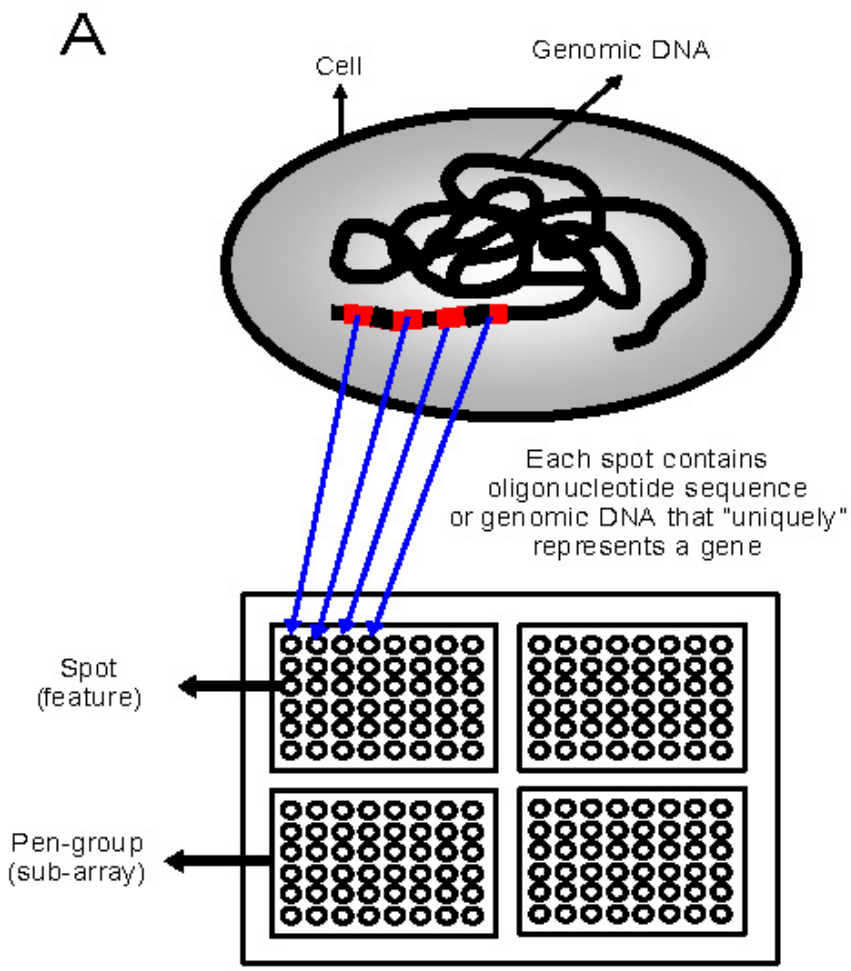
Introduction to Bioinformatics

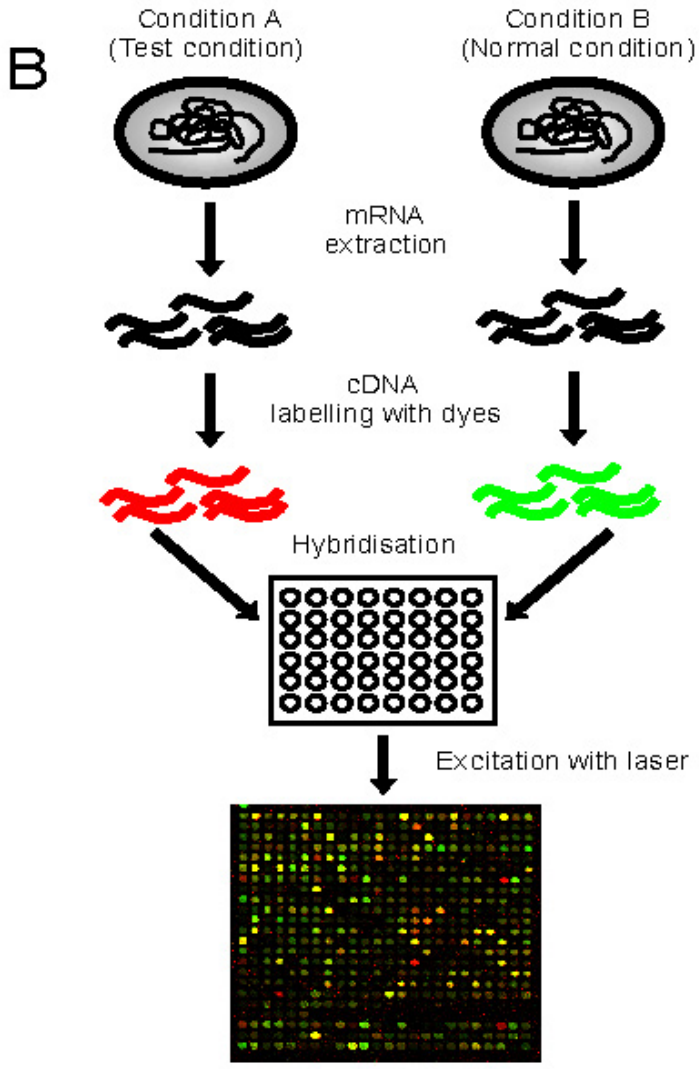
Microarray analysis

part 15

By: Mahdi Vasighi

DNA Microarray

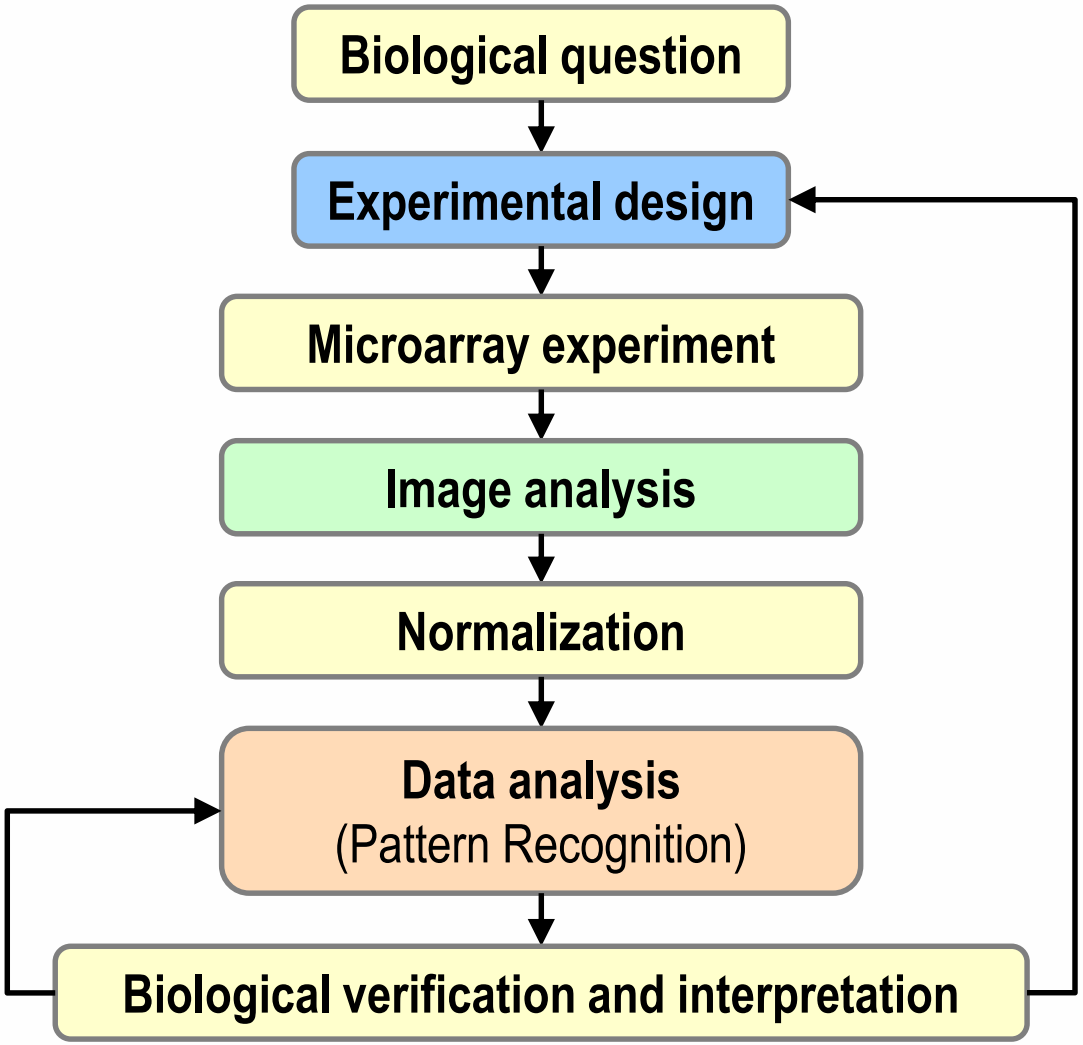




Final image stored as a file

DNA Microarray

The microarray data analysis process



DNA Microarray
Normalization

When one compares the expression levels of genes that should not change in the two conditions (say, housekeeping genes), what one quite often finds is that an average expression ratio of such genes deviates from 1.

may be due to various reasons:

- different dye absorption
- spatial heterogeneity in the chip
- different amounts of starting mRNA material

Normalization is a general term for a collection of methods that are directed at resolving the systematic errors and bias introduced by the microarray experimental platform to allow appropriate comparison of data

DNA Microarray
Normalization

Removing Flagged Features

- **Bad feature**: The pixel standard deviation is very high relative to the pixel mean.
- **Negative feature**: The signal of the feature is less than the signal of the background
- **Dark feature**: The signal of the feature is very low.
- **Manually flagged feature**: The user has flagged the feature using the image processing software.

