Introduction to Bioinformatics

Fabian Hoti 6.10.

Analysis of Microarray Data

- Introduction
- Different types of microarrays
- Experiment Design
- Data Normalization
- Feature selection/extraction
- Clustering

Microarrays: Introduction

- A method to measure the activity of genes in
  - different phaces of the cell cycle,
  - different phases of development,
  - different invironments (treatments)
- Sets of genes that are over/under expressed in certain conditions are likely to have a related biological function or regulatory relationship
- Microarray technolgy enables measuring the activity/expression level of thousands of genes simultaneously

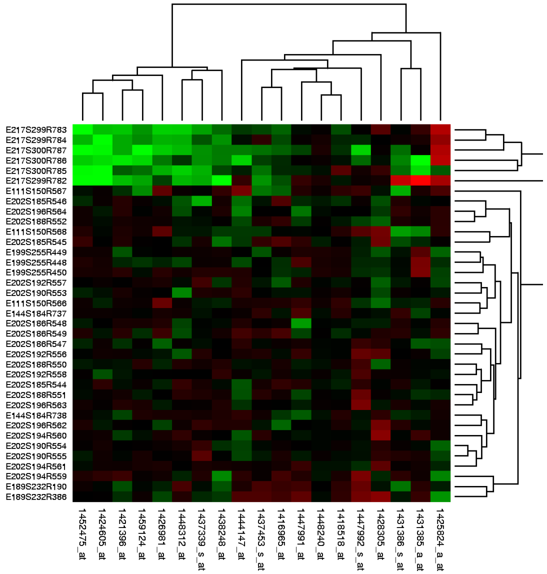
Microarrays: Introduction

In short

- DNA sequences corresponding to genes of the organism are attached to a support medium (for example a glass slide)
- mRNA extracted from tissues or cells is copied into cDNA and labeled with fluorescent dye
- The labeled cDNA is hybridized to the support medium and scanned with a microscope to measure the amount of label at each gene position
- The amount of label is assumed to be proportional to the amount of RNA in the original biological sample

Microarrays: Introduction

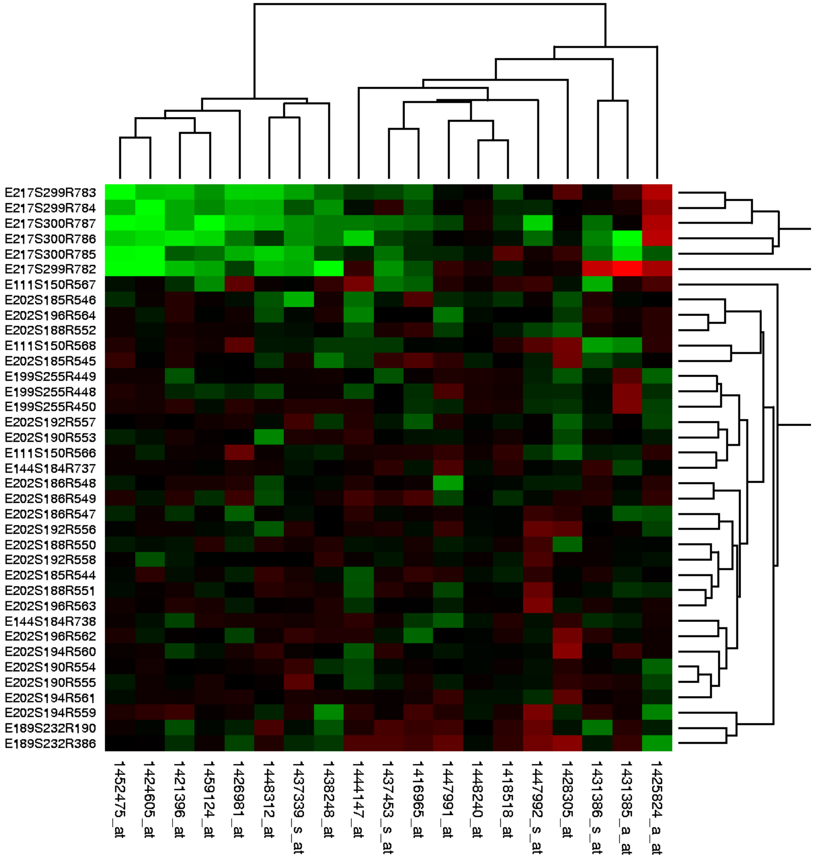
- Output of a microarray experiment



- Samples (vertical axis), genes (horizontal axis)
- Entry (i, j) gives the expression level of a gene (j) in sample i
- Under expressed (green), over expressed (red)
- Samples and genes are ordered using a clustering algorithm

