Machine learning in high-throughput genomics and proteomics

(Part I)

Xuegong Zhang, Ph.D.
Professor of Pattern Recognition and Bioinformatics
Tsinghua University
zhangxg@tsinghua.edu.cn

Outline

- Background
  - Microarray and mass-spectrometry technologies
  - Typical bioinformatics problems
  - Basic concepts of machine learning
- Microarray data mining with learning machines
  - Classification, clustering and gene selection
  - Problems, solutions, and possible pitfalls
- Further topics
  - Important considerations and open problems
  - Systems biology study after microarray data mining
  - New types of arrays

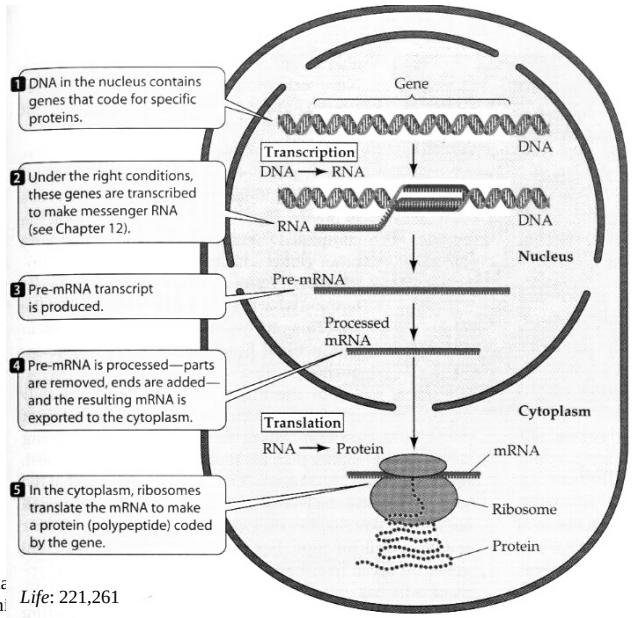
Target Audience:

- Researchers/students in information science, math or statistics who are interested in biological applications and bioinformatics
- Biologists who want to learn more about the underlying concepts behind the widely-used bioinformatics tools

Part I
Biological Backgrounds

- Microarray and mass spectrometry: technologies and data
- Typical bioinformatics topics
- Basic concepts of Machine Learning

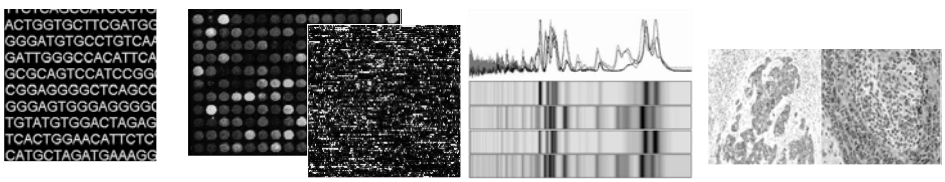
The Central Dogma



Life: 221,261

The Central Dogma

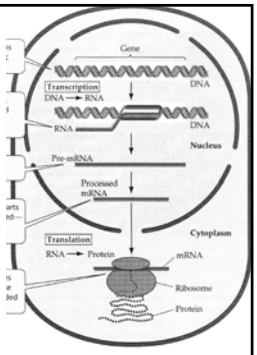
DNA → mRNA → Protein → Phenotype



DNA Microarrays

Gene Expression and Microarrays

Gene Expression



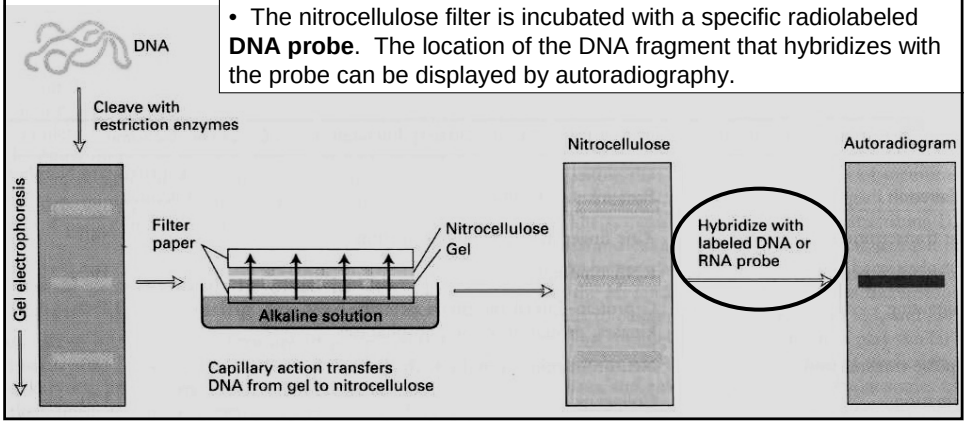
- Gene → Primary Transcript → nuc. mRNA → cytosolic mRNA → protein → protein activity
- Basic cellular processes are realized by tightly regulated gene expression programs
- Different cell types in a multicellular organism express different sets of genes at different time and with different quantities

Measuring mRNA abundance in cytosol *(gene expression levels)*

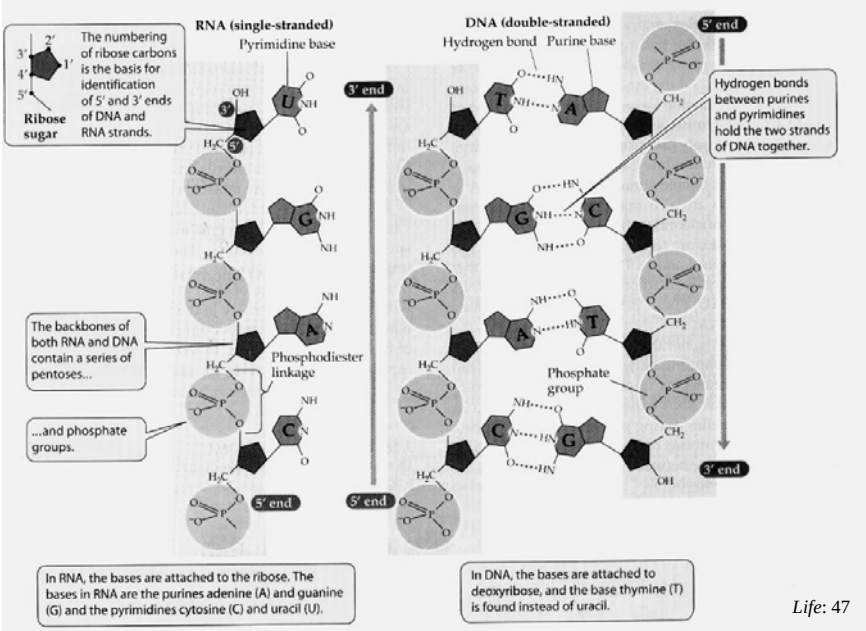
- Northern blotting
  - The traditional method
  - **Considered as the gold standard**
  - Gene by gene
  - Very time consuming
- Microarray

Southern Blotting [E.M. Southern, 1975, JMB, 98:508] for detecting the presence of specific DNA sequences following gel electrophoresis of a complex mixture of restriction fragments.