Background Subtraction

- contribution of non-specific hybridization of labelled target to the glass,
- natural fluorescence of the glass slide itself.

For Affymetrix Data

- Discard these genes from the analysis
- Replace the negative numbers with the smallest possible positive number

DNA Microarray
Normalization

Total intensity normalization

The basic assumptions in a total intensity normalization are

- The total quantity of RNA for the two samples is the same.
- Same number of molecules of RNA from both samples hybridize to the microarray.

Then, the total hybridization intensities for the gene-sets should be equal.

$$N_{total} = \frac{\sum_{k=1}^{N_{gene-set}} R_k}{\sum_{k=1}^{N_{gene-set}} G_k} \qquad T_k^{'} = \frac{R_k^{'}}{G_k^{'}} = \frac{R_k}{G_k \times N_{total}} = \frac{T_k}{N_{total}}$$

DNA Microarray
Normalization

Mean log centering

In this method, the basic assumption is that the mean $\log_2$(expression ratio) should be equal to 0 for the gene-sets.

$$N_{mlc} = \frac{\sum_{k=1}^{N_{gene-set}} \log_2\left(\frac{R_k}{G_k}\right)}{N_{gene-set}}$$

The intensities are now rescaled such that

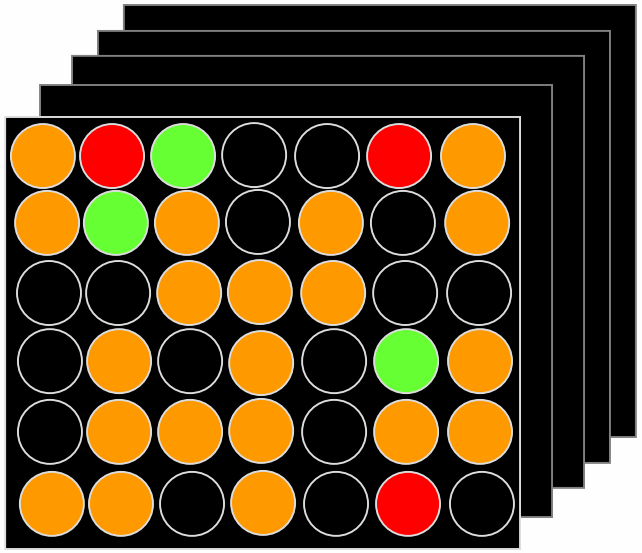
$$G_k^{'} = G_k \times (2^{N_{mlc}}) \qquad R_k^{'} = R_k$$

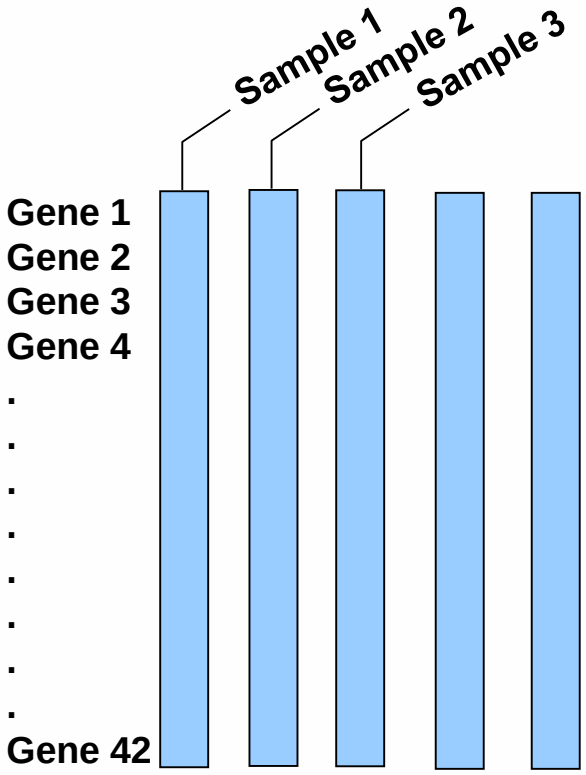
$$T_k^{'} = \frac{R_k^{'}}{G_k^{'}} = \frac{R_k}{G_k \times (2^{N_{mlc}})} = \frac{T_k}{2^{N_{mlc}}}$$

$$\log_2 (T_k^{'}) = \log_2 (T_k) - \log_2 (2^{N_{mlc}}) = \log_2 (T_k) - N_{mlc}$$

MA plot

DNA Microarray

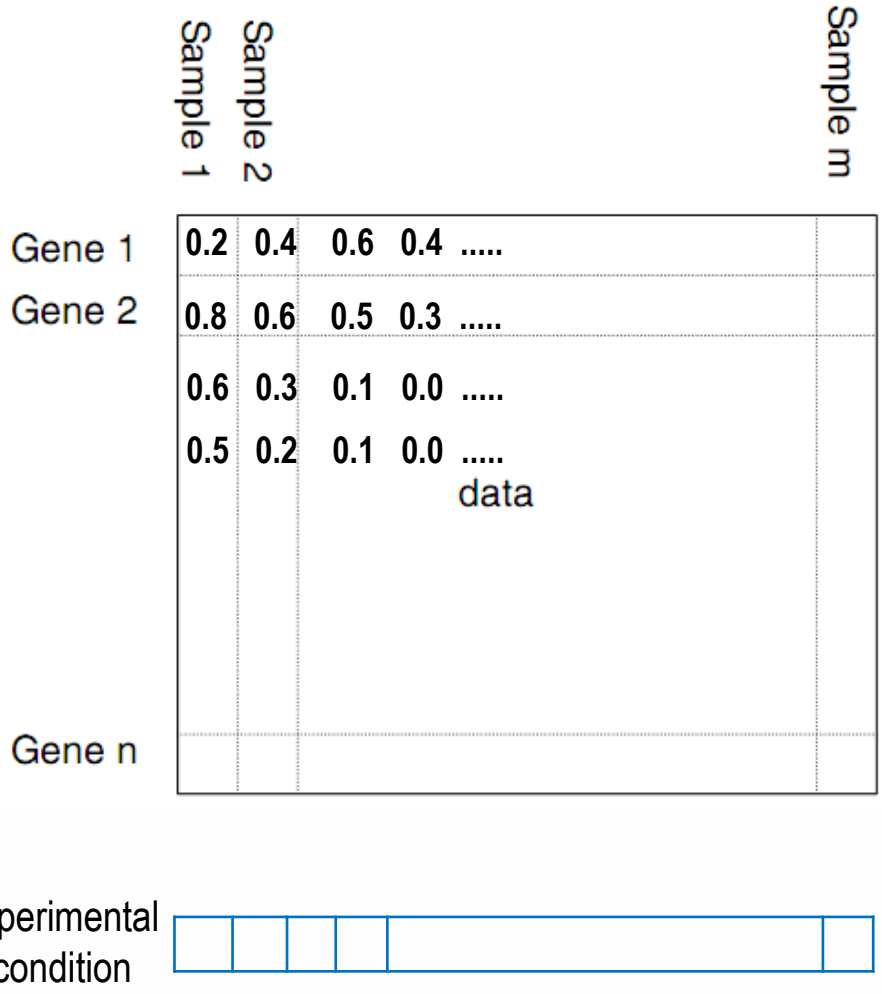




The processed data, after the normalization procedure, can then be represented in the form of a matrix, often called gene expression matrix

DNA Microarray

Each row in the matrix corresponds to a particular gene and each column could either correspond to an experimental condition or a specific time point at which expression of the genes has been measured.



- Supervised learning
- Unsupervised learning

Example: ‘disease state’ or ‘normal state’ → [1,0]

DNA Microarray
Analysis of data

Similarity (Distance)

It is common to describe the similarity between two profiles in terms of the distance between them in the high-dimensional space of gene expression or sample measurements.

Correlation Coefficient

This is a statistical concept that quantifies the level of relationship between two sets of measurements

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

r takes a value from between −1 and +1. A value of −1 represents strong

$d(X,Y) = 1 - abs(r(X,Y))$

$d(X,Y) = 1 - r(X,Y)^2$

correlation. A value of 0 represents uncorrelated variables.

