Please email if you want to chat
fraenkel@mit.edu

Ernest Fraenkel

Structure of the Unit

- Wet lab and computational lab focus on measuring and understanding response to a drug (etoposide) in cell culture
- In the computational labs, you will compare etoposide changes in 20.109 data and a published dataset.
- We will use concepts you have seen in 6.0002 to analyze these data.
- Computational assignments will give you the building blocks for your written report.
- The lab will be conducted in the programming environment called “R”.

Why R?

Lecture Schedule

Date	Topic
March 10th	Clustering and PCA
March 12th	Analyzing RNA-Seq
March 17th	Big Data for BE
March 19th	Transcriptional Regulation
SPRING BREAK	
March 31st	Molecular Networks
April 2nd	Single-cell Analysis

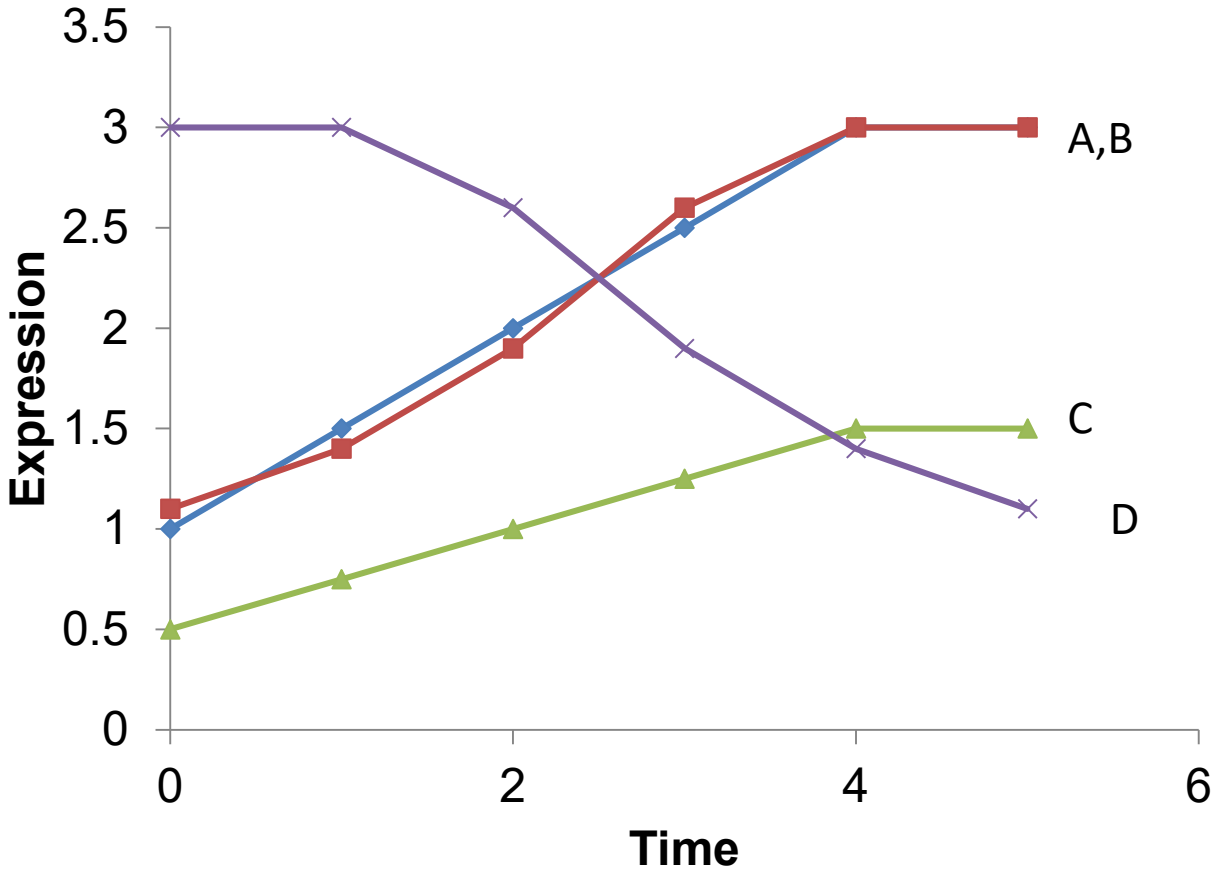
Write on Board Before Class:

Learning Objectives

- See the big picture of this unit
- Choose the right distance metric to compare the expression of two genes
- Describe why you would cluster expression by genes or experiments
- Manually cluster small vectors using hierarchical or k-means clustering
- Read a dendrogram
- Describe the results of Principal Component Analysis (PCA)