Microarrays: Introduction

- Clustering reveals groups of samples/genes with similar profiles



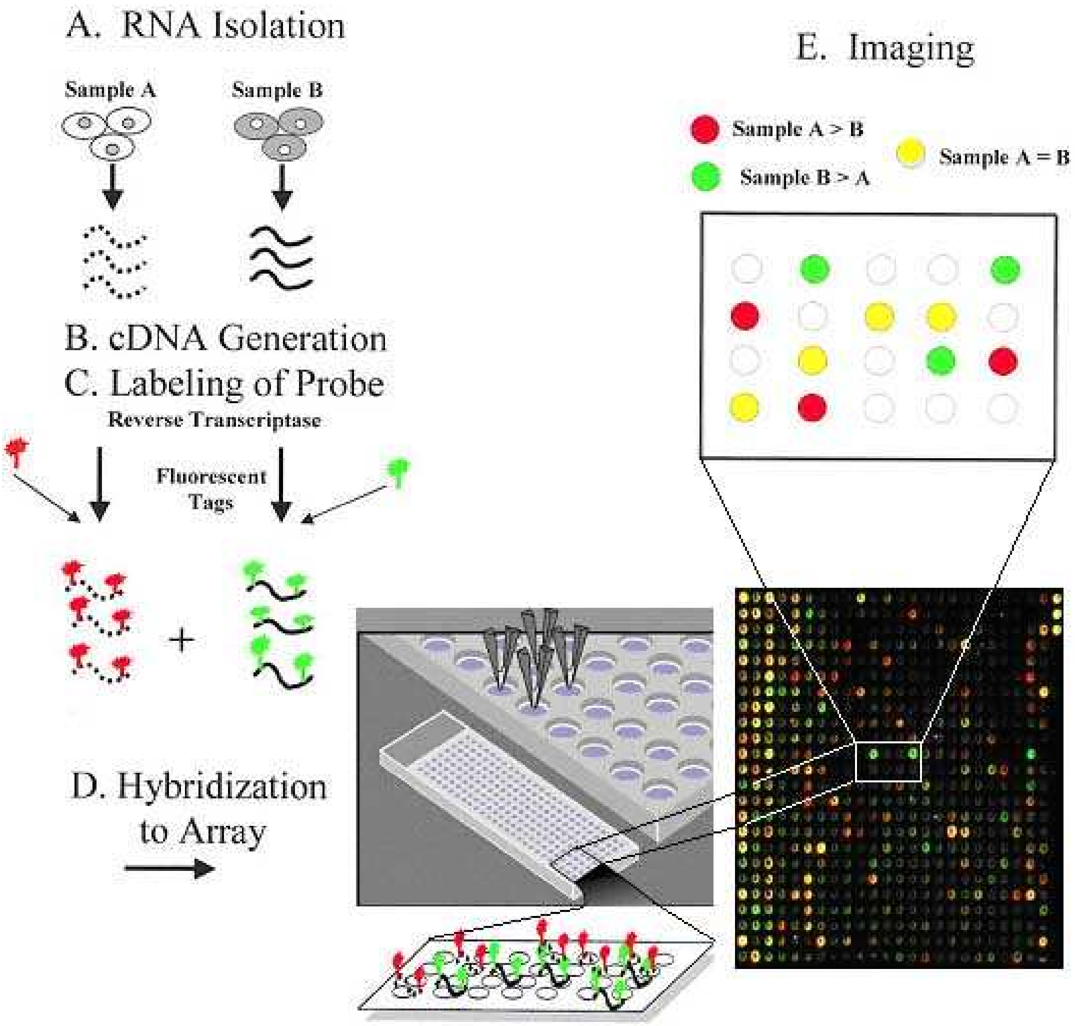
Two Main Types of Microarrays

- cDNA, or spotted arrays
  - mRNA from two biological samples are compared
- High-density oligonucleotide arrays
  - mRNA from one biological sample
- Their require different experimental designs
- and different types of analysis

Spotted Arrays

1. DNA (single-strand) fragments representing exons of genes of an organism are spotted on a high-density grid pattern on a glass slide
2. mRNA are obtained from two biological samples
3. cDNA (single-strand complement DNA) copies of the two mRNA samples are fluorescently labeled with two dyes, Cy3 and Cy5
4. A mixture of the two labled cDNAs is hybridized to the slide (complement sequences bind)
5. The slide is scanned to measure the amount of label over each spot
6. The ratio of the labels is used as an indicator of the ration of the mRNA levels in the biological samples

