Analysis of Microarray Data with Methods from Machine Learning and Network Theory

Summer Lecture 2015

Prof. Dr. A. B. Cremers
Dr. Jörg Zimmermann

Overview of Lecture

1. Introduction to Microarray Data, Machine Learning, and Network Theory: What are the possibilities and problems?

2. Basic notions and methods from Statistics: parameter estimation, unbiasedness, consistency, ...

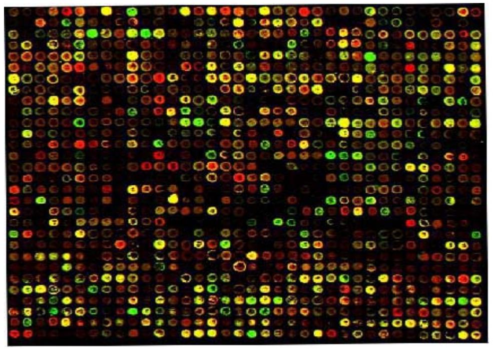
3. Basic notions and methods from Machine Learning: supervised and unsupervised learning, classification, overlearning, validation, ...

4. Application of statistics and ML to Microarray Data: Clustering and Prediction

5. Network Theory: How to extract and analyse biologically relevant networks from microarray data?

DNA Microarray Data

- Genome Chips containing a collection of microscopic DNA spots
- Simultaneous determination of > 10^5 Gene Expression Levels
- Dramatic acceleration of data aquisition
- New possibilities for disease diagnosis, treatment studies, network analysis, …



DNA Microarray Data

The resulting data have the following format:

$$\begin{array}{lllll} x_{11} & x_{12} & \ldots & x_{1p} & (L_1 \ \ldots \ L_p) \\ \cdot & & & \cdot & \\ \cdot & & & \cdot & \\ x_{n1} & x_{n2} & \ldots & x_{np} & \end{array}$$

n = number of measured cell states (e.g. gene expression levels)

p = number of samples

x_{ij} = real number, e.g. representing expression level of gene *i* in sample *j*

L_j = Label of sample *j* (e.g. “diseased” / “not diseased”)

Challenges for Data Analysis

- **Cleaning (removing systematic measurement effects)**
- **Dimensionality Reduction**
- **Large sample effects:**

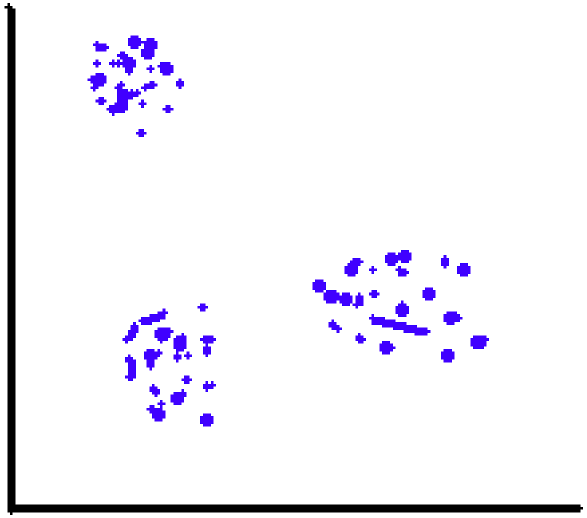
→ **Type I and Type II errors (false positives / false negatives)**

