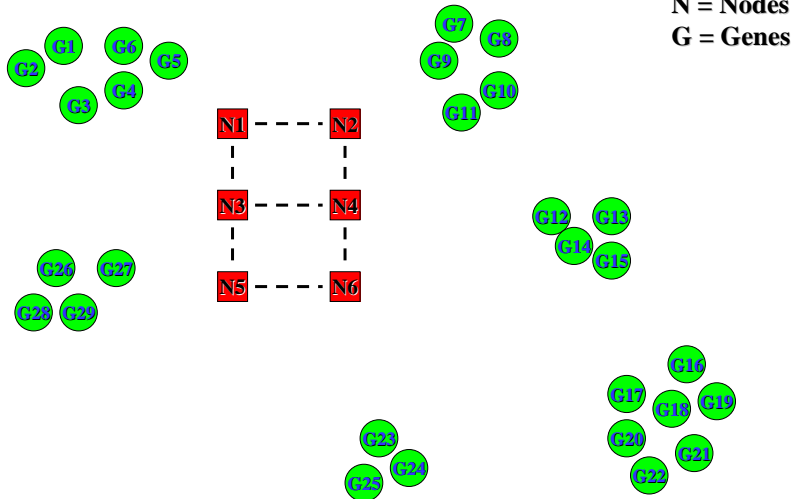
	<u>Control</u>				<u>Drug</u>			
Expression	9	12	14	17	18	21	23	26
Average	13				22			





### Self-Organizing Maps (SOM)

1. Specify the number of nodes (clusters) desired, and also specify a 2-D geometry for the nodes, e.g., rectangular or hexagonal

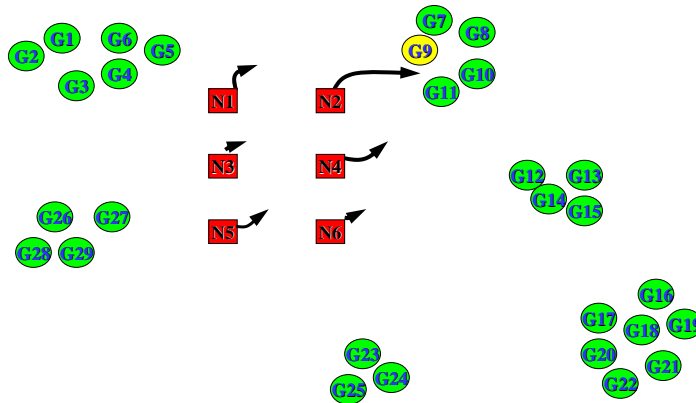


Armidale Animal Breeding Summer Course, UNE, Feb. 2006



### Self-Organizing Maps (SOM)

2. Choose a random gene, e.g., G9
3. Move the nodes in the direction of G9. The node closest to G9 (N2) is moved the most, and the other nodes are moved by smaller varying amounts. The farther away the node is from N2, the less it is moved.

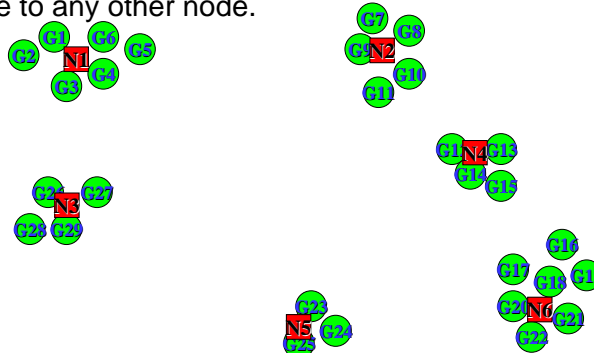


Armidale Animal Breeding Summer Course, UNE, Feb. 2006



### Self-Organizing Maps (SOM)

- Steps 2 and 3 (i.e., choosing a random gene and moving the nodes towards it) are repeated many (usually several thousand) times. However, with each iteration, the amount that the nodes are allowed to move is decreased.
- Finally, each node will “nestle” among a cluster of genes, and a gene will be considered to be in the cluster if its distance to the node in that cluster is less than its distance to any other node.