Spotted Arrays



http://userwww.sfsu.edu/ infoarts/links/isea2006bioartf/art.bio.isea.06web.html

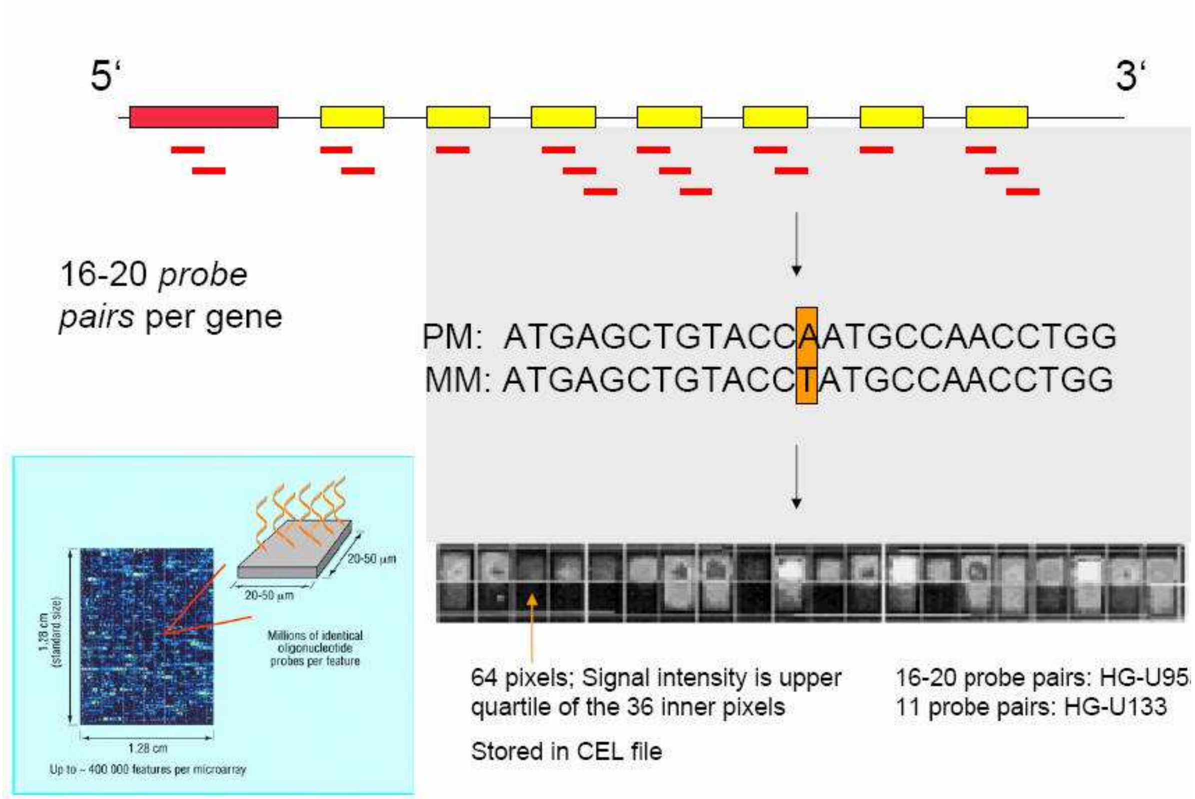
HD Oligonucleotide Arrays

For example the Affymetrix chip

1. For each gene a series (11-16) of short oligonucleotides (25 bp in length) are synthesized at very high densities onto a support medium
   - These sequences named perfect match (PM) probes match perfectly exon sequences at different positions of a gene
2. Another series of mismatching (MM) probes are produced that have the same sequence as the PM probes except for a missmatch base at the middle
3. A cDNA copy of a mRNA sample is produced and labeled with biotin and hybridized to the array

HD Oligonucleotide Arrays

- The amount of DNA at each position is measured as the amount of label affixed to the PM minus the amount of label affixed to MM


5'
3'
20-50 µm
20-50 µm
1.28 cm
(standard size)
Millions of identical oligonucleotide probes per feature
1.28 cm
Up to ~ 400 000 features per microarray

16-20 *probe pairs* per gene

PM: ATGAGCTGTACCAATGCCAACCTGG
MM: ATGAGCTGTACCTATGCCAACCTGG

64 pixels; Signal intensity is upper quartile of the 36 inner pixels

Stored in CEL file

16-20 probe pairs: HG-U95
11 probe pairs: HG-U133

Geeleher et al , Web-based Tools for the Analysis of DNA Microarrays

Design of Microarray Experiments

There are several sources of variation

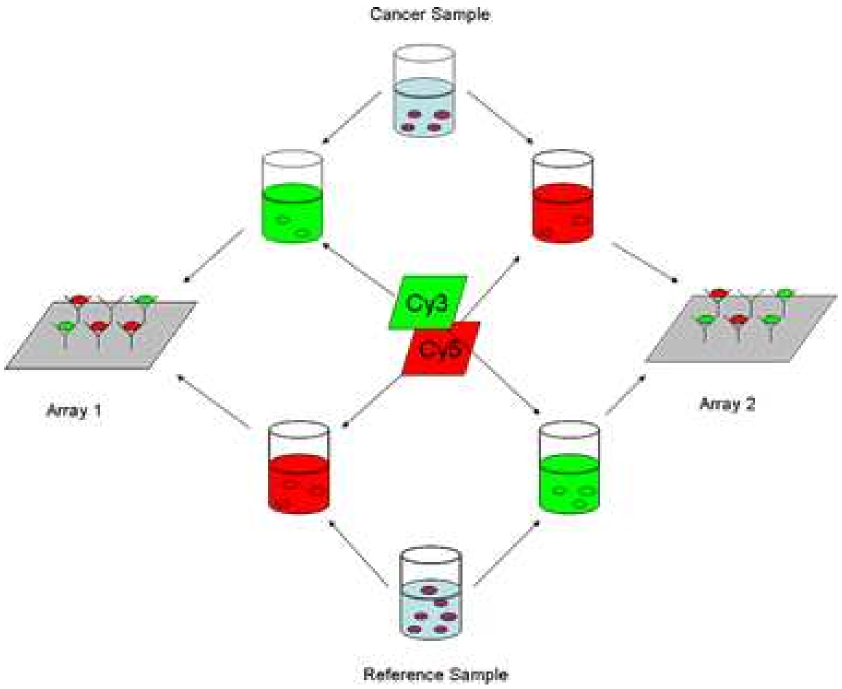
- Biological sample
  - The state of a cell is not constant
  - The effect of a treatment may vary
- Slide
  - Overall gene intensities may vary between slides
  - Differences may be due to variation in scanning
- Dye
  - Differences on effectiveness of labeling

Design of Microarray Experiments

- Biological replication
    - Measurement of different mRNA samples from the same cell or tissue
    - Captures biological variation of the mRNA levels in the cell or tissue
    - The use of multiple biological replicants reduces bias due to biological variability
- Technical replication
    - Preparation of multiple slides from the same biological sample
    - Addresses variation due to the technical measurement prosedure
    - Using multiple technical replicants reduces bias due to variability in the measurment