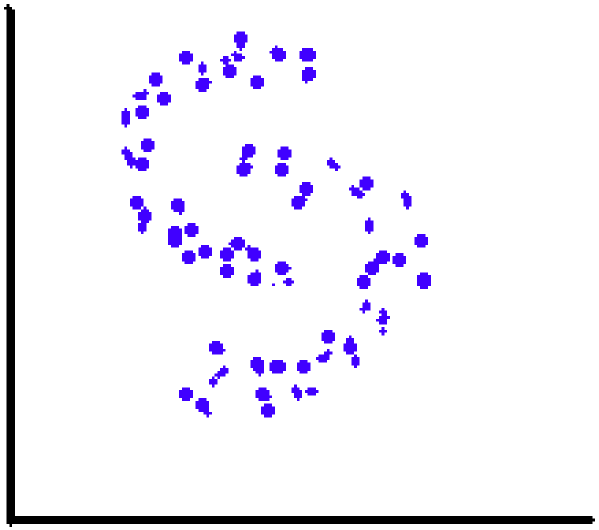
- **Variable Selection (Identification of relevant Variables)**
- **Identification of new disease classes**
- **Classification of data into known disease classes**

Cluster Analysis

Finding structure in data without labels:

→ **Unsupervised Learning**





Does a cluster characterize a (new) disease type?

Prediction Problem (Classification)

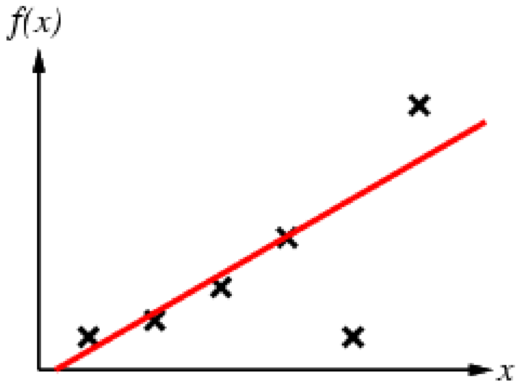
- Classify data into known disease classes:

→ **Supervised Learning**

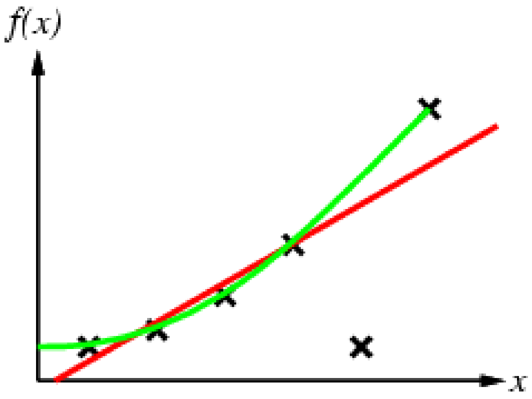
- Split data in Training and Test set
- Learn a model on the training set
- Validate model on the test set
- If validation was successful, use model to predict disease classes on new data sets

Prediction Problem (Regression)

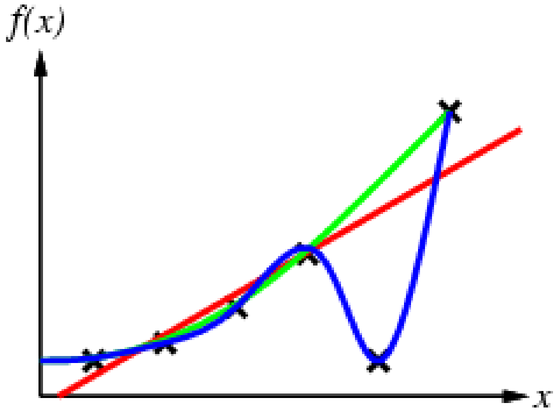
Under- and Overlearning: the problem of generalization