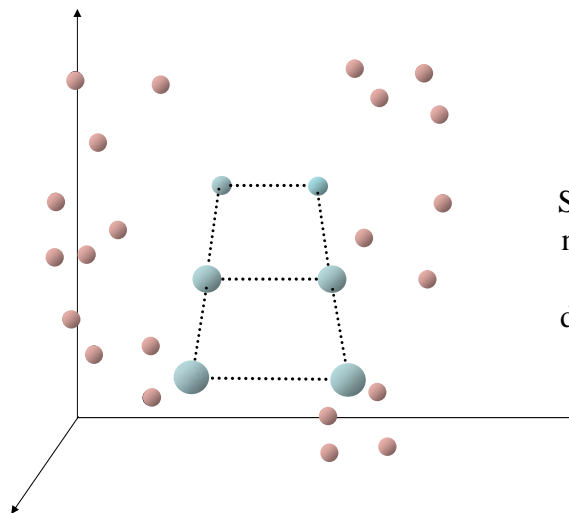


Armidale Animal Breeding Summer Course, UNE, Feb. 2006



### Self-Organizing Maps (SOM)

Perhaps a better view...



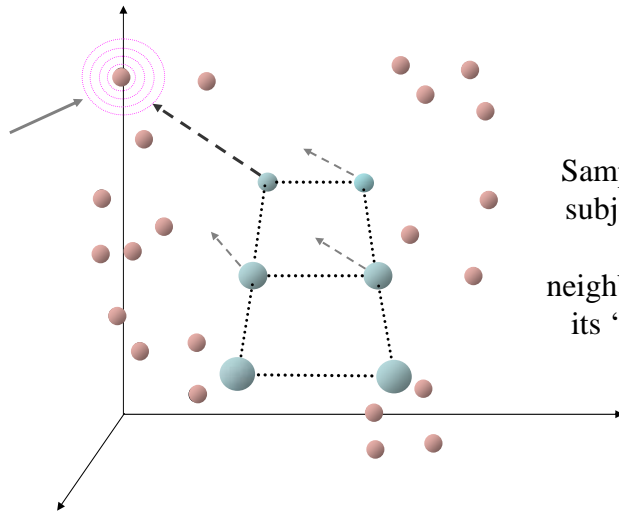
Situate grid of nodes along a plane where datapoints are distributed

Armidale Animal Breeding Summer Course, UNE, Feb. 2006



### Self-Organizing Maps (SOM)

Perhaps a better view...

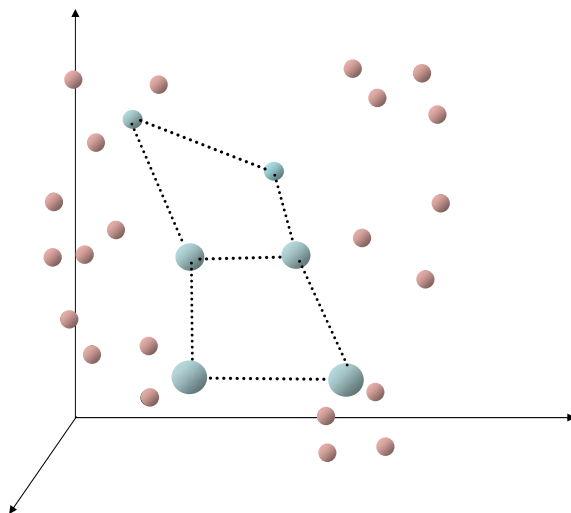


Sample a gene and subject the closest node and neighboring nodes to its 'gravitational' influence



### Self-Organizing Maps (SOM)

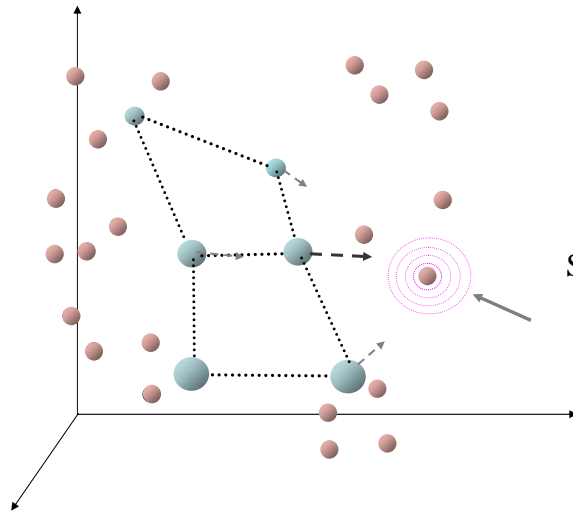
Perhaps a better view...





