Bioinformatics

Ahmet Sacan

Microarrays

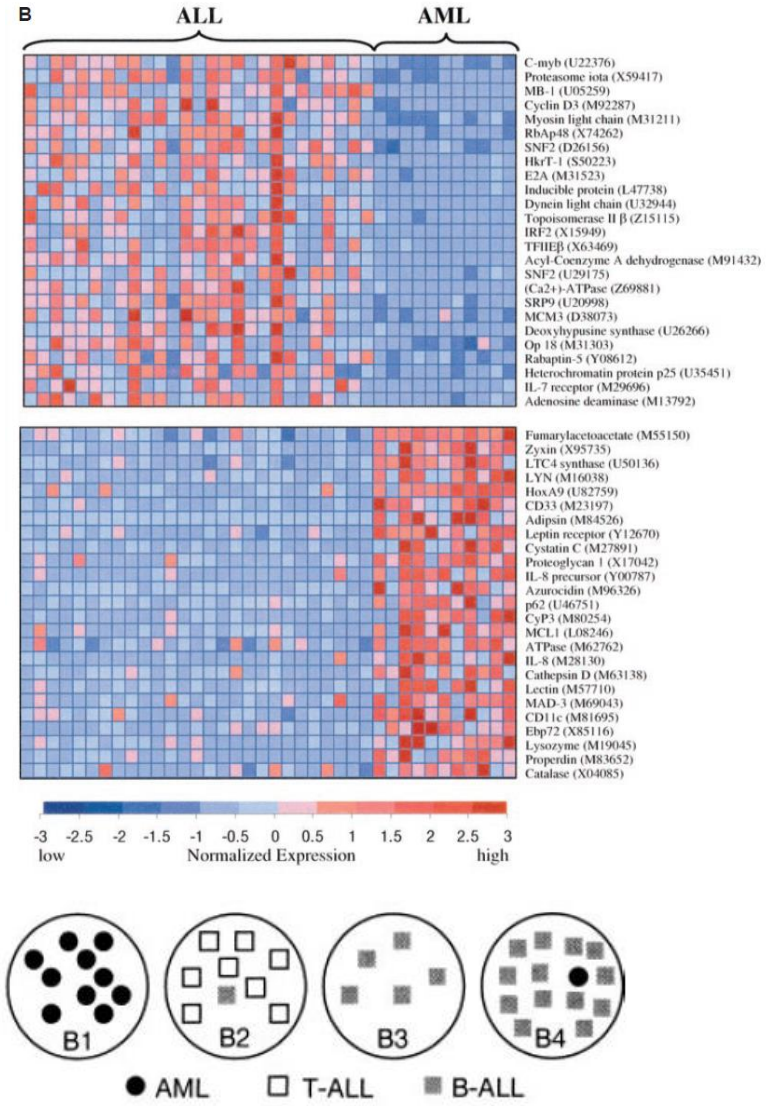
Images are from "Understanding Bioinformatics" and Wikipedia.

Golub '99

- Identification of cancer subtypes is important for proper treatment
  - Acute myeloid leukemia (AML) vs. acute lymphoblastic leukemia (ALL)
- Classification used to be based primarily on morphological appearance of the tumor.

Golub '99

- Collected RNA from 38 leukemia patients.
- Identified genes that were correlated with classification.
- Additionally refined the ALL into B-cell and T-cell derived tumors (using SOM).

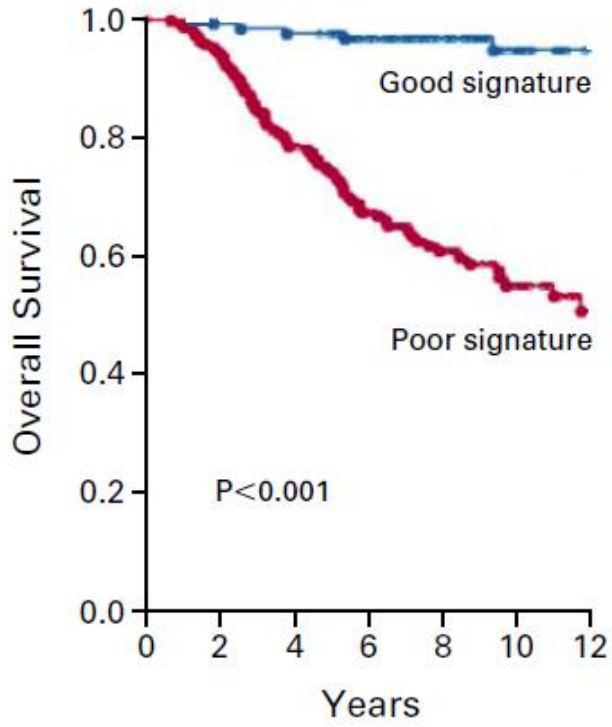


Golub '99

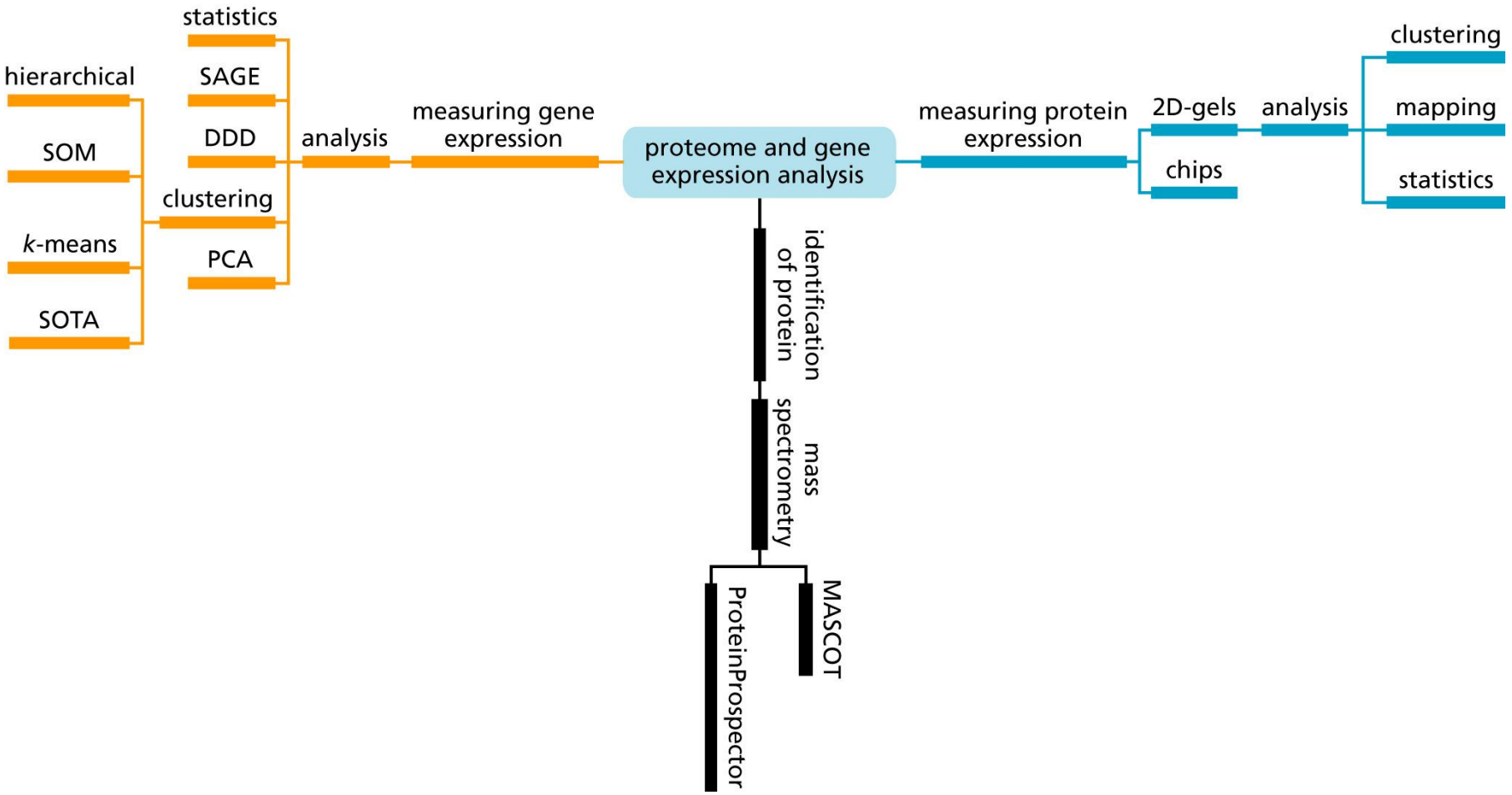
- Class prediction: assigning tumors to known classes.
- Gene discovery: identification of genes that differ from one tumor class to another
- Class discovery: identification of new cancer classes.

van de Vijver '02

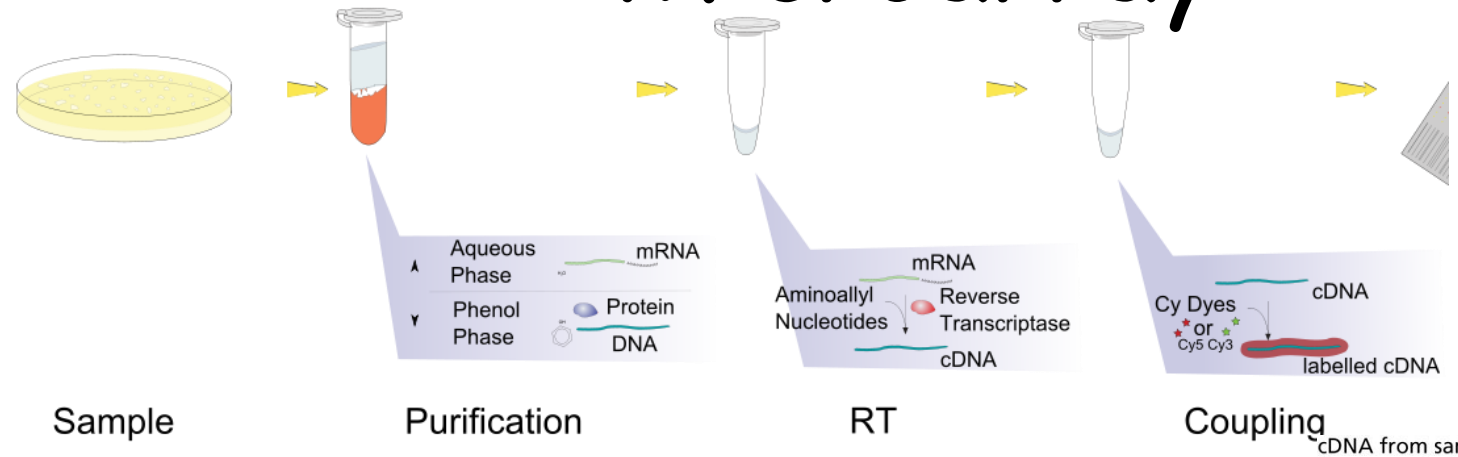
- Analyzed tumors from frozen-tissue bank (295 samples)
- 70-gene prognosis profile for primary breast carcinomas