Underfitting



Good Fit



Overfitting

Data Analysis Methods

Dimension Reduction

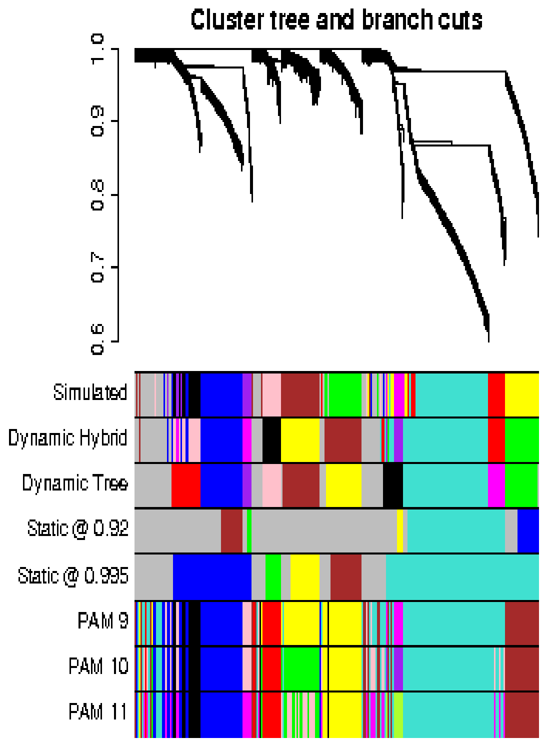
- PCA (Principle Component Analysis)
- ICA (Independent Component Analysis)
- Multidimensional Scaling

Unsupervised Learning

- K-Means / K-Medoid
- Hierarchical Clustering Algorithms

Supervised Learning

- Linear Discriminant Analysis
- Maximum Likelihood Discrimination
- Nearest Neighbor Methods
- Decision Trees
- Random Forests
- Bayesian Networks



Why so many methods?

- There are many different questions one can ask
- One needs model assumptions in order to get meaningful results (Futility of bias free learning)
- Approximative methods for reasons of computational efficiency
- There are still many open foundational questions in statistics and machine learning (quantification of uncertainty, incorporation of prior knowledge, ...)
- Therefore empirical validation of methods is of great importance

Gene Network Analysis

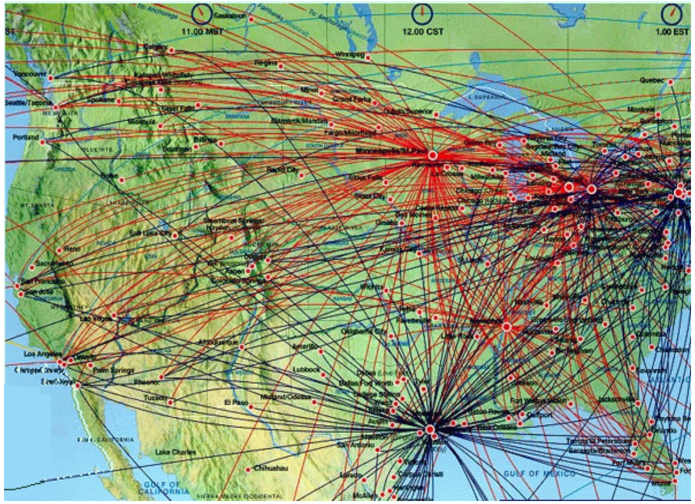
- Understand the “system” instead of reporting a list of individual parts
- Focus on modules as opposed to individual genes
  - this greatly alleviates meaningful biological interpretations
- Network terminology is intuitive to biologists

The Network Perspective

Does this map tell you which cities are important?

St. Paul
Sioux Falls
Milwaukee
Buffalo
Detroit
Cedar Rapids
Chicago
Omaha
Cleveland
Peoria
Indianapolis
Philadelphia
Kansas City
St. Louis
Washington
Wichita
Lexington-fayette
Richmond
Tulsa
Nashville-davidson
Durham
Memphis
Chattanooga
Charlotte
Birmingham
Dallas
Shreveport
Jackson
Columbus
Savannah