- The DNA to be analyzed is digested with restriction enzymes and then separated by agarose gel electrophoresis.
- The DNA fragments in the gel are denatured with alkaline solution and transferred onto a nitrocellulose filter or nylon membrane by blotting, preserving the distribution of the DNA fragments in the gel.
- The nitrocellulose filter is incubated with a specific radiolabeled **DNA probe**. The location of the DNA fragment that hybridizes with the probe can be displayed by autoradiography.



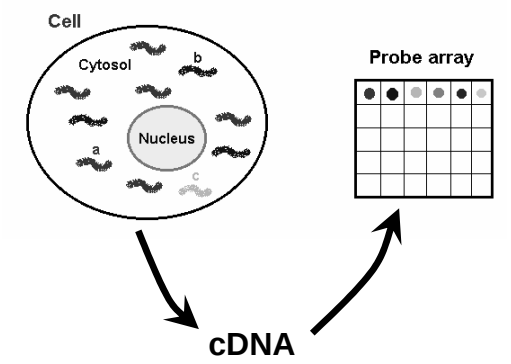
The Key Principle: Complementary Base-Pairing



Life: 47

- Northern Blotting
  - Detecting RNA fragments, instead of DNA fragments
  - RNA fragments are treated with formaldehyde to ensure linear conformation
  - The amount of a specific RNA in a sample can be estimated from a Northern blot
- Western Blotting
  - Detecting a particular protein in a mixture
  - The probe used is antibodies
  - Also called *immunoblotting*

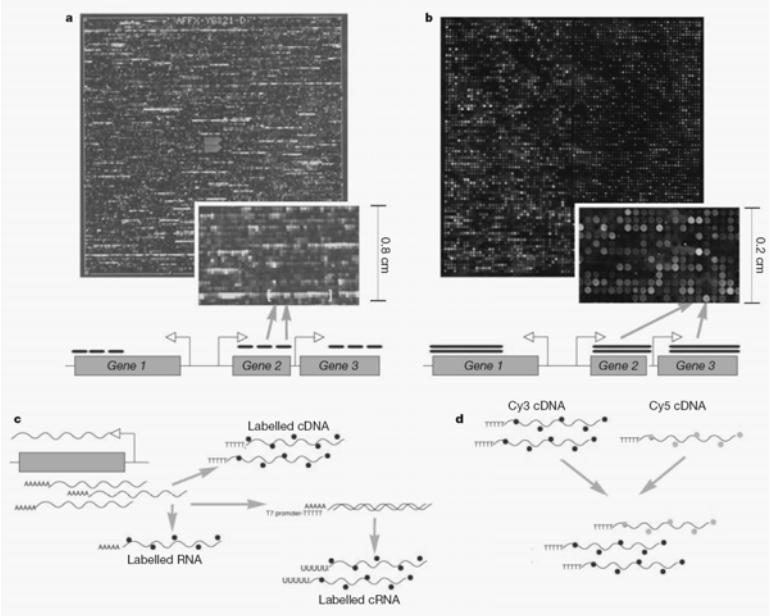
Going high-throughput: the basic idea of microarrays



Expression Levels:

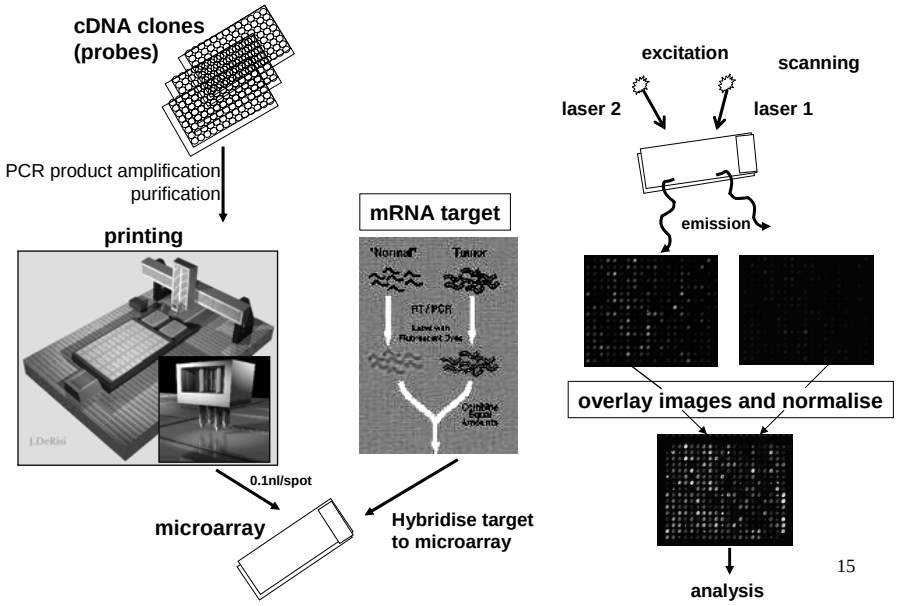
gene a	**high**
gene b	**medium**
gene c	**low**

Two ways of making microarrays



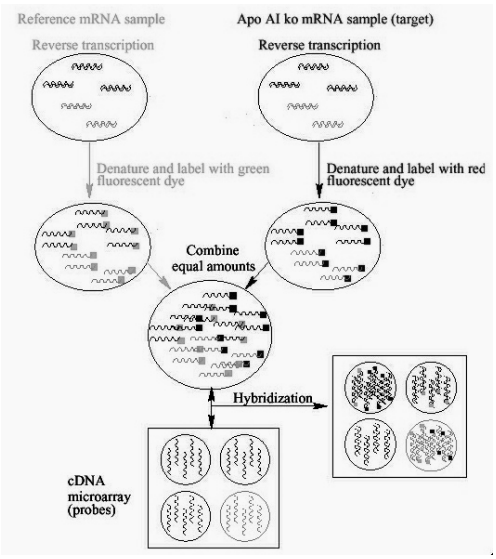
From D.J.Lockhart & E.A. Winzeler, Genomics, gene expression and DNA arrays, *Nature*, 405(15): 827-836, 2000