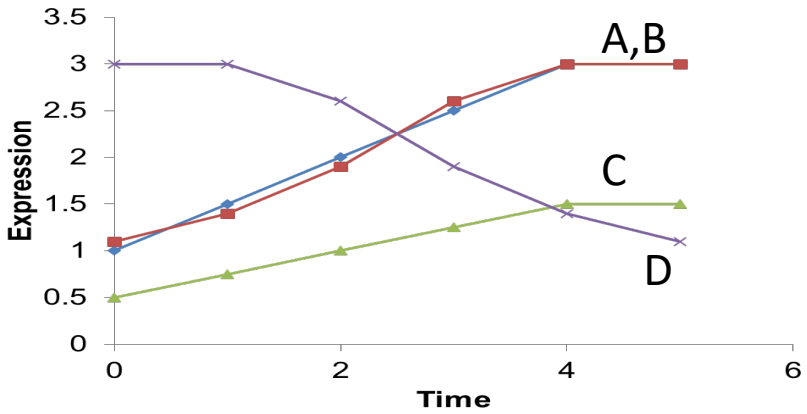
Comparing
the Expression of Genes

Draw on LEFT Board and keep



Comparing gene expression

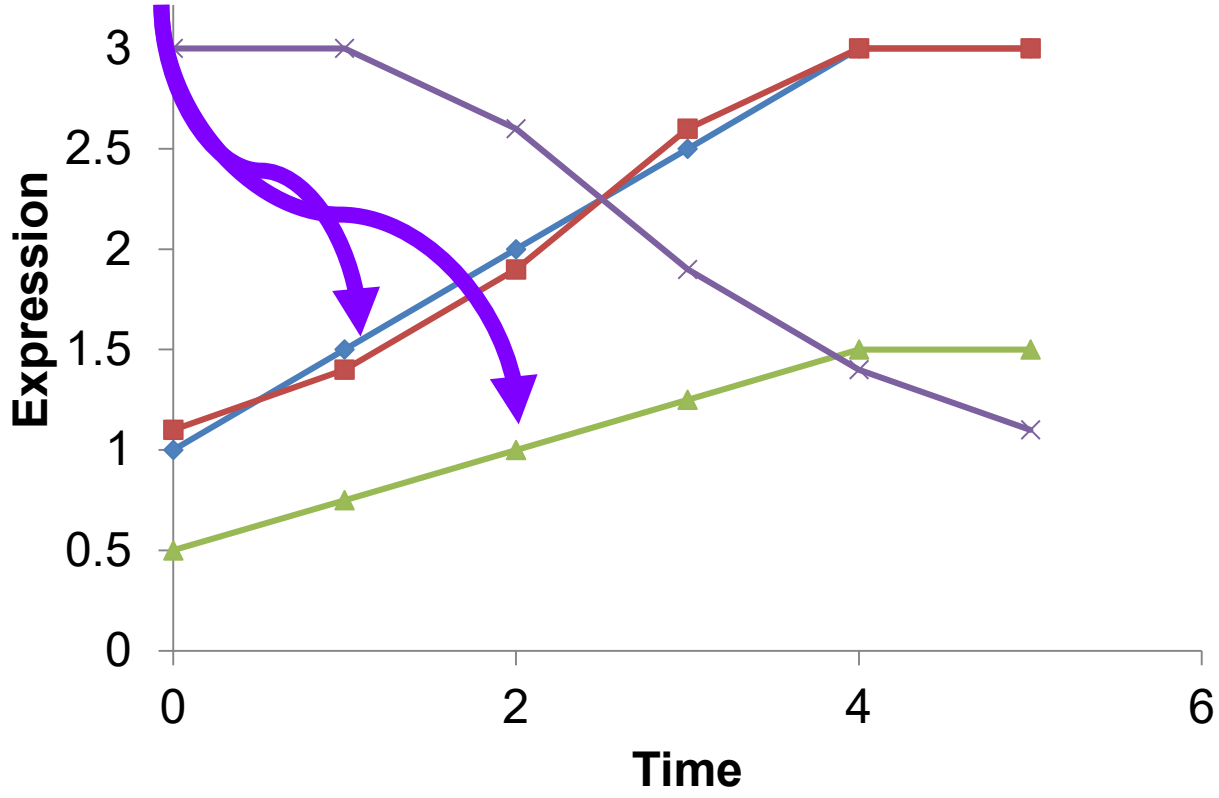
- Draw gene expression patterns on board
- Which of the genes on this plot are most similar?
- How do we quantify similarity of expression?
- Let's consider the simplest description first.
  - A and B are most similar.
  - Euclidean distance would describe this type of similarity.



Distance Metrics

Which other pairs of genes might be co-regulated?

Can we capture the similarity of these patterns?



Euclidean distance provides an intuitive description:

In our timecourse:

$X_A=(x_{A1}, x_{A2},\ldots x_{AN})$

$X_b=(x_{B1}, x_{B2},\ldots x_{BN})$

$$d(X_A, X_B) = \sqrt{\sum_{i=1}^{N}(x_{Ai} - x_{Bi})^2}$$

Pearson Correlation

- To understand Pearson Correlation, we need to define a Z-score
- Z_{Ai}= z-score of gene *A* in experiment i:

Z-score

$$Z_{Ai} = \frac{X_{Ai} - \bar{X}_A}{\sigma}$$

Standard deviation

$$\sigma = \sqrt{\frac{\sum(X_{Ai} - \bar{X}_A)^2}{N}}$$

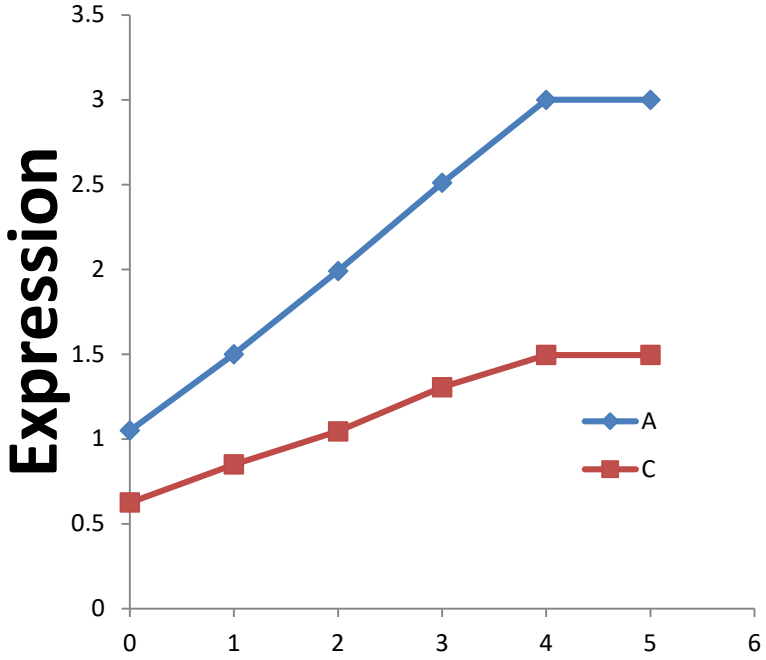
over all experiments

$$r_{A,B} = \frac{\sum_{i=1}^{N_{expt}} Z_{Ai} Z_{Bi}}{N}$$

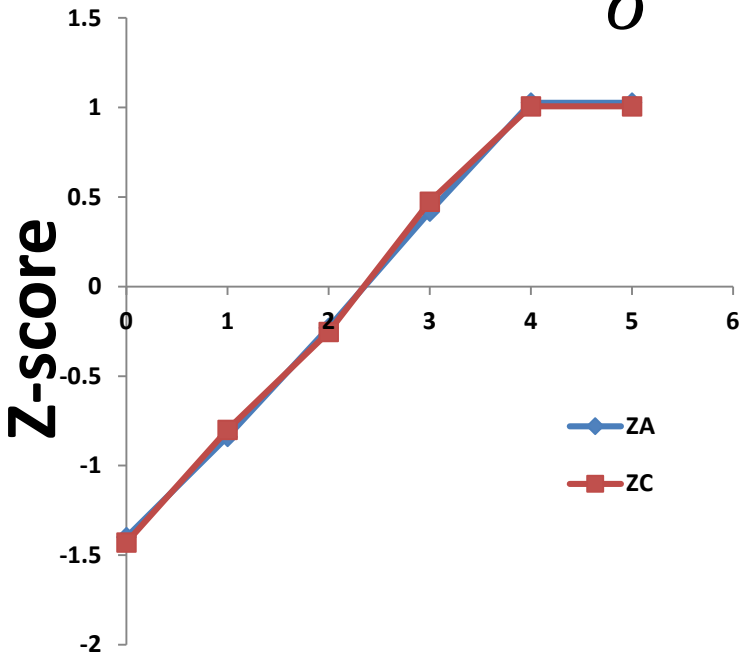
Pearson correlation

from +1 (perfect correlation) to -1 (anti-correlated)