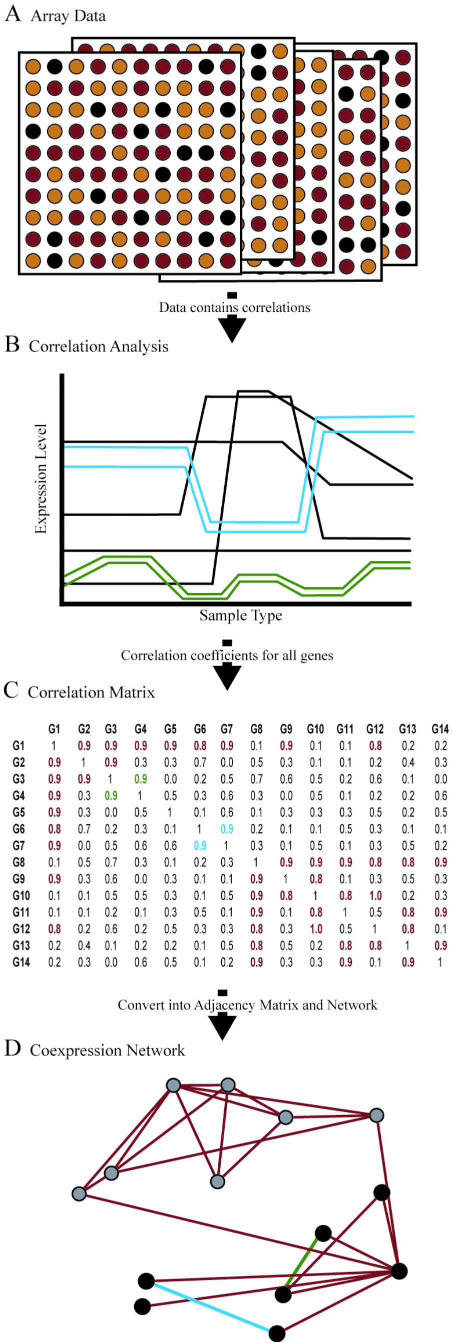
This one does!



The nodes with the largest number of links (connections) are most important!

Slide of Paul Mischel and AL Barabasi

Figure 1


A Array Data
Data contains correlations
B Correlation Analysis
Expression Level
Sample Type
Correlation coefficients for all genes
C Correlation Matrix
Convert into Adjacency Matrix and Network
D Coexpression Network

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.9	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