Printed cDNA microarrays



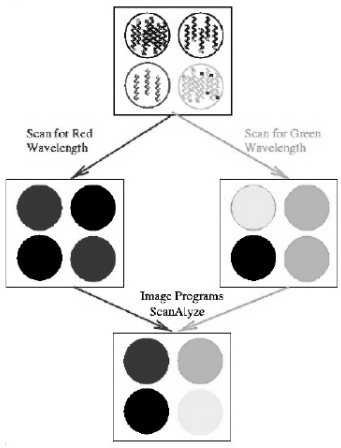
15

Printed cDNA microarrays

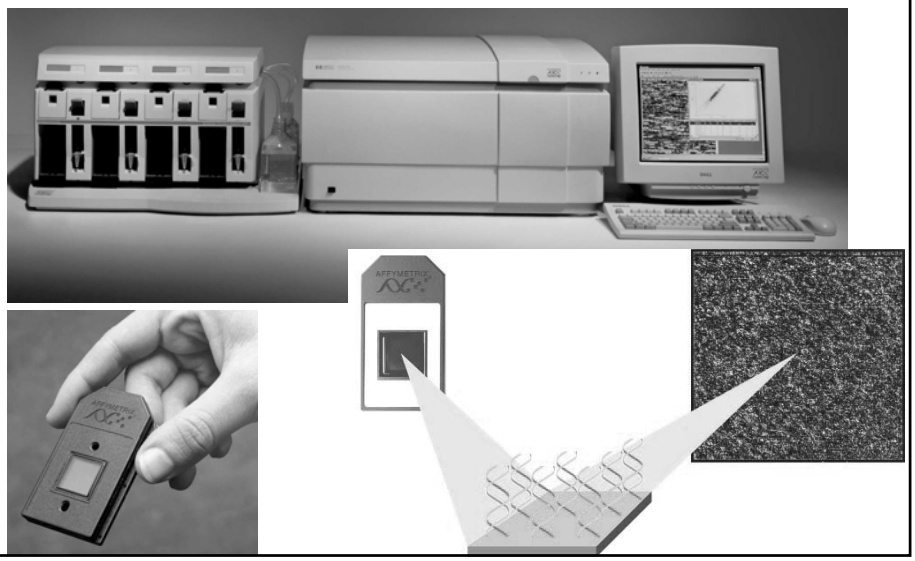


DEMO

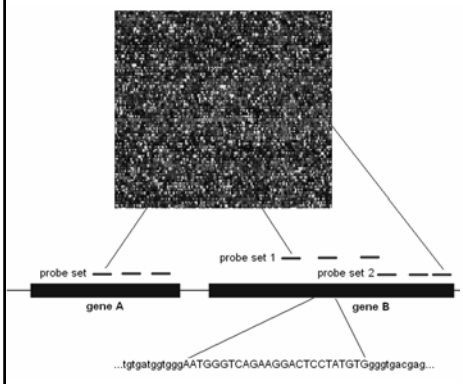
http://www.bio.davidson.edu/courses/genomics/chip/chip.html

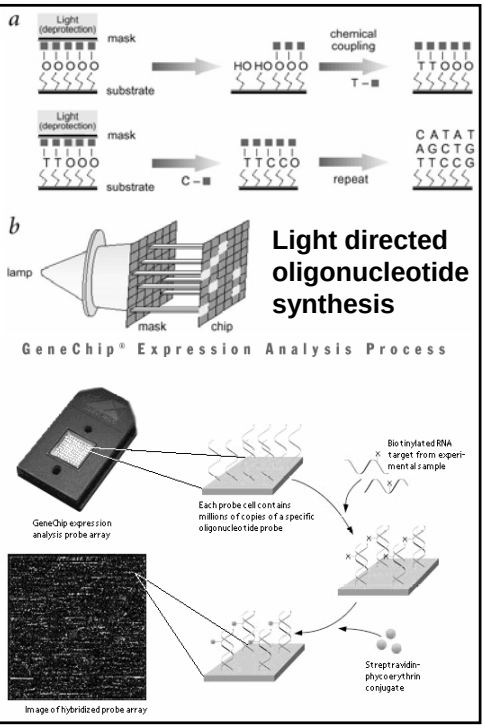


Oligonucleotide microarrays



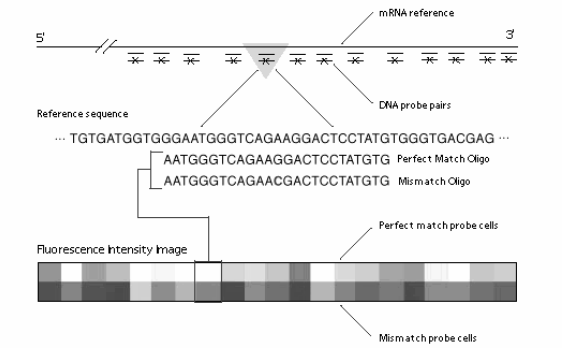
The GeneChip® microarray

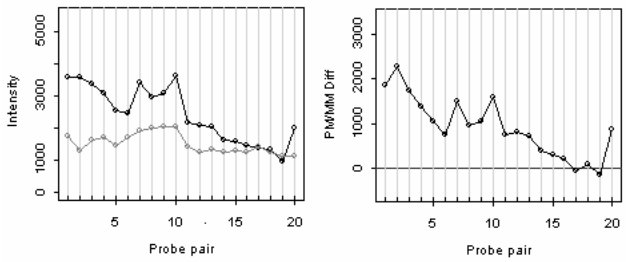




Xuegong Zhang
Tsinghua University

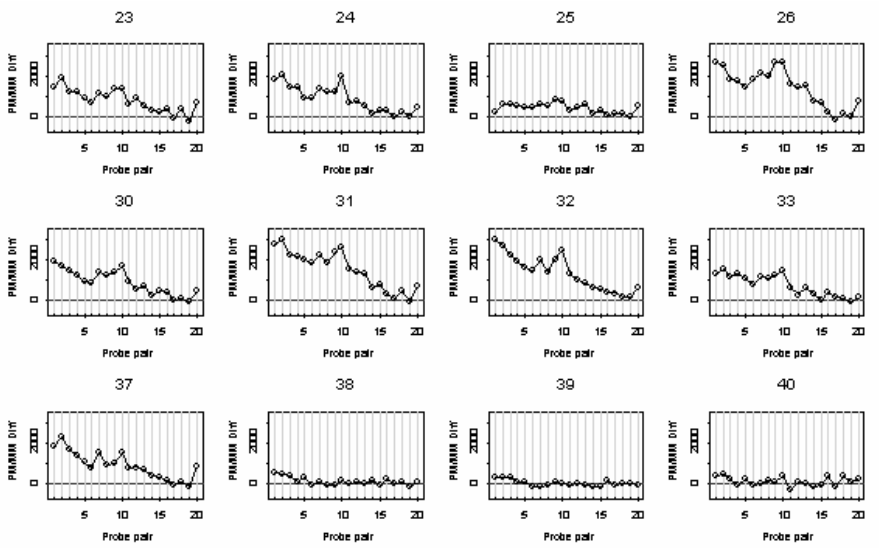
PM/MM Probes, Probe-pairs, Probe-sets





Xuegong Zhang
Tsinghua University

Example: Data for one gene in many arrays



20

Getting the expression from the probes: the Li-Wong Model

An multiplicative model for one gene :

$$Y_{ij} = PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}$$

θ_i : expression level index for array i ,
comparable to PM/MM differnece level,
assumed to be proportional to true expresion level

ϕ_j : probe sensitivity index for probe pair j ,
subject to constraint $\sum_j \phi_j^{\,2} = J$ for identifiability