Design of Microarray Experiments

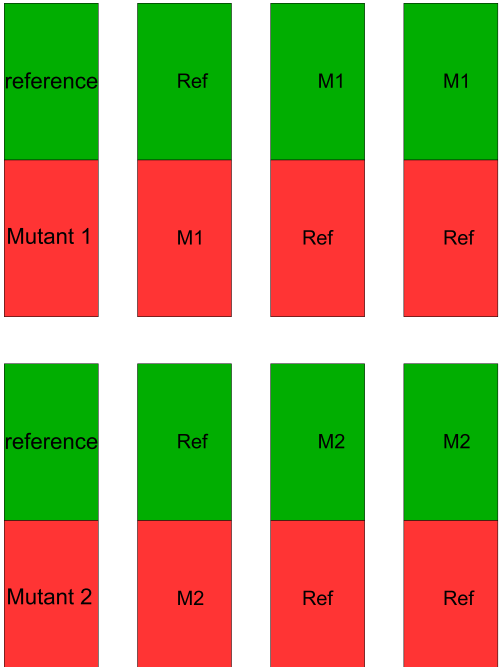
- Correcting for the dye effect in Spotted Microarrays



- The dye effect is a so called block factor
- The effect of block factors can be taken into account by assigning each block level to each treatment

Miller at al, Antibody microarray profiling of human prostate cancer sera: Antibody screening and identification of potential biomarkers. Proteomics, 2003

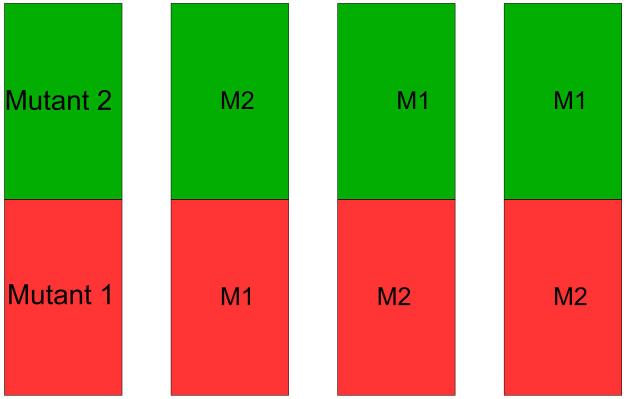
Design of Microarray Experiments



reference	Ref	M1	M1
Mutant 1	M1	Ref	Ref
reference	Ref	M2	M2
Mutant 2	M2	Ref	Ref

- **Reference design** for two treatments
  - Each treatment sample is compared to a common reference sample
  - This design includes a dye swap feature and two biological or technical replicates

Design of Microarray Experiments



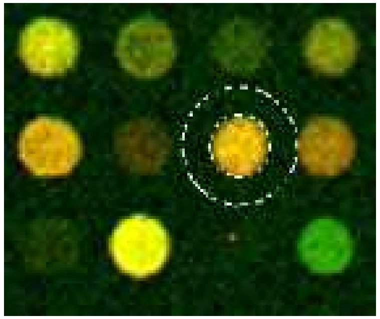
- **Loop design** for two treatments
  - Treatment samples are compared with each other (No reference sample needed)
  - This design includes a dye swap feature and two biological or technical replicates
  - As the number of treatments and blocking factors becomes large the reference design becomes more efficience than the loop design

Design of Microarray Experiments

- HD oligonucleotide arrays use single-dye techniques a have thus a different blocking structure
- Technical replication is needed to detect when the variation between samples (treatments) is greater than that between slides

Data Quality and Background Subtraction

- Data should be analysed for quality and poorer quality data should be removed
  - dirt on the slide or variation within the signal within one "spot" leads to unreliable points
- The signal for each array position is examined by breaking the point into pixels (8×8) and examing the variation among these pixels
- Background correction is done by extracting the background signal based on measurements made in the neighborhood of the DNA spot



Normalization of Microarray Data

- The need for normalization arises when there exsists uneven variation between dyes or slides
- For example, if the same biological sample is hybridized on two slides, the resulting expression values should be scattered along a straight line with a 45^o degree slope
- The same should also occur when comparing a treatment that should only effect a small proportion of the genes

