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

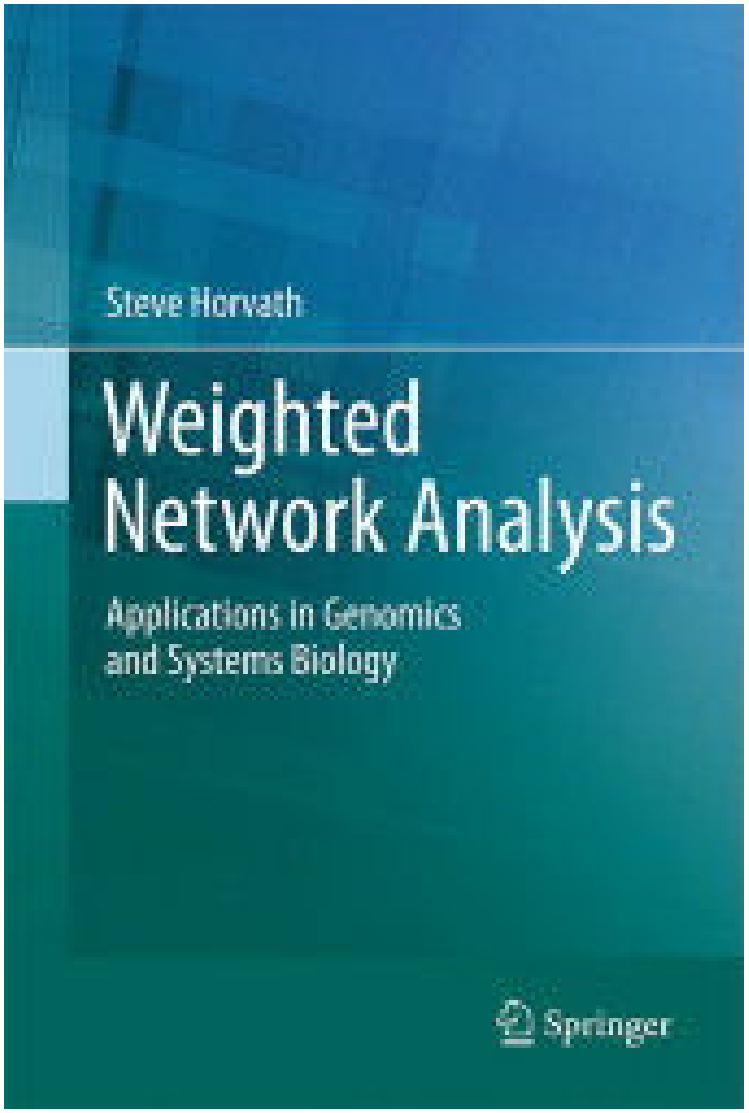
Summer Lecture 2015

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

Weighted Network Analysis

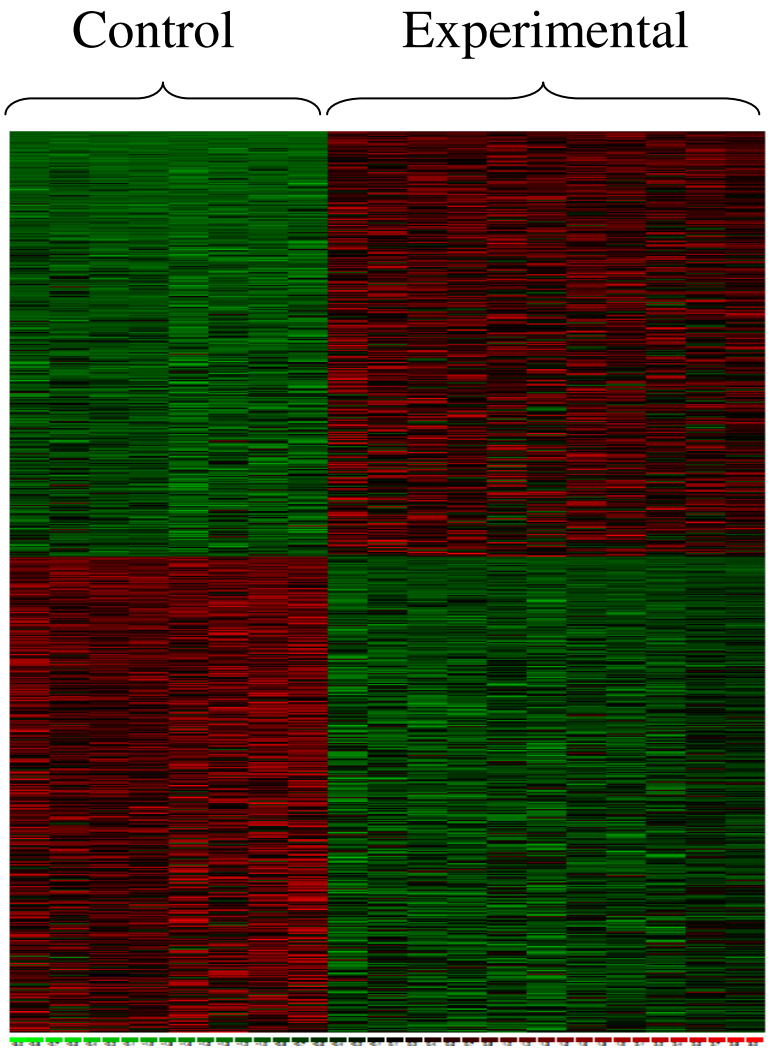
- How to construct a weighted gene co-expression network?
- Why use soft thresholding?
- How to detect network modules?
- How to relate modules to an external clinical trait?
- What is intramodular connectivity?
- How to use networks for gene screening?
- How to integrate networks with genetic marker data?
- What is weighted gene co-expression network analysis (WGCNA)?