Distance = $1-r_{A,B}$

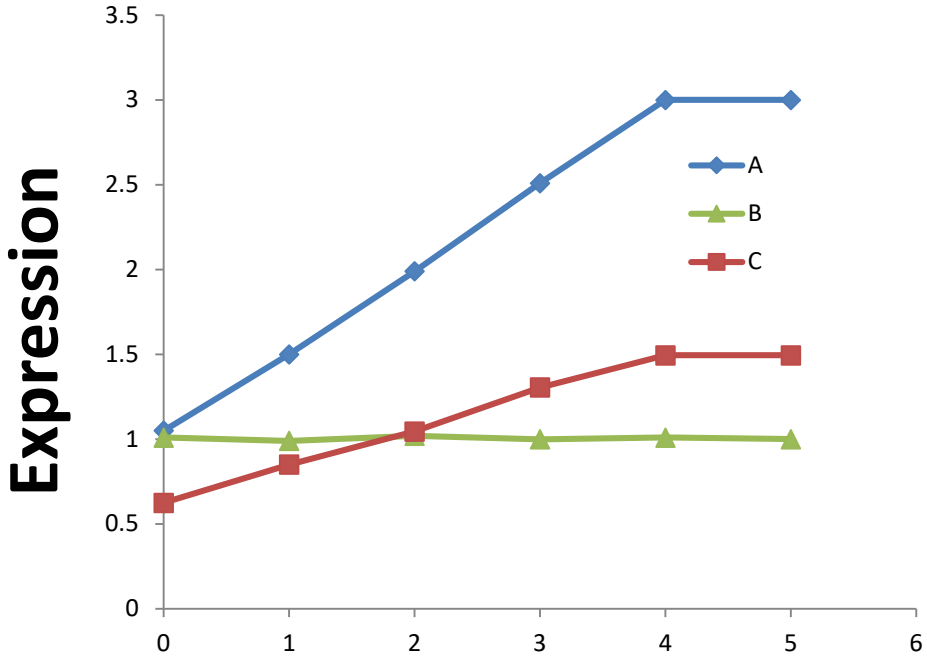


$$Z_{Ki} = \frac{X_{Ki} - \bar{X}_K}{\sigma}$$

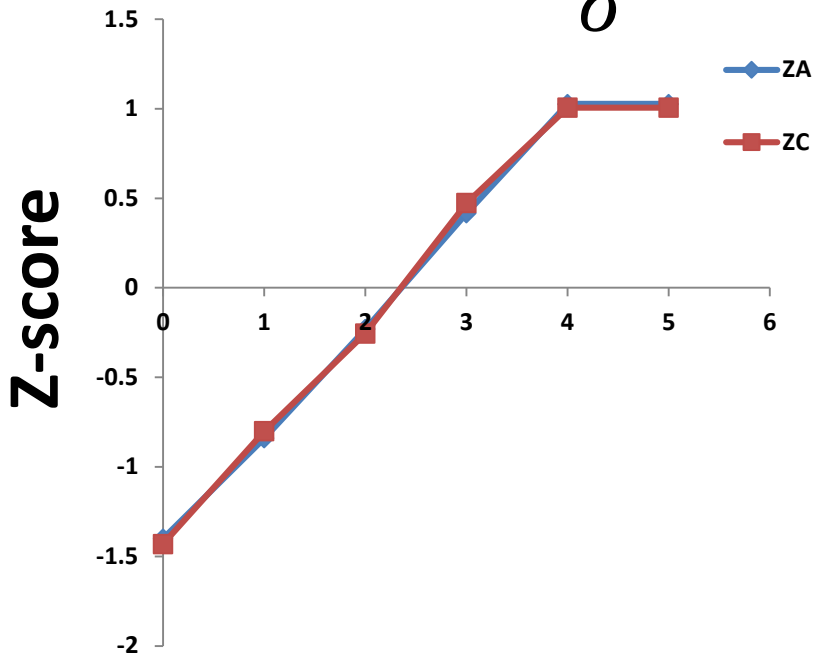


$r_{A,C}$= 0.999

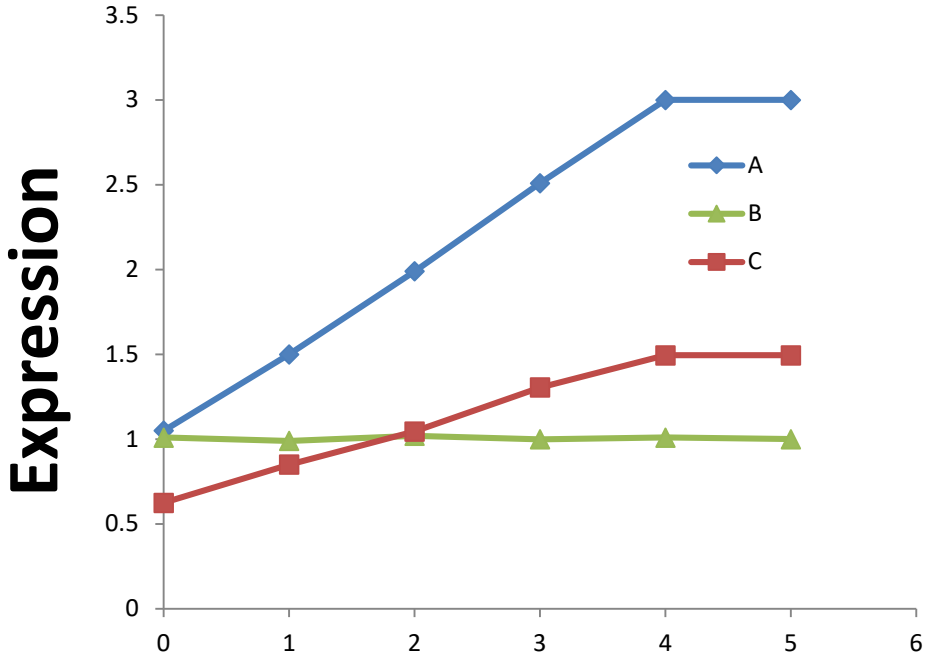
$$r_{A,B} = \frac{\sum_{i=1}^{N_{expt}} Z_{Ai} Z_{Bi}}{N}$$