ε_{ij} : random normal errors

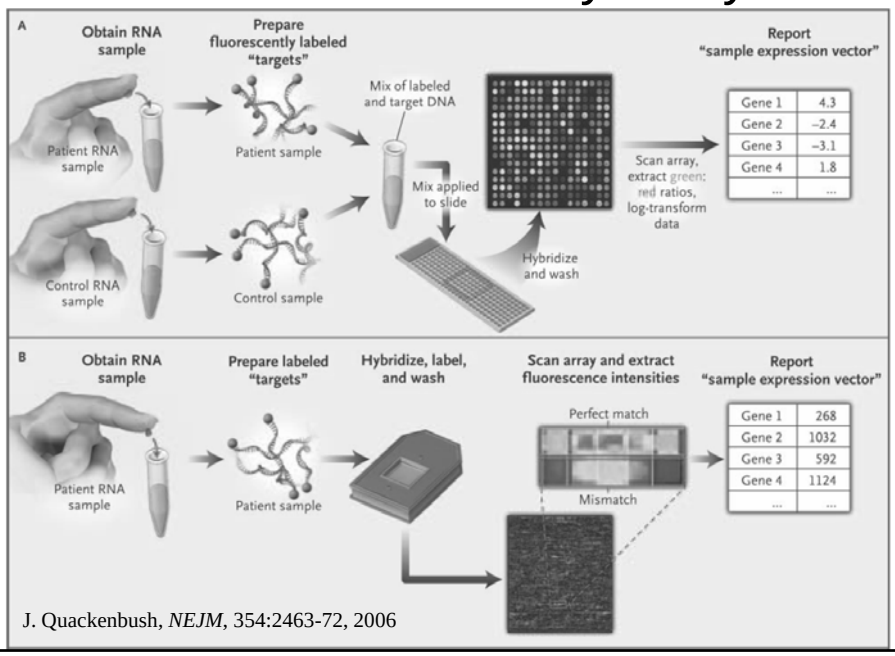
Cheng Li & Wing H. Wong, Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *PNAS*, vol.98, no.1, pp.31-36, 2001

dChip (DNA-Chip Analyzer)

www.dchip.org

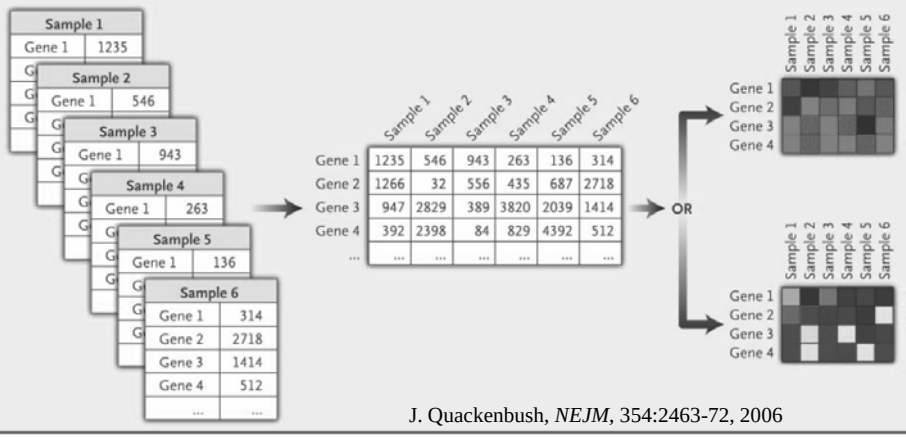
- Data Processing
  - …
- Probe-level Analysis
  - Normalize arrays
  - View data
  - Model-based expression values
  - Outlier array
  - Exon and tiling array analysis
  - …
- High-level Analysis
  - Clustering
  - Time course analysis
  - Classification
  - Significant genes
  - Gene mapping
  - Pathway analysis
  - …
- SNP Array Analysis
  - LOH and copy number
  - Linkage analysis
  - …

Overview of microarray analysis



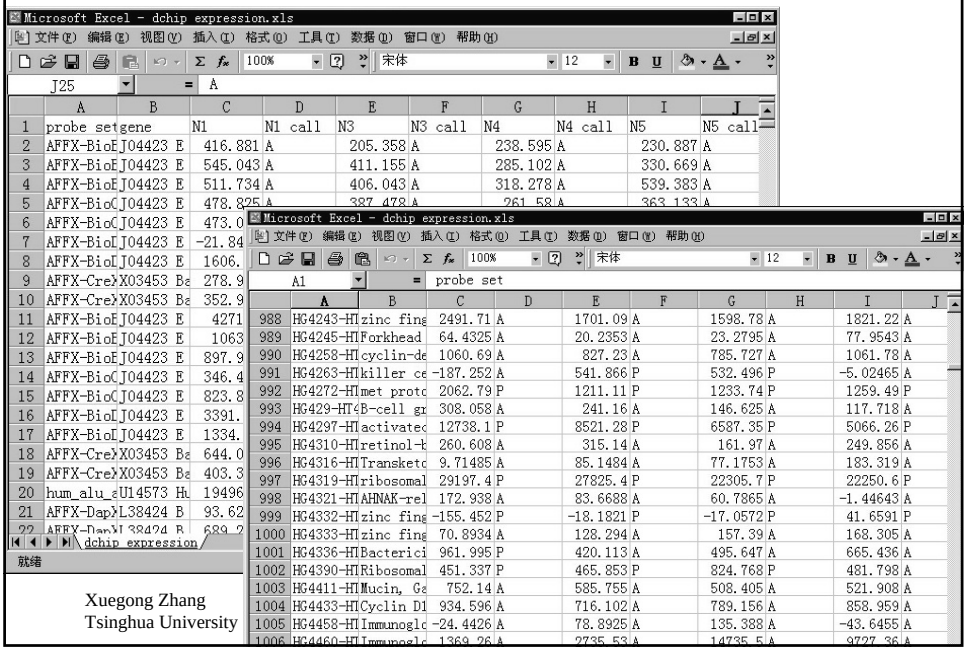
J. Quackenbush, *NEJM*, 354:2463-72, 2006

The Expression Matrix



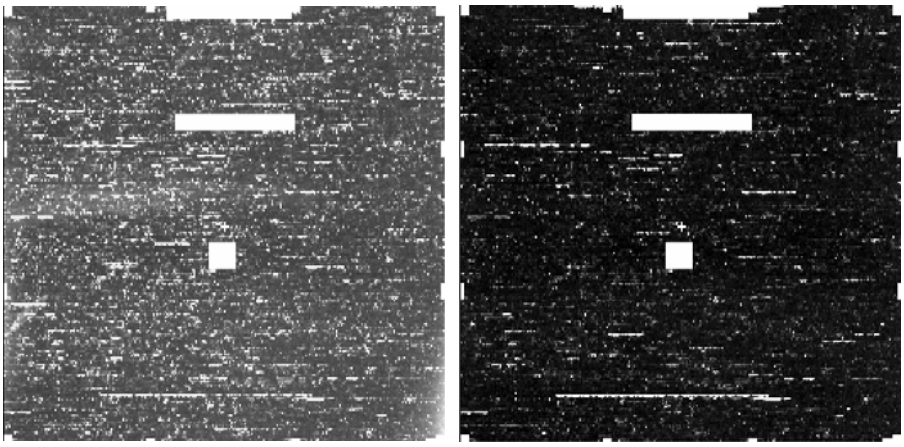
J. Quackenbush, *NEJM*, 354:2463-72, 2006