### Self-Organizing Maps (SOM)

Perhaps a better view...

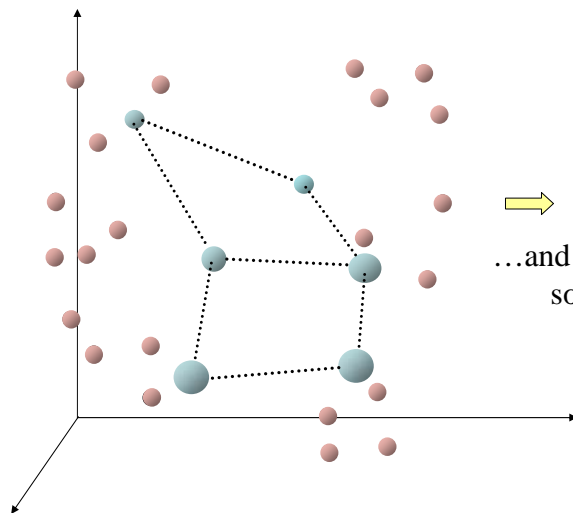


Sample another gene...



### Self-Organizing Maps (SOM)

Perhaps a better view...

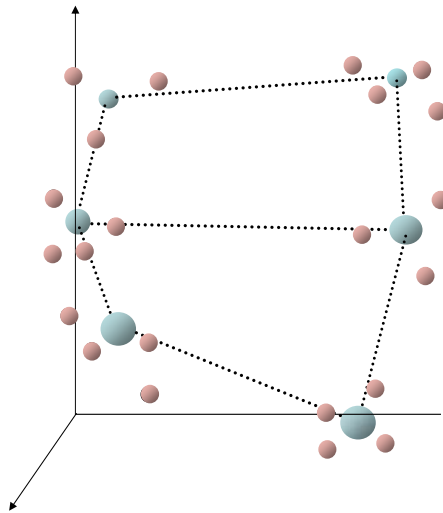


...and so on, and so on...



### Self-Organizing Maps (SOM)

Perhaps a better view...



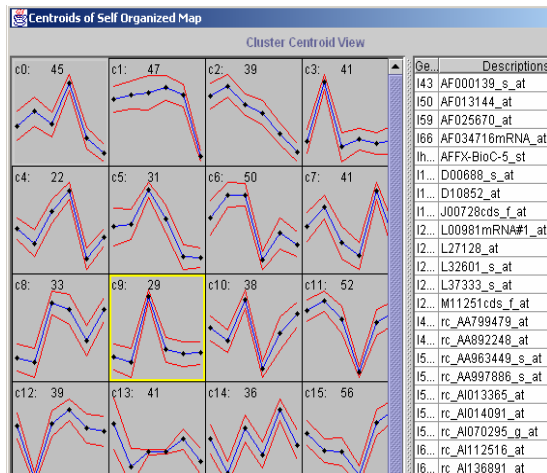
...until all genes have been sampled several times over. Each cluster is defined with reference to a node, specifically comprised by those genes for which it represents the closest node.