Network Inference: Steps for constructing a correlation network

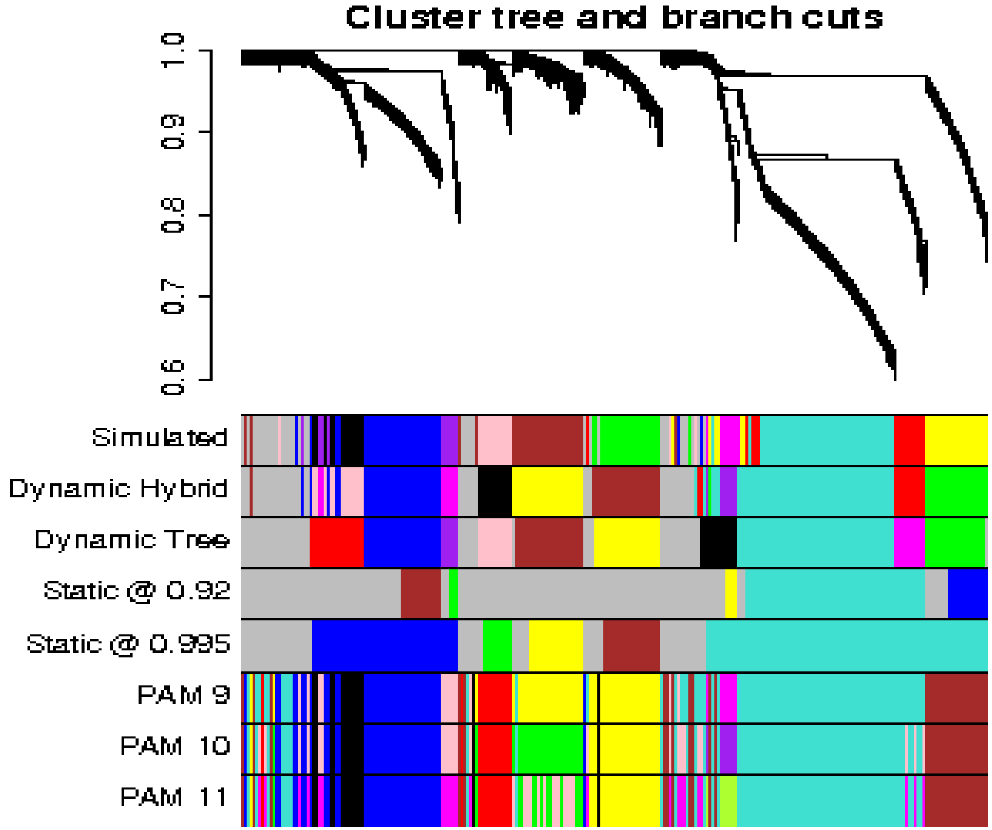
A) Gene expression data

B) Measure concordance of gene expression with a Pearson correlation

C) The Pearson correlation matrix is either dichotomized to arrive at an unweighted adjacency matrix → unweighted network

Or transformed continuously with the power adjacency function → weighted network

Network Analysis: Gene Modules as branches of a cluster tree

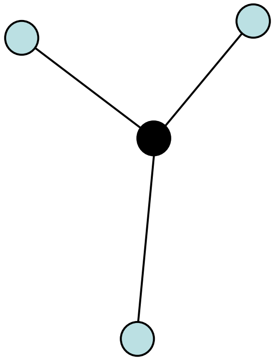


Module=branch of a cluster tree

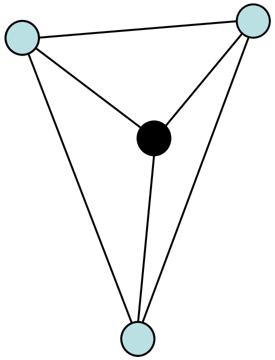
Module genes are assigned the same color

Network Concepts: Clustering Coefficient

Measures the cliquishness of a particular node:
A node is cliquish if its neighbors know each other



Clustering Coef of the black node = 0

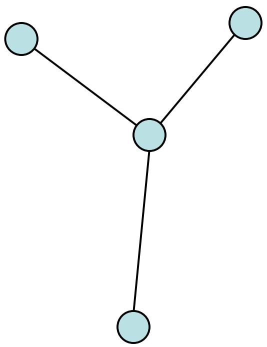


Clustering Coef = 1

Network Concepts: Centralization

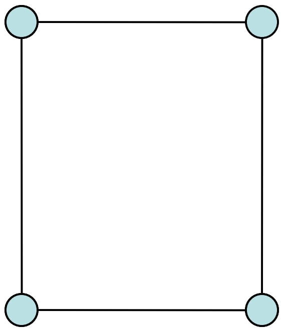
Centralization = 1 if the network has a star topology