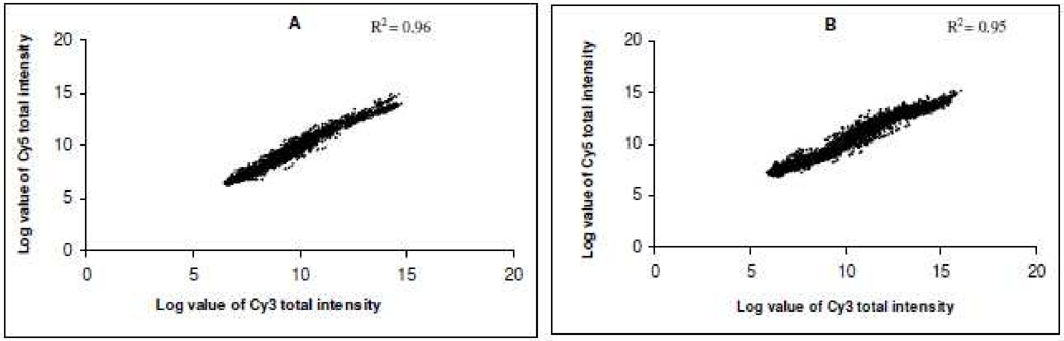


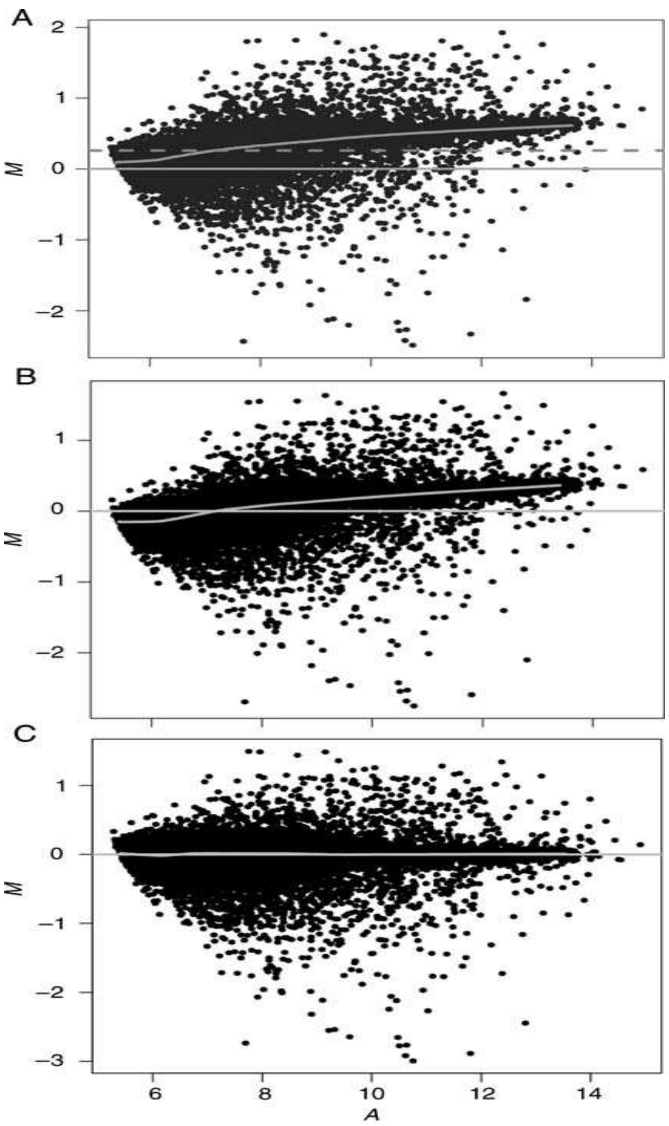
Figure 1
Scatter plot illustrating log value of Cy3, Cy5 total intensities for the biological and dye swap (technical) replicates. Each gene is represented by a point of Cy3 and Cy5 log values, A: represents target hybridizations B: represents dye swap hybridizations.

Ghanem et al. BMC Developmental Biology 2007 7:90

Normalization of Microarray Data

- The variation is commonly viewed by the M vs. A scatter plot (Next slide)
  - $A = 0.5(\log(G) + \log(R)$ is the average "dye" intensity between two samples
  - $M = \log(G/R) = log(G) - log(R)$ is the intensity ratio of two samples
  - If the intensity is equal, than $M = 0$
- Global normalization transfers the data so that the overall average of M is zero (see next slide)
- Local normalization transfers the data so that the average of M is zero locally also (see next slide)

Normalization of Microarray Data



Normalization of Microarray Data

- Assume intensity values for n genes and two dyes $(\log(G_i), \log(R_i))$, for $i = 1, \dots, n$
- In the following slides we will show how to estimate the local mean $\bar{M}_i$ of the values $M_i = \log(G_i) - \log(R_i)$,
- The local mean $\bar{M}_i$ is then subtracted from M_i
- Thus the normalized data is $log(G_i) - \bar{M}_i$ and $log(R_i)$

Normalization of Microarray Data

- Lowess normalization
  - A straight line is fitted at each point using data within a local window
  - Given data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ the local mean at point x_i is given $\bar{y}_i = a + bx_i$ where a and b minimize

    $$\sum_{i=1}^{n} K_\lambda(x - x_i)(y_i - a - bx_i)^2$$

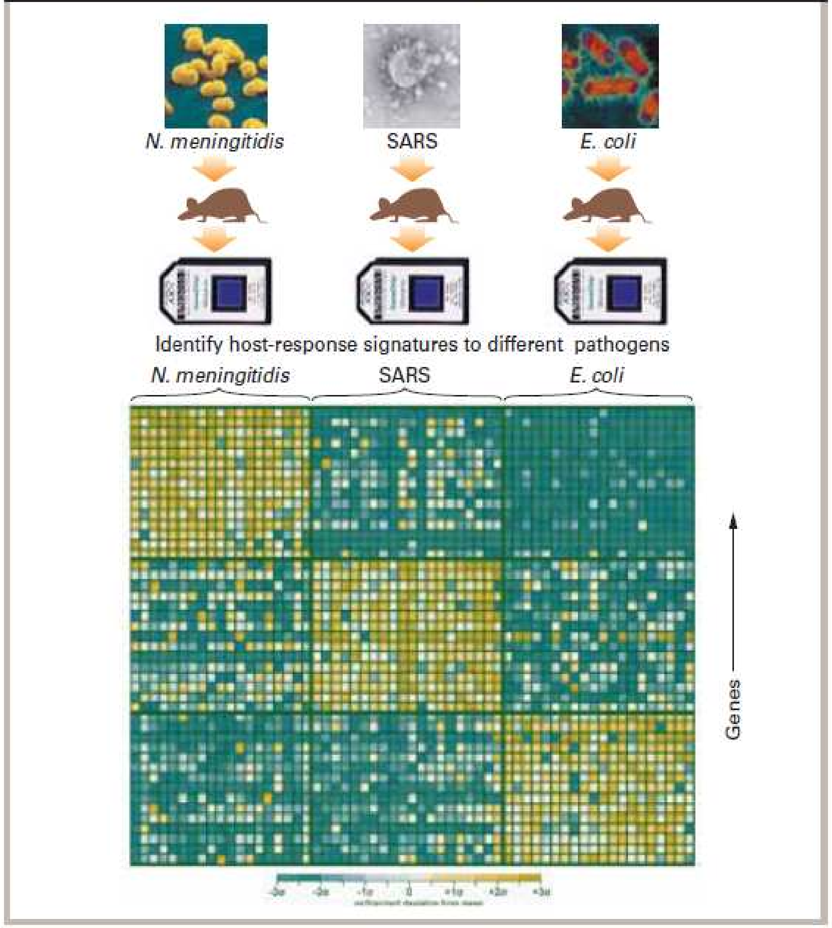
    where $K_\lambda(x - x_i)$ gives weight/contribution of measurement (x_i, y_i)
  - At each point a separate problem is solved