Example Data Sheets



Xuegong Zhang
Tsinghua University

The Problem of Normalization

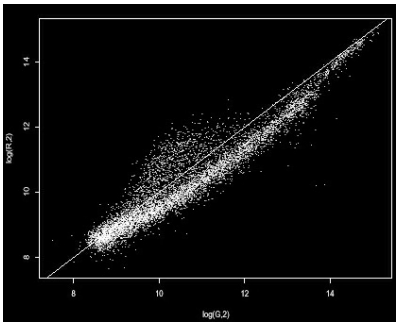


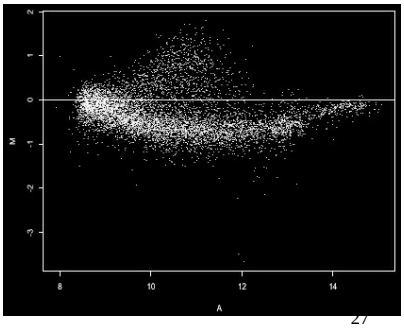
The M-A plot

- LogR vs. LogG plot
- M vs. A plot (M-A plot)

A=(log(Cy5)+log(Cy3))/2
M=log(Cy5/Cy3)

- An MA-plot amounts to a 45° couterclockwise rotation of a logR vs. logG plot followed by scaling





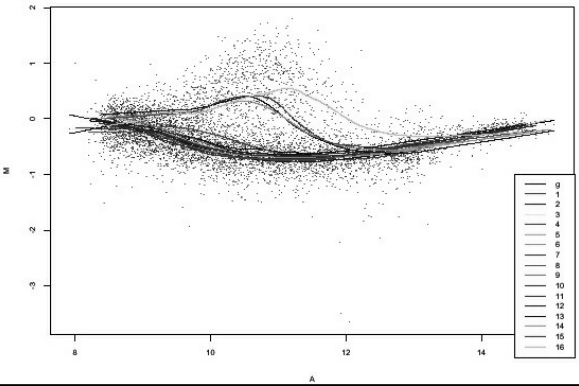
References for Normalization

- Housing keeping genes / spot-in genes
- Average or Median
- Rank-invariant genes

Example: Pin-wise Normalization (Dudoit et al. 2000)

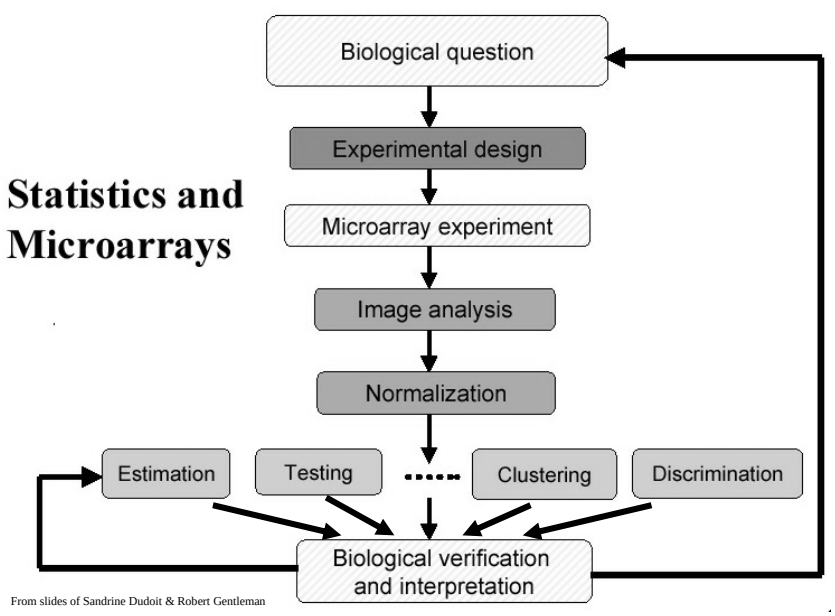
- Pin-wise normalization using all the genes
- Requires the assumption that up- and down-regulated genes with similar average intensities are roughly cancelled out or otherwise most genes remain unchanged
- Lowess (locally weighted least squares) smooth

