

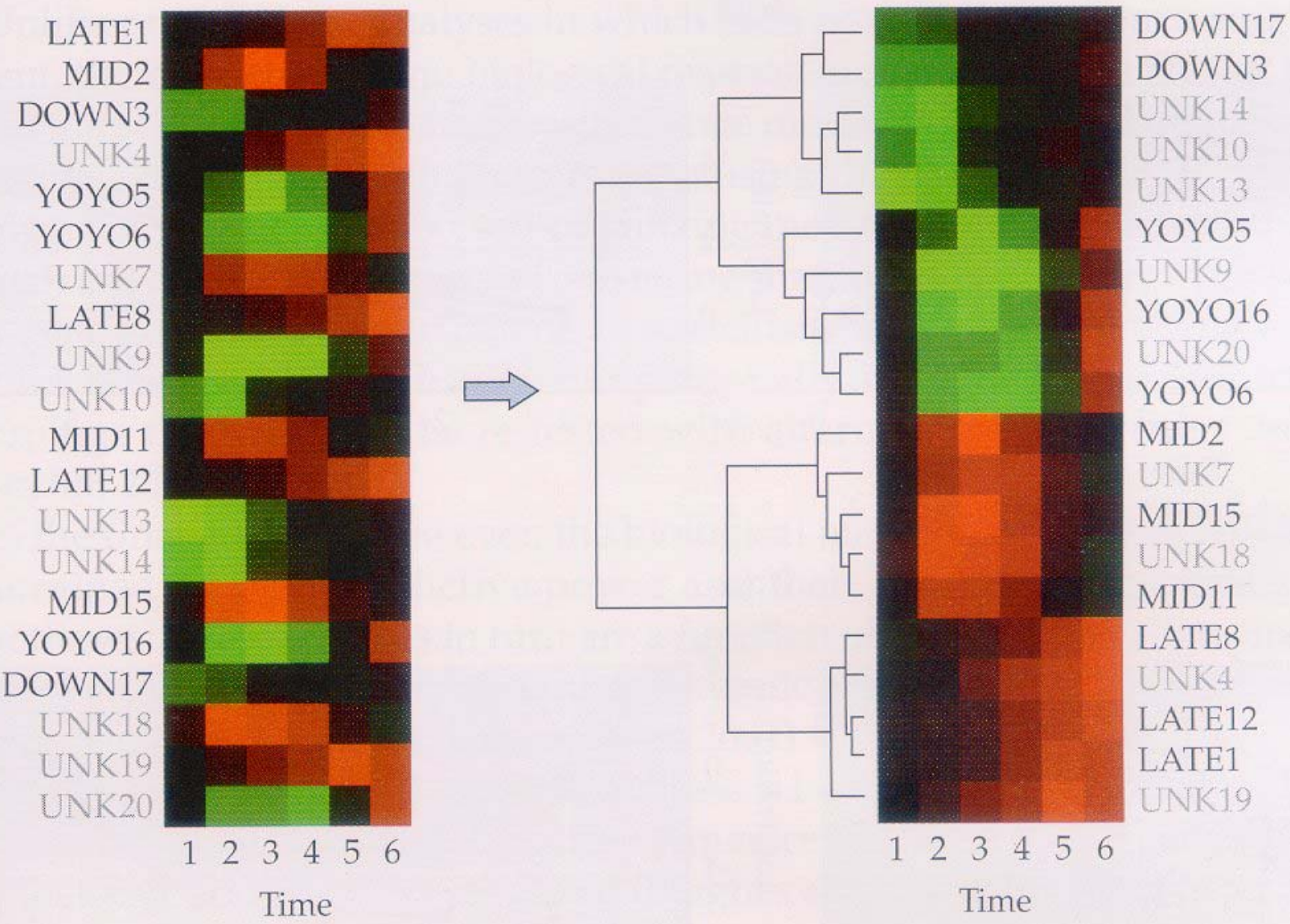
# Cluster Analysis of Gene Expression Microarray Data

BIOL 495S/ CS 490B/ MATH 490B/  
STAT 490B Introduction to Bioinformatics

April 8, 2002

# Data representations

- Data are relative measurements  $\log_2(\text{red} / \text{green})$
- Data are usually collections of gene expressions in many experiment conditions.
- A tabular representation of data, for example rows are genes and columns are different experiment conditions.
- Graphical representation with color-coding. High ratios of expression samples are coded red and low ratios are coded green, with the brightness of the color proportional to the magnitude of the differential expression. A ratio of 1 is black.