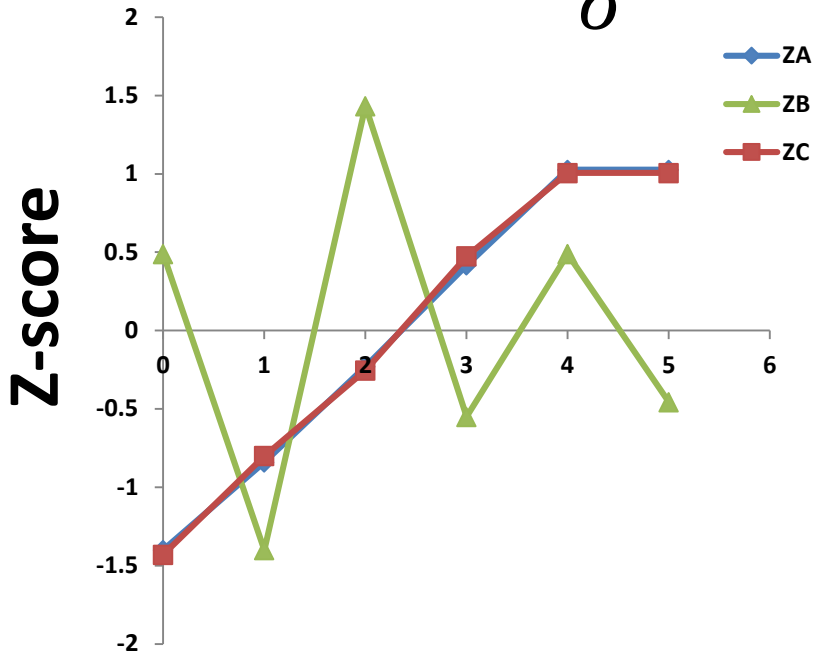
$$Z_{Ki} = \frac{X_{Ki} - \bar{X}_K}{\sigma}$$