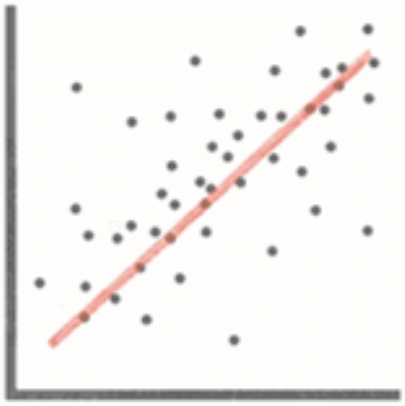
DNA Microarray

Analysis of data

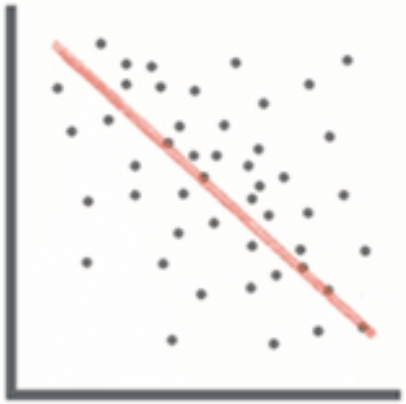
Similarity (Distance)

It is common to describe the similarity between two profiles in terms of the distance between them in the high-dimensional space of gene expression or sample measurements.

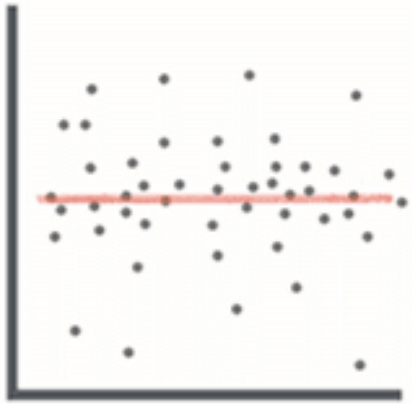
Correlation Coefficient



Positive Correlation



Negative Correlation



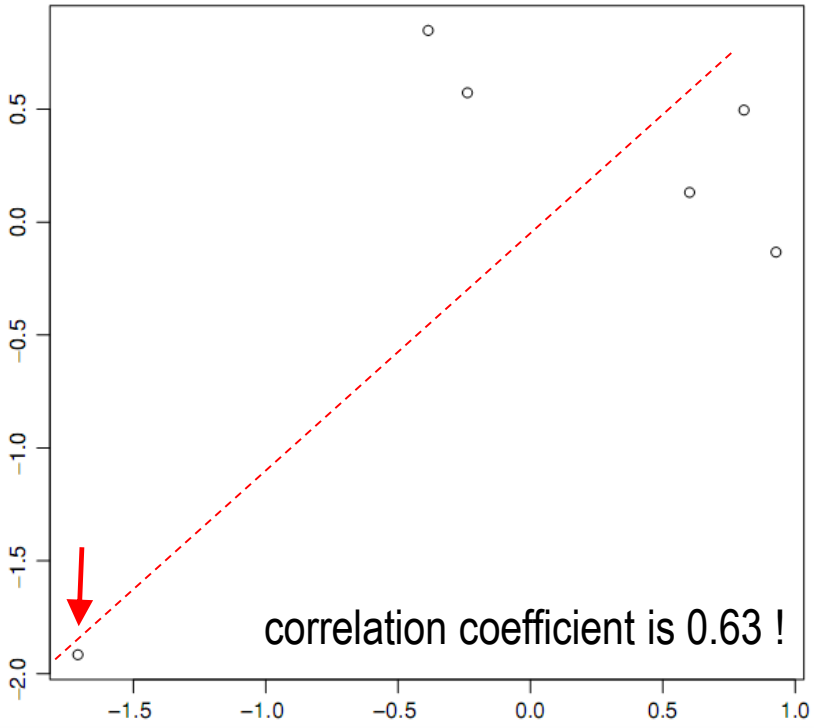
No Correlation

DNA Microarray
Analysis of data

Spearman Correlation

Spearman correlation is a measure of correlation that is robust to outliers and, because of this, it is often more appropriate for microarray analysis.

Time	X	Y
0.5	−0.76359	−4.05957
2	2.276659	−1.7788
5	2.137332	−0.97433
7	1.900334	−1.44114
9	0.932457	−0.87574
11	0.761866	−0.52328



DNA Microarray

Analysis of data

Spearman Correlation

Spearman correlation is a measure of correlation that is robust to outliers and, because of this, it is often more appropriate for microarray analysis.

Time	X	Y	X Rank	Y Rank
0.5	−0.76359	−4.05957	1	1
2	2.276659	−1.7788	6	2
5	2.137332	−0.97433	5	4
7	1.900334	−1.44114	4	3
9	0.932457	−0.87574	3	5
11	0.761866	−0.52328	2	6

Calculate *r* between ranks

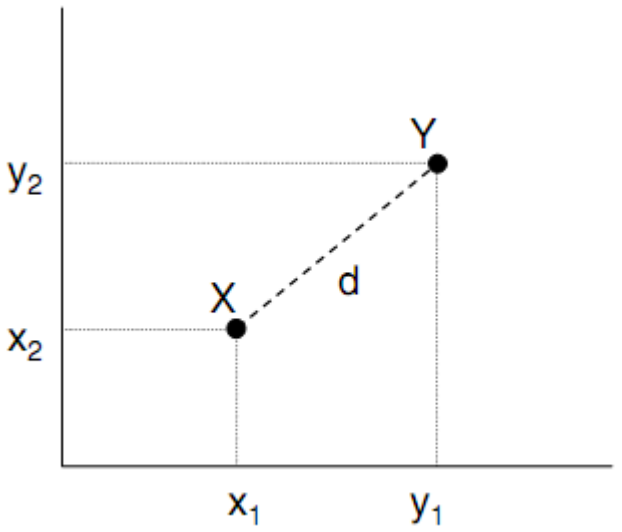
Now correlation coefficient is -0.09

DNA Microarray

Analysis of data

Euclidean Distance

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_j)^2}$$



A more general form of the Euclidean distance is called the *Minkowski distance*, calculated as:

$$D_{Min}(A,B) = \sqrt[p]{\sum_{i=1}^{n} (a_i - b_i)^p}$$

One of the key problems with Euclidean distance is that it is not scale invariant: two gene expression profiles with the same shape but different magnitudes will appear to be very distant.

DNA Microarray

Analysis of data

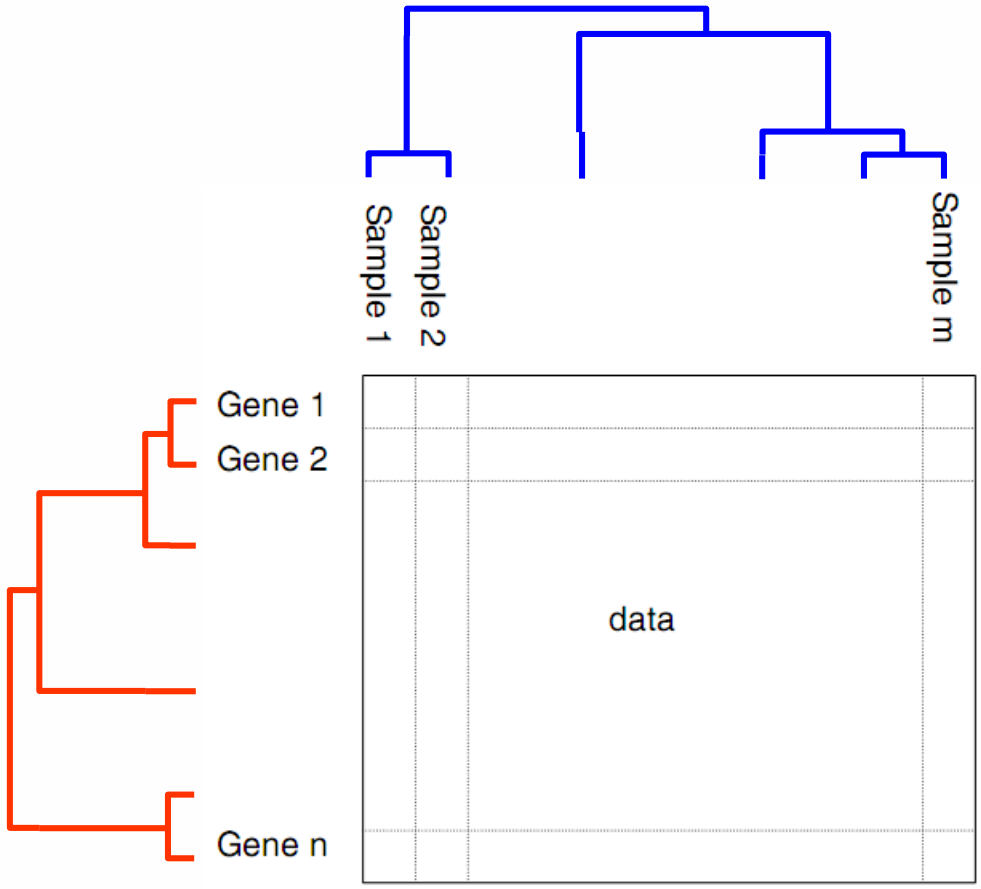
Clustering of microarray data

There are two straightforward ways to study the gene expression matrix:

1. Comparing expression profiles of genes by comparing rows in the expression matrix.

2. Comparing expression profiles of samples by comparing columns in the matrix.

If we find that two rows are similar, we can hypothesize that the respective genes are co-regulated and possibly functionally related.