$$\log R/G \rightarrow \log R/G - c_j(A) = \log(k_j(A)\, R/G)$$

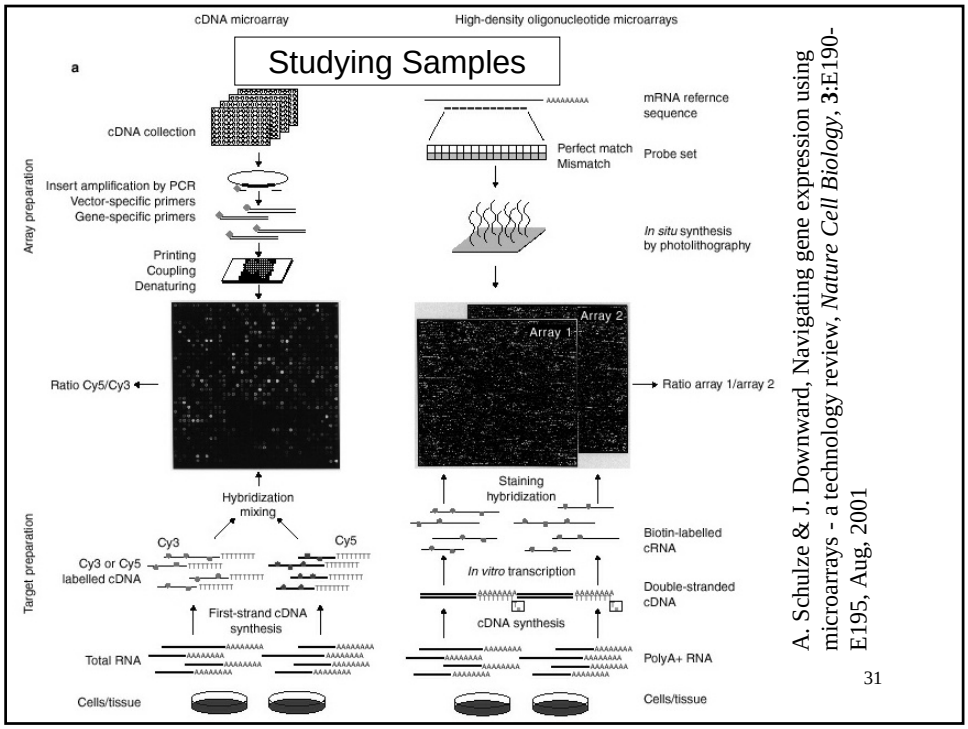


Xuegong Zhang
Tsinghua University

Bioinformatics analysis of microarrays

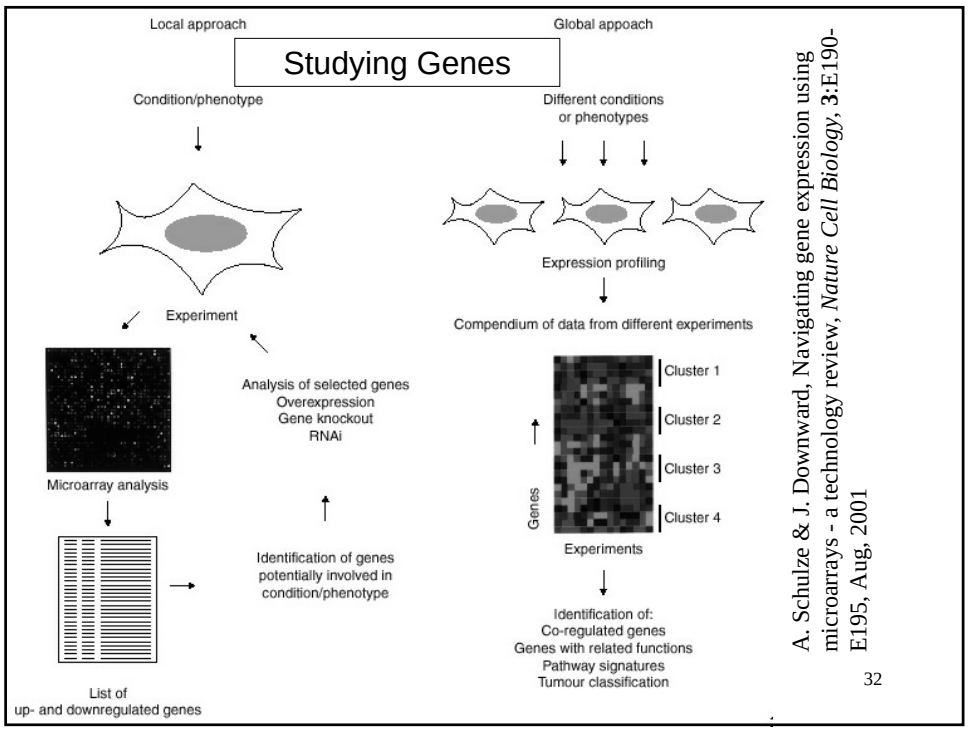


From slides of Sandrine Dudoit & Robert Gentleman



A. Schulze & J. Downward, Navigating gene expression using microarrays - a technology review, *Nature Cell Biology*, **3**:E190-E195, Aug, 2001

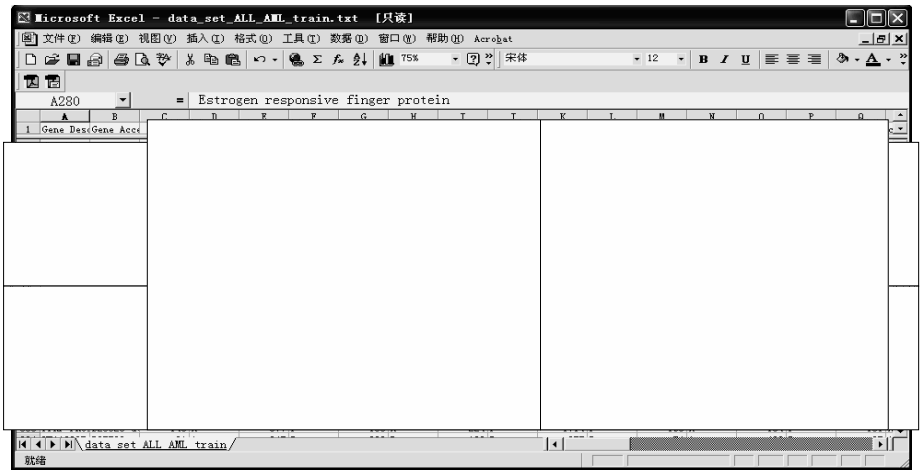
31



A. Schulze & J. Downward, Navigating gene expression using microarrays - a technology review, *Nature Cell Biology*, **3**:E190-E195, Aug, 2001