### Self-Organizing Maps (SOM)

Results

- X-axis is time after dose
- Y-axis is normalized gene expression level
- Group ~1000 genes into 24 categories

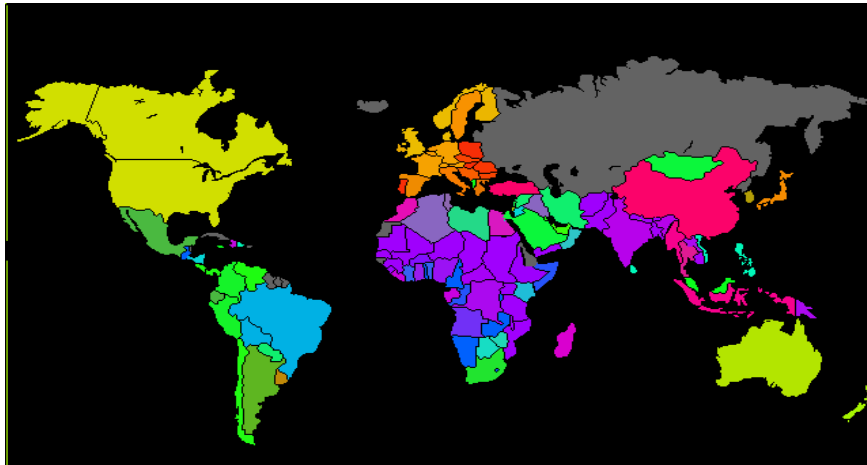




### Self-Organizing Maps (SOM)

Example <http://www.cis.hut.fi/research/som-research/worldmap.html>

Using 39 indicators of poverty and well-being from the World Bank, a map of the world where countries have been colored with the color describing their poverty type:

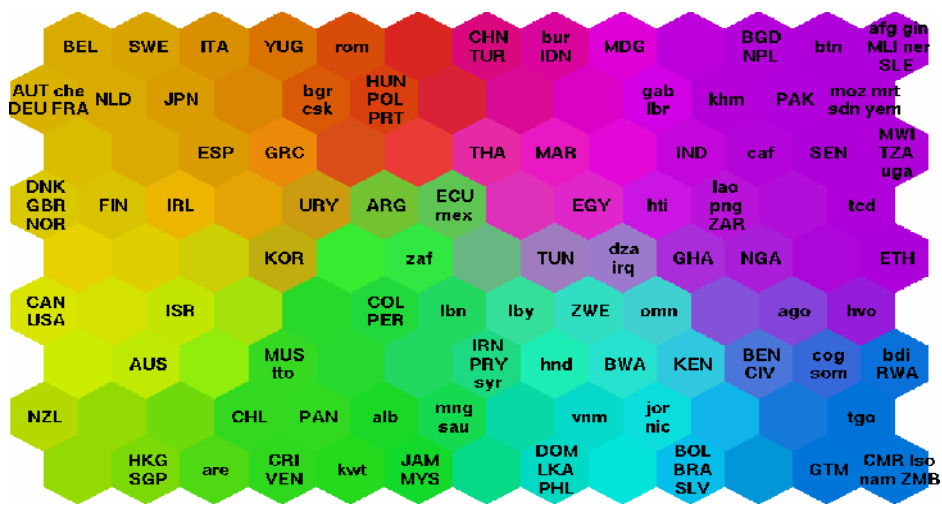


Armidale Animal Breeding Summer Course, UNE, Feb. 2006



### Self-Organizing Maps (SOM)

Example <http://www.cis.hut.fi/research/som-research/worldmap.html>



Armidale Animal Breeding Summer Course, UNE, Feb. 2006



## Self-Organizing Maps (SOM)

### Details to Consider

- Several methods exist for choosing initial data points for clusters.
- How to choose the initial number of clusters.
- Method of recalculating cluster center after adding a new data point can be varied. How much weight is given to new data point?
- Routines for merging and dividing clusters and detecting outliers can be added at each iteration.

### Advantages

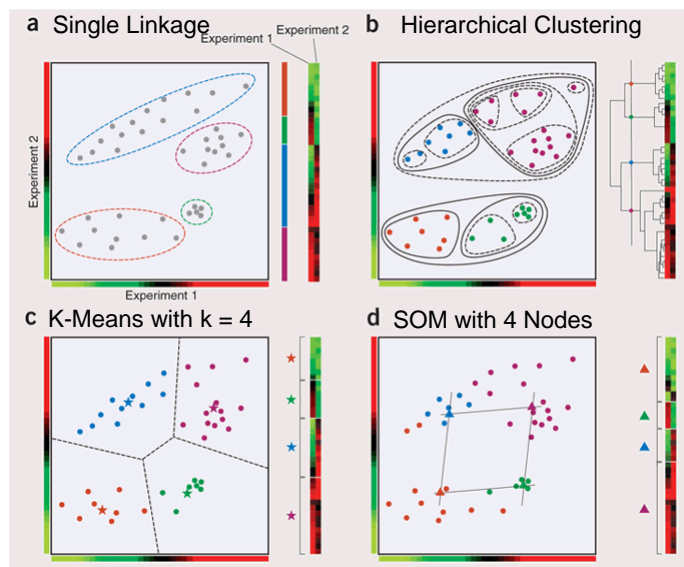
- Able to come closer to 'optimal' clustering through iterations.
- Doesn't force a tree-structure on data