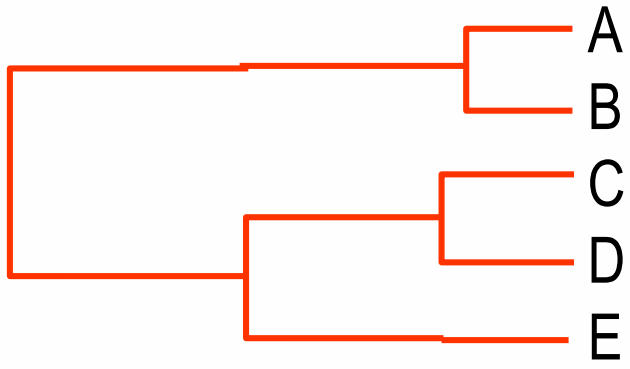
DNA Microarray

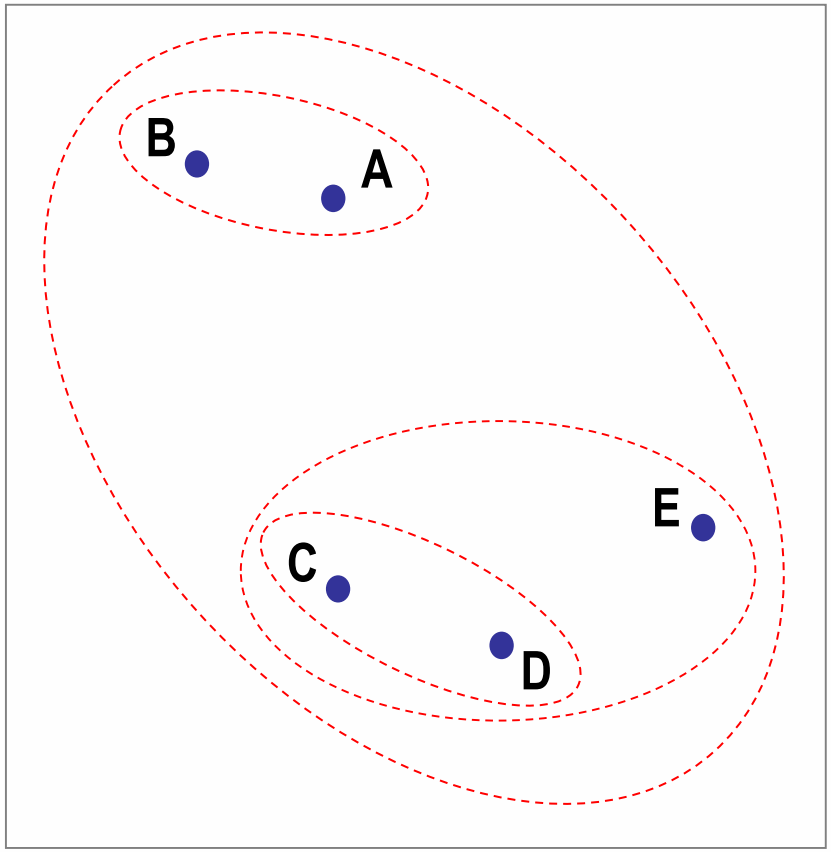
Analysis of data

Clustering of microarray data

Hierarchical cluster analysis

HCA arranges the gene or sample profiles into a tree so that similar profile appear close together in the tree and dissimilar profiles are farther apart.





DNA Microarray

Analysis of data

Clustering of microarray data | **Hierarchical cluster analysis**

- Calculate a distance matrix between all points (d)
- While there is more than one cluster
  1. Find the two closest clusters C_1 and C_2
  2. Merge C_1 and C_2 into new cluster C_{new}
  3. Compute distance from C_{new} to all other clusters
  4. Remove rows and columns of d corresponding to C_1 and C_2
  5. Add a row and column to d corresponding to the new cluster C
  6. Go to step 1

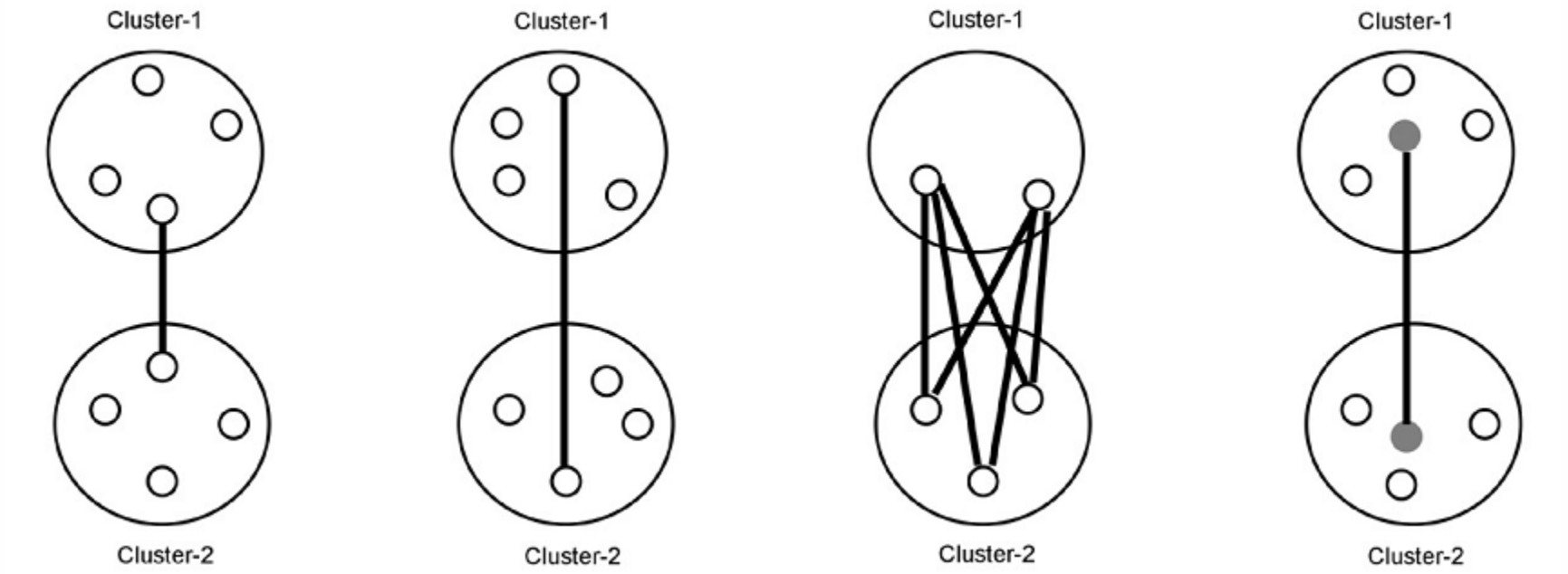
There are a number of different methods for doing this, and the trees formed by these methods frequently look different.

DNA Microarray

Analysis of data

Clustering of microarray data | **Hierarchical cluster analysis**

Linkage methods:



Single linkage | **Complete linkage** | **Average linkage** | **Centroid linkage**

DNA Microarray

Analysis of data

Clustering of microarray data | **Hierarchical cluster analysis**

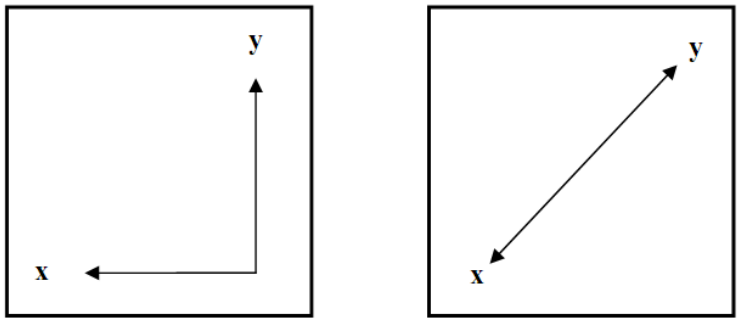
Distance measures:

Euclidean distance :

$$d_E(\mathbf{x},\mathbf{y}) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Manhattan distance :

$$d_M(\mathbf{x},\mathbf{y}) = |x_1-y_1|+|x_2-y_2|+\ldots+|x_n-y_n| = \sum_{i=1}^{n}|x_i-y_i|$$



Manhattan Euclidean

DNA Microarray
Analysis of data

Clustering of microarray data | **Hierarchical cluster analysis**

Distance measures:

Chebychev distance:

$$d_{max}(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i|$$

Correlation distance:

$$d_R(\mathbf{x},\mathbf{y}) = 1 - r_{xy}$$

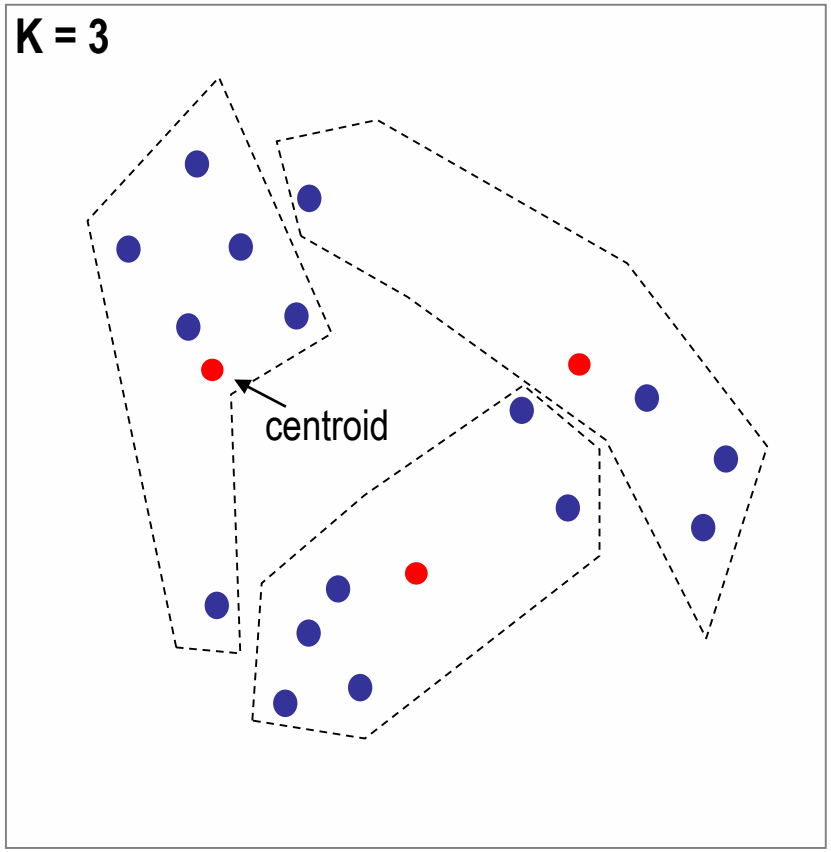
Mahalanobis distance:

DNA Microarray

Analysis of data

Clustering of microarray data | **K-mean clustering**

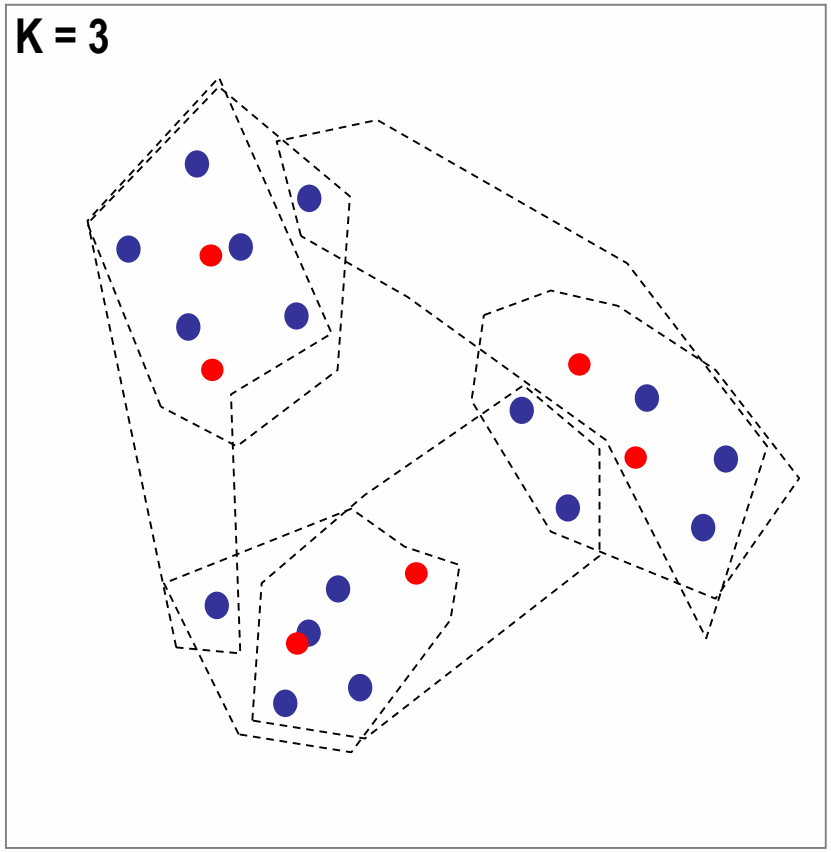
1. First, the user tries to estimate the number of clusters (K)
2. Randomly choose N points into K clusters.
3. Calculate the centroid for each cluster.
4. For each point, move it to the closest cluster.
5. Repeat stages 3 and 4 until no further points are moved to different clusters.



DNA Microarray
Analysis of data

Clustering of microarray data | **K-mean clustering**

1. First, the user tries to estimate the number of clusters (K)
2. Randomly choose N points into K clusters.
3. Calculate the centroid for each cluster.
4. For each point, move it to the closest cluster.
5. Repeat stages 3 and 4 until no further points are moved to different clusters.



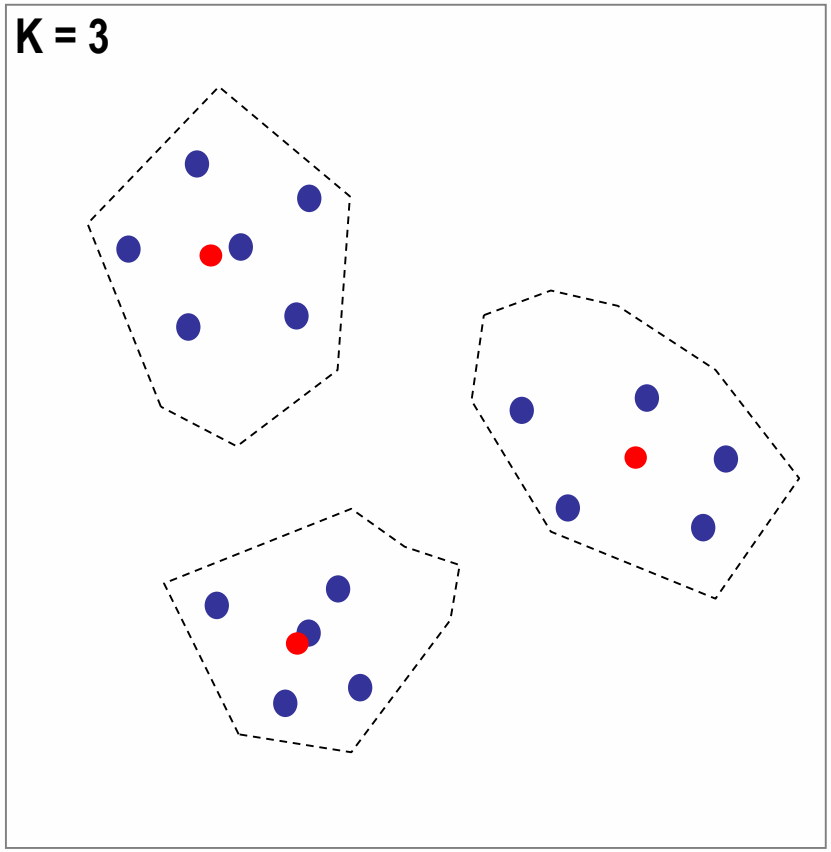
DNA Microarray
Analysis of data

Clustering of microarray data | **K-mean clustering**

This is one of the simplest and fastest clustering algorithms.

However, it has a major drawback: The results of the k-means algorithm may change in successive runs because the initial clusters are chosen randomly.

- How can we assess the quality?



DNA Microarray

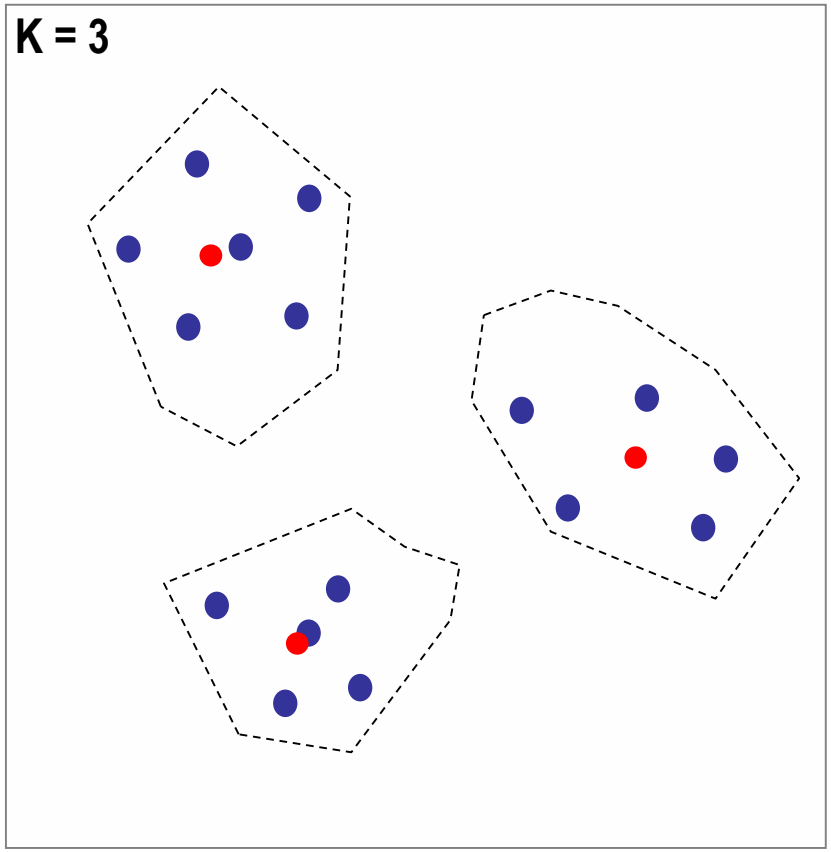
Analysis of data

Clustering of microarray data | **K-mean clustering**

we can assess the quality by:

- Repeating the clustering and check the stability of cluster members

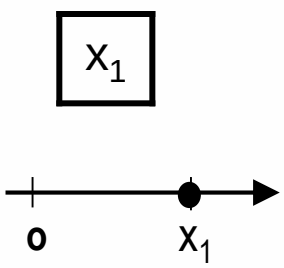
- Average of distance between members of a cluster and cluster center.