$$r_{A,B} = \frac{\sum_{i=1}^{N_{expt}} Z_{Ai} Z_{Bi}}{N}$$



$$Z_{Ki} = \frac{X_{Ki} - \bar{X}_K}{\sigma}$$



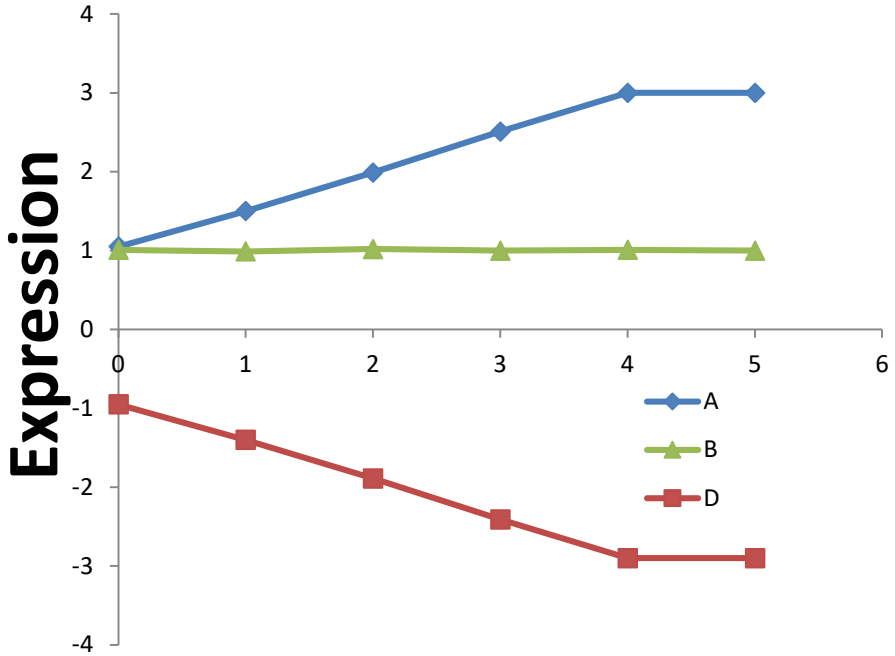
$r_{A,B}$= -0.01

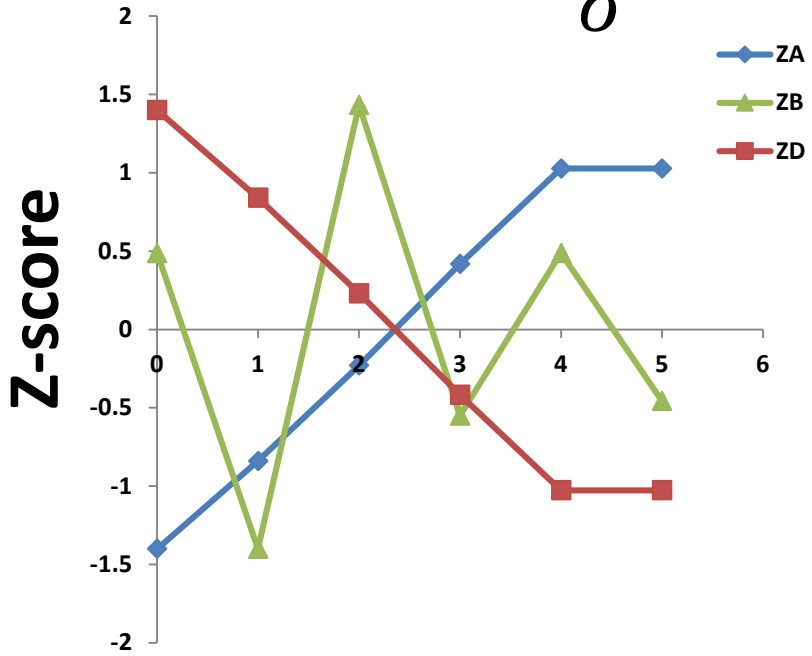
$r_{A,C}$= 0.999

$r_{B,C}$= -0.03

$$r_{A,B} = \frac{\sum_{i=1}^{N_{expt}} Z_{Ai} Z_{Bi}}{N}$$

$$Z_{Ki} = \frac{X_{Ki} - \bar{X}_K}{\sigma}$$





$r_{A,B}$= -0.01

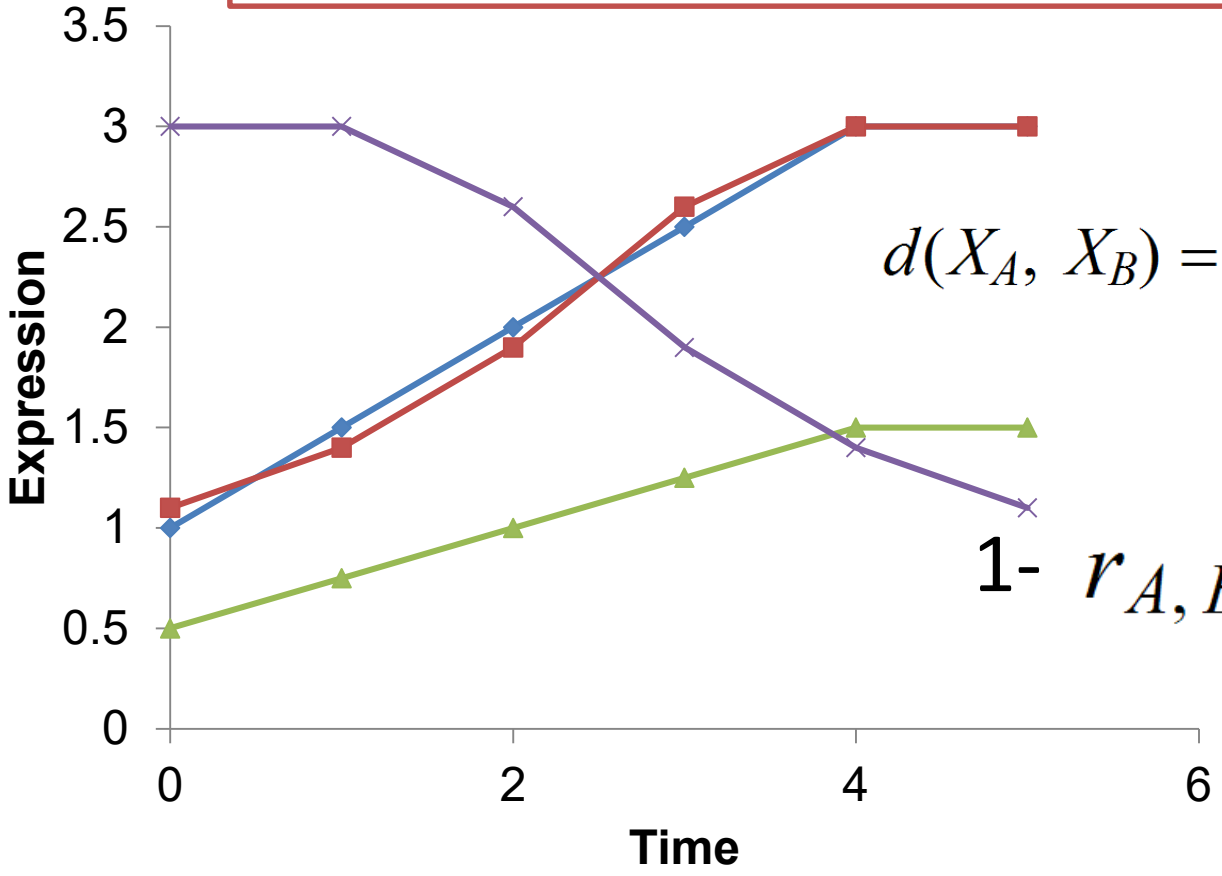
$r_{A,D}$= -1.0

$r_{B,D}$= 0.007

$$r_{A,B} = \frac{\sum_{k=1}^{N_{expt}} Z_{kA} Z_{kB}}{N}$$

Distance Metrics

Which would you use to find co-regulated genes?



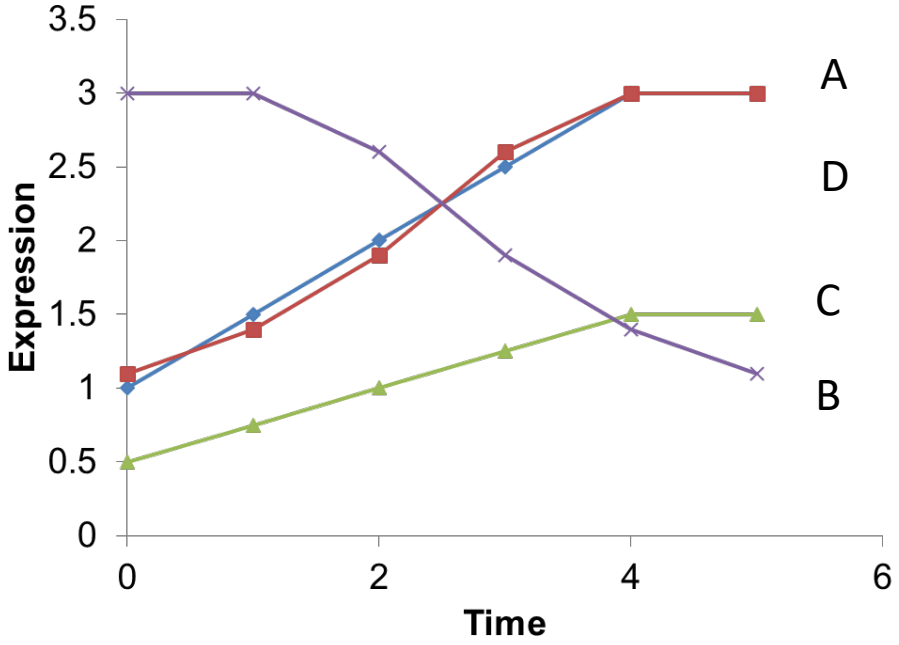
$$d(X_A,\ X_B) = \sqrt{\sum_{k=1}^{N} (X_{A,k} - X_{B,k})^2}$$