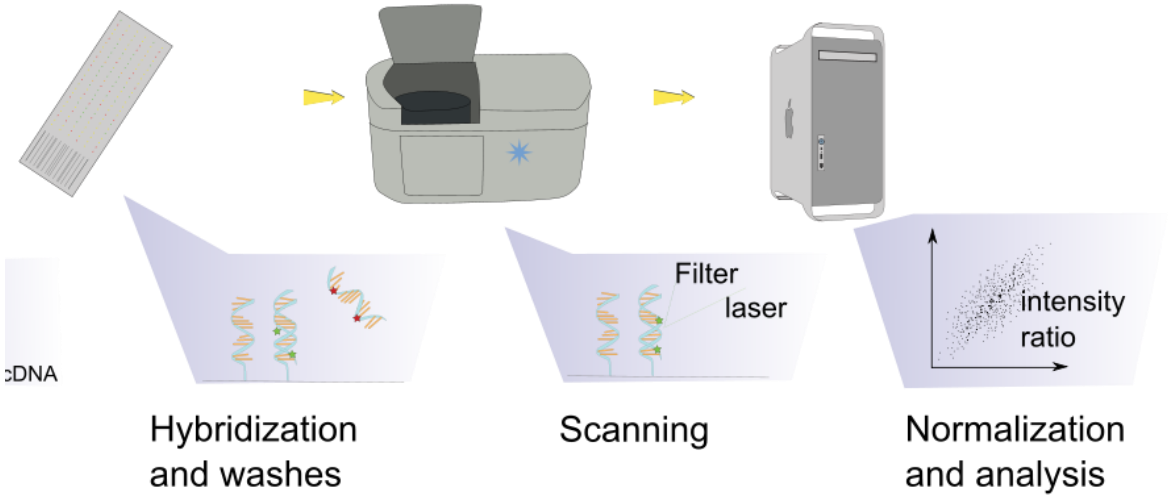


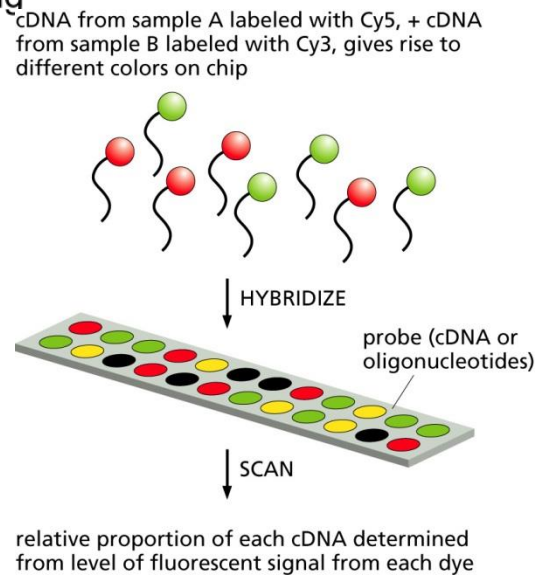
Expression analysis



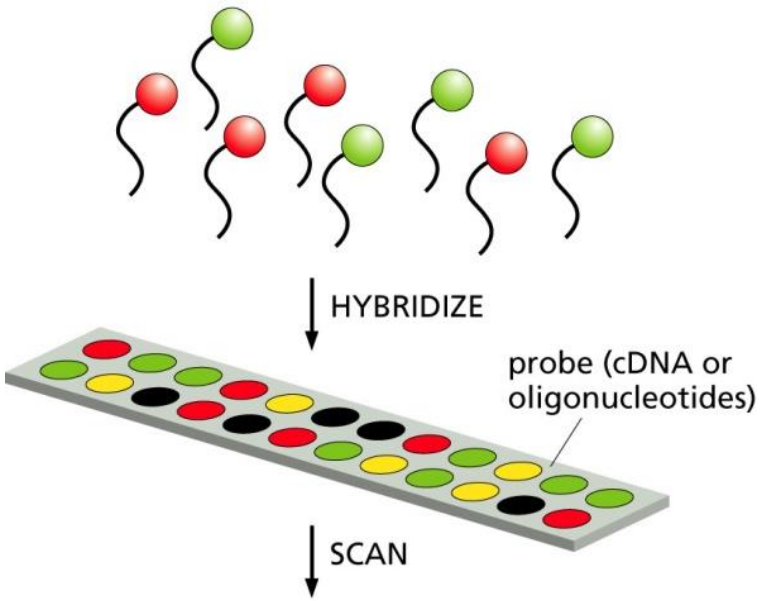
Microarray



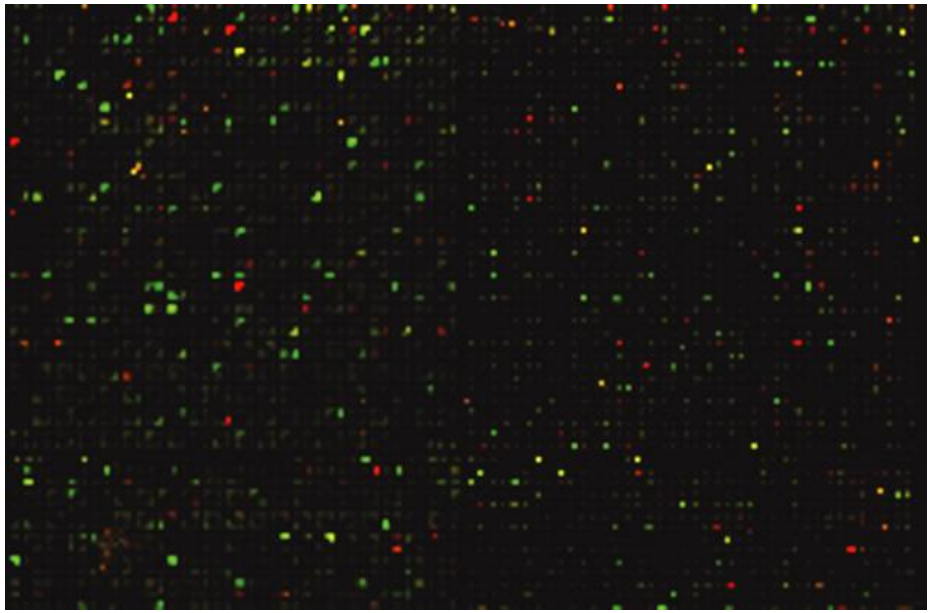


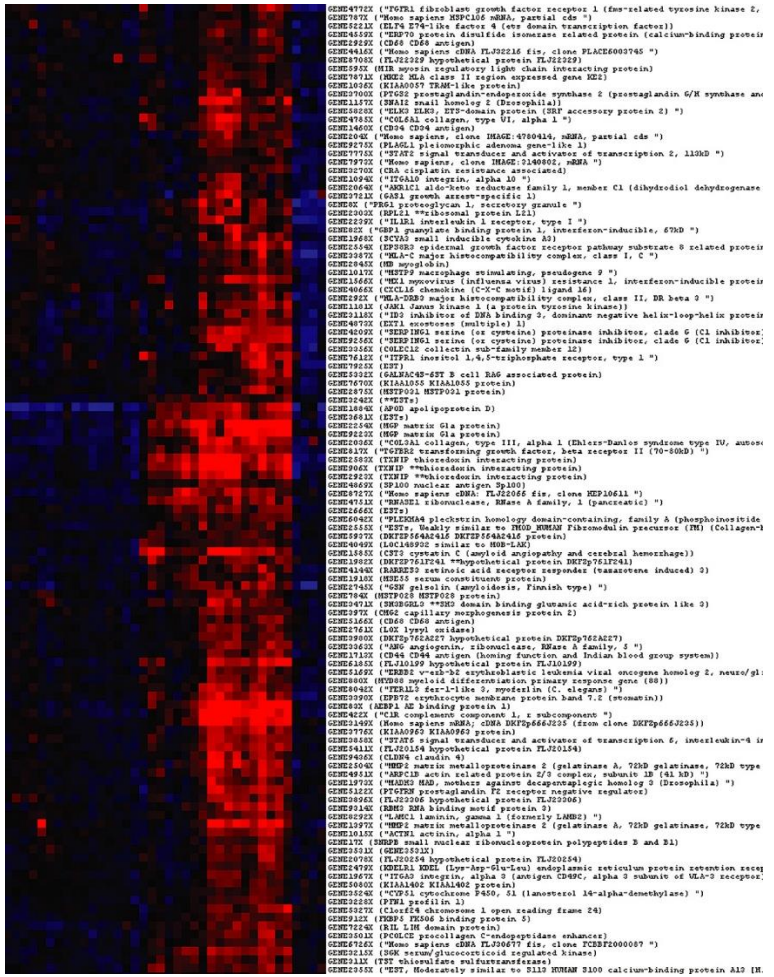


- See also:
  - http://www.bio.davidson.edu/courses/genomics/chip/chip.html


HYBRIDIZE
probe (cDNA or
oligonucleotides)
SCAN

Microarray



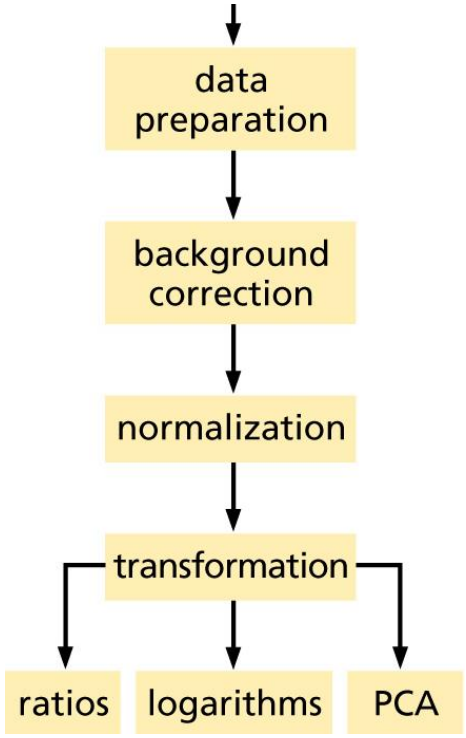


Microarray data normalization

- Normalization removes systematic experimental errors
- The measured expression:

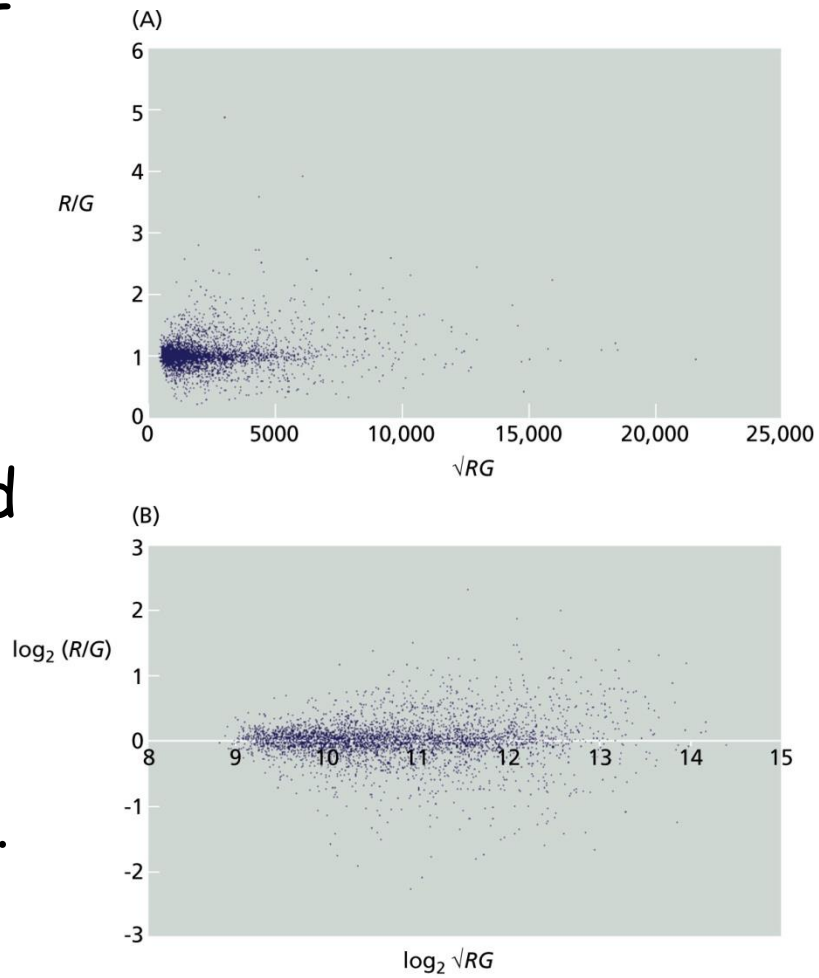
$$X = \gamma e^{\eta} + \epsilon$$

  - γ: actual expression level
  - ϵ: additive error with distribution $N(0, \sigma_\epsilon)$
  - e^{η}: multiplicative error term that is proportional to the level of expression, with distribution $N(0, \sigma_\eta)$.