# Objectives

- Identify clusters of genes that share an expression profile.
- Genes in the same group are co-expressed, which gives hints that the unknown genes may have functions of the respective groups they cluster in.

# Clustering methods

- Clustering methods work for expression data of  $G$  genes under  $E$  experiments, for example, the expression data from  $G=2000$  genes at  $E=8$  time intervals.
- One type of approaches called *bottom-up* begins with each gene in its own cluster.
- Another type of approaches called *top-down* begins by selecting a predetermined number of clusters. Genes are then assigned to those clusters to minimize variation within clusters and maximize variation between them.

# Hierarchical clustering method

- It is a bottom-up approach.
- It begins with each gene in its own cluster. Clusters are then recursively clustered based on similarities, creating a hierarchical, treelike organization.
- Many of the same methods are used for both gene expression clustering and phylogeny reconstruction.

## Three steps of the hierarchical clustering method by Eisen et al. (1998)

- Construct a matrix of similarity measures between all pairs of genes;
- Recursively cluster the genes into a treelike hierarchy;
- Determine the boundaries between individual cluster.

# Correlation coefficient is a measure of similarity of a pair of genes

Let  $x_{ge}$  be the relative measure  $\log_2(\text{red/green})$  for gene  $g$  in experiment  $e$ . Generally,  $x_{ge}$  is a centered value for a given array  $e$  to remove the possible fluorescence bias. This centering step makes the mean value of  $x_{ge}$  over all genes in an array equals 0, or the mean ratio equals 1.

Let  $\bar{x}_g$  be the average (mean) expression level for gene  $g$  over the  $E$  experimental conditions,

$$\bar{x}_g = \frac{1}{E} \sum_{e=1}^E x_{ge}$$

Let  $s_g$  be the standard deviation of the measures in  $E$  experiments for gene  $g$ ,

$$s_g = \sqrt{\frac{1}{E} \sum_{e=1}^E (x_{ge} - \bar{x}_g)^2}$$

Then for a pair of gene  $i$  and  $j$  the correlation coefficient  $r_{ij}$  is

$$r_{ij} = \frac{1}{E} \sum_{e=1}^E \left( \frac{x_{ie} - \bar{x}_i}{s_i} \right) \left( \frac{x_{je} - \bar{x}_j}{s_j} \right)$$

Genes with similar expression profiles will have values of  $r_{ij}$  near unity. Correlation coefficient is a measure of the similarity between pairs of genes.

Correlation coefficient only measures the similarity of change patterns of a pair of genes, but ignores the scales of the changes by standardizing expression over experiments.

We calculate correlation coefficients for each of the  $[G(G-1)]/2$  pairs of genes  $i$  and  $j$ . With the matrix of correlations in hand, we proceed to the clustering algorithm in the following recursion:

1. Find the pair of clusters with the highest correlation and combine the pair into a single cluster.
2. Update the correlation matrix using the average values of the newly combined clusters.
3. Repeat step 1 and 2  $G-1$  times until all genes have been clustered.