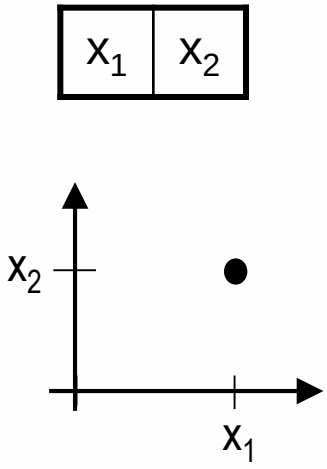


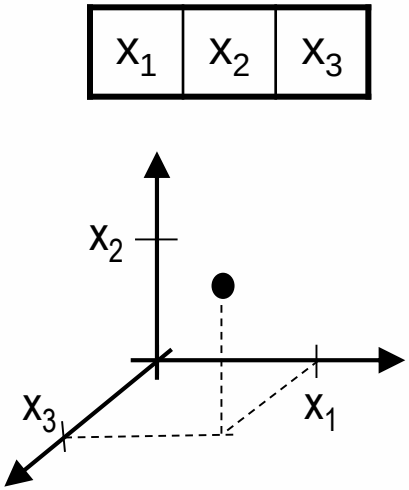
DNA Microarray

Analysis of data

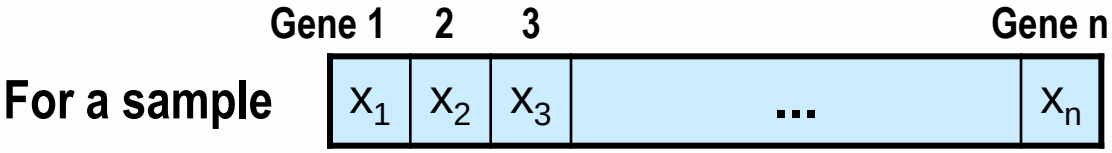
Dimensionality reduction







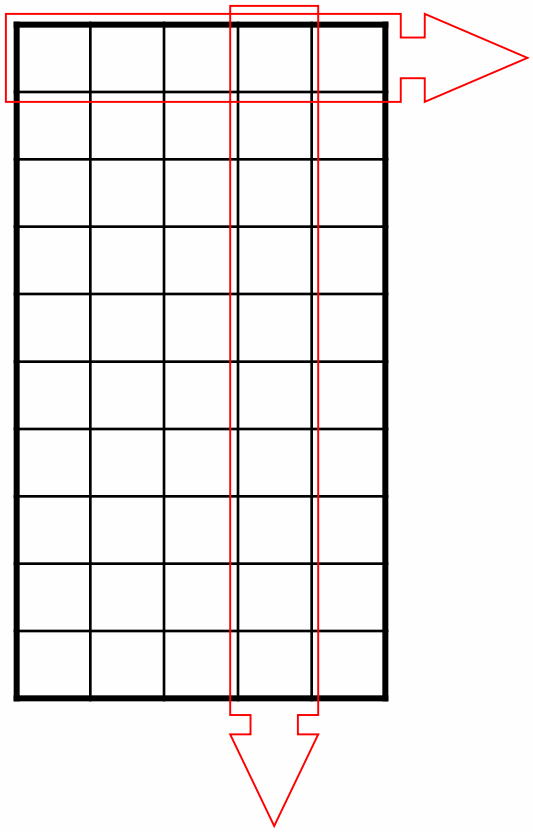
One of the central features of microarray data is that there is a lot of it!



This is the coordination of a point in n-dimensional space

DNA Microarray
Analysis of data

Dimensionality reduction



A point in 5 dimensional space

A point in 10 dimensional space

For example, we have expression values of 12000 genes for 20 sample

Data matrix (12000×20)

DNA Microarray
Analysis of data

Dimensionality reduction

Often, we want to visualize microarray data, either:

- as an aid to visual analysis or
- as a precursor to the application of more sophisticated algorithms.

we are trying to represent very high-dimensional data in the two or three dimensions described. We will describe two methods for visualizing microarray data in two or three dimensions:

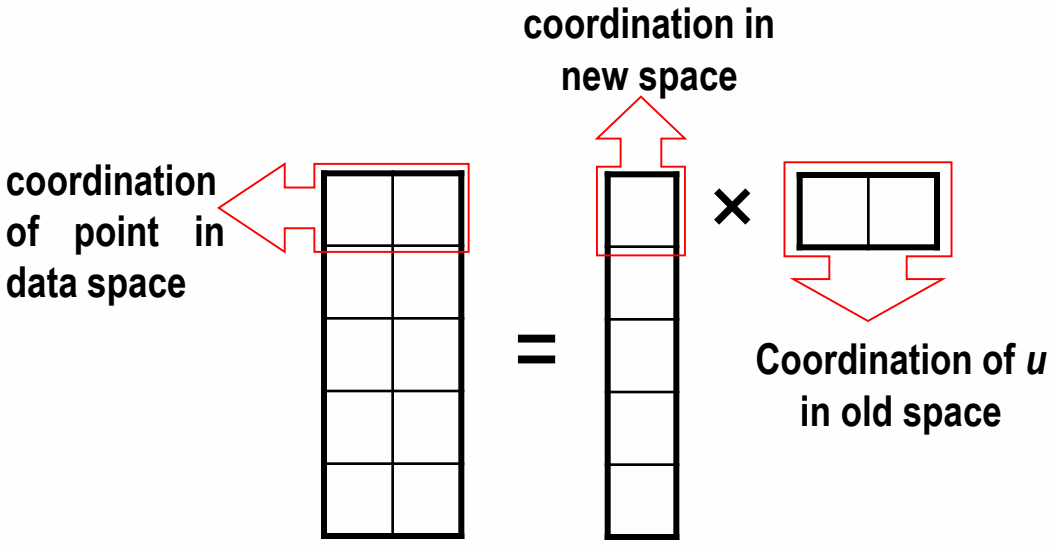
- Principal Component Analysis (PCA)
- Multi-Dimensional Scaling (MDS)

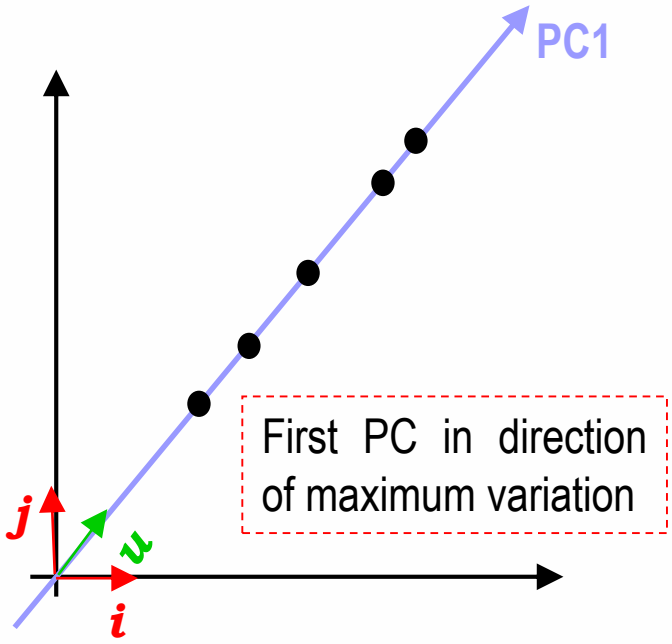
DNA Microarray

Analysis of data

Dimensionality reduction | **Principle component analysis**

Principal component analysis (PCA) is a method that projects a high-dimensional space onto a lower dimensional space.