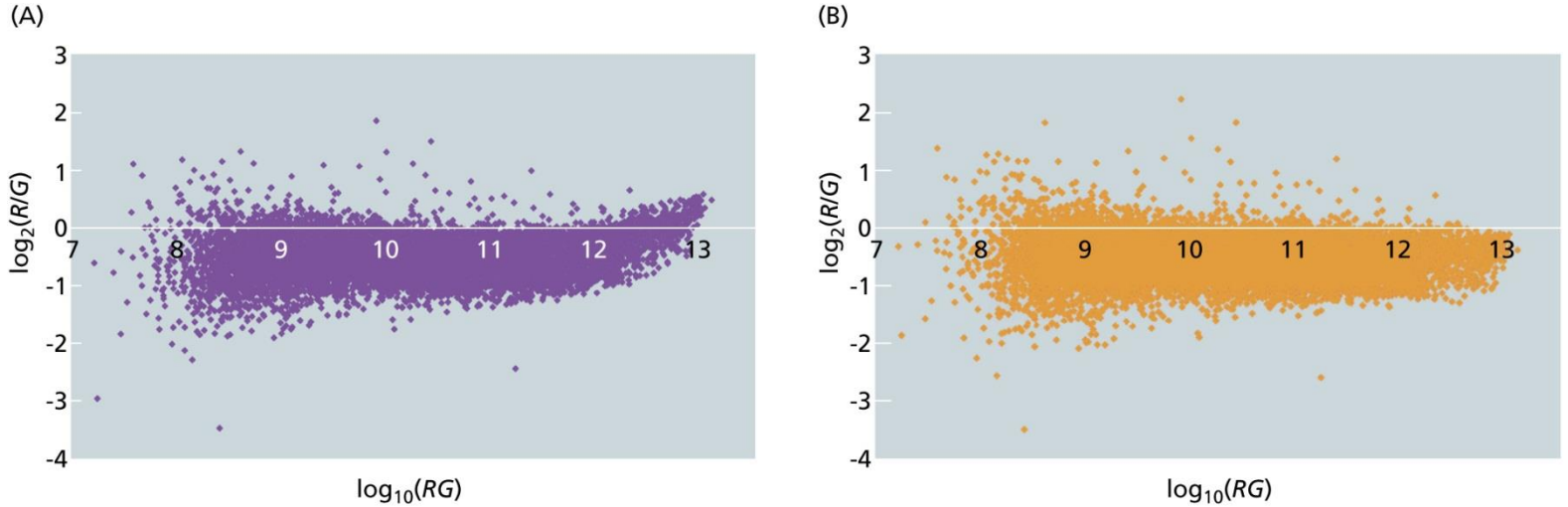
Problems with R/G

- Plot shows how distribution of ratios varies with the level of expression.
- The distribution of R/G is skewed upward for all expression values.
- Taking logarithms solves this problem. Data is now centered and more equally distributed around 0.
  - log(X)=-log(1/X)
  - Increases and decreases in expression are treated equally.



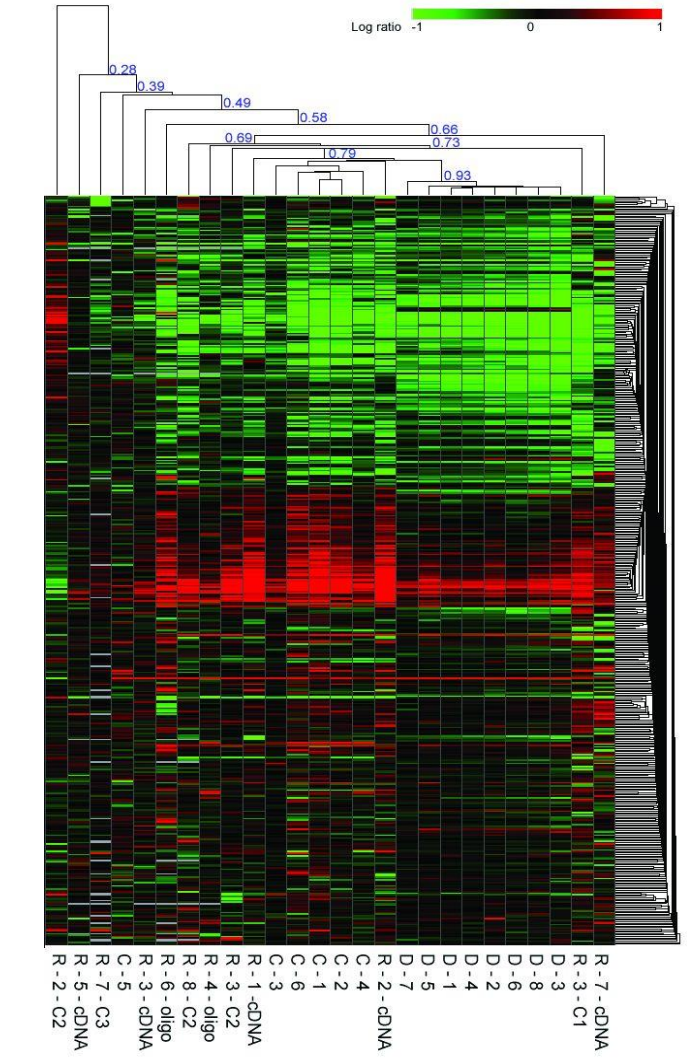
Lowess normalization

- Suitable when expression ratios have a curvature dependent on the expression levels
- Lowess: LOcally WEighted Scatterplot Smoothing
- Applies regression analysis in small windows.



Clustering

Cluster to detect gene clusters and regulatory networks

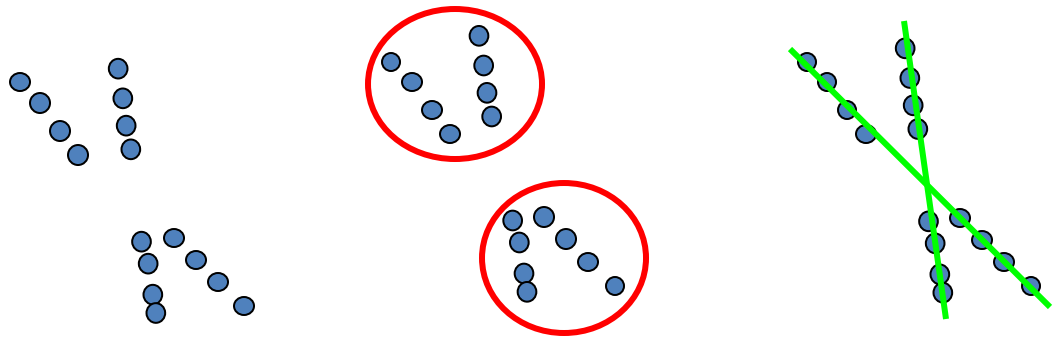


Supplementary Figure 1: Clustering of laboratory/platform combinations using log ratio values of commone genes

Cluster to detect patient subgroups

Clustering Methods

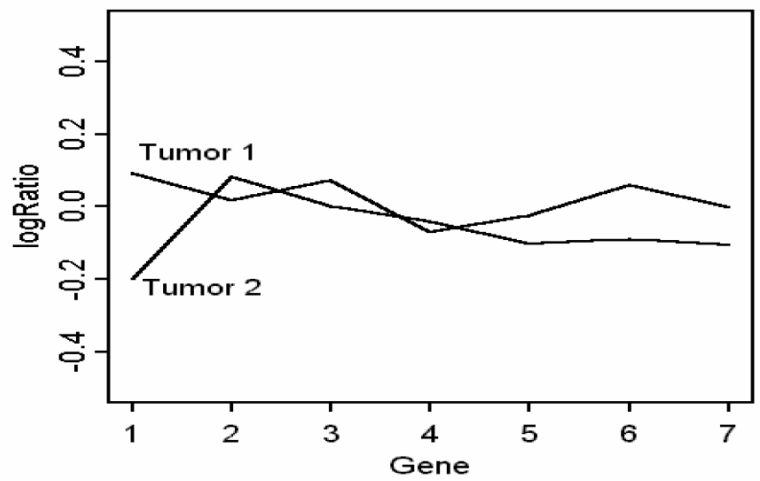
- Hierarchical Clustering
- k-means clustering
- Self-organizing maps

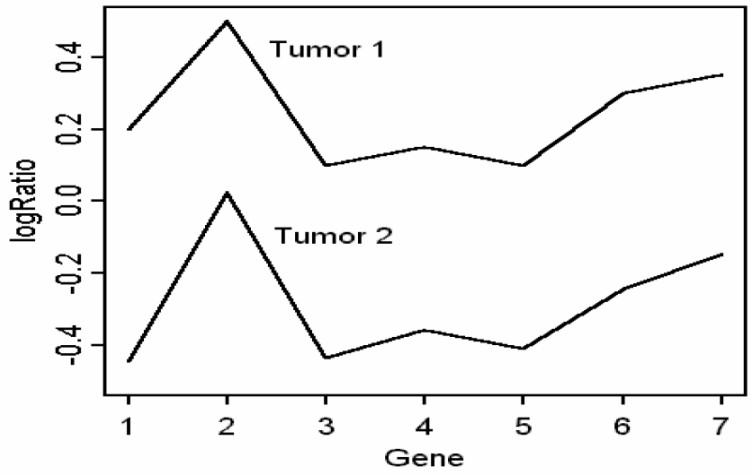


Distance measures

- Euclidean
- Pearson correlation
- Cosine angle (aka. Uncentered Pearson)

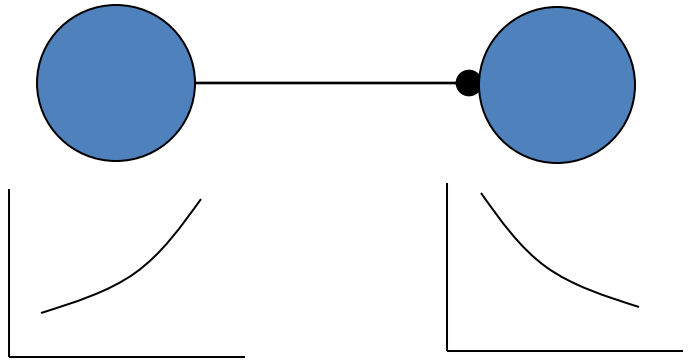
  - Pearson and cosine are invariant to scaling
  - Pearson is also invariant to translation, e.g.:
    - Pearson(X1,X2) == Pearson(X1, 2*X2+5)





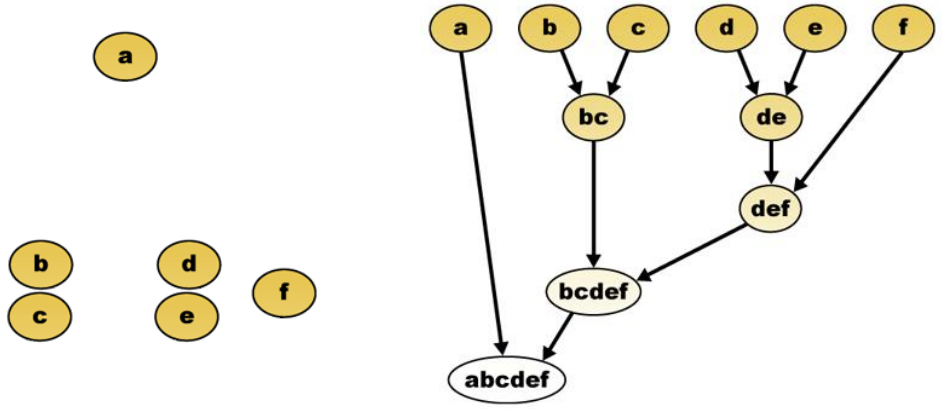
Pitfalls in Distance measures

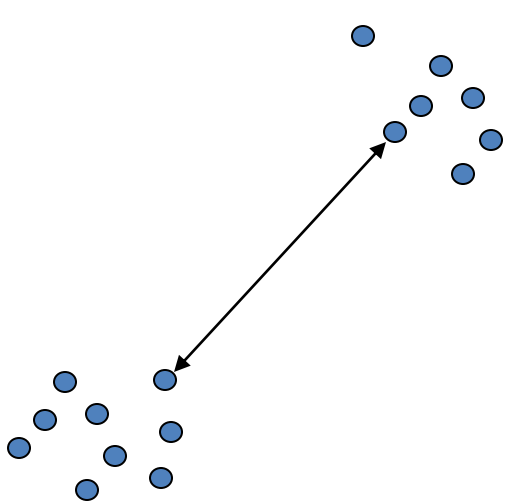
- Negative correlation may also be of interest (e.g., closely related on the signaling or regulatory networks).