$$1\text{-}\ r_{A,B} = 1\text{-}\frac{\sum Z_A Z_B}{N}$$

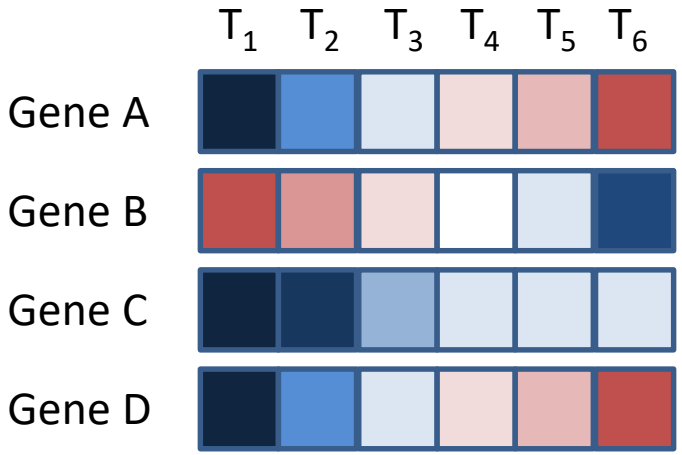
Write on Board: Learning Objectives

- Choose the right distance metric to compare the expression of two genes
- Describe why you would cluster expression by genes or experiments
- Manually cluster small vectors using hierarchical or k-means clustering
- Read a dendrogram
- Describe the results of Principal Component Analysis (PCA)