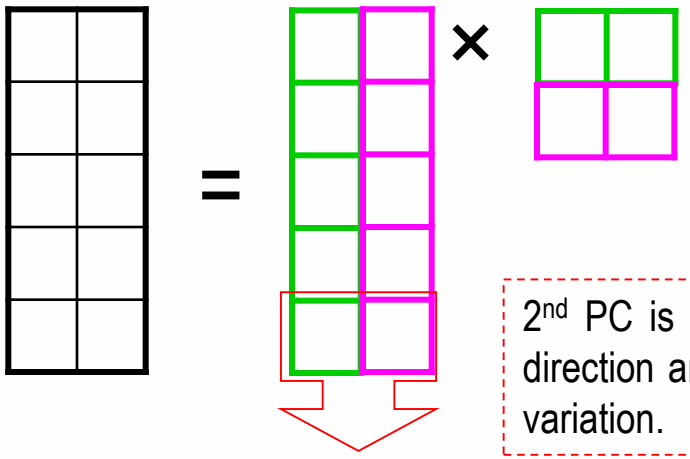


DNA Microarray

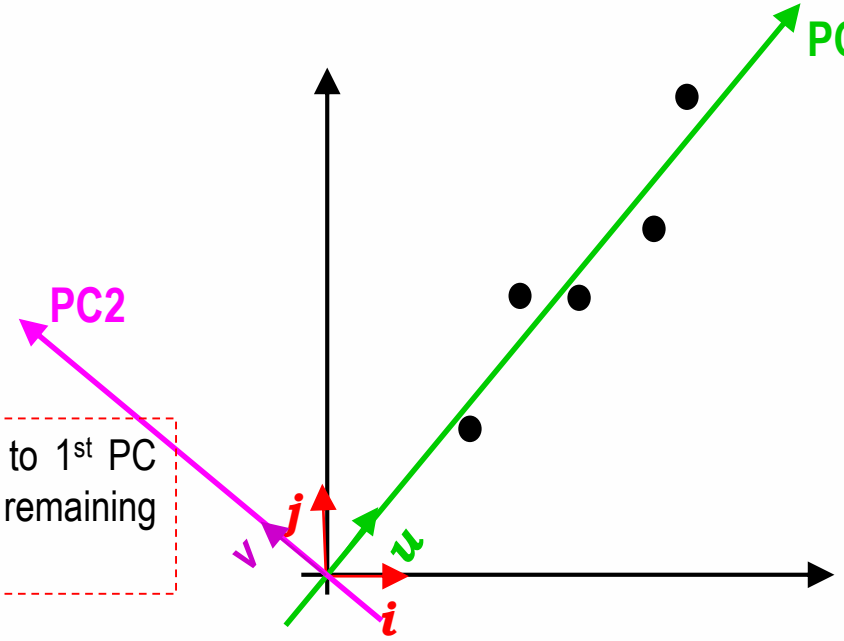
Analysis of data

Dimensionality reduction | **Principle component analysis**

Principal component analysis (PCA) is a method that projects a high-dimensional space onto a lower dimensional space.



2nd PC is orthogonal to 1st PC direction and models remaining variation.



DNA Microarray

Analysis of data

Dimensionality reduction | **Principle component analysis**

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}$$

X
Data matrix

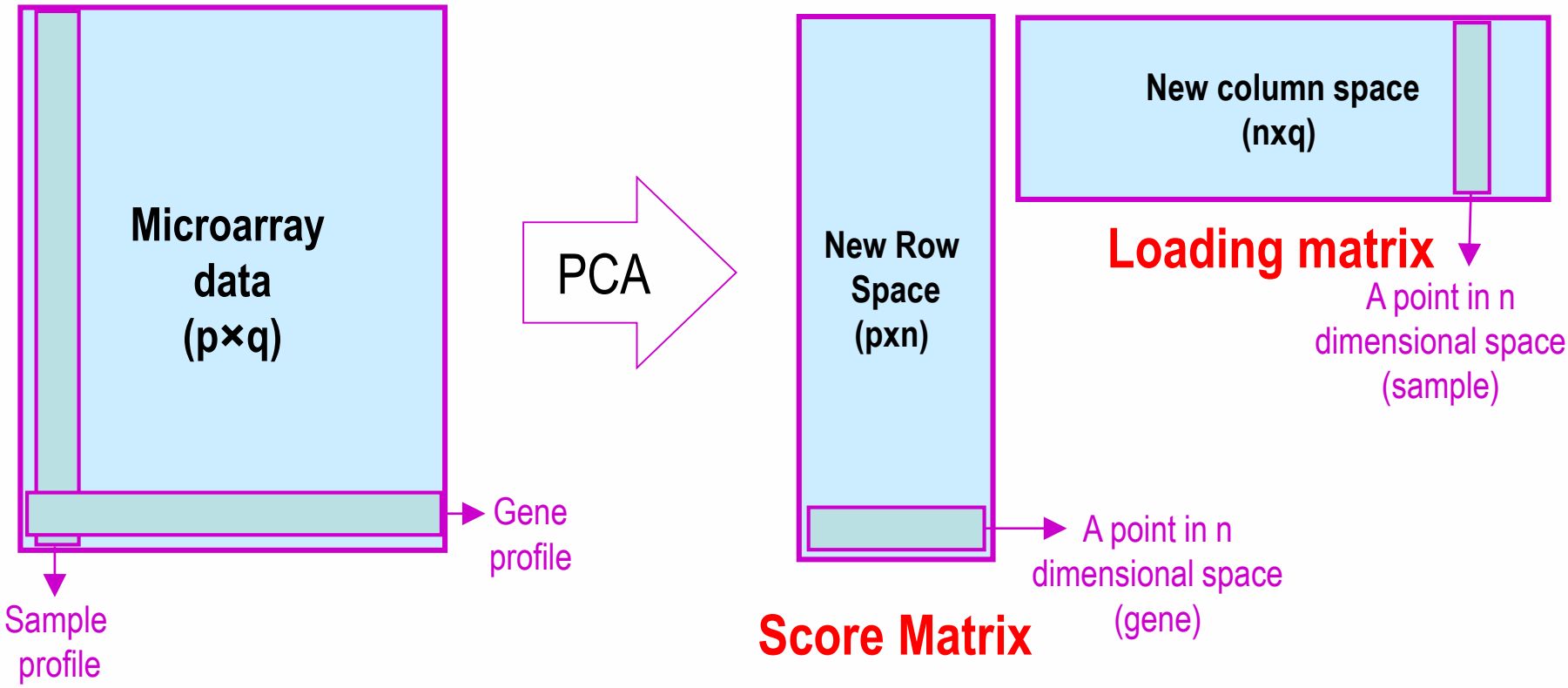
T
Score matrix

P^T
Loading matrix

DNA Microarray

Analysis of data

Dimensionality reduction | **Principle component analysis**

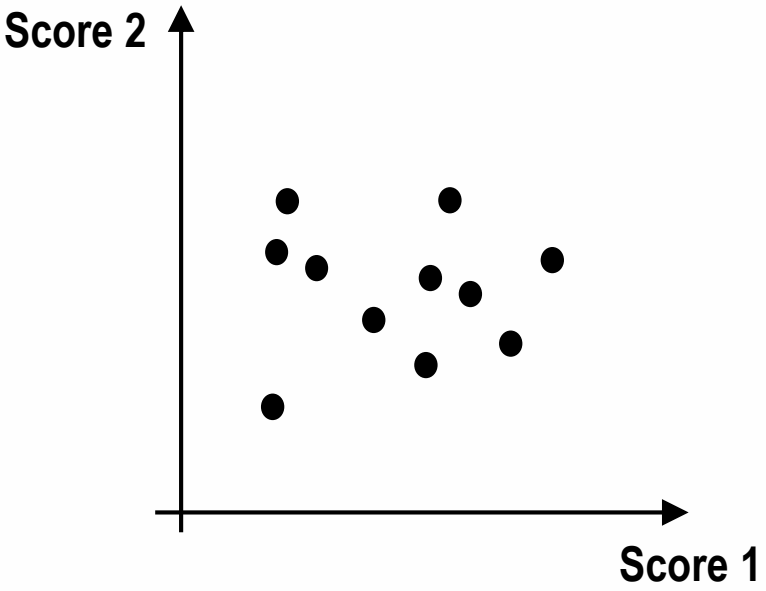


DNA Microarray

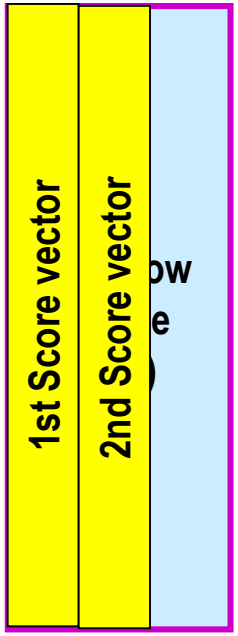
Analysis of data

Dimensionality reduction

Principle component analysis



Score plot



Score Matrix

New column space
(nxq)

Loading matrix

DNA Microarray
Analysis of data

Dimensionality reduction | **Principle component analysis**

Nature Biotechnology 26, 303-304 (2008)

_computational BIOLOGY

PRIMER

What is principal component analysis?

Markus Ringnér

Principal component analysis is often incorporated into genome-wide expression studies, but what is it and how can it be used to explore high-dimensional data?