### Disadvantage


- Larger number of options for clustering means that details of process may be hidden.



## Comparing Clustering Methods

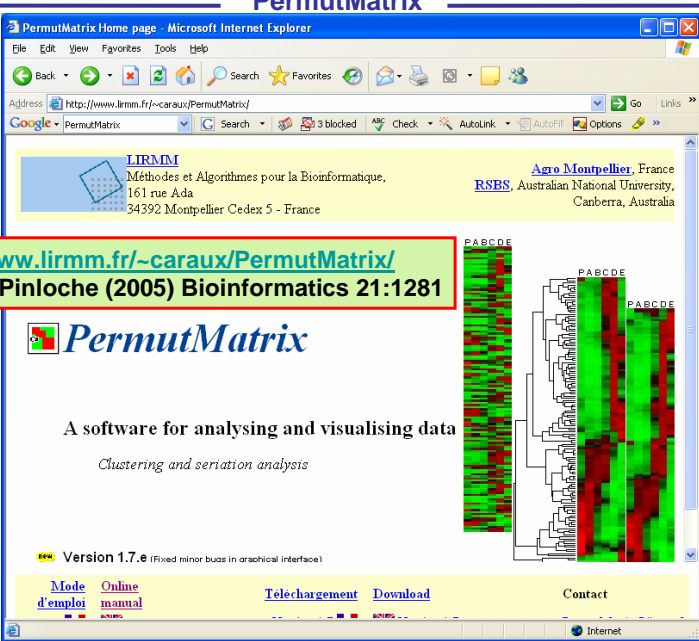


A Quantitative Overview to Gene Expression Profiling in Animal Genetics




## PermutMatrix

<http://www.lirmm.fr/~caraux/PermutMatrix/>  
 Caraux and Pinloche (2005) Bioinformatics 21:1281

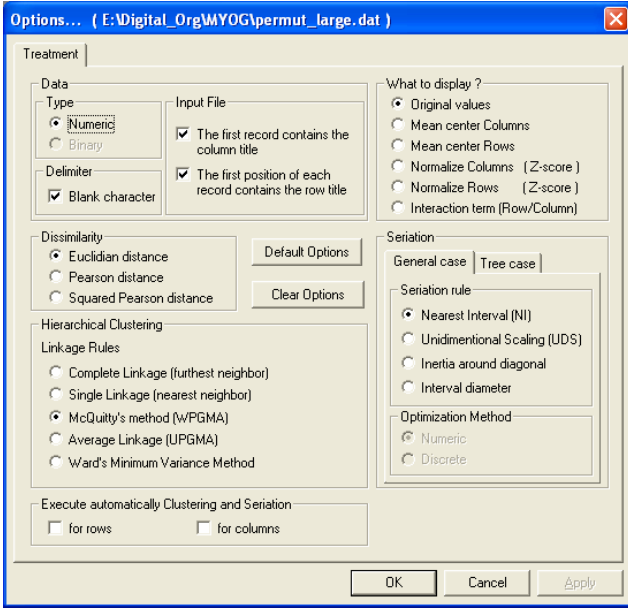


Armidaile Animal Breeding Summer Course, UNE, Feb. 2006

A Quantitative Overview to Gene Expression Profiling in Animal Genetics



## PermutMatrix



Armidaile Animal Breeding Summer Course, UNE, Feb. 2006



PermutMatrix

Example

Pregnancy vs Lactation vs Involution