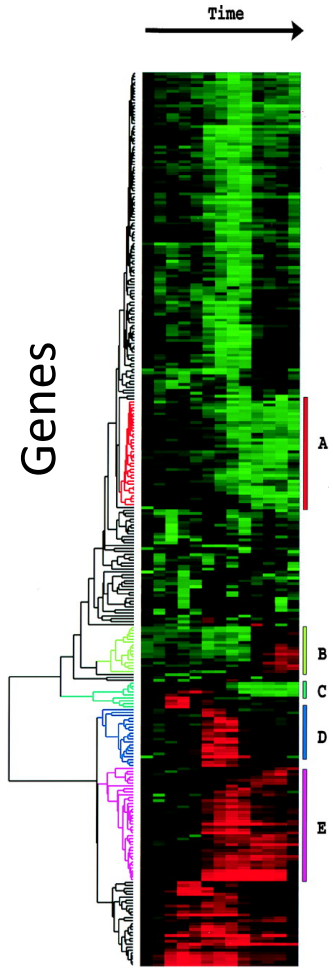
Many ways to plot expression



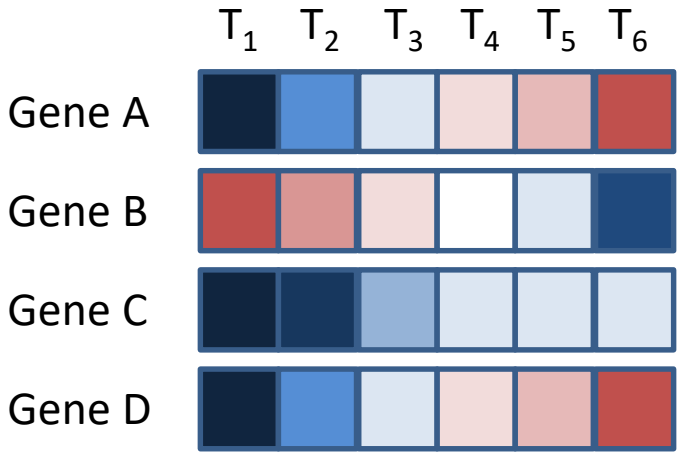
Heatmap



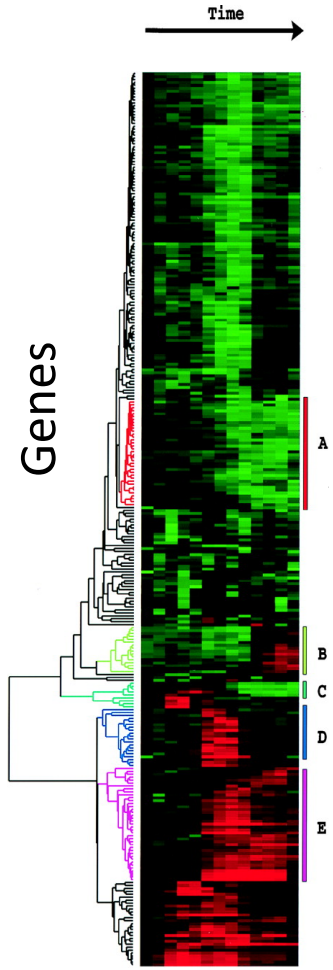
Many ways to plot expression

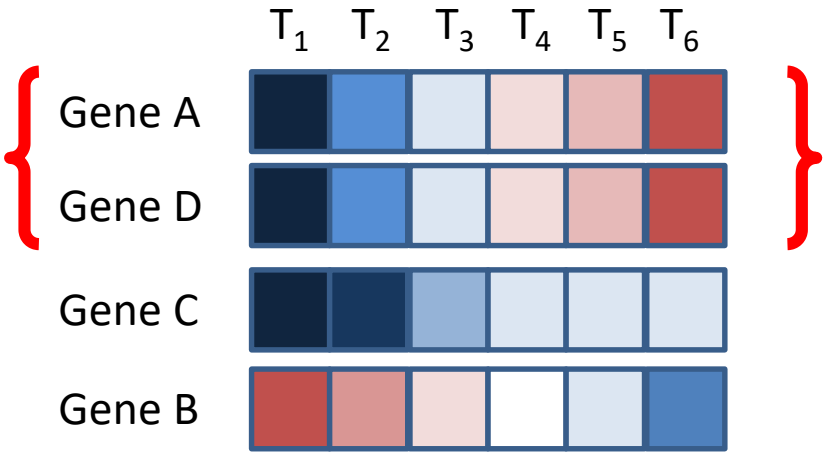


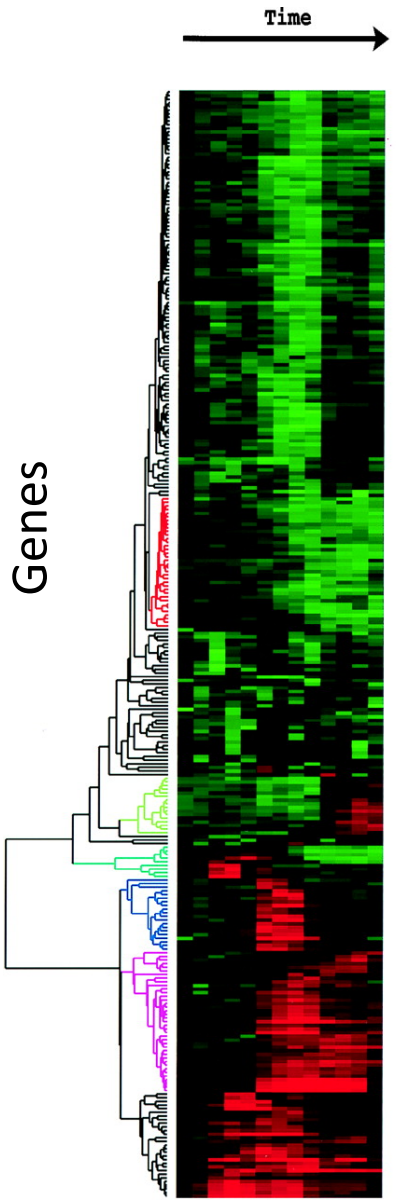
Heatmap



Clustering







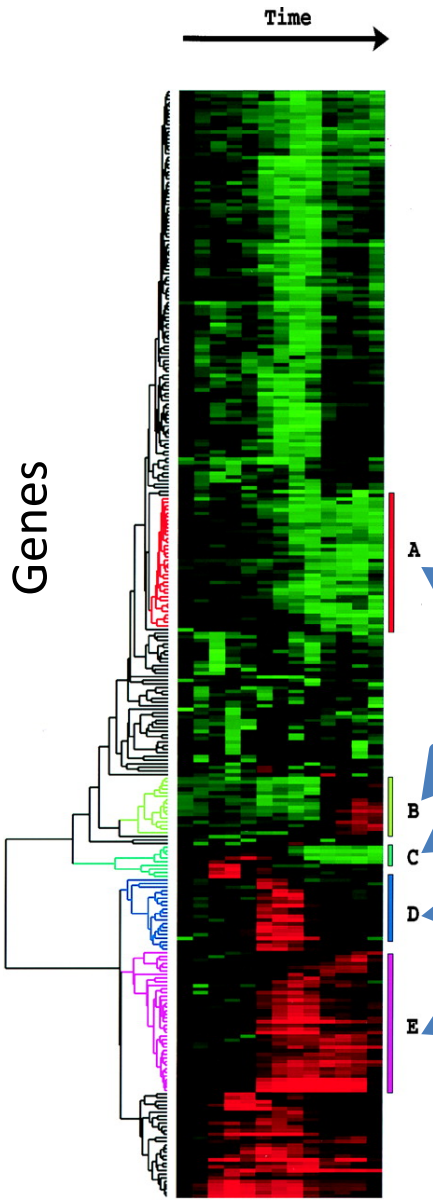
Clustering 8600 human genes based on time course of expression following serum stimulation of fibroblasts

Key: Black = little change Green = down Red = up

(relative to initial time point)

What can you learn from the clustering genes?

Iyer et al. *Science* 1999



Clustering 8600 human genes based on time course of expression following serum stimulation of fibroblasts

Key: Black = little change Green = down Red = up

(relative to initial time point)

Why might you cluster experiments?

(A) cholesterol biosynthesis

(B) the cell cycle

(C) the immediate-early response

(D) signaling and angiogenesis

(E) wound healing and tissue remodeling

Iyer et al. *Science* 1999

Why cluster?

- Cluster genes (rows)
  - Measure expression at multiple time-points, different conditions, etc.

Similar expression patterns may suggest similar functions of genes

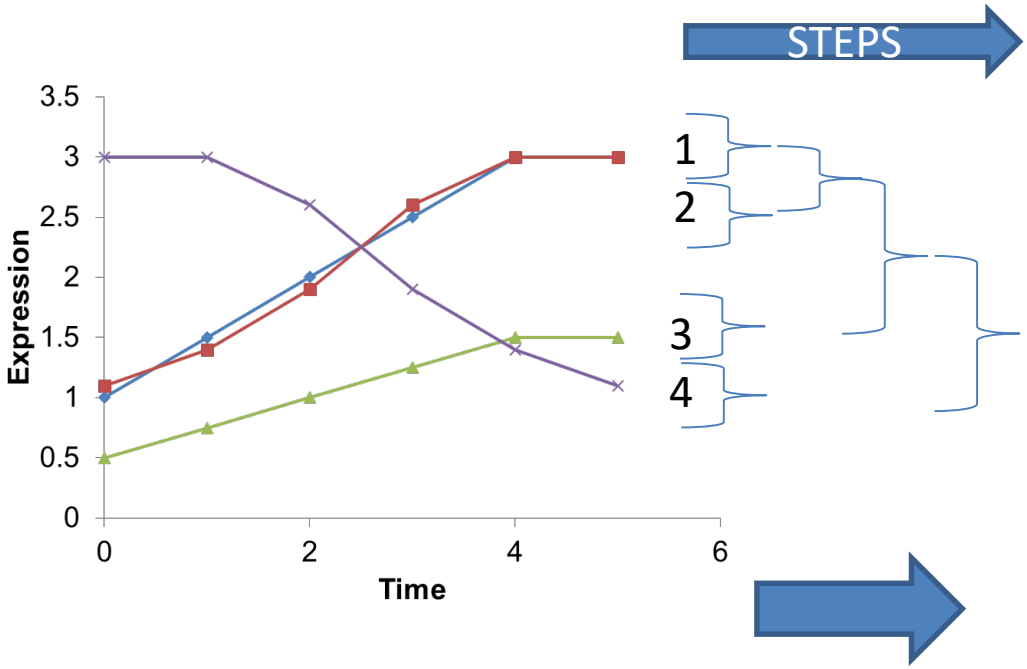
- Cluster samples (columns)
  - e.g., expression levels of thousands of genes for each tumor sample