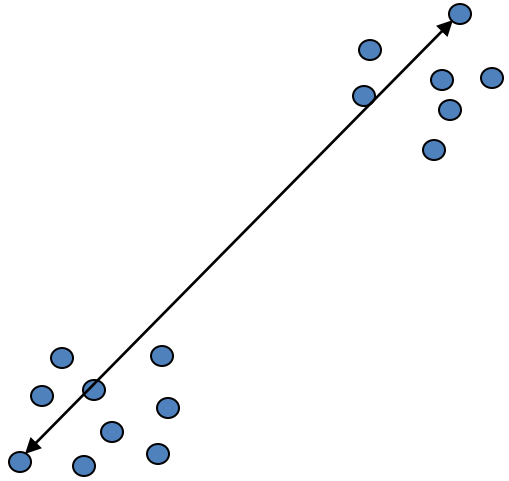
- Workaround: use absolute value of correlation, e.g. (1-|Pearson|)

Hierarchical Clustering

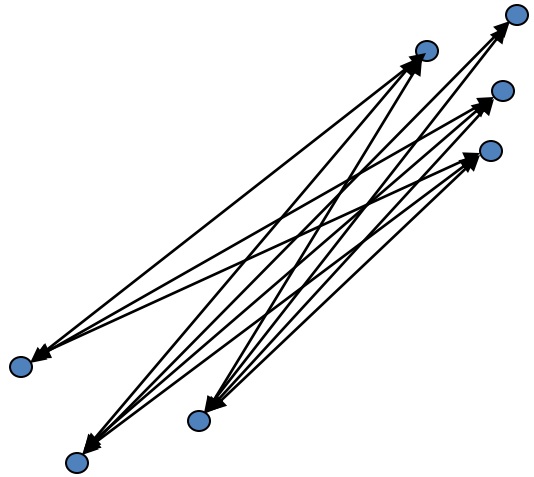




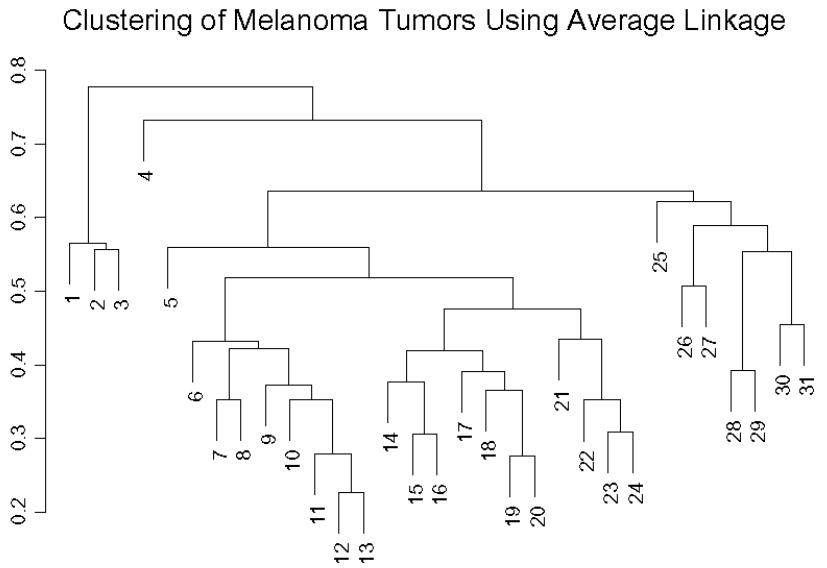
Single linkage

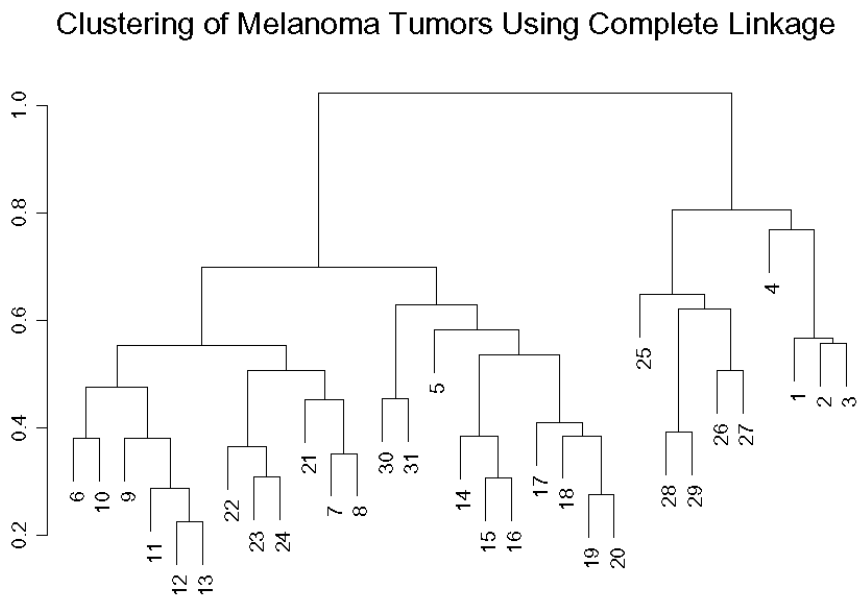


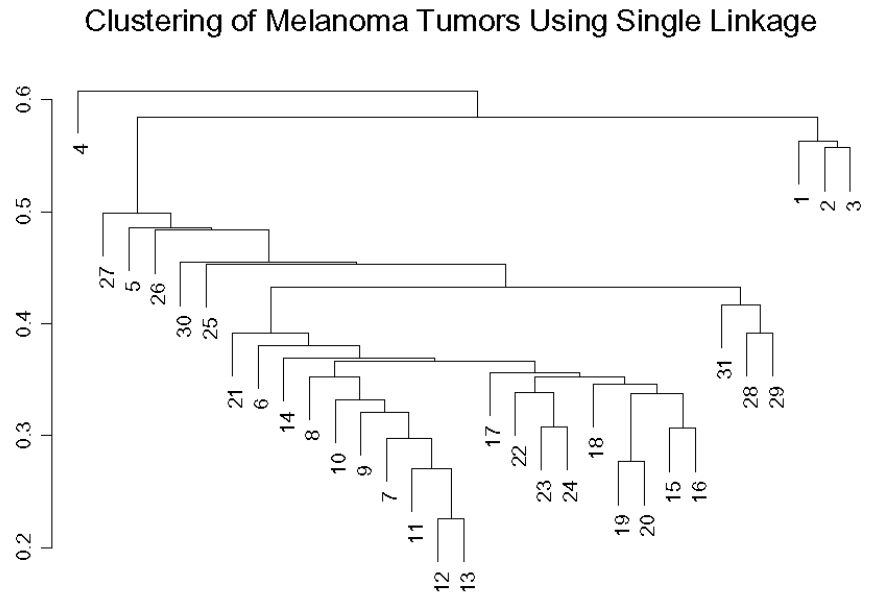
Complete linkage



Average linkage (UPGMA)



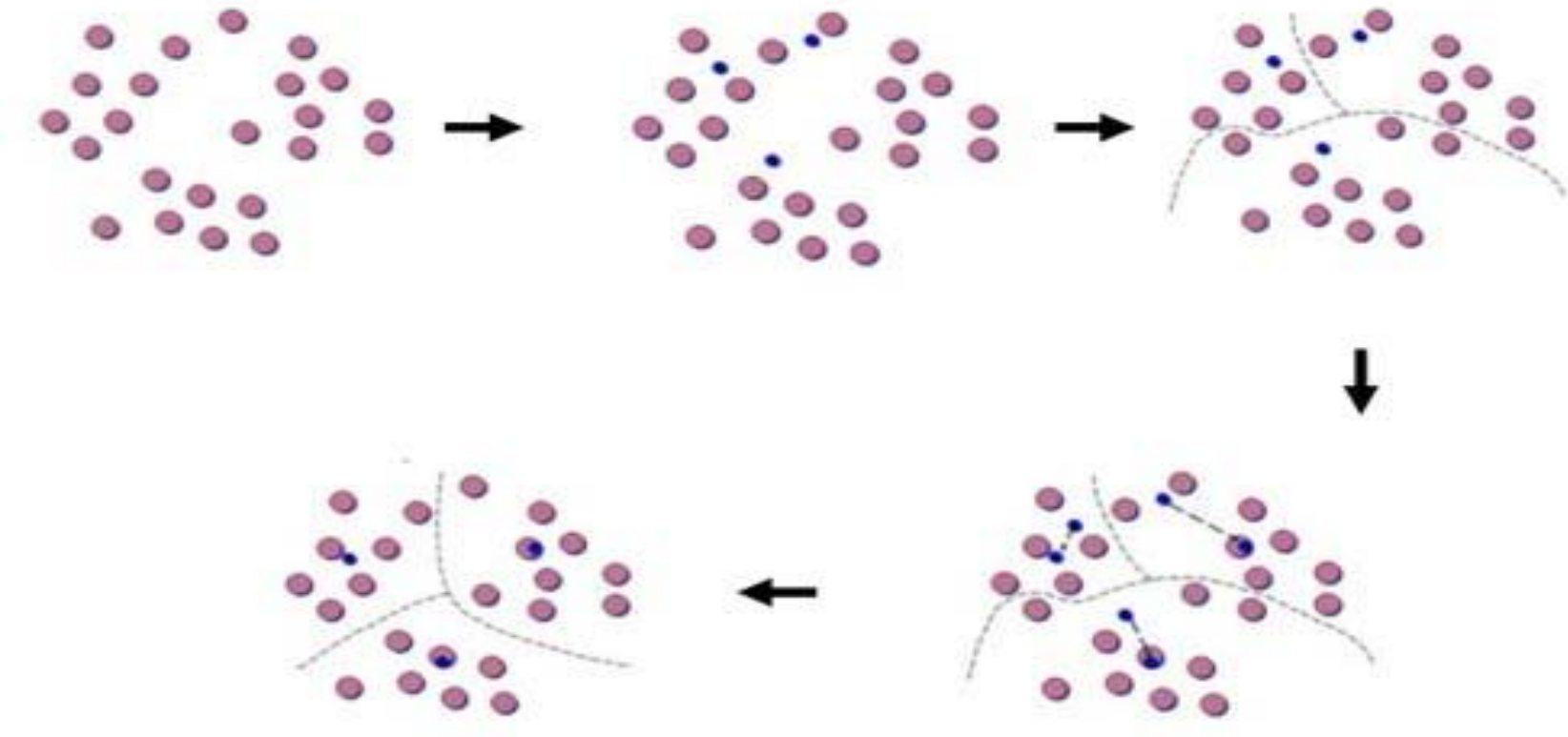




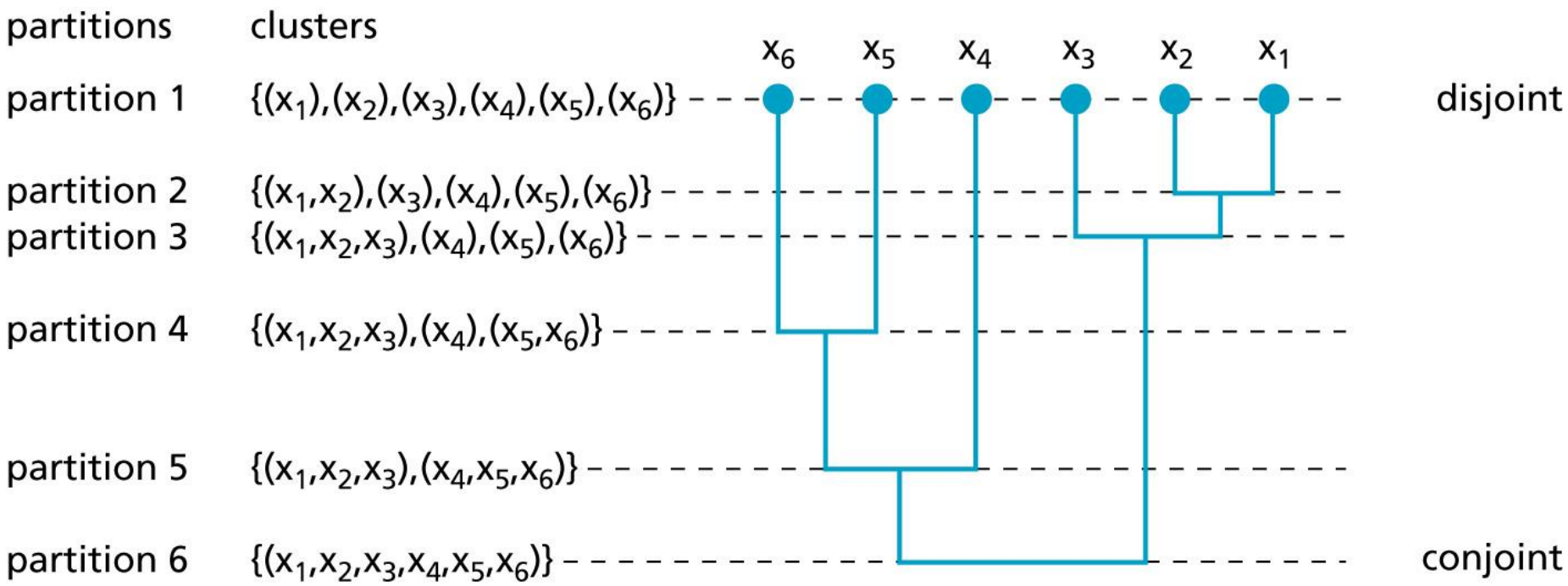
Dendrograms using 3 different linkage methods,
distance = 1-correlation