= 0 if all nodes have the same connectivity



Centralization = 1

because it has a star topology



Centralization = 0

because all nodes have the same connectivity of 2

Network Science

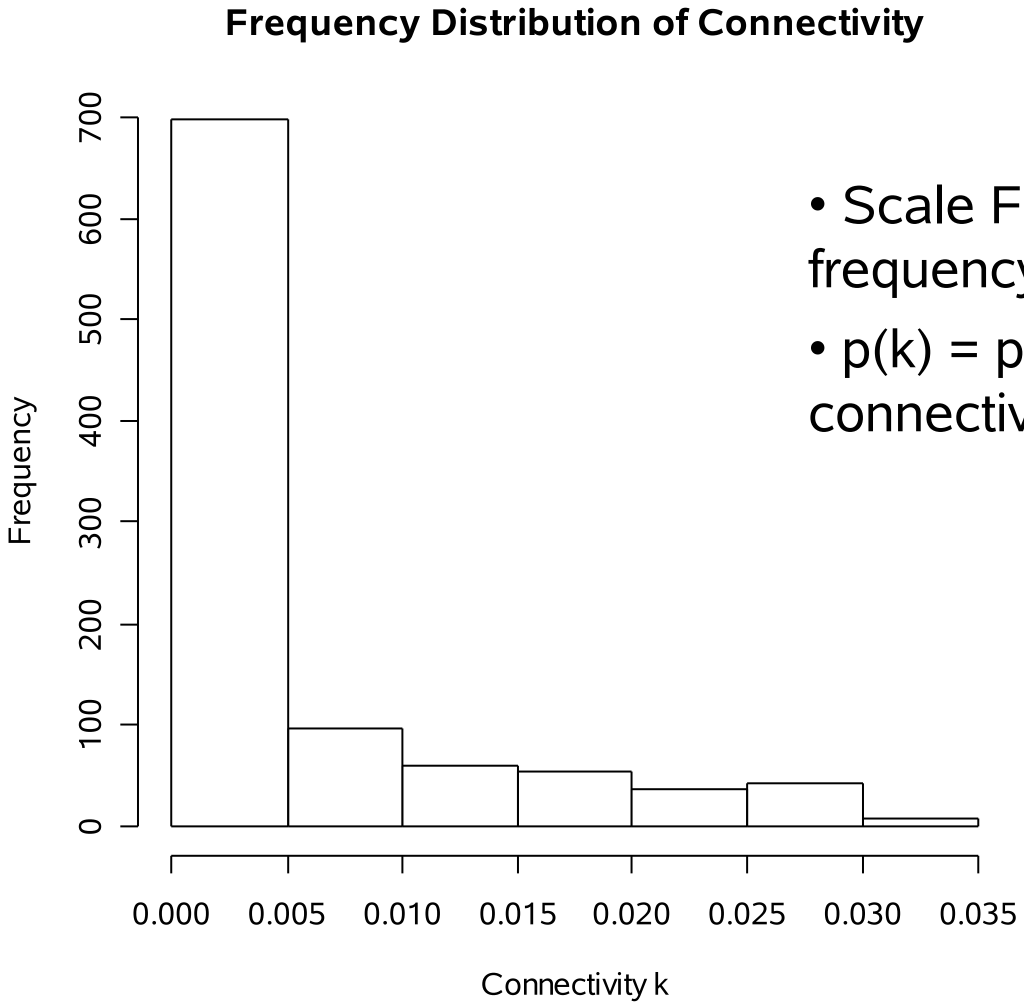
- **Network based methods have been found useful in many domains:**
  - protein interaction networks
  - the world wide web
  - social interaction networks
  - OUR FOCUS: genetic networks

Network Concepts: Scale free topology

SFT is a fundamental property of many real world networks:

- It entails the presence of hub nodes that are connected to a large number of other nodes
- Such networks are robust with respect to the random deletion of nodes
- It has been demonstrated that metabolic networks exhibit scale free topology at least approximately.

P(k) vs k in scale free networks



- Scale Free Topology refers to the frequency distribution of the connectivity k
- p(k) = proportion of nodes that have connectivity k

Literature:

The Elements of Statistical Learning: Data Mining, Inference, and Prediction
T. Hastie, R. Tibshirani, J. Friedman, Springer, 2001

Pattern Recognition and Machine Learning
C. M. Bishop, Springer, 2006

Computational Genome Analysis
R. Deonier, S. Tavare, M. Waterman, Springer, 2005

The Structure and Dynamics of Networks
M. Newman, A.-L. Barabasi, D. J. Watts, Princeton University Press, 2006

Weighted Network Analysis - Applications in Genomics and Systems Biology
S. Horvath, Springer, 2011

All of Statistics - A Concise Course in Statistical Inference (Winner of the 2005 DeGroot Prize)
L. Wasserman, Springer, 2004

Download of Lecture Slides

Lecture slides can be downloaded from the following webpage:

http://www.informatik.uni-bonn.de/~jz/lectures/mada2015.html

Username: mada2015

Password: regnos

If you encounter difficulties, please send an email to:

jz@iai.uni-bonn.de