CS 2001 Intro to research in CS

Machine Learning in Bioinformatics

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Who am I?

- **Milos Hauskrecht**
- **An assistant professor at CS department**
- Secondary affiliations: **ISP, CBMI and UPCI**
- **Research: AI**
  - Planning, optimization and learning in the presence of uncertainty
- **Applications:**
  - Biomedical informatics.
  - Modeling and control of large stochastic network systems.
  - Decision-making for patient-management.
  - Anomaly detection in clinical databases.

Bioinformatics

- **Bioinformatics**
  - **Application of CS and informatics to biological and clinical sciences**
- Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned
- **Machine Learning**
  - **An area that studies methods and the design of computer programs that let us learn from past experience**

Machine Learning in Bioinformatics

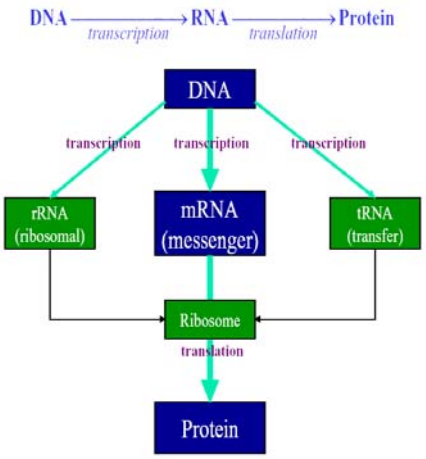
- **Why do we need machine learning in bioinformatics?**
  - A large volumes of data generated in medicine
    - patient records with clinical data
    - special studies with new high-throughput data sources:
      - DNA microarray data, Protemic profiles, multiplexed arrays
- **Use the data to:**
  - Learn a model that let us screen for the presence of a disease for new patients or predicts a good therapy choice
  - Identification of disease biomarkers

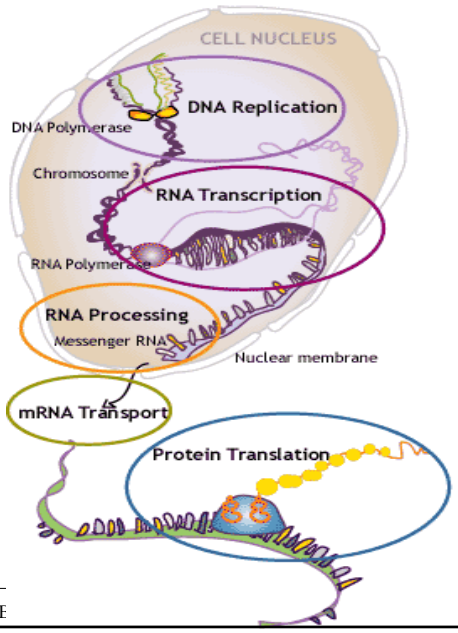
This talk

- **Overview of data sources**
  1. DNA microarray data
  2. SELDI-TOF-MS proteomic data profiles
  3. Luminex arrays

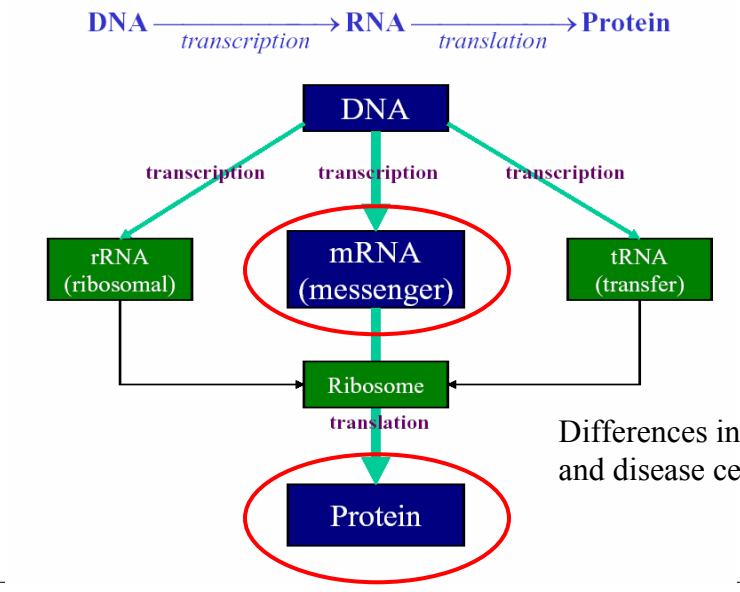
- **Basics about the methods of analysis**
  - Classification
  - Evaluation