Normalization of Microarray Data

- Loess normalization
  - Similar to Lowess normalization except that a higher order polynomial is fitted at each point
- Quantile normalization
  - The intensity values in two samples (dyes or slides) are ordered
  - The intensity values of one slide are replaced with the intensity value of same rank order in the other sample (Master sample)
  - Now both samples will have the same distribution of intensity values
  - Easy too apply to multiple slides

Diagnostic Prediction

Figure 2: Host response signatures for pathogen identification. In this animal model, mice are infected with different pathogens and whole-genome expression is measured by microarray analysis, resulting in a specific host-response signature. This type of study can identify specific patterns for microarray diagnostic applications.



- *Affymetrix: Infectious Disease Application Note*

Testing for differential Expression

- Question:How to determ is the observed difference in the mean expression level of a gene between two groups significant
- Answer: Perform a statistical test (t-test)
- The t-statistics is

$$t_{1,2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

where $\bar{x}_i$ and s_i^2 are the mean and the variance of the gene in group i, respectively

Testing for differential Expression

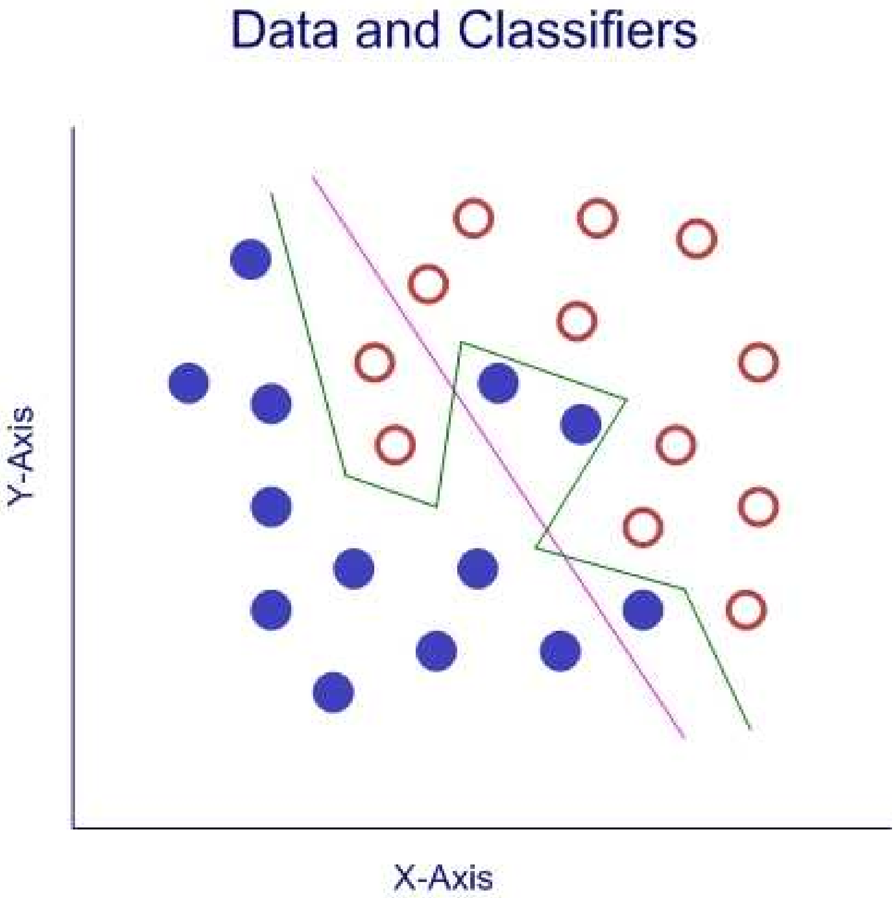
- The null hypothesis is that there is no difference in the means
- Calculate the p-value of the t-statistics for each gene
- Those genes whoms p-value is smaller than 0.05 are significantly different (null hypothesis rejected)
- When testing multiple genes the p-value must be recalculated
- One solution is the $Bonferroni$ correction ($=$ independency assumption)
- For p genes the corresponding significance is $0.05/p$

Feature Selection/Extraction

- If we want to predict the outcome (for example cancer type) we need to choose which of the thousands of genes to use
- Feature selection
  - Choose the genes with the highest t-statistics value
  - Choose genes with the highest variance
- Feature extraction
  - Choose combination of the original genes ”features”
  - For example the mean of a cluster of genes or several principle components (PCA)
- In all cases the dimension of the chosen feature space is much smaller than the number of genes