Suppose the initial correlation matrix for  $G=5$  genes is

	2	3	4	5
1	0.3	0.2	0.8	0.1
2		0.9	0.1	0.8
3			0.2	0.7
4				0.1

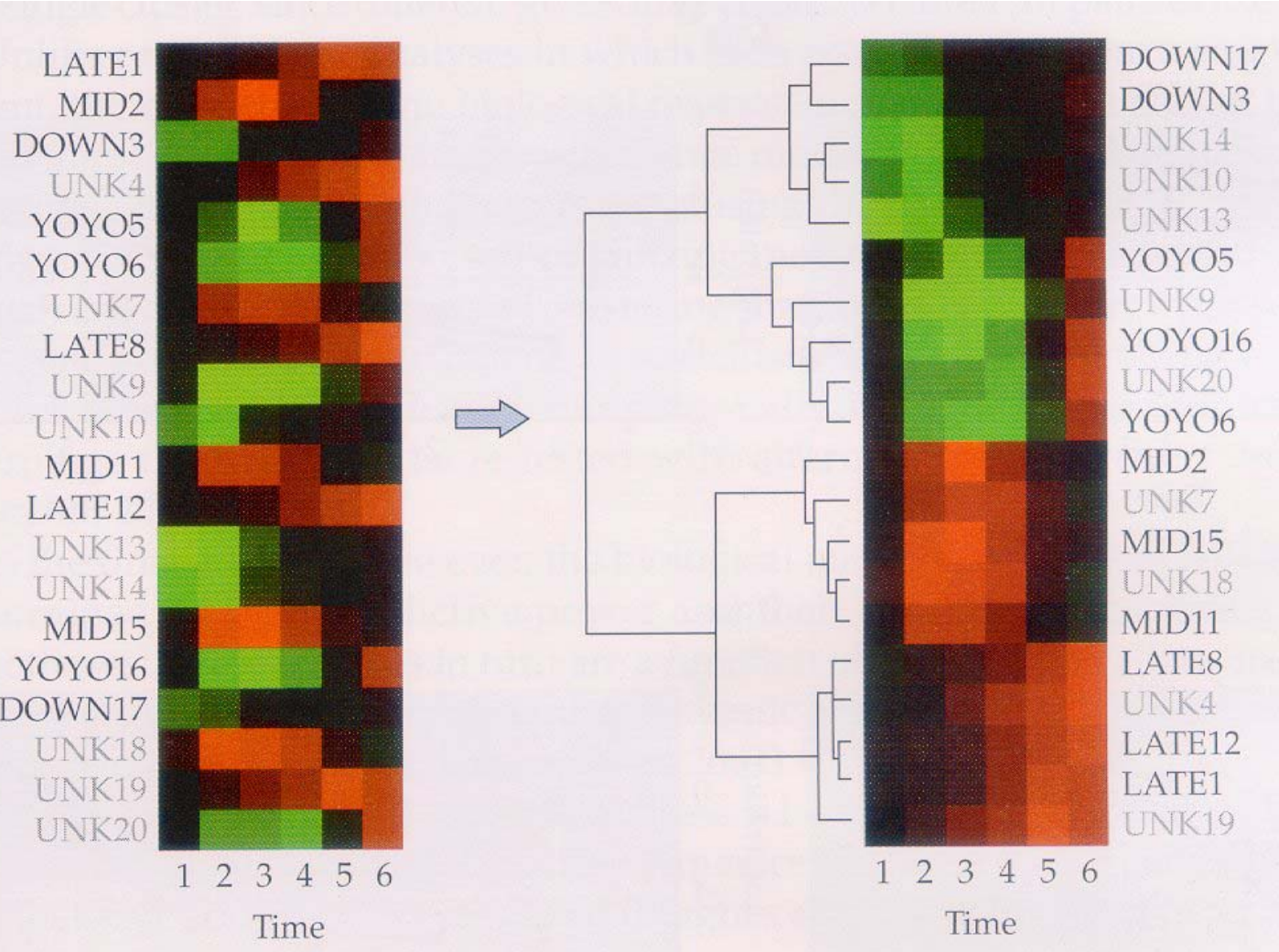
The highest correlation is the 0.9 observed between genes 2 and 3, so we combine them into a cluster and recompute the correlation matrix:

	23	4	5
1	0.25	0.8	0.1
23		0.15	0.75
4			0.1



The final step in the clustering process is to determine the boundaries of individual clusters. In the above example, do 2 and 3 form a cluster to the exclusion of 5, or is there a single 235 cluster? In some cases, we define clusters based on shared biological function. For instance, if genes 2, 3, and 5 were all heat-shock proteins, it might be sensible to include all of them in a single cluster.

An initially disordered set of gene expression profiles can be converted into an immediately intelligible set of clusters by hierarchical clustering. The clusters also suggest functions for those unknown (UNK) genes.



- Hierarchical clustering can be performed both on the genes and the treatments, allowing detection of patterns in two dimensions.
- There are numerous different types of clustering algorithms that may or may not give the same tree structure, and there is often no solid statistical support for the clustering produced by these methods.
- The biological meaning of clusters will be determined by their predictive power and their capacity for generating hypotheses. We can align cluster data with gene ontology and other genomic annotation, including linkage to literature database.

# References

- Textbook p141-145.
- Eisen, M. B., P. Spellman, P. Brown and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. (USA)* 95: 14863-14868.