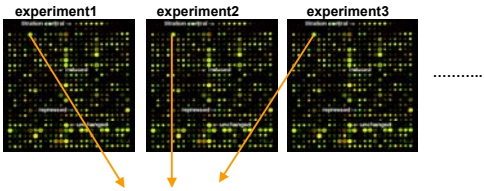
Microarray analysis 2

CS426, week 12
Golan Yona

2) Analysis of co-expression

Search for similarly expressed genes



Gene i: (x1,x2,x3,…)

This is an expression profile of a gene

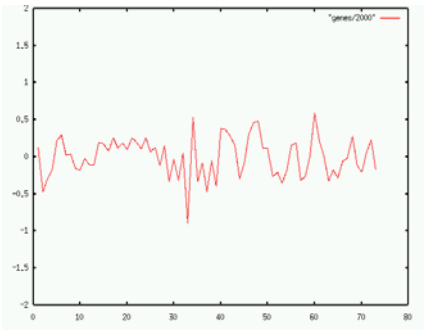
x1,x2,x3,.. are usually in log scale

Example: Yeast cell-cycle data

Expression at time t

$$\log(\frac{Et}{Ec})$$

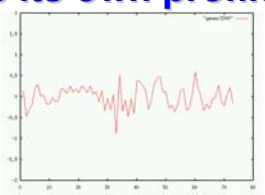
Background Expression (control)



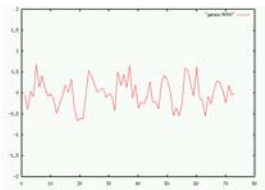
Time

Each gene has its own profile

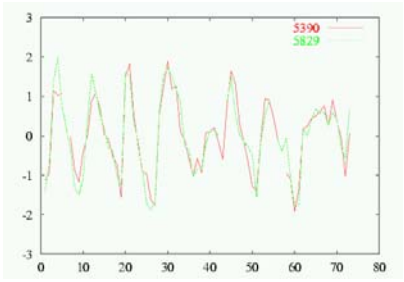
Gene X



Gene Y

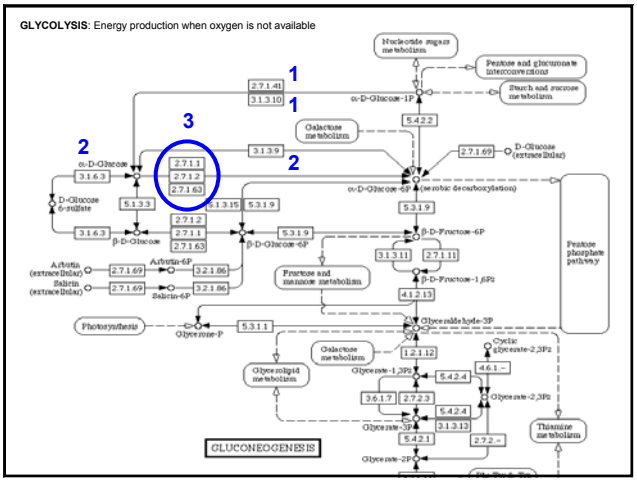


Genes can have similar expression profiles..

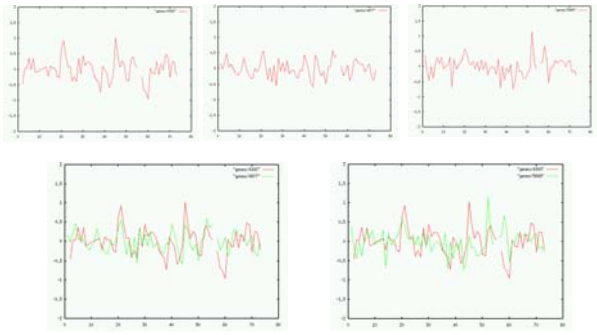


..because

1. They can be part of the same complex as interacting proteins (strong constraint)
2. They can be part of the same pathway without interacting directly
3. They can have similar regulatory elements (not necessarily functionally related)
4. They can have similar regulatory elements and similar sequences -> similar functions (fail-safe mechanisms through redundancy by gene duplication -> robustness, immunity, concurrency)

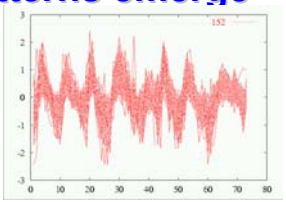


But similarity is not obvious

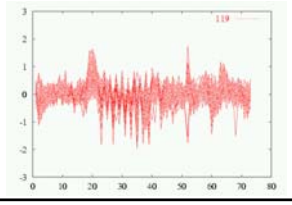


..yet strong patterns emerge

Cluster of 88 proteins involved in DNA replication and repair, mitosis and meiosis



Cluster of 141 proteins involved in translation initiation, ribosomal proteins, tRNA synthetases



How to measure similarity?

Given two expression vectors U,V of dimension d (number of features)

- The normalized Euclidean metric

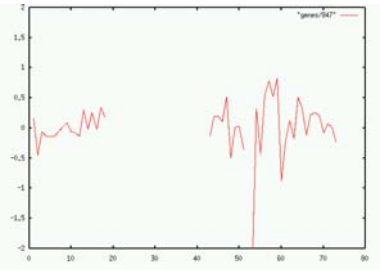
$$Dist_{euc}(\mathbf{V},\mathbf{U}) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(V_i - U_i)}$$

- Pearson correlation

$$Corr(\mathbf{V},\mathbf{U}) = \sum_{i=1}^{d}\frac{(V_i - < V >)(U_i - < U >)}{\sigma_V \sigma_U}$$

$\langle v \rangle = 1/d \sum_i v_i$ $\sigma_v = 1/d \sum_i (v_i - \langle v \rangle)$

Problem with missing values



Possible solutions

Averaging: replace every missing feature with the average over all genes.