Prediction/Classification

- The chosen set of features (genes) can be used to construct a classifier — a function that maps the feature space into the possible classes (for example type of tumor)



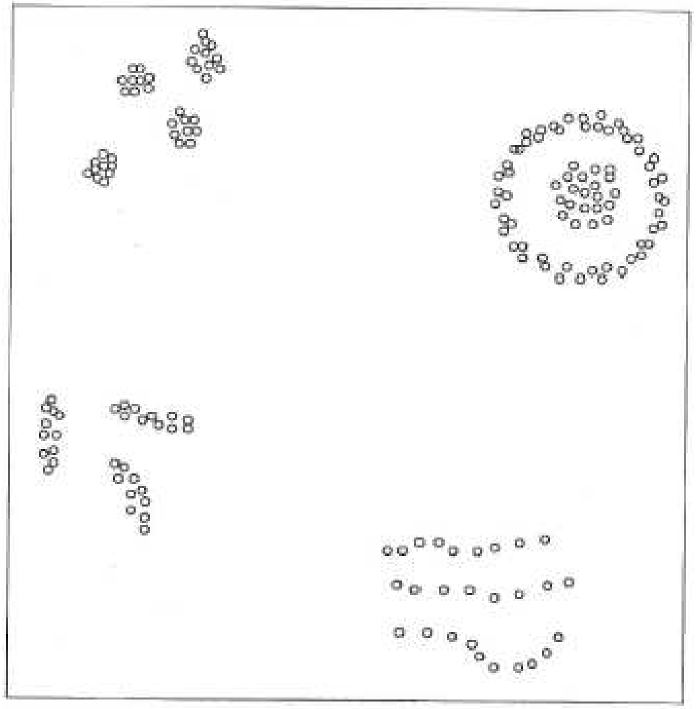
Prediction/Classification

- Linear discriminant analysis (LDA)
  - Assume normally distributed classes with equal covariance structure (estimated using teaching samples)
- Quadratic discriminant analysis (QDA)
  - Assume normally distributed classes with different covariance structure
- Nearest-neighbor classifier (NNC)
  - A new sample is classified to the class of the nearest ”teaching” sample

Clustering

- Genes that respond the same way to a given treatment may share a regulatory or functional relationship
- A method is needed to compare the intensity changes between genes and group together those genes that are responding in the same manner
- Clustering methods based on a distance measure can be applied

Clustering



- How is a cluster defined ?
- Are the ”nearest” samples in the same cluster?

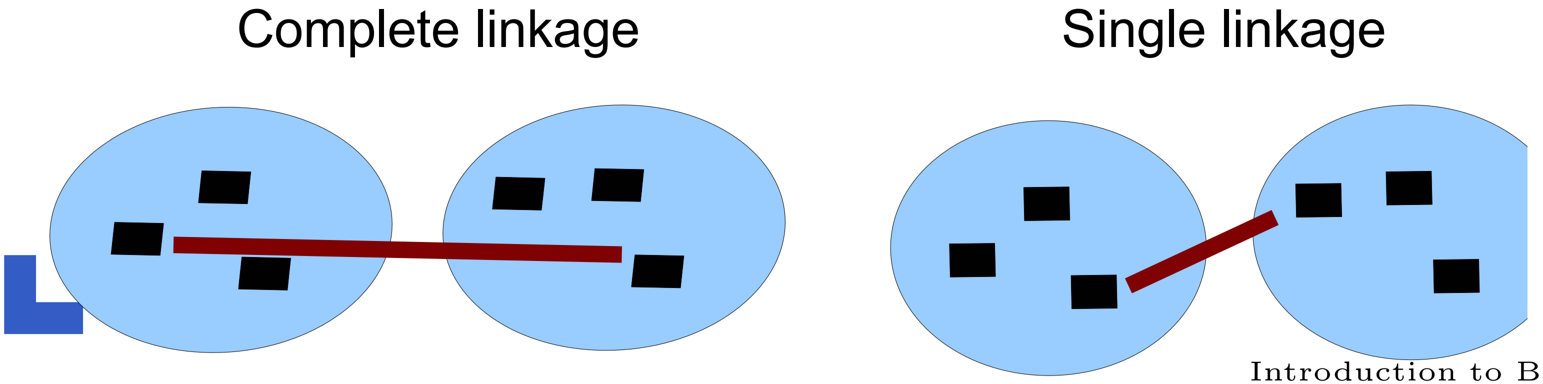
Clustering

Common clustering algorithms

- Hierarchical clustering (bottom-up approach)
  - All samples are initially in separate clusters
  - Eventually all samples are merged into one cluster
- K-means clustering
  - The number of clusters is predefined
  - The composition of the cluster vary during the clustering process
- Probalistic modeling based clustering
  - Samples are assigned to clusters according to probabilities

Hierarchical Clustering

- The most used clustering approach in microarrays
  1. A between genes distance measure (Euklidiean measure, Correlation coefficient, etc...)
  2. A between clusters distance measure
     - Complete linkage - the maximum pairwise distance between genes in two clusters
     - Single linkage- the minimum pairwise distance between genes in two clusters
     - Average linkage - the average pairwise distance between genes in two clusters