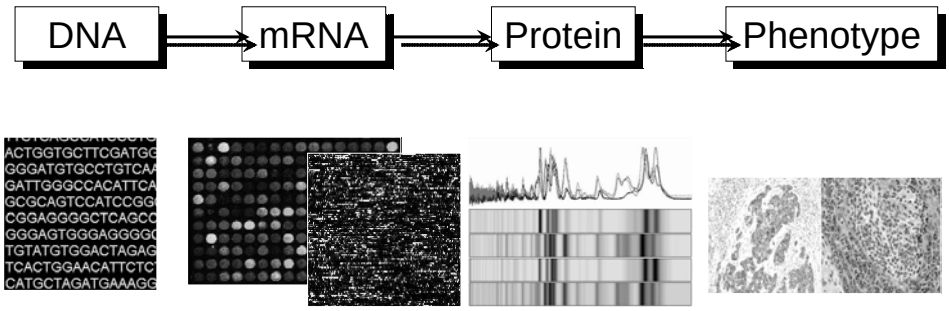
32

Classification study with multiple arrays

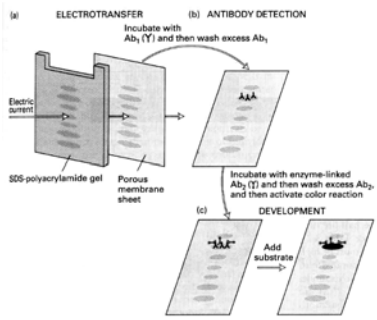


Proteomics and Mass Spectrometry

Revisit the Central Dogma

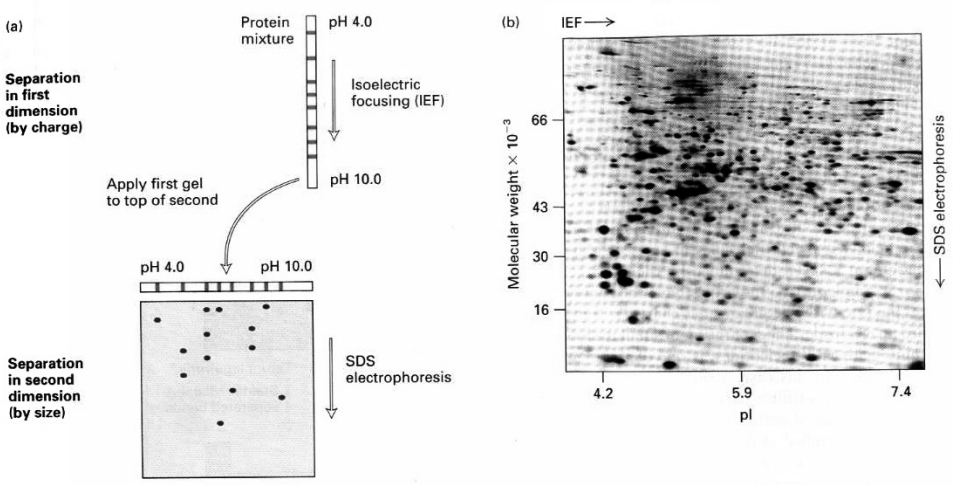


Detecting Proteins



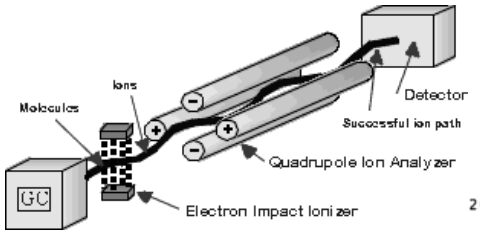
- Southern Blotting
  - Detecting DNA sequences
- Northern Blotting
  - Detecting RNA fragments
- Western Blotting
  - Detecting a particular protein in a mixture
  - The probes used are antibodies
  - Also called *immunoblotting*

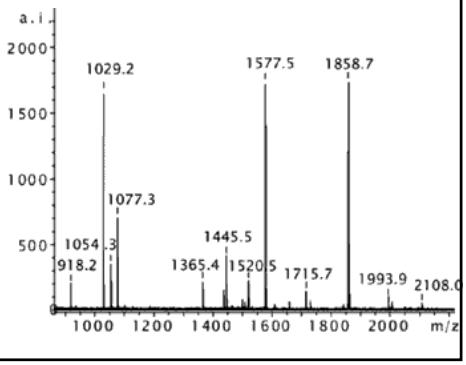
2D gel electrophoresis



37

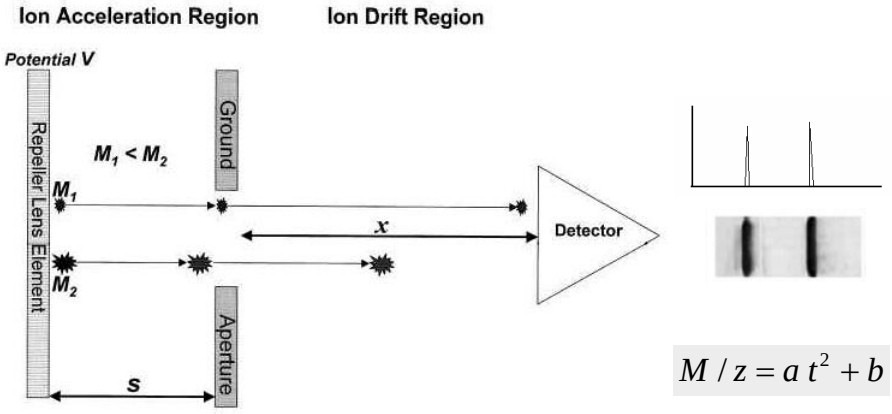
Mass spectrometry





Xuegong Zhang
Tsinghua University

TOF-MS (Time-Of-Flight Mass Spectrometry)

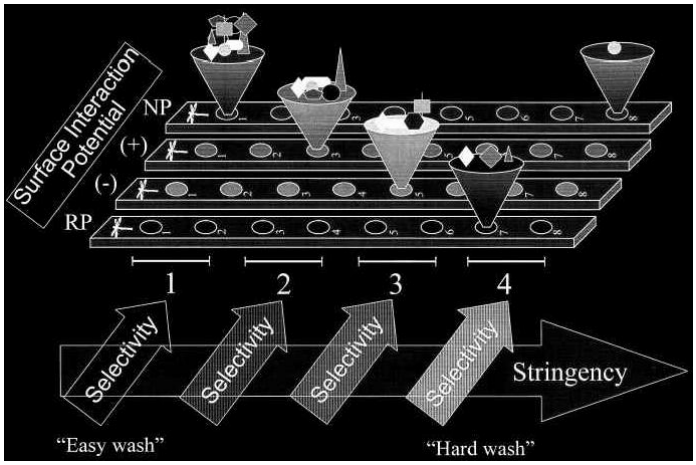


A mass spectrometry system uses 3 components – an ionization source, an analyser and a detector

39

SELDI-TOF-MS: ProteinChip® from Ciphergen

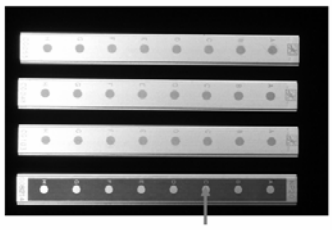
Surface Enhanced Laser Desorption/Ionization - Time Of Flight - Mass Spectrometry



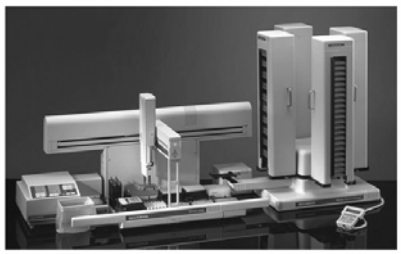
40

SELDI-TOF-MS

Serum Sample Loading



One Microliter of Serum



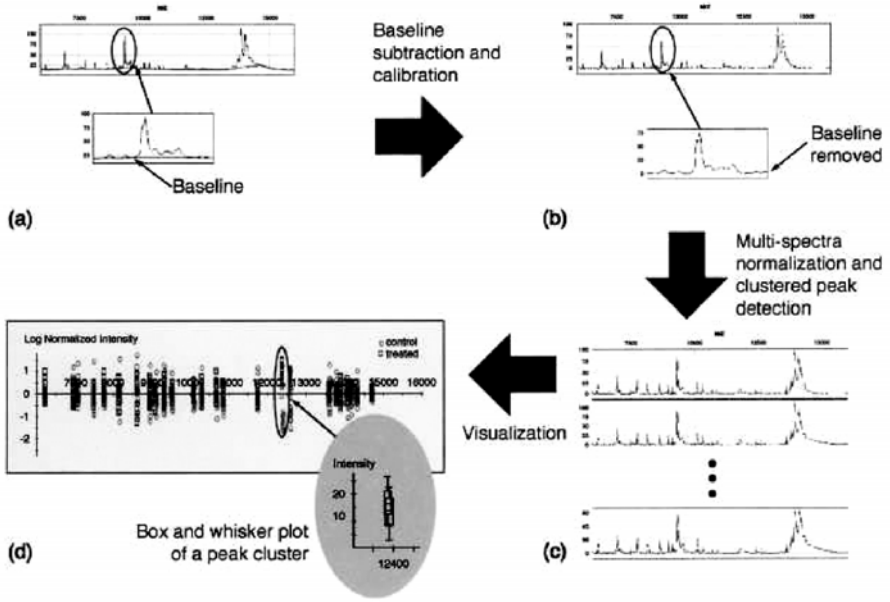
Robotic handling

- Improved Reproducibility
- Better Sample Handling
- Increased Throughput
- Reduce Cross Contamination