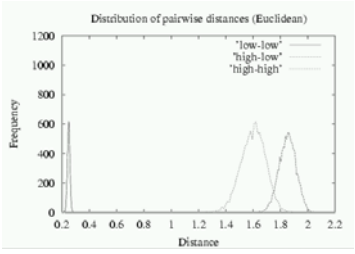
Nearest-neighbor approach (better): use only similar genes to define the values for the missing features (with weights)

EM approach (even better)– iteratively define classes based on the current approximation, and re-estimate the missing features based only on genes in these groups. Repeat until convergence.

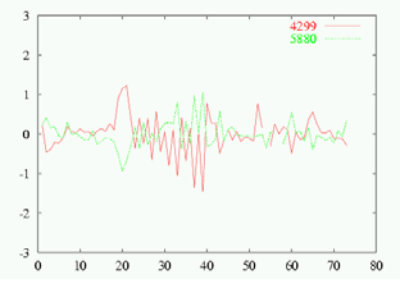
Or use just the available data – less noise but have to normalize properly

How to measure significance of expression similarity

Effective solution (see paper in the website): permute the vectors multiple times and compute the distance between the permuted vectors to generate a background distribution. Use that distribution to compute the zscore of the original distance value.



Correlation and anti-correlation



How to detect patterns in expression arrays

Clustering

Dimensionality reduction

Statistical and graphical models (e.g. Bayesian networks)

Clustering

Unsupervised learning technique for data exploration and pattern discovery

"Clustering is finding a natural grouping in a set of data, so that samples within a cluster will be more similar to each other than they are to samples in other clusters."

There are many clustering algorithms. We will focus on two.

The goal: finding groups of correlated genes ("signature groups") and extract features of groups.

Clustering can also be done for experiments

Hierarchical clustering

Clusters are composed of subclusters that are composed of subclusters that are composed of subclusters..

- The bottom level - each point (gene) is a cluster (n clusters).
- Next level - two clusters are merged (n-1 clusters)
- Next level - two clusters (from the previous level) are merged (n-2 clusters)
- …
- Top level – all n points are in one cluster

Usually represented in dendogram

Two types: divisive and agglomerative

Divisive

Top-down

Start with all samples in one cluster and successively split into separate clusters

Agglomerative

Bottom-up approach

Less computationally intensive

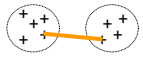
Start with n singletons and successively merge clusters

Outline:
Place each sample (gene) in a separate cluster
Repeat:
Merge the two most similar clusters into one cluster
Until all samples are in one cluster

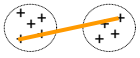
Types of agglomerative clustering

The types differ in the way we measure similarity/distance between clusters

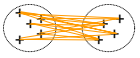
Single linkage – the distance between two clusters is the minimal distance between points in these clusters



Maximal linkage - the distance between two clusters is the maximal distance between points in these clusters



Average linkage - the distance between two clusters is the average distance between points in these clusters



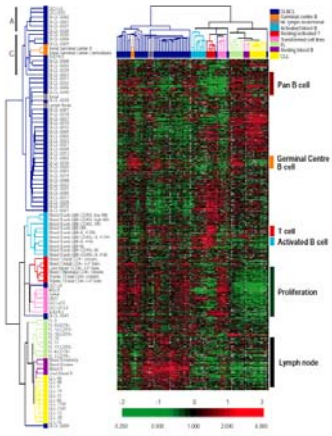


Figure 1

Figure 2:

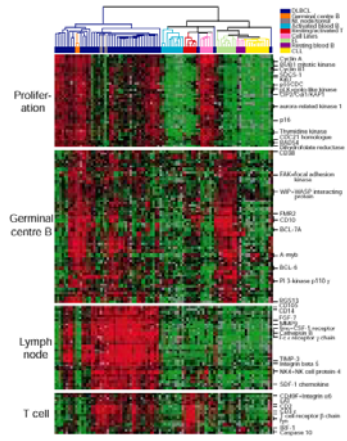
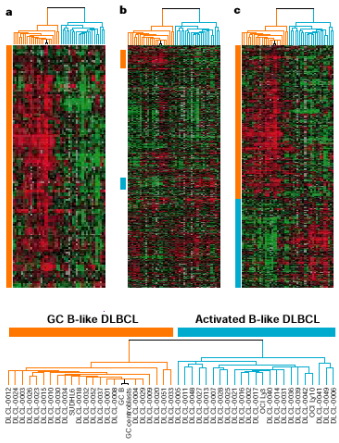


Figure 3:



Partitional clustering: k-means

Another popular clustering algorithm

Set the number of clusters k. Select centroids for these clusters (at random, or based on PCA)

Repeat:

- **Classify each sample to the closest class (closest centroid)**
- **Recompute the centroid of each class as the average of the samples classified to that class**

Until means (centroids) converge

Open issues

Unstable properties of clustering algorithms

Interpretation of results (there are other factors)