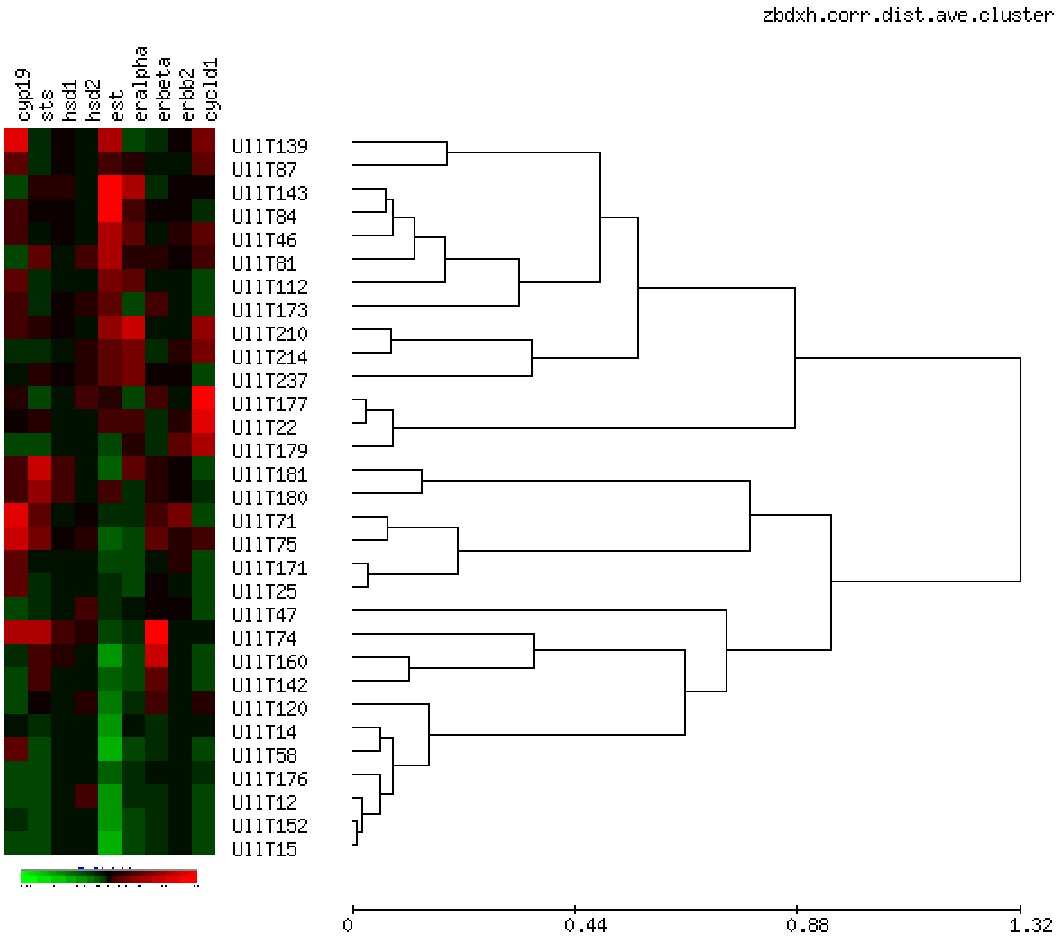
Complete linkage

Single linkage

Hierarchical Clustering

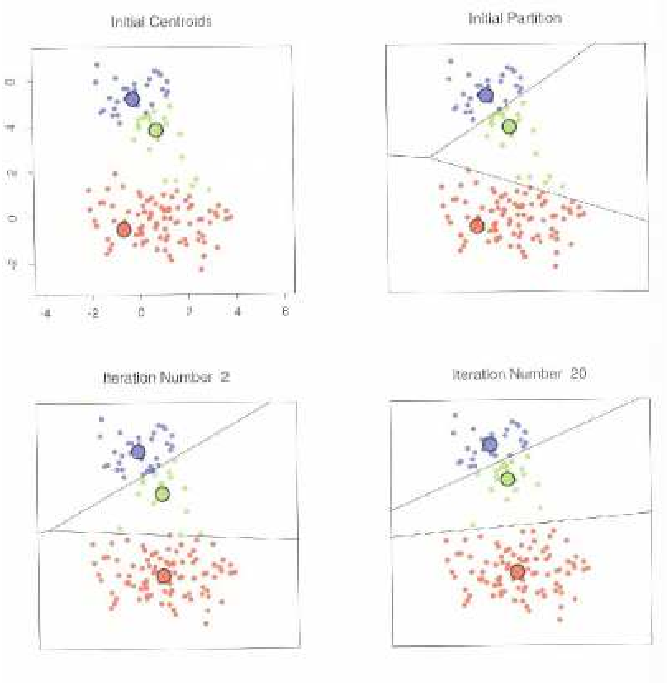
- After choosing the distance measures the algorithm is
  1. Initially each gene forms its own cluster
  2. The two nearest clusters are merged
- Step 2 is repeated until a predefined number of clusters is achieved
- A tree structure showing the merging path and distances is obtained
- Different number of clusters can be obtained by cutting the tree at different levels

Hierarchical Clustering



Yoshimura et al, Intratumoural mRNA expression of genes from the oestradiol metabolic pathway and clinical and histopathological parameters of breast cancer

K-mean Clustering



1. K initial centroids are chosen randomly
2. All samples are assigned to the nearest centroid (cluster)
3. New centroid positions are calculated as the average of the samples in the cluster

- Steps 2 and 3 are repeated for N rounds

Model Based Clustering

- We assume our data has been generated from a mixture of K-distributions

$$p(x) = \frac{1}{K} \sum_{j=1}^{k} p_j(x)$$

- We can use maximum likelihood methods to assign each sample to the one of the K distributions