(Data from Bittner *et al.*, *Nature*, 2000)

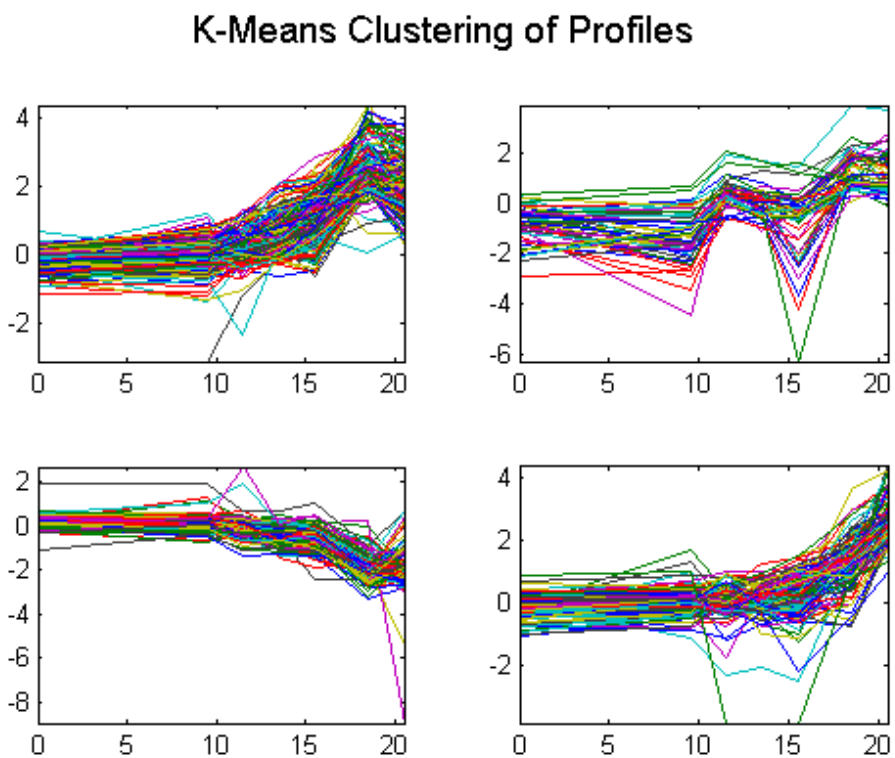
K-means clustering

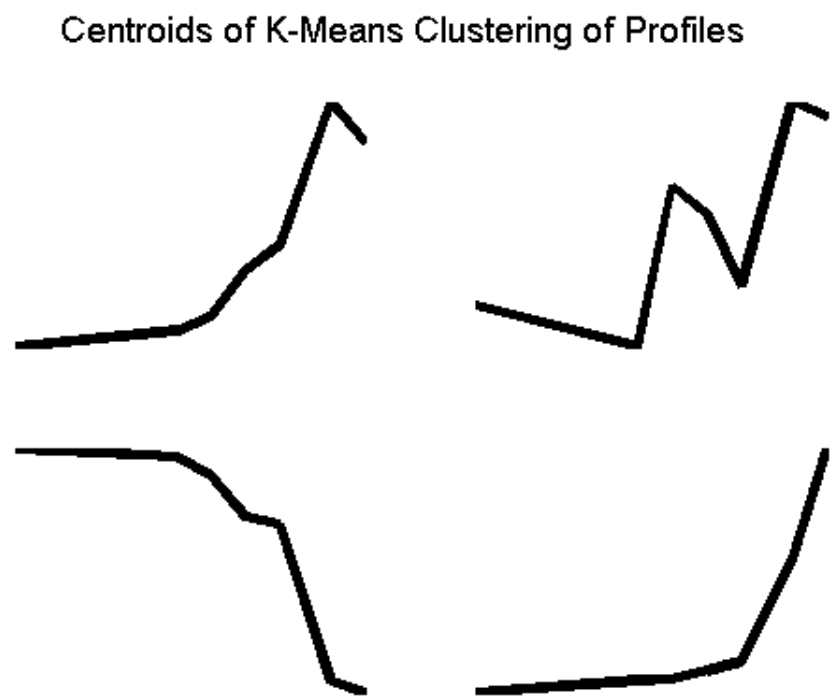


Converting hierarchical clustering to partitions



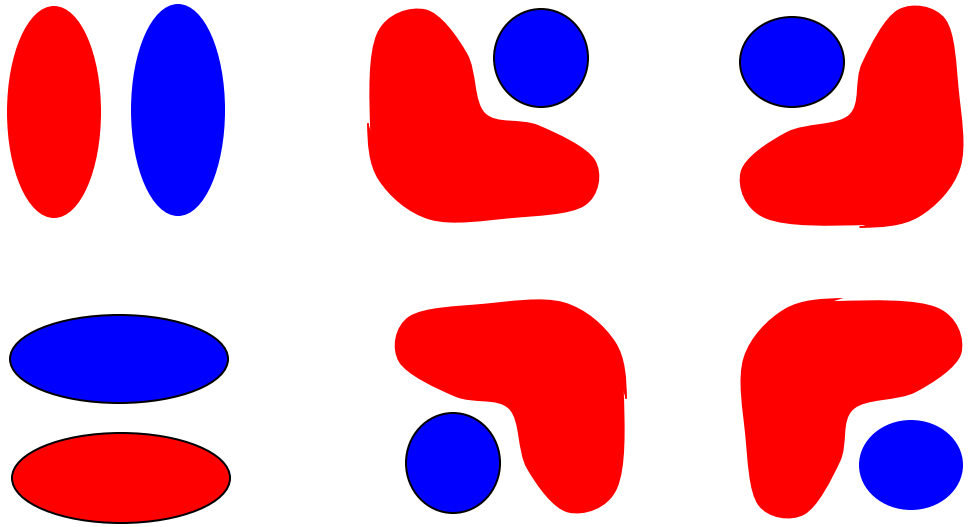
K-means clustering



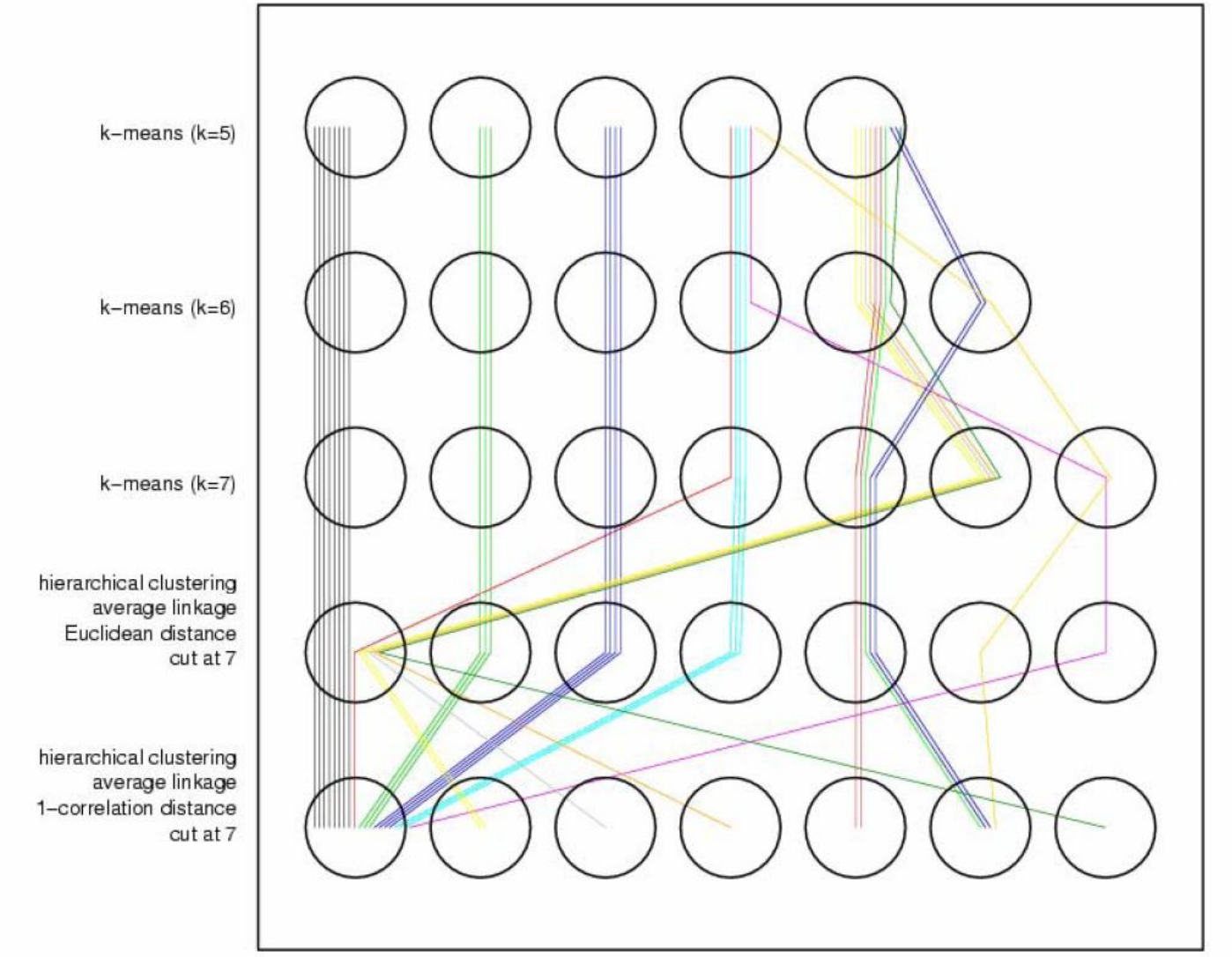


K-means clustering

- Simple, ~fast
- Selection of a good k: trial & error
- Does not always converge
- Sensitive to initialization