Several measurement techniques used in the life sciences gather data for many more variables per sample than the typical number

geometrical interpretations of the data. To allow for such interpretations, imagine that the microarrays in our example measured the

Dimensional reduction and visualization

We can reduce the dimensionality of our two-dimensional expression profiles to a single

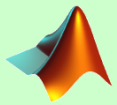
nature.com/naturebiotechnology

DNA Microarray

Analysis of data

Dimensionality reduction | **Principle component analysis**

princomp

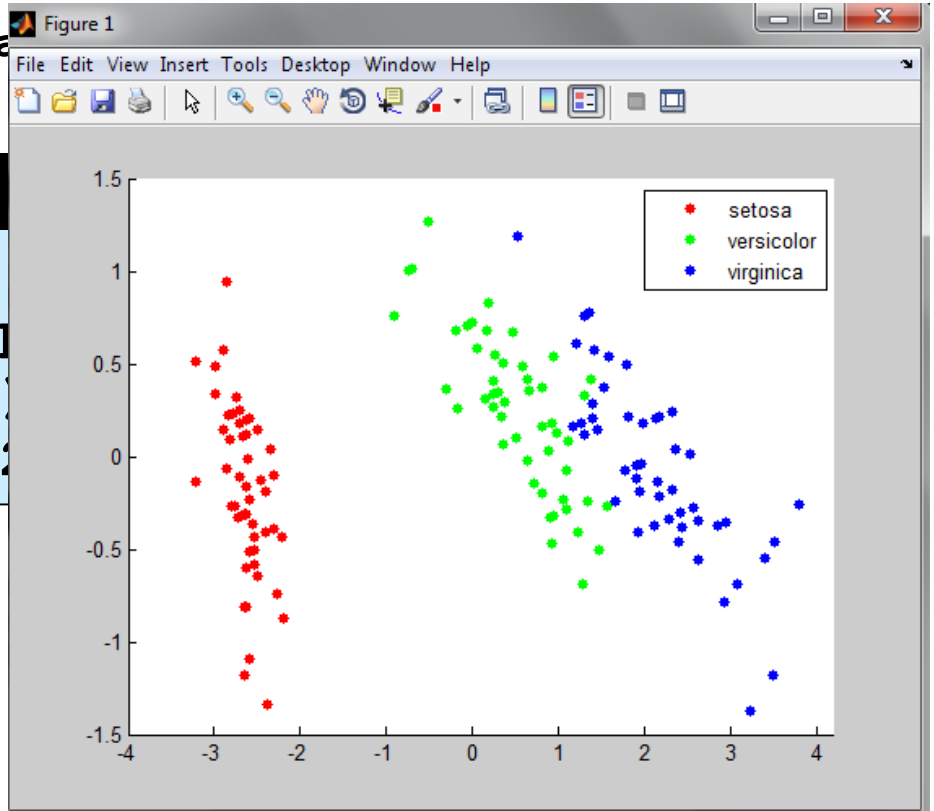


Purpose: Principal component ana

Syntax: [COEFF,SCORE,latent

Example

```
load fisheriris;
[pc,score,latent,tsquare] = p
scatter(score(:,1),score(:,2)
gscatter(score(:,1),score(:,
```

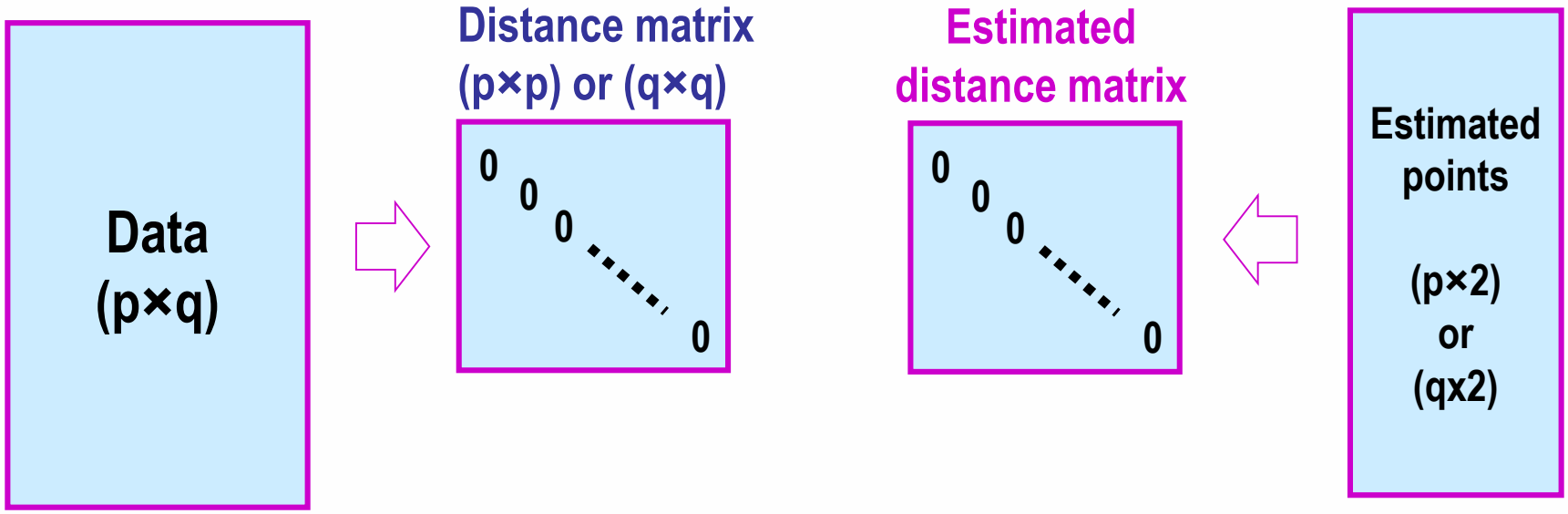


DNA Microarray

Analysis of data

Dimensionality reduction | **Multi-Dimensional Scaling**

Multidimensional scaling (MDS) is a different approach to dimensionality reduction and visualization. It starts from from the measurements of distance between the samples or profiles being compared.

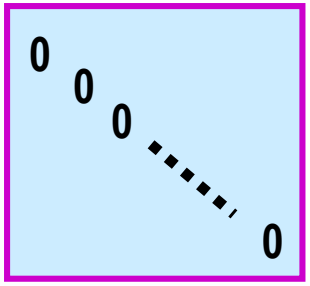


DNA Microarray

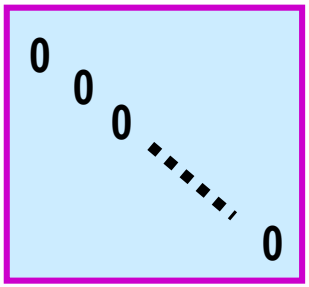
Analysis of data

Dimensionality reduction | **Multi-Dimensional Scaling**

Distance matrix (p×p) or (q×q)



Estimated distance matrix



estimated Coordination

(p×n) or (qxn)

Changing estimated coordination will cause changing in estimated distances and if we have a good estimation, then stress should be minimized.

$$Stress = \sum_{i,j} \mathbf{D}_{i,j} - \mathbf{D}_{(est)_{i,j}}$$

DNA Microarray

	Software	URL
1	**Express Yourself** - An automated, online microarray data processing platform, where you can upload image files and carry out data processing and data analysis.	http://array.mbb.yale.edu/analysis/
2	**Expression Profiler** - A set of tools for clustering, analysis and visualization of gene expression and other genomic data. Tools in the Expression Profiler allow to perform cluster analysis, pattern discovery, pattern visualization, study and search Gene Ontology categories, generate sequence logos, extract regulatory sequences, study protein interactions, as well as to link analysis results to external databases.	http://ep.ebi.ac.uk/EP/
3	**Cluster & Treeview** - **Cluster** performs a variety of types of cluster analysis and other types of processing on large microarray datasets. Currently includes hierarchical clustering, self-organizing maps (SOMs), K-means clustering, principal component analysis. **Treeview** can be used to graphically browse results of clustering and other analyses from Cluster.	http://rana.lbl.gov/EisenSoftware.htm
4	**Xcluster** - cross platform software for analysing microarray data.	http://genetics.stanford.edu/~sherlock/cluster.html
5	**J-Express** - A Java implementation of hierarchical clustering, self organized maps, and principal component analysis, with several different viewing options and output formats.	http://www.microarrays.org/software.html
6	**TM4** - A package of Open Source software programs for microarray analysis	http://www.tigr.org/software/
7	**GeneXPress** - A visualization and analysis tool for gene expression data, integrating clustering, gene annotation, and sequence information.	http://genexpress.stanford.edu/
8	**GEPAS** - Gene Expression Pattern Analysis Suite.	http://gepas.bioinfo.cnio.es/tools.html
9	**GenMAPP** - A computer application designed to visualize gene expression data on maps representing biological pathways, and other biologically meaningful groups of genes.	http://www.genmapp.org/
10	**OligoArray** - An application which computes gene specific oligonucleotides for genome-scale oligonucleotide microarray construction.	http://berry.engin.umich.edu/oligoarray/