X **Vincent Fusaro and Sally Ross**
T

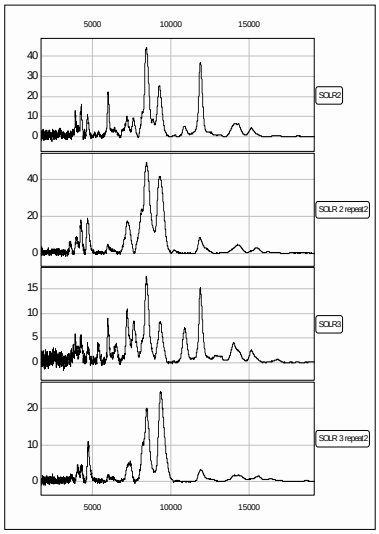
1

Typical Data Processing Steps



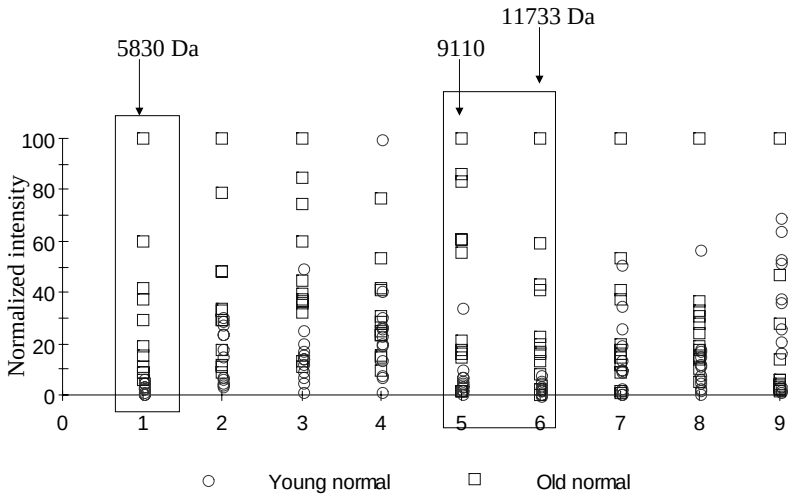
Data Processing

- Baseline subtraction
- Filtering
- Calibration
- Normalization
- Peak Detection
- Peak Alignment and Biomarker Detection
- Assessment of the Expression Indexes



Xuegong Zhang
Tsinghua University

Biomarker discovery



Tsinghua University

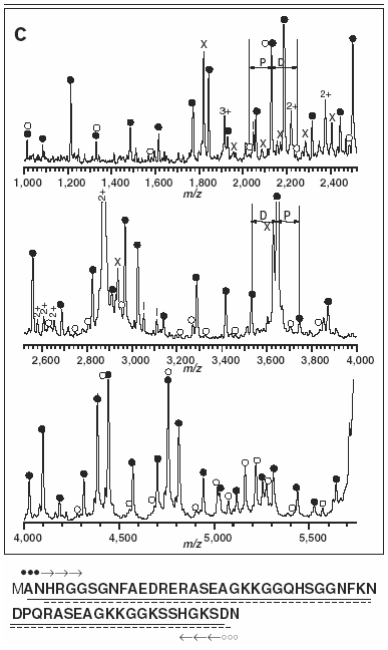
Mass-spectrometry data sheets

- Mass Spectrums
  - Time serieses
  - Could be discretized as expression profiles
- Biomarkers
  - Protein profiles at certain "meaningful" molecular weights
  - Still looks like gene expression data in format

marker #	Case 1	Case 2	Case 3	…
1	123.4	345.6	23.4	
2	23.1	1234.6	3245.2	
3	123.2	987.4	654.3	
…				

m/z	Case 1	Case 2	…
2345.6	123.4	345.6	
2347.2	23.1	1234.6	
2349.1	123.2	987.4	
…			

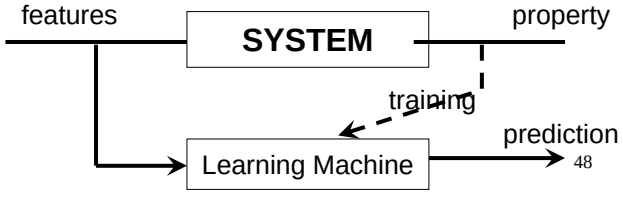
Peptide Identification and AA Sequencing with MS/MS



Basic Concepts of Machine Learning

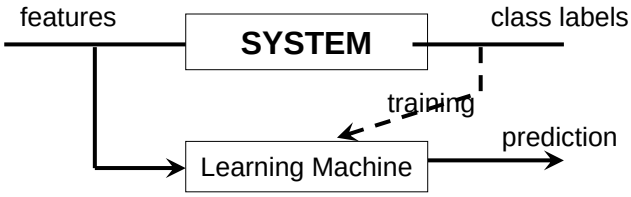
Machine Learning (ML): basic concepts

- A system (unknown but existent)
- A set of observable features (input)
- An interested property (output)
- The Learning Machine
- Predictions
- Training Samples
- The Learning Algorithm

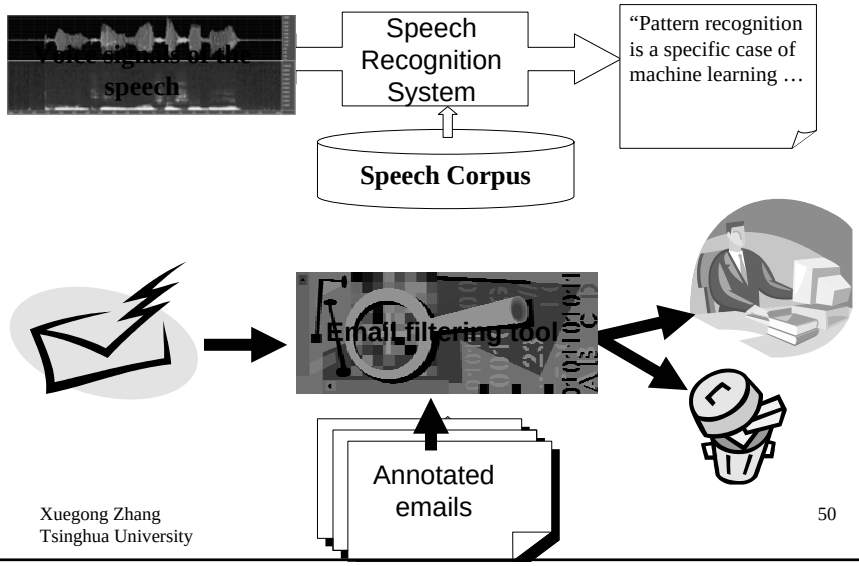


Pattern Recognition (PR)

- ML for classification properties,
  e.g., normal vs. cancer, good vs. bad prognosis,
  subtypes of a disease, ……
  → Pattern Classification



Examples of PR systems



Let Machines to Learn the Central Dogma