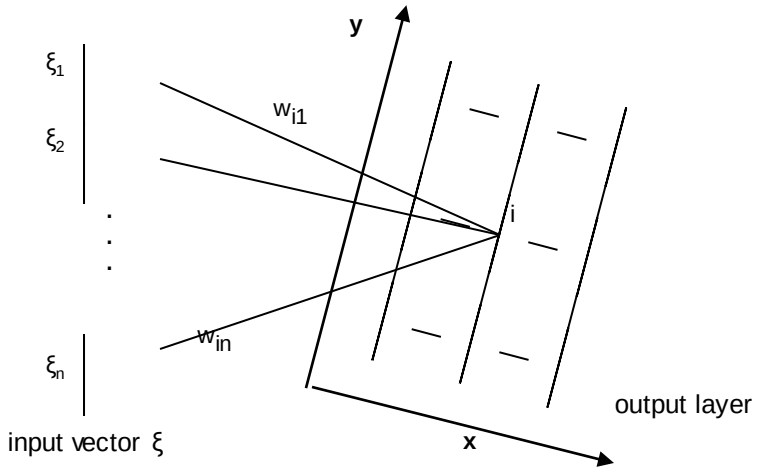


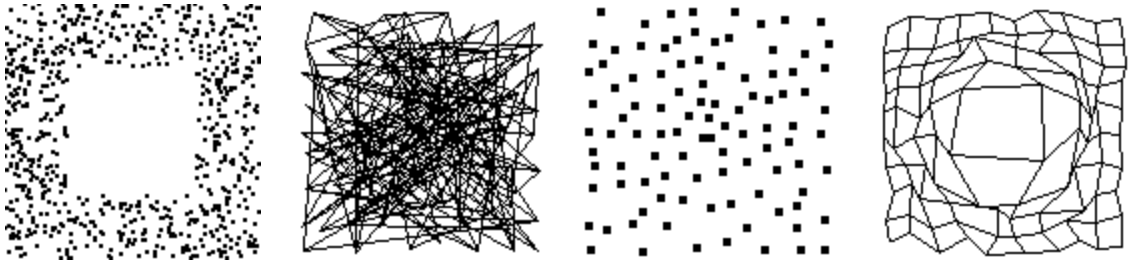
K-means clustering



Self Organizing Maps

- Dimension and data reduction
- Identifies spread/distributio n





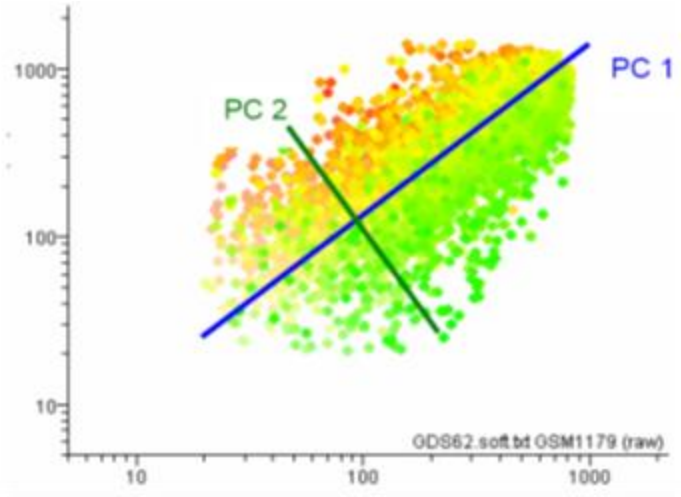
input SOM, t=0 SOM, t=0 SOM, t=1000

Dimensionality Reduction

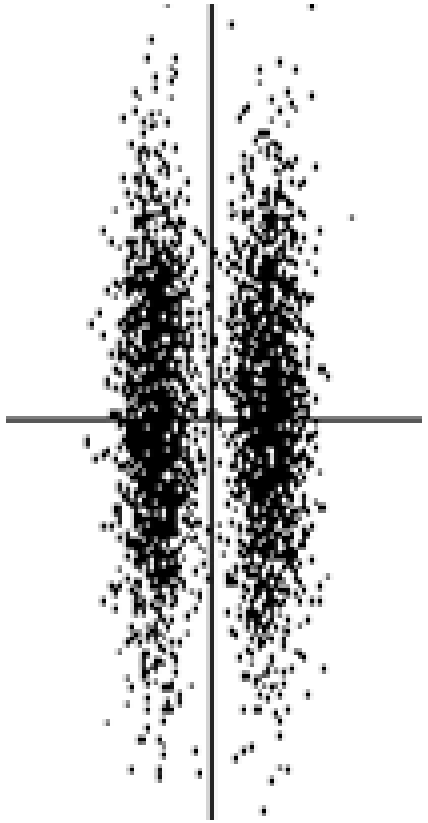
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- (Classical) Multidimensional scaling (MDS)
- + ~30 others

Principal Component Analysis

- Measure 10,000 genes in 8 different patients
  - A matrix of 10,000x8 measurements
- Imagine each 10,000 gene is plotted in a multi-dimensional on a scatter plot consisting of 8 axes.
  - Results in a cloud of values in multi-dimensional space.
- PCA extracts directions where the cloud is most extended

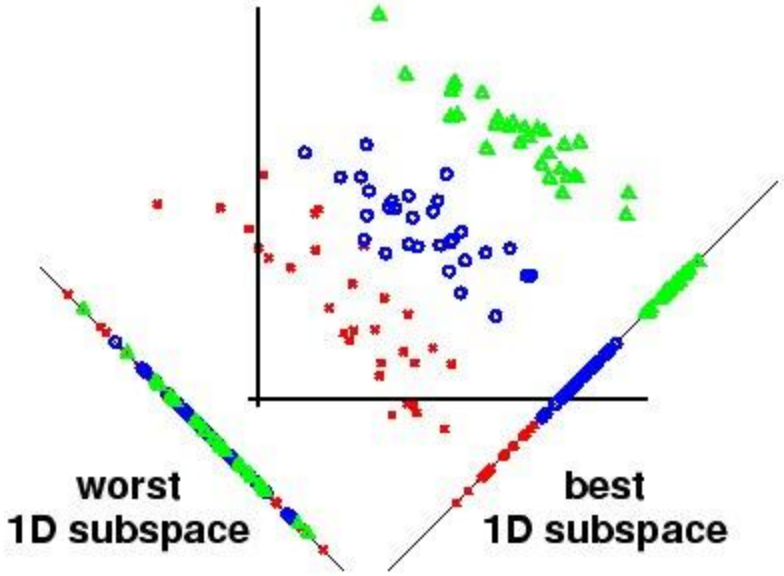


Linear Discriminant Analysis



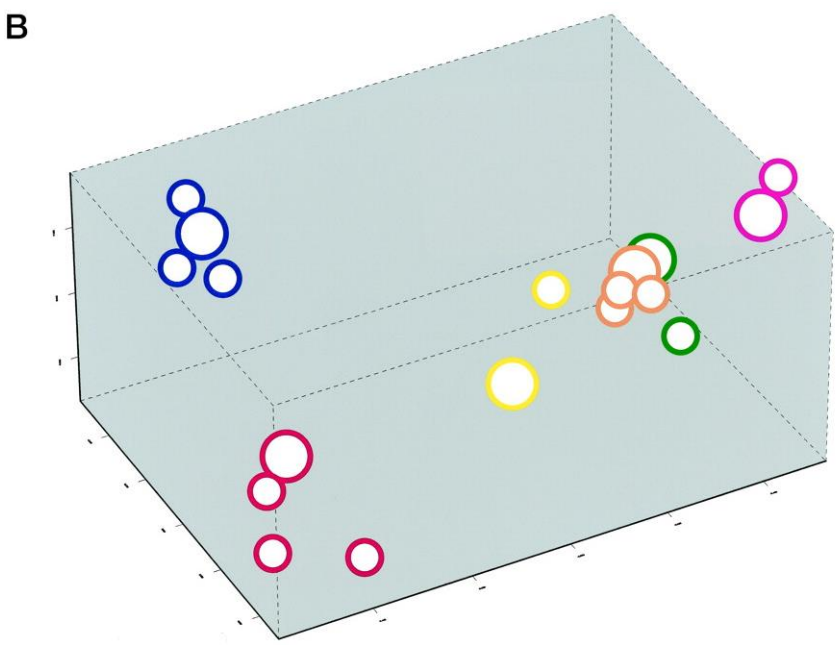
PCA

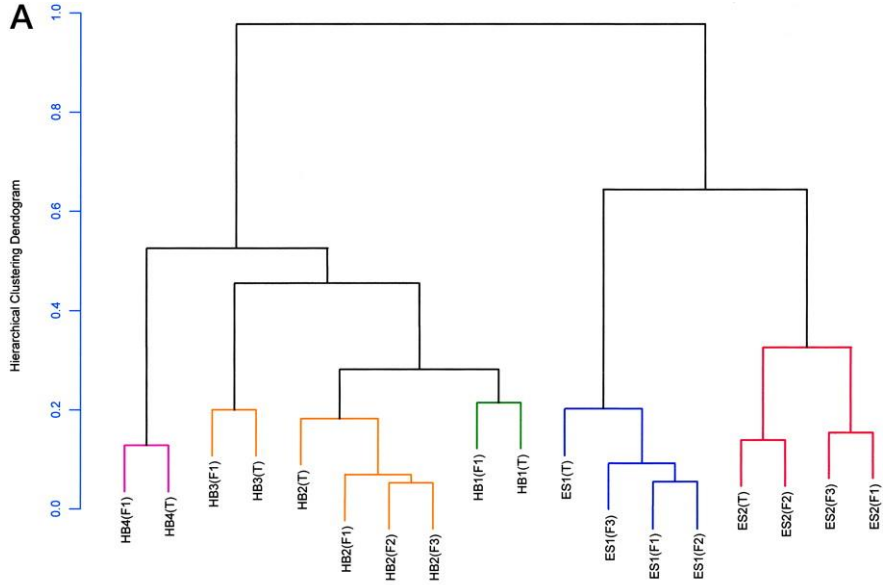
- LDA: Redefine "interesting" projections using class separability



Multidimensional Scaling (MDS)

- Assersohn, 2002
  - Samples from fine needles aspirates (FNA) and from tumors in breast cancer.
  - Color: patient, Large circle: tumor, Small circle: FNA



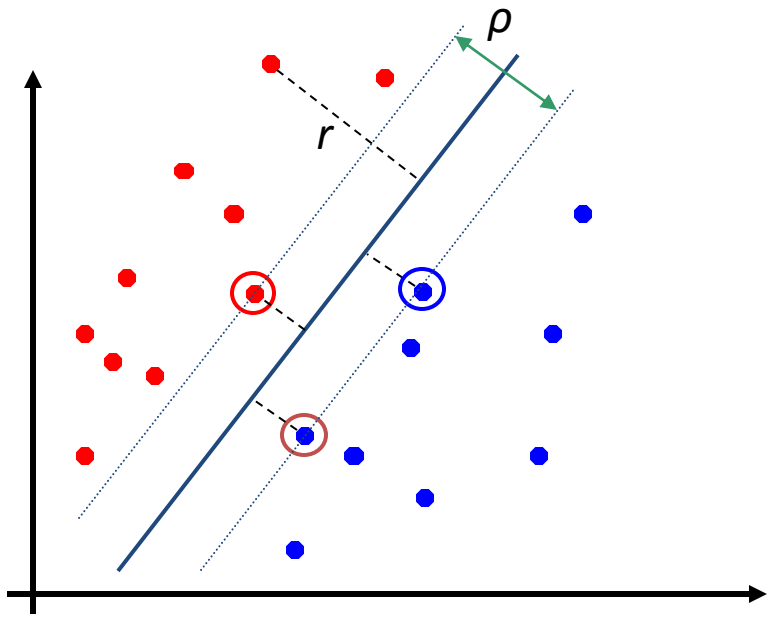


Classification

- Neural Network
- Support Vector Machines (SVM)
- Decision Trees
- + many others