Similar expression patterns may suggest biological relationship among samples

Two types of approaches: Agglomerative & Divisive

Agglomerative:

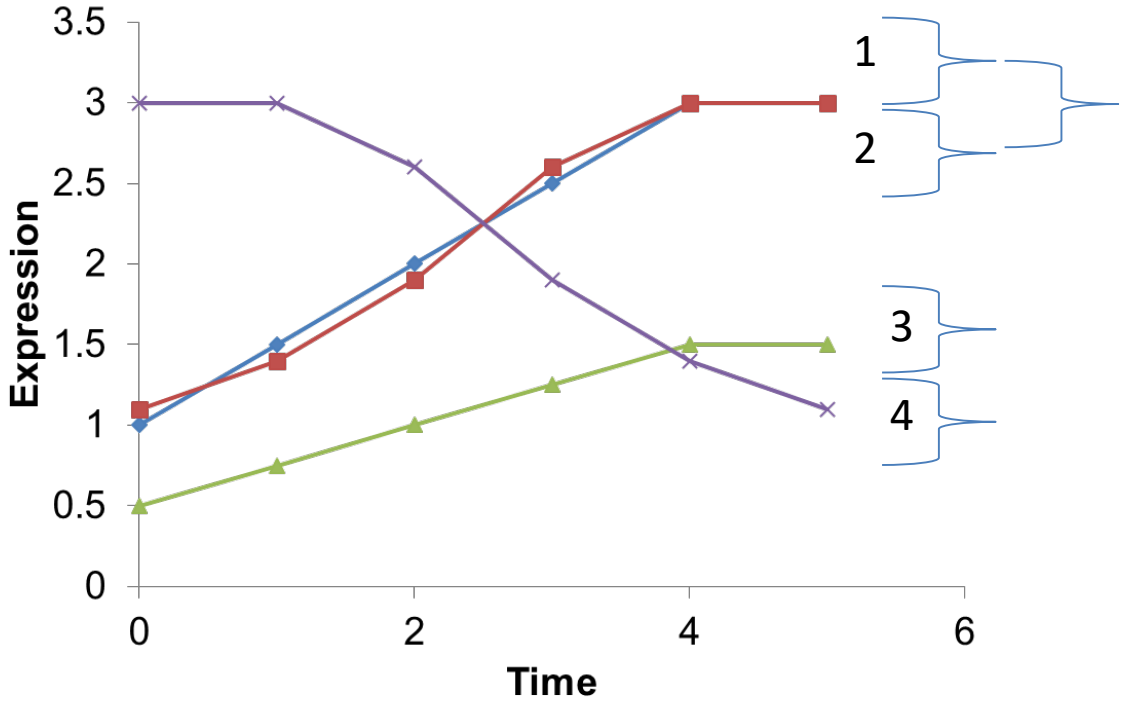
- Initialize: Each vector is in its own cluster
- Repeat until there is only one cluster:
  - Merge the two most similar clusters.



Step 1: each gene is its own cluster
Step 2: combine the two most similar genes
Step 3: find the two most similar clusters

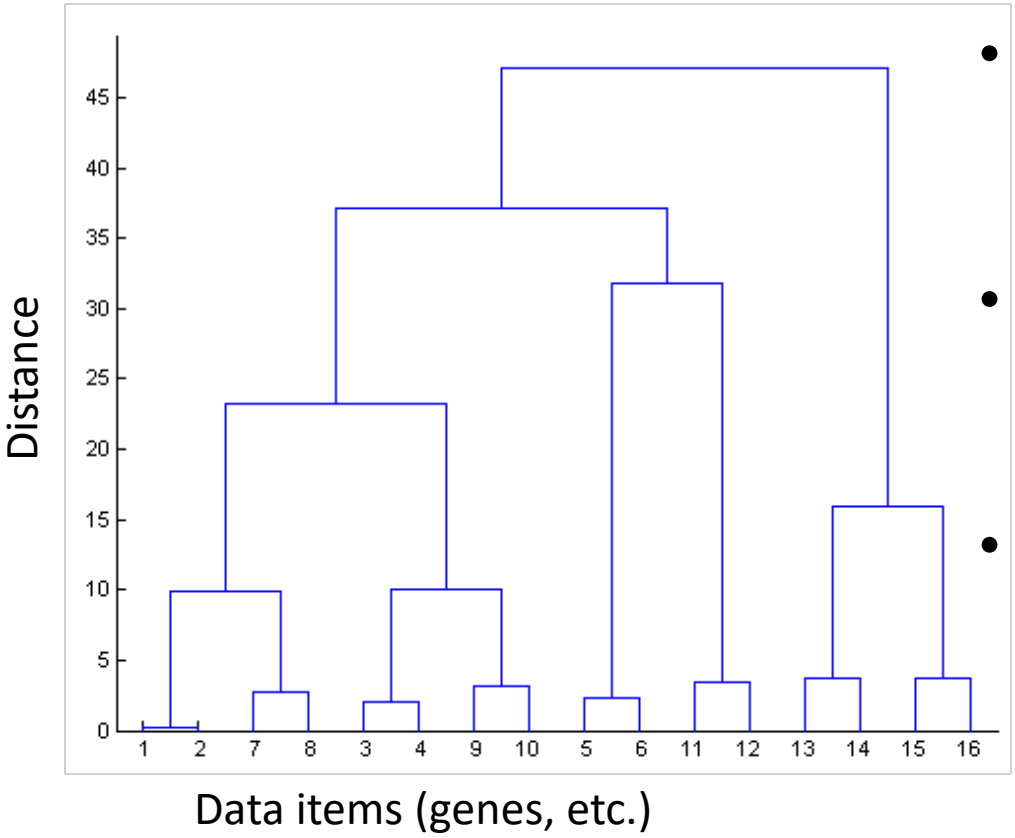
Several options:

minimum distance between members of cluster A,B
maximum distance between members of cluster A,B
average distance between members of cluster A,B



… but I have not told you how to compute distance between the two genes in the new cluster with individual genes

Dendrograms



- The final cluster is the root and each data item is a leaf
- The heights of the bars indicate how close the items are
- Can ‘slice’ the tree at any distance cutoff to produce discrete clusters
- The results will always be hierarchical, even if the data are not.
- The order of the leaf nodes is not meaningful