Book on weighted networks



Also available as E-book.

Standard microarray analyses seek to identify ‘differentially expressed’ genes



- Each gene is treated as an individual entity
- Often misses the forest for the trees: Fails to recognize that thousands of genes can be organized into relatively few modules

Philosophy of Weighted Gene Co-Expression Network Analysis

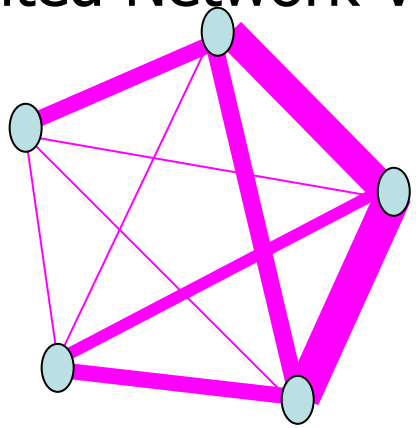
- Understand the “system” instead of reporting a list of individual parts
  - Describe the functioning of the engine instead of enumerating individual nuts and bolts
- Focus on modules as opposed to individual genes
  - this greatly alleviates multiple testing problem
- Network terminology is intuitive to biologists

Network=Adjacency Matrix

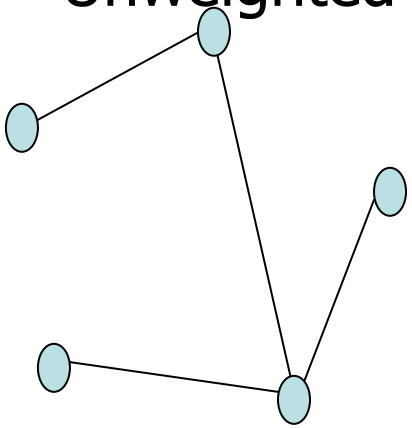
- A network can be represented by an adjacency matrix, $A=[a_{ij}]$, that encodes whether/how a pair of nodes is connected.
  - A is a symmetric matrix with entries in [0,1]
  - For <u>unweighted</u> network,
    - entries are either 1 or 0
      - Encoding the presence of a link (edge) between nodes
  - For <u>weighted</u> networks,
    - entries are real numbers
    - reports the connection strength between nodes

`Holistic' view of a weighted network

Weighted Network View

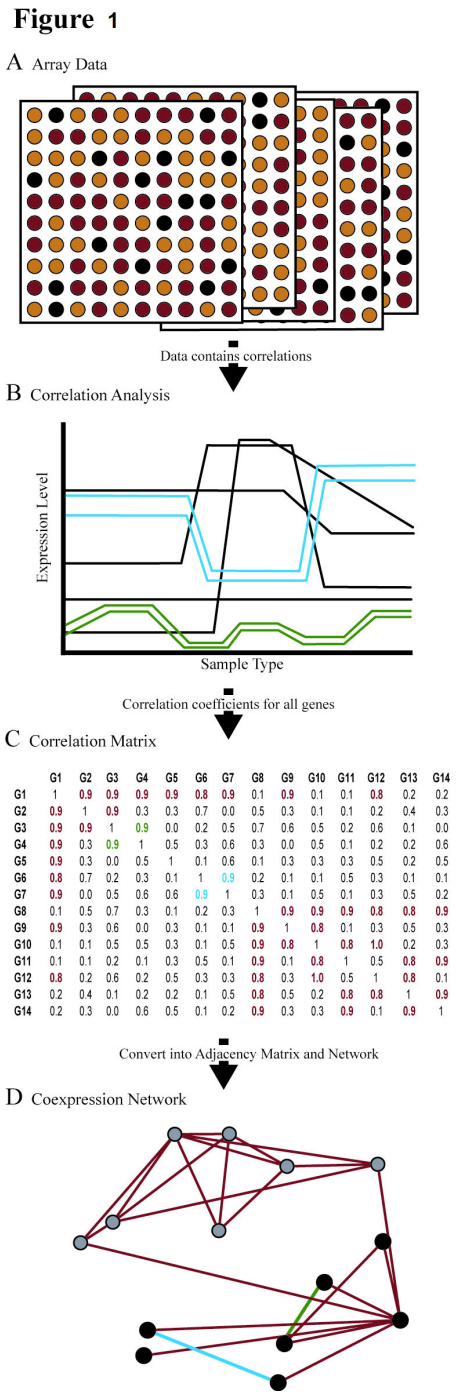


Unweighted View



- All nodes are connected
- Connection Widths=Connection strenghts

Some nodes are connected
All connections are equal



Steps for constructing a co-expression network

A) Microarray 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 adjacency matrix → unweighted network

Or transformed continuously with the power adjacency function → weighted network

Power adjacency function for constructing unsigned and signed weighted gene co-expr. networks