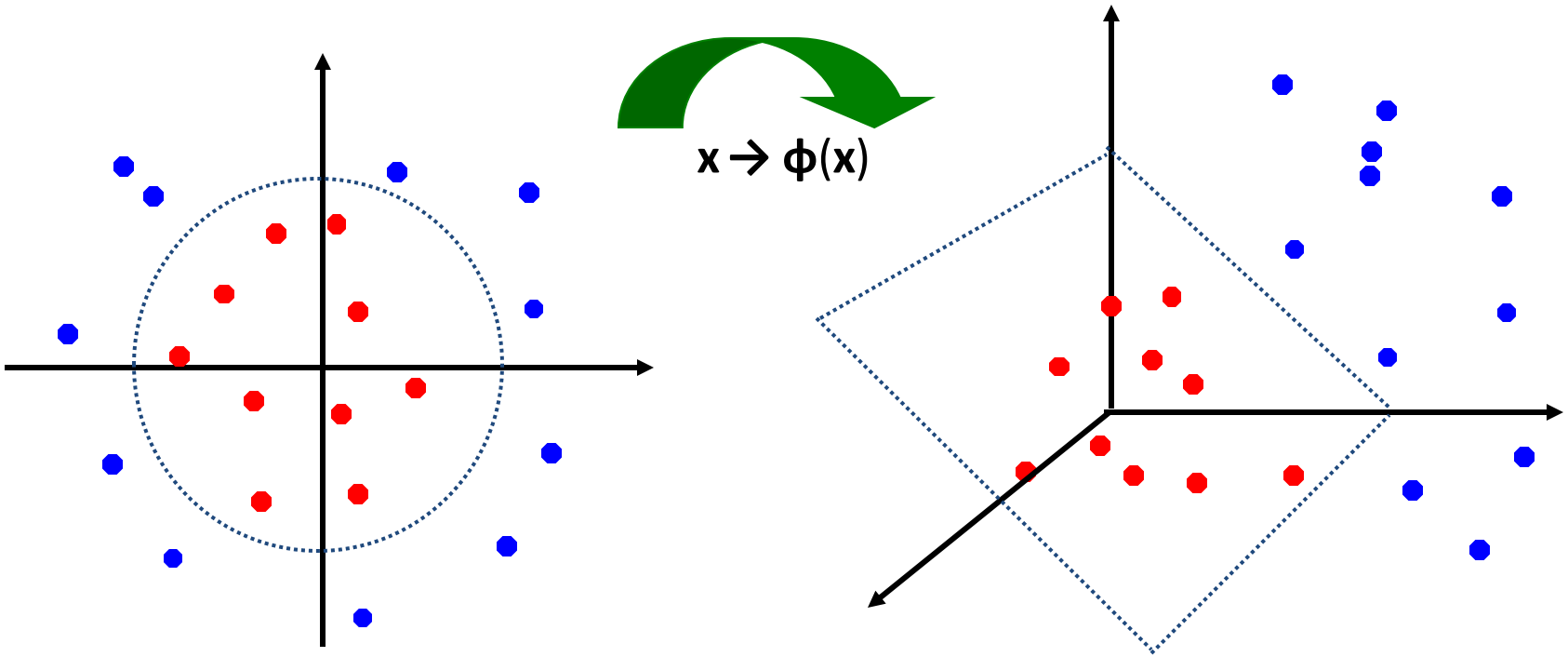
Support Vector Machines

- r: distance from each sample to the separator
- Samples closest to the hyperplane are the support vectors
- ρ : the margin (distance) between support vectors



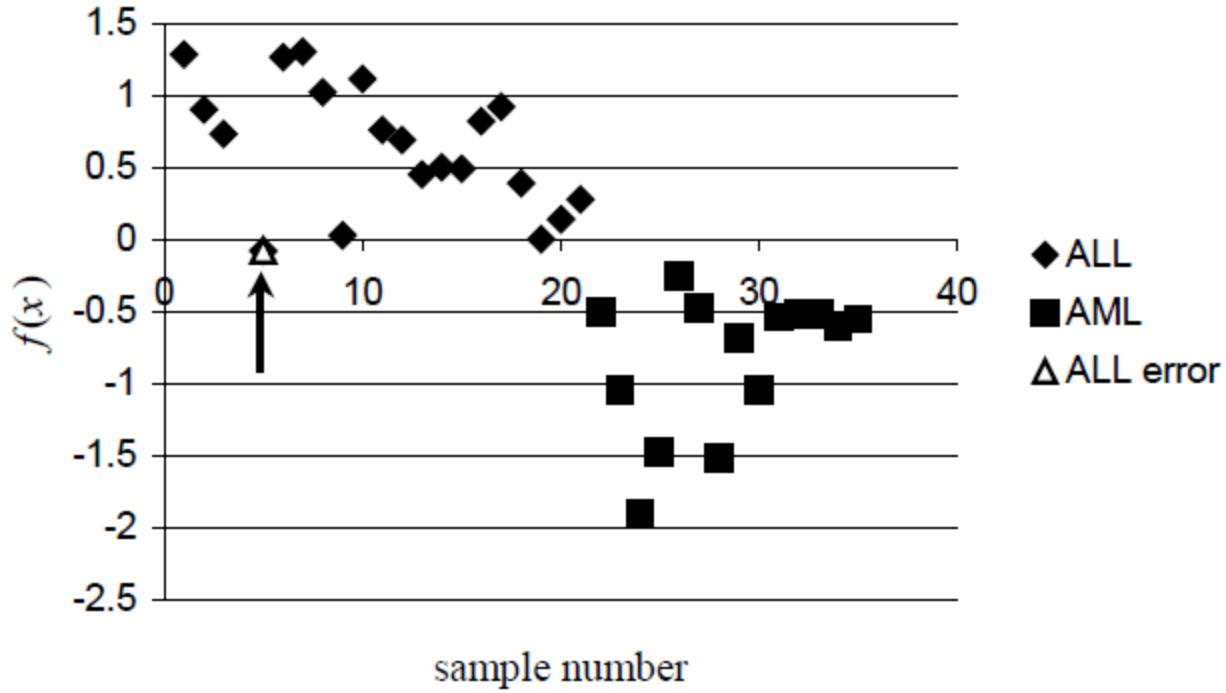
Non-linear SVM (Kernel trick)

- Map to a higher dimension so the classes become separable.



SVM application

- Mukherjee '03 applied SVM to Leukemia data from Golub '99.
  - f(x) is the distance to separating plane

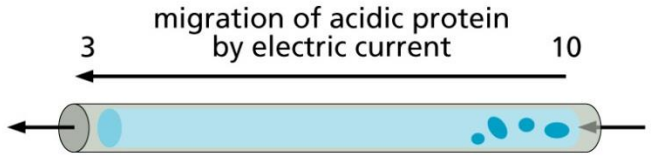


Other types of expression

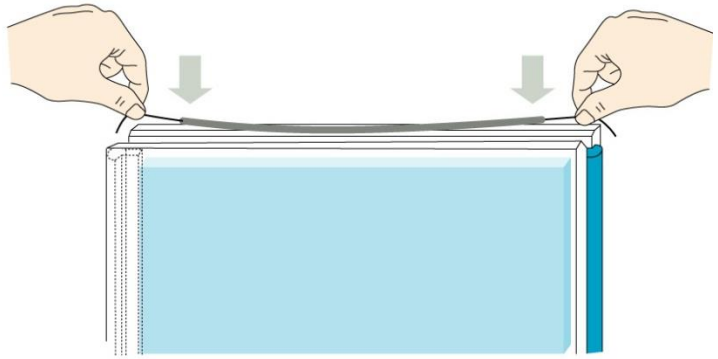
- micro-RNA
  - small (~23) regulatory RNA molecules
  - quantitative RT-PCR
- Protein
  - Gel electrophoresis
  - Liquid chromatography
  - Mass spectrometry

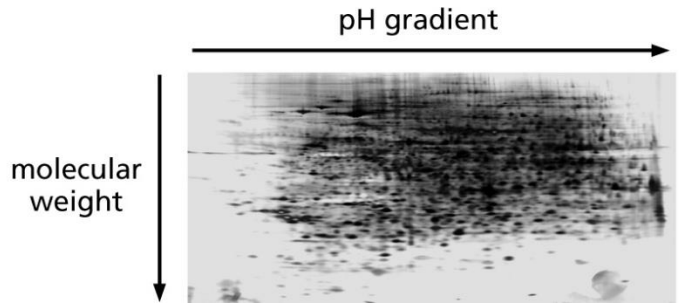
2D gel electrophoresis

(A) first dimension electrophoresis separation



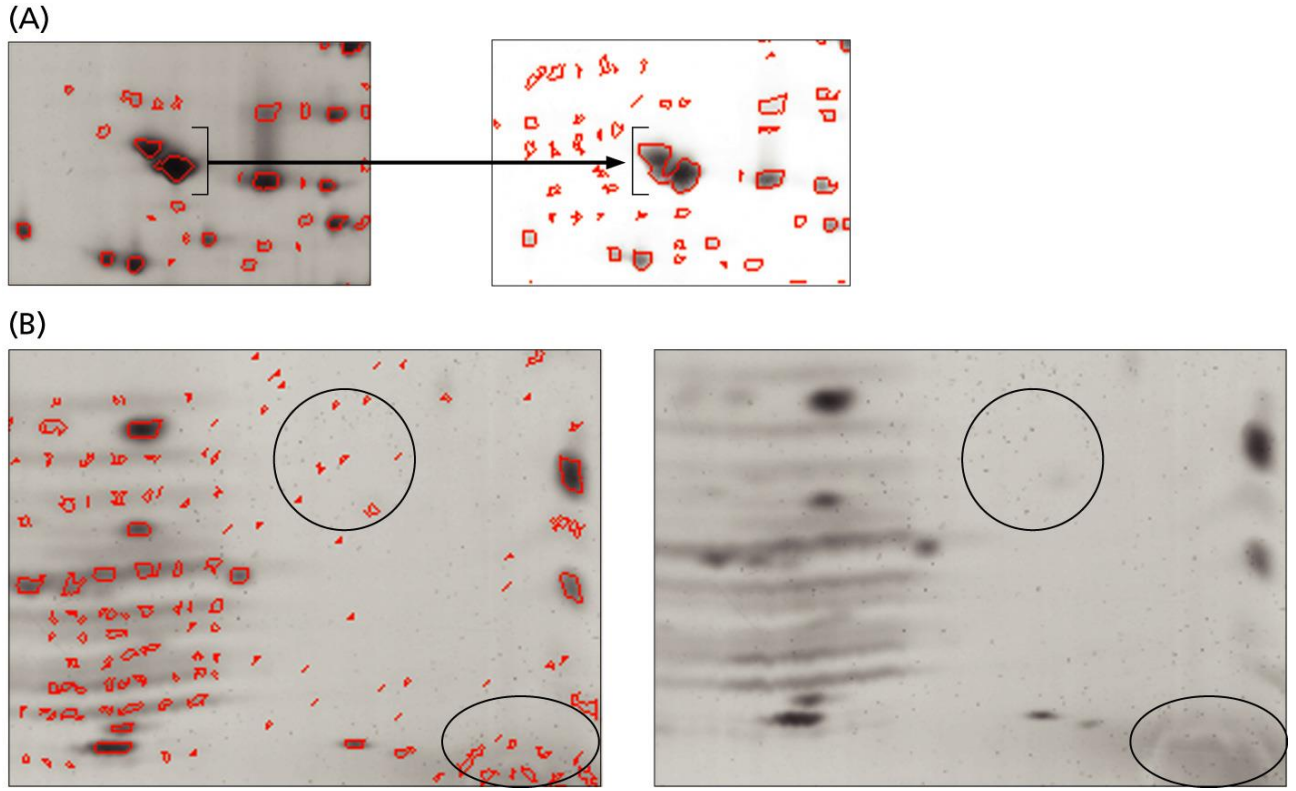
(B) second dimension separation



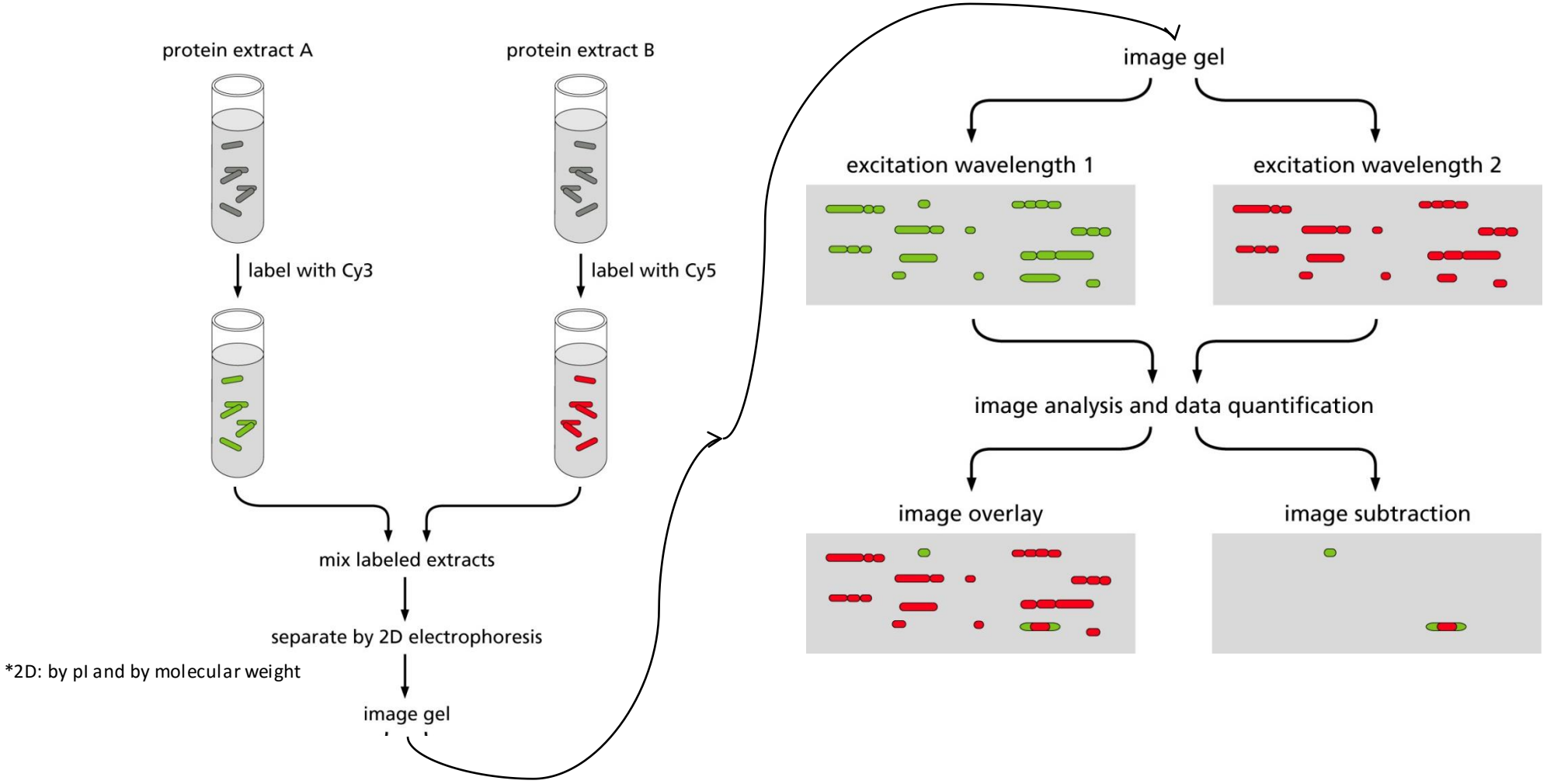


2D gels

- Spot-identification can be problematic
  - (A) two spots detected as one.
  - (B) dust and smudge



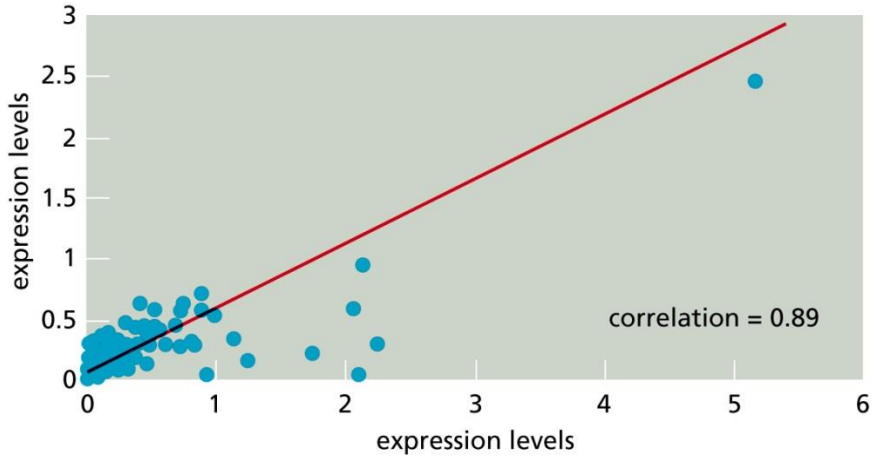
2-color 2D gel electropheresis



*2D: by pI and by molecular weight

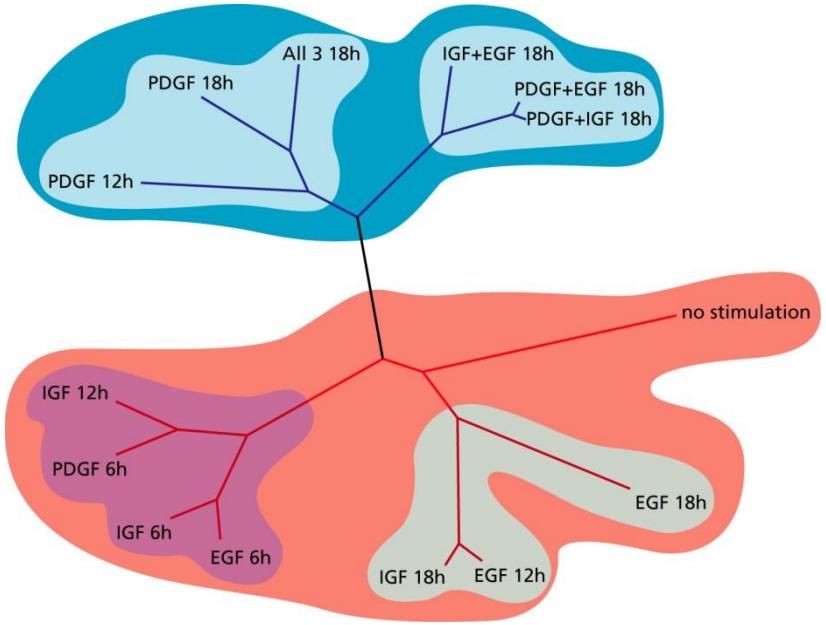
Example: Growth Factor treatment

- Cells stimulated with different growth factors
  - Epidermal growth factor (EGF)
  - Insulin growth factor (IGF)
  - Platelet derived growth factor (PDGF)
- Scatterplot and regression line helps compare 2 (or 3) samples
  - Outlier proteins = very different expression between samples.



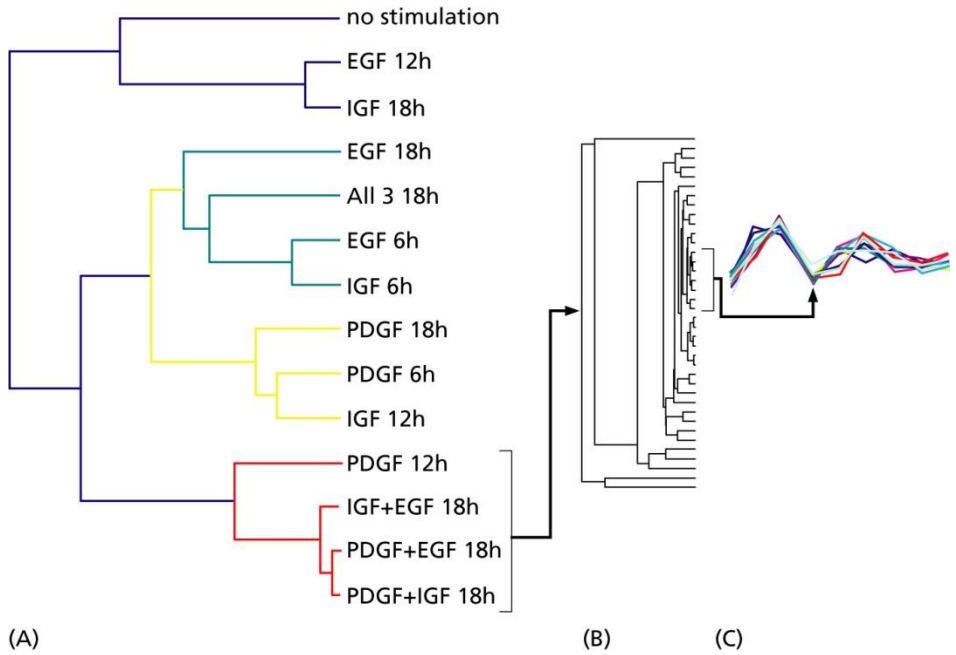
Clustering

- Two groups are identified
  - Red: similar to no stimulation
  - Blue: Longer duration of stimulation, or multiple growth factors

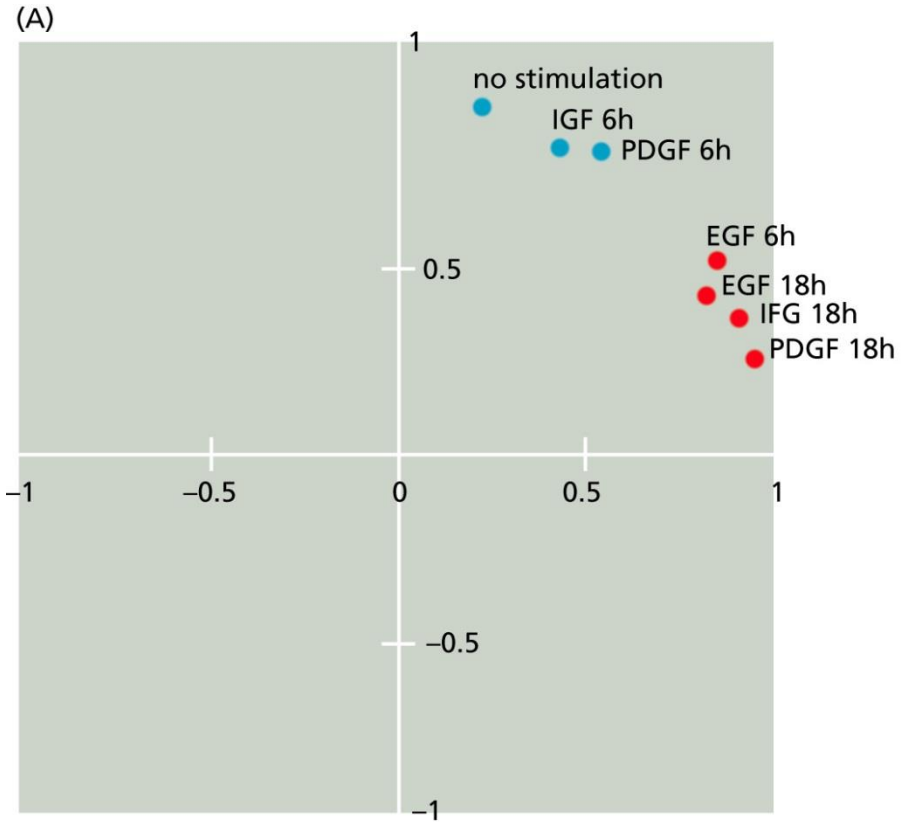


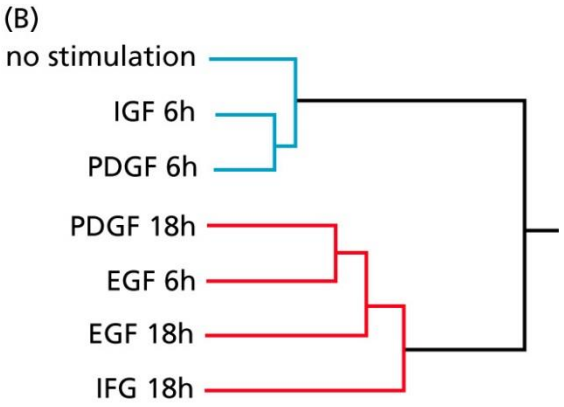
Re-clustering

- (A) Re-cluster samples using a subset of the spots
- (B) Cluster proteins from a sub-group of samples
- (C) Analyze the individual expression patterns.



Principal Component Analysis





Mass Spectrometry

- MS gives mass-charge ratio of each ion fragment, from which peptide mass can be calculated.
- Identifying protein(s) that could've produced these peptides is the computational challenge.
  - Mutations and post-translational modifications need to be handled.