Basics



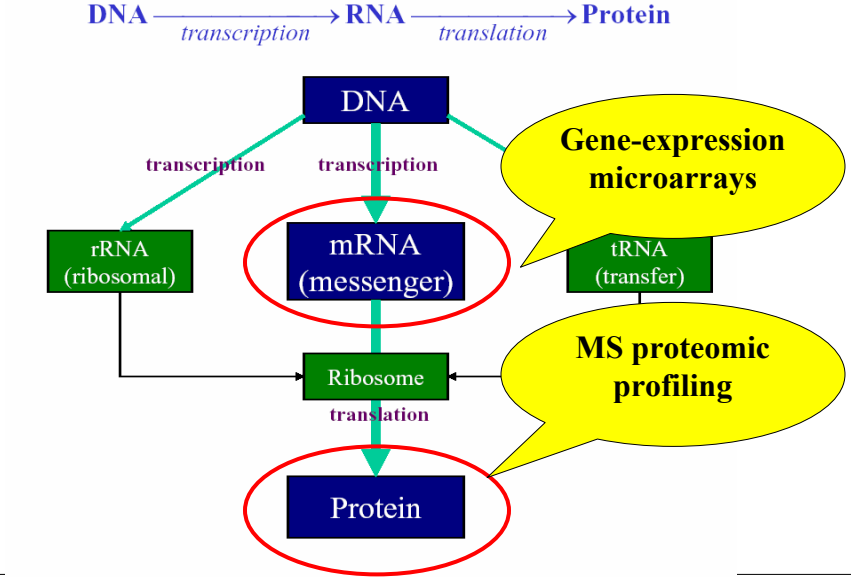


Measuring RNA and Protein levels



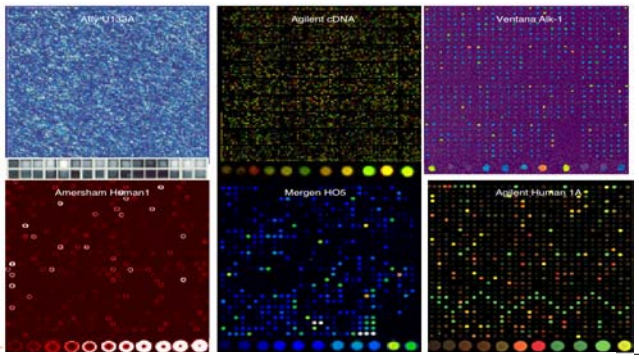
Differences in normal and disease cells

Measuring RNA and Protein levels



DNA microarrays

- Are small, solid supports onto which the sequences or subsequences from thousands of different genes are attached. The supports are usually **glass microscope slides**, or **silicon chips** or **nylon membranes**. The DNA is printed, spotted, or actually synthesized directly onto the support. The spots can be DNA, cDNA, or **oligonucleotides**

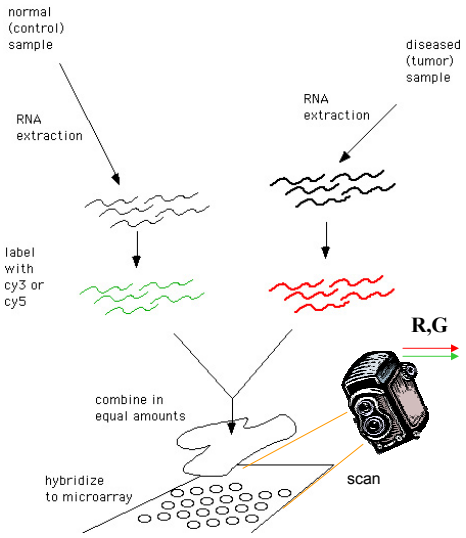


Extracting the information from microarrays

- **Hybridization probing**: a technique that uses fluorescently labeled nucleic acid molecules as "**mobile probes**" to identify **complementary molecules**, sequences that are able to base-pair with one another on the microarray chip.
- A single-stranded DNA fragment is made up of **four different nucleotides:**
  - **adenine (A), thymine (T), guanine (G), and cytosine (C),** that are linked end to end.
  - Pairs or complements: **Adenine (A) – Thymine (T) Guanine (G) -- Cytosine (C)**
- Two complementary sequences always find each other and lock together or hybridize:
  - immobilized target DNA on the DNA microarray hybridizes with mobile probe DNA, cDNA, or mRNA,

Extracting the information from microarrays

- Assume two types of cells:
  - Normal cells
  - Cancer cells
- Extracted mRNA for each type is labeled with a different florescent dye (Cy 3 or Cy 5)
- Combine them in equal amounts
- Hybridize them on the microarray
- Scan the array: the color tells if the one cell type is over-expressed as compared to the other type



Sources of variation

- **The signal we obtain from a microarray is not perfect**
  - If we take two microarrays for the same samples we see a lot of variation
- **Why?**
  - Technical artifacts: uneven spotting
  - Variation in RNA purity
  - Different labeling efficiencies of flourescently labeled nucleotides
  - Variations in expression level measurements (scan), …
- **Preprocessing:**
  - Attempts to remove unwanted sources of variation while preserving a useful information
  - Eliminate systematic biases

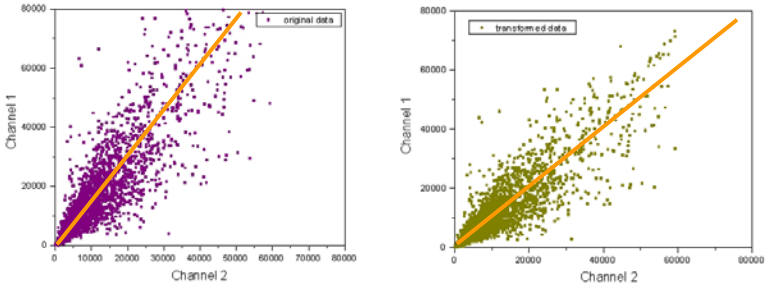
Preprocessing

- **Cy3 and Cy5 are incorporated into DNA with different efficiencies**
  - Genes that are expressed at comparable levels come with different than 1:1 dye mixing ratio
  - This creates a background signal
- **Example solution methods:**
  - **Global normalization:** assure that expression level for the array is 1 on average.
  - **Housekeeping genes:** genes that are known to be stable across different conditions (e.g. beta-actin). Correct the signal using the mean of expressions at housekeeping genes.

Preprocessing

Example:

- correct the slope so that 1:1 on average



Preprocessing

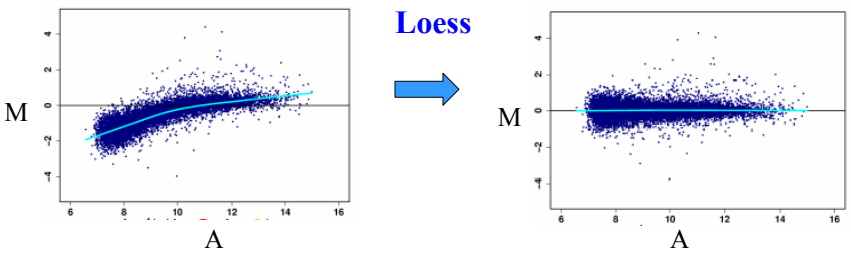
Example:

- Loess normalization (locally weighted polynomial regression)

Traditional analysis relies on log-ratios: **M=log(R/G)**

and averages: A=1/2 log(RG) as the primary data

- M makes gene expression measurements relative.

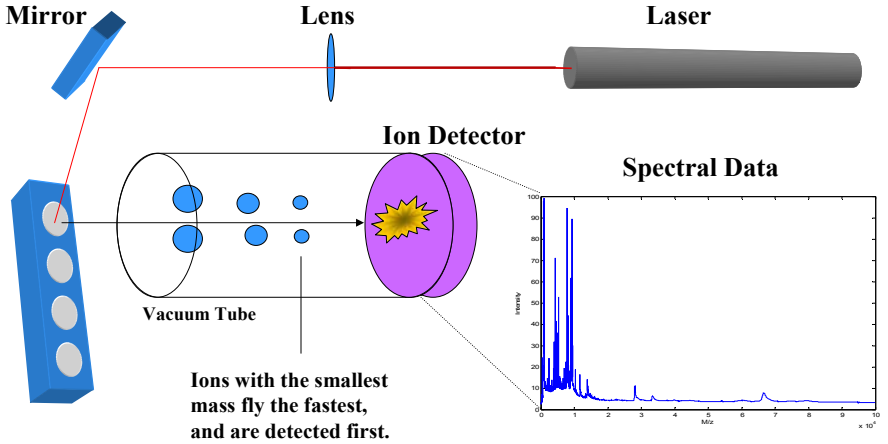


Mass Spectrometry proteomic profiling

- **Surface Enhanced Laser Desorption/Ionization Time of Flight Mass Spectrometry** (SELDI-TOF-MS)
- A technology recently developed by Ciphergen Biosystems
- A mass profile of a sample (serum, urine, cell lysates) in the range of 0-200K Daltons
- A profile is believed to provide a rich source of information about a patient. Potentially useful for:
- **Detection of a disease**
  - Many promising results on various types of cancer
- **Discovery of disease biomarkers**
  - Identify species (protein/peptides) responsible for the differences

SELDI-TOF MS

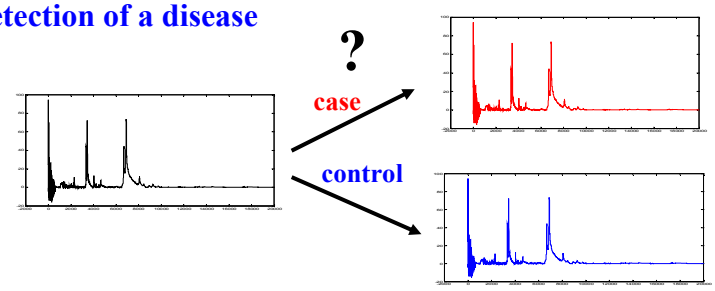
Surface Enhanced Laser Desorption/Ionization Time of Flight Mass Spectrometry



Importance of SELDI-TOF MS data

SELDI profiles are believed to be useful for:

- **Detection of a disease**



  - Promising results on various types of cancer
- **Discovery of disease biomarkers**
  - Identify species (protein/peptides) responsible for the differences

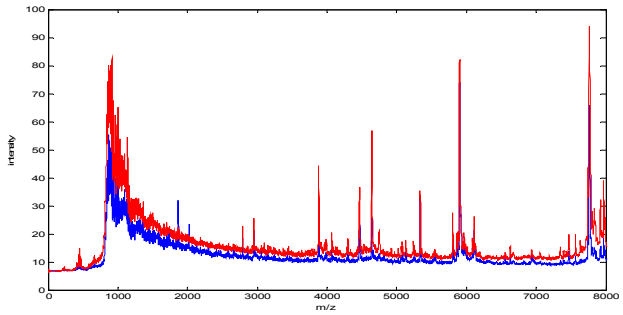
Challenges in MS data analysis

Many sources of variation

- Sample collection, sample storage & sample processing
- Instrument conditions

+ natural variation in protein expression levels among individuals

Example: profiles of two patients with pancreatic cancer

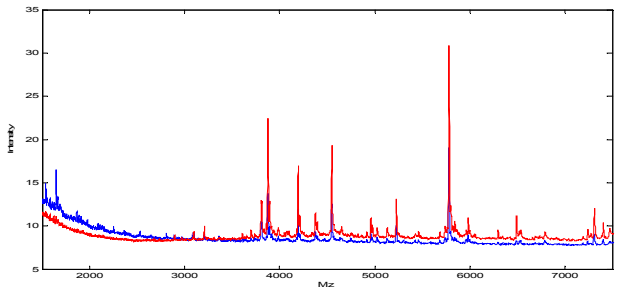


Challenges in data analysis

Three types of systematic instrument errors

- Baseline shift
- Intensity measurement error
- Mass inaccuracy

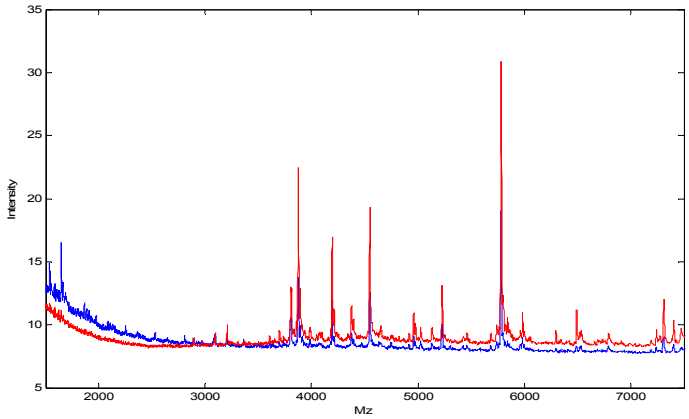
Example: two profiles for the same reference serum



Baseline Shift

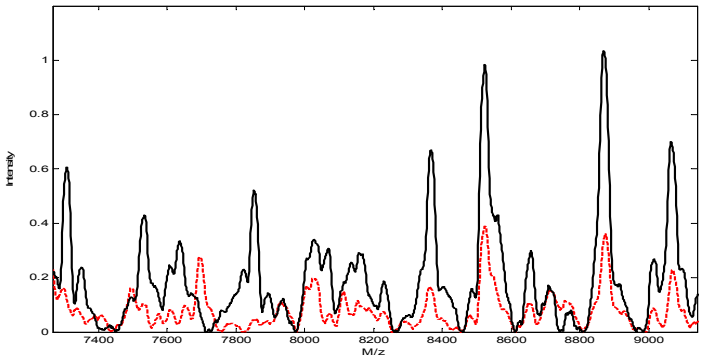
- A systematic error where intensity measurements differ from 0
- **Example:**

  2 profiles of the same pooled reference (QA/QC) serum



Intensity measurements errors

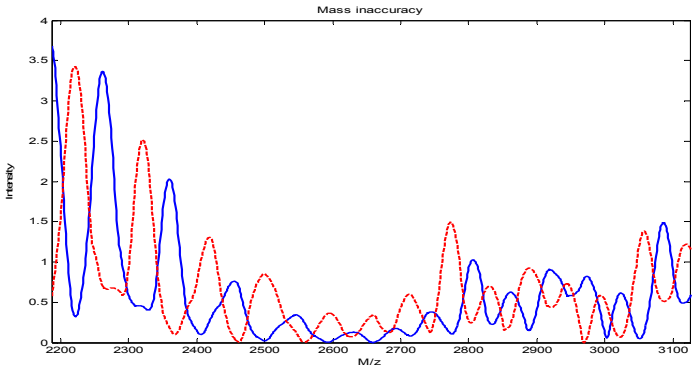
- Intensity measurements fluctuate randomly
- **Example:** 2 profiles of the same pooled reference (QA/QC) serum



- **A** higher intensity values come with a higher variance

Mass inaccuracy

- Misalignment of readings for different *m/z* values.
- **Example:**
  - 2 profiles of the same pooled reference (QA/QC) serum
  - A clear phase shift

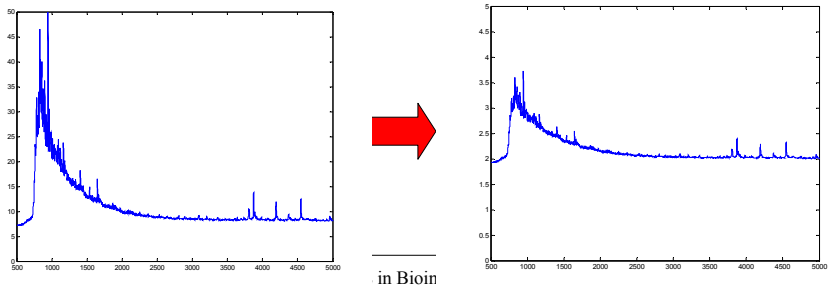


Profile preprocessing

- **Aims is to remove the noise and systematic biases in the signal while preserving useful information**
- Typical preprocessing steps:
  - **Smoothing:**
    - Aims to eliminate the noise in the signal
  - **Calibration/rescaling :**
    - Attempts to eliminate differences in intensity among profiles
  - **Profile transformations (variance reduction)**
    - Reduces the multiplicative noise
  - **Baseline correction:**
    - Reduces the systematic baseline error
  - **Profile alignment:**
    - Correct for mass inaccuracy

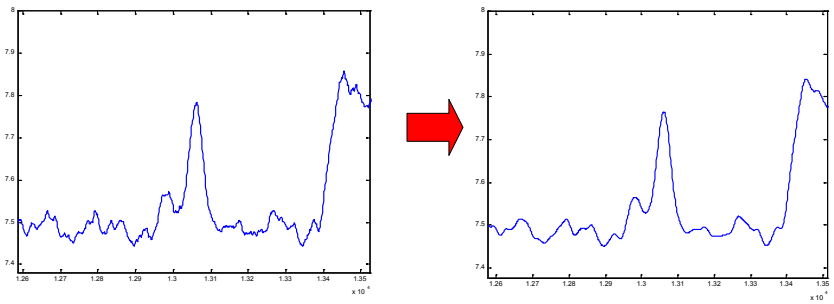
Preprocessing: variance stabilization

- **Aims is to eliminate** the dependency of the noise on the intensity of the signal;
- **Example transforms used in variance stabilization:**
  - Log transform
  - Square-root transform
  - Cube-root transform - appears to be the best



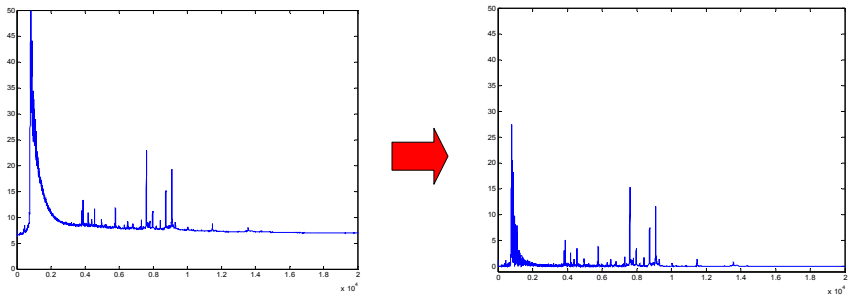
Preprocessing: Smoothing

- **Aims is to** eliminate a high-frequency noise in the signal
- **Threat:** a loss of information in the high frequency signal



Preprocessing: Baseline correction

- **Aims is to** eliminate a systematic intensity bias by which the profile readings differ from 0



Disease detection

How to use the high throughput information about gene expression levels or protein (peptide) expressions?

- **Detection of a disease**
  - Many promising results on various types of cancer
- **Discovery of disease biomarkers**
  - Identify species (protein/peptides) responsible for the differences

- Here: the problem of building **a classification model** that is capable to determine with a high accuracy the presence or absence of the disease **in future patients**

Disease detection

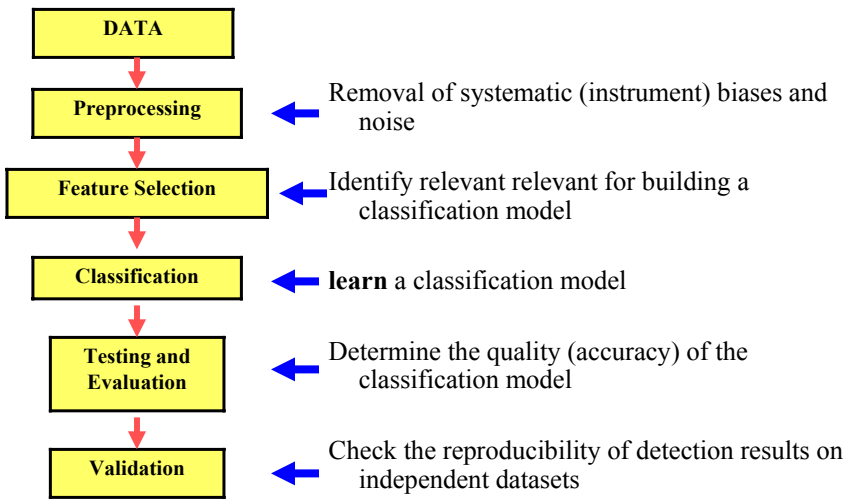
Objective:

- Build **a classification model** that is capable to determine with a high accuracy the presence or absence of the disease **in future patients**

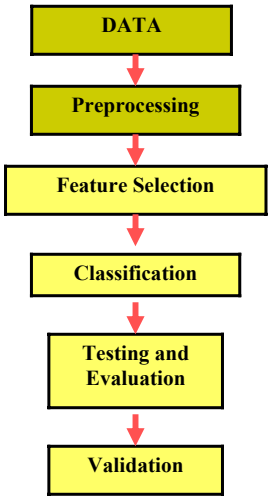
Typical steps:

- **Data preprocessing:**
- **Feature selection**
- **Building of a classification model**
- **Evaluation**
- **Validation**

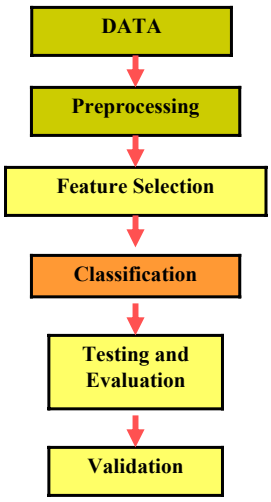
Analysis of expression profiles



Analysis of expression profiles



Analysis of expression profiles



Detection of a disease

Objective:

- Find a classification model that assigns with a high accuracy correct disease/no-disease labels to profiles generated for individual patients

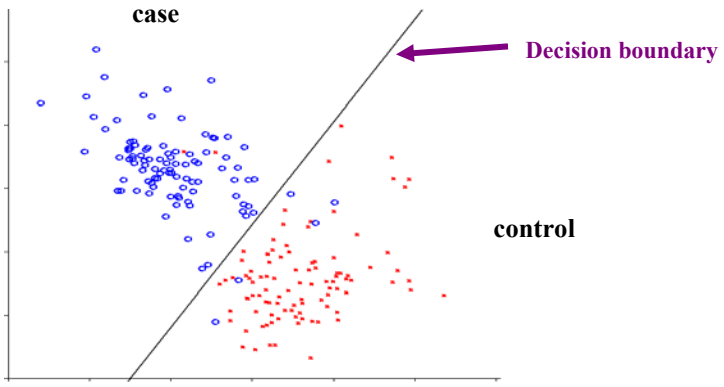
Profile → **Classifier model** → **Class label (case or control)**

The approach:

- **use past data to learn a classification model**
- **variety of machine learning approaches exist:**
  - Logistic regression
  - Linear and quadratic discriminant analysis
  - Classification and regression trees (CART)

Learning a classification model

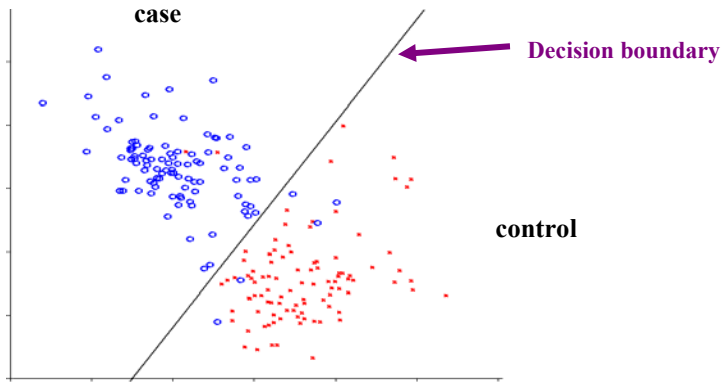
- **Example: a dataset with 2 features (e.g. expression levels of two genes)**



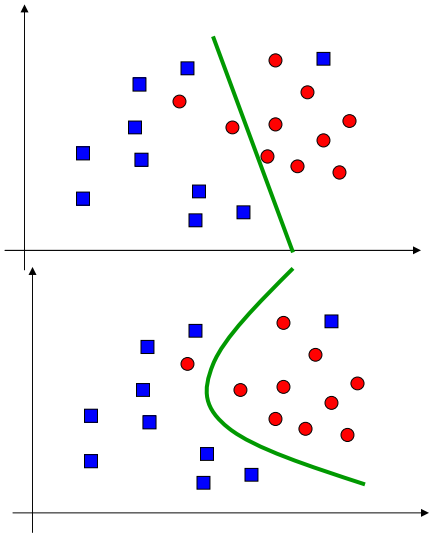
We learn **the decision rule** that achieves the best separation between case and control samples

Learning a classification model

- **Logistic regression and linear SVM methods give the linear decision boundary**



Non-linear boundaries

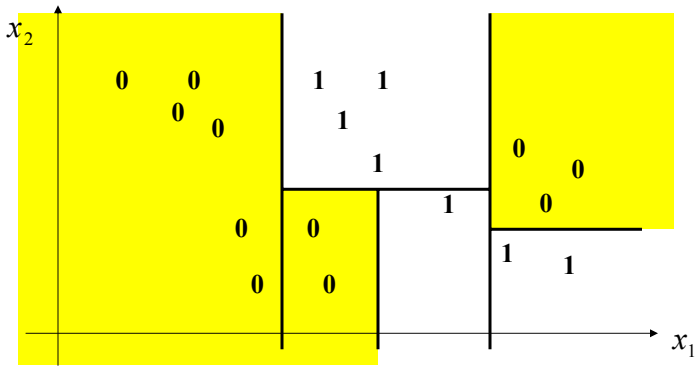


Linear separator
in the input space

Non-linear separator
in the input space

Decision trees

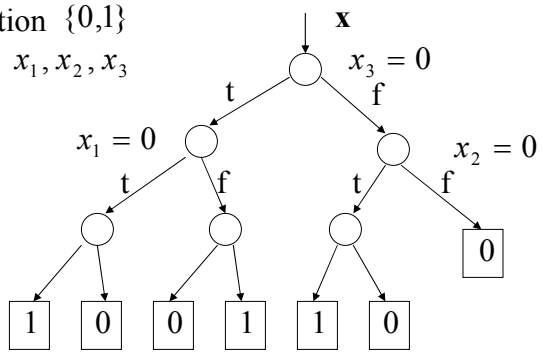
- An alternative approach to what we have seen so far:
  - **Partition the input space to regions**
  - **classify independently in every region**



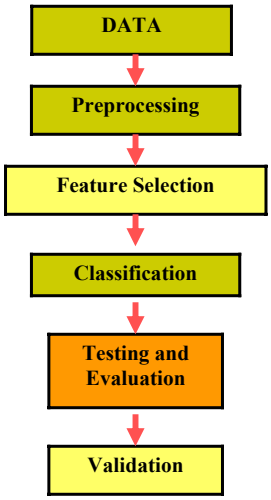
Decision trees

- The partitioning idea is used in the **decision tree model**:
  - Split the space recursively according to inputs in **x**
  - Regress or classify at the bottom of the tree

Example:
Binary classification {0,1}
Binary attributes x_1, x_2, x_3

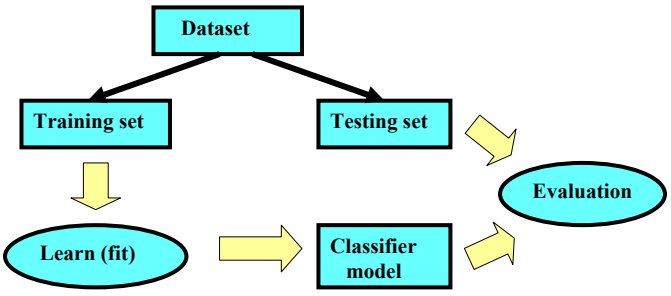


Analysis of expression profiles

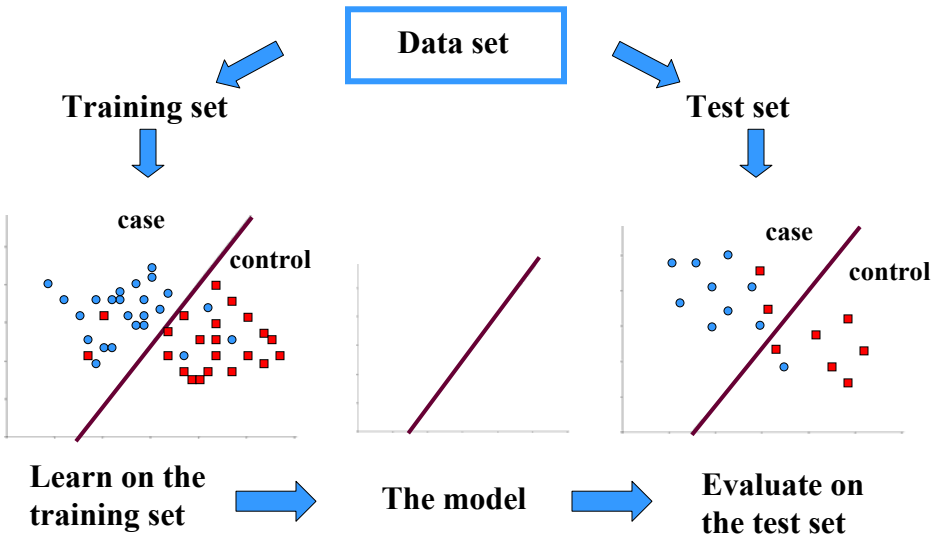


Evaluation framework

- We want our classifier to generalize well to future examples
- **Problem:** But we do not know future examples !!!
- **Solution:** evaluate the classifier on **the test set** that is withheld from the learning stage



Evaluation framework



Evaluation metrics

Confusion matrix: Records the percentages of examples in the testing set that fall into each group

		Actual: Case	Actual: Control
Prediction	Case	TP 0.3	FP 0.1
	Control	FN 0.2	TN 0.4

Misclassification error:

$$E = FP + FN$$

Sensitivity:

$$SN = \frac{TP}{TP + FN}$$

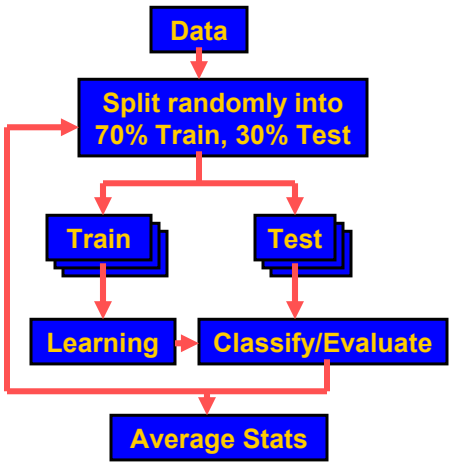
Specificity:

$$SP = \frac{TN}{TN + FP}$$

Evaluation

- To limit the effect of one lucky or unlucky train/test split PDAP supports averaging through:
  - **bootstrap**
  - **random sub-sampling**
  - **k-fold cross-validation**

Example:
- random subsampling



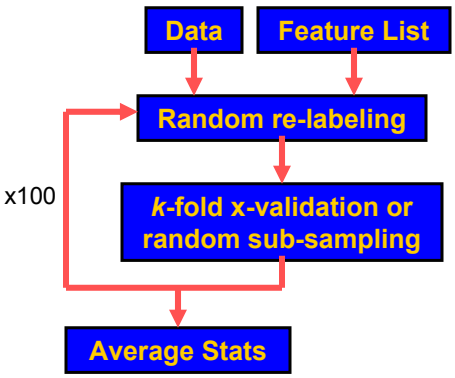
Validation of the predictive signal via PACE

- Additional assurance on the existing model: Is the signal real? Does it differ significantly from a random process?

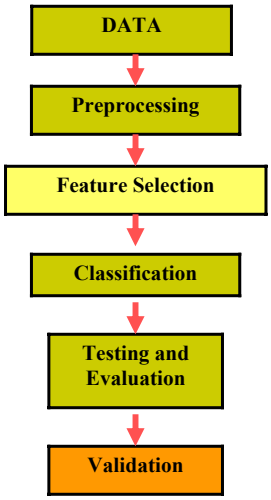
Permutation test

Use **a random re-labeling** of profiles' classes and evaluate the classifiers learned using such data

Idea: Reject the null hypothesis that the achieved classification error (ACE) on the true data is obtained purely by chance

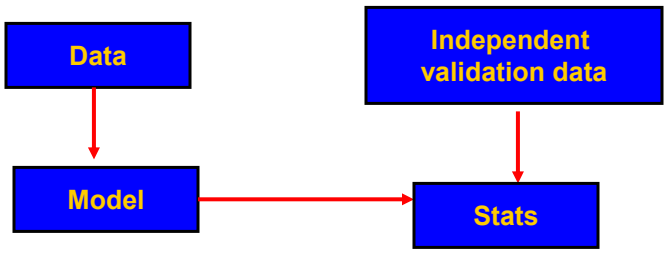


Analysis of expression profiles

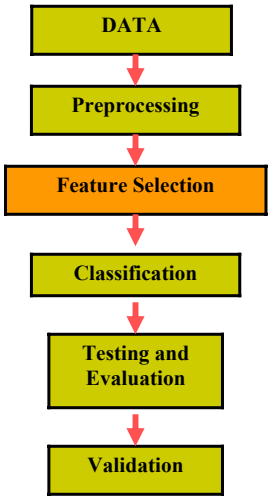


Validation

- **Keep an independent test validation set aside**
  - Taken out from the original data at the beginning (the researcher holds the decoder ring)
  - New data from the follow-up study
  - New data from different site



Analysis of expression profiles



Dimensionality problem

One important thing before we can learn a classifier

- **The dimension of the profile for proteomic data is very large** (~60,000 M/z measurements)

- **A large number of model parameters** must be to learned (estimated)

- **The total number of cases in the study is relatively small** (typically < 100)

- Leads to a **large variance of parameter estimates** (overfit)

ML dimensionality reduction

Goal: identify a small subset of features that achieve a good discriminatory performance

Solution: apply ML dimensionality reduction methods and their combinations

- **Feature selection**
  - Filter methods
  - Wrapper methods
  - Regularized classifiers
- **Feature aggregation:**
  - PCA
  - ICA
  - Clustering
- **Heuristics**

Feature filtering

Solution: identify a small set of features that are good individual discriminators

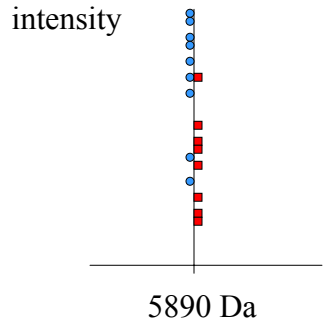
Univariate scores in PDAP :

- Fisher score
- t-test score
- Wilcoxon rank-sum score
- etc.

- **Univariate Scores:**
  - can be ordered relative to each other
  - Are often related to some statistic (p-value)
- **Features picked using a variety of criteria:**
  - Best k features
  - p-value or False discovery rate (FDR) thresholds

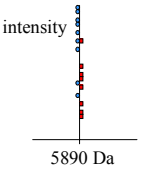
Differentially expressed features

- Is a specific profile position a good individual discriminator between case and control?



- **How to measure a separation?**

Differentially expressed features



- Example: **Measures based on the t-test**
- Used in statistics in hypothesis testing:
- **H0 - Null hypothesis:**
  - the mean of one group (case) **is not different** from the mean of the other group (controls) $(\mu_- = \mu_+)$
- H1 alternative hypothesis:

$$(\mu_- \neq \mu_+)$$

- A t-test (see if we can reject a null hypothesis):
- Relies on the *t* statistic, that represents a normalized distance measurement between populations

$$t_i = (\mu_- - \mu_+) \Big/ \sqrt{\frac{\sigma_-^2}{n_-} + \frac{\sigma_+^2}{n_+}}$$

Follows Student distribution

- We can compute a p-value (the significance for the null reject)

Differentially expressed features

Solution: Pick only features that are differentially are good individual discriminators for the case versus control

Scores based on univariate statistical analysis :

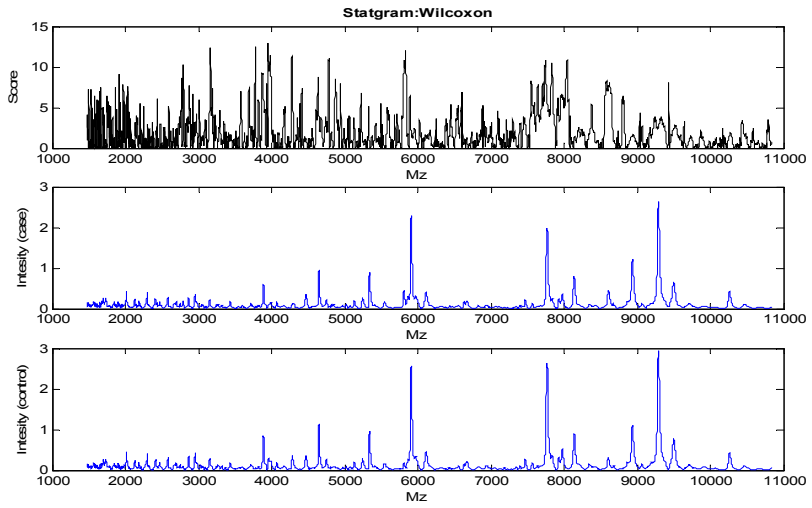
- Fisher score and its variants
- AUC (Area Under ROC Curve)
- T-test score (Long & Baldi)
- Wilcoxon rank-sum score
- etc.

- Additional criteria may control **significance levels** and **False discovery rate** (Benjamini & Hochberg)

Differentially expressed features

- **A score based on Wilcoxon rank-sum test**

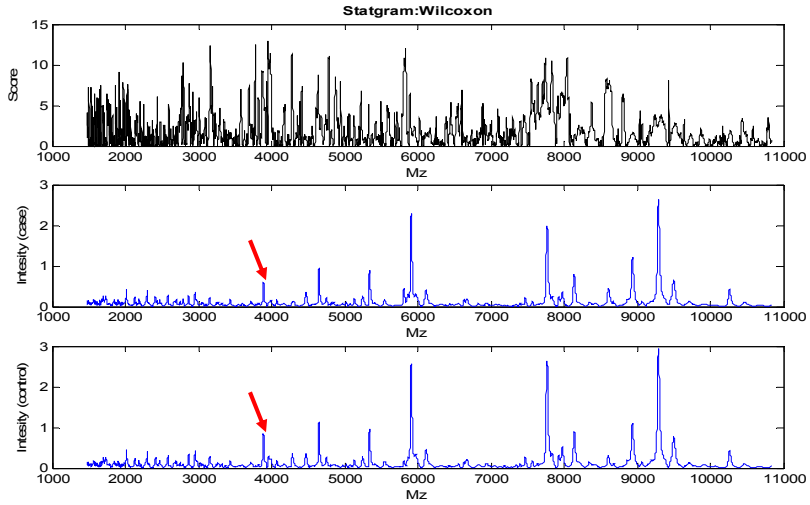
Pancreatic cancer data



Differentially expressed features

- **A score based on Wilcoxon rank-sum test**

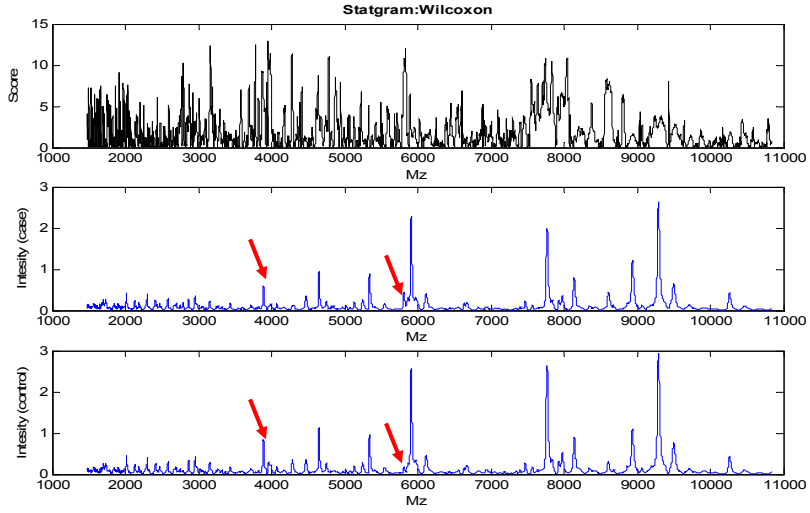
Pancreatic cancer data



Differentially expressed features

- **A score based on Wilcoxon rank-sum test**

Pancreatic cancer data



Multivariate feature selection

Univariate methods

- each feature (profile position) is considered individually
- May not be sufficient to capture differences in disease/control variability

Multivariate methods

- a biomarker panel (a more than one feature) is needed to define the differences in between case/control
- Subsets of features need to be considered

Multivariate feature selection

Example of a multivariate method:

- **Wrapper methods**
- uses a classification model that is tried with different subsets of features

- **Idea:**
  - Learn a classifier on subset of samples in the training set
  - Obtain a measure of accuracy on the remaining train samples (internal test set)
  - Repeat 1 and 2 multiple times using cross-validation schemes
  - Pick sets of features leading to the best average cross-validation measure

Wrapper methods

The problem with wrapper methods

- Combinatorial optimization
- To find the best combination of k features they need to evaluate $\binom{n}{k}$ configurations
- Recall: n is huge

Solutions:

- **Greedy wrappers:** pick features one by one greedily
- **Apply heuristics to make the search more efficient**

Conclusions

- **High-throughput data sources measure the expression of genes, proteins, and their fragments**
  - Potentially useful information for disease detection
  - Biomarker (biomarker panel) discovery
- **Problems: Many sources of variability**
  - Steps need to be taken before the profiles can be analyzed
  - Data collection, storage and sample processing confounding and reproducibility is an issue
- **Multivariate analysis:**
  - More than one feature (gene, protein expression level) used to detect the disease
- **Encouraging:** good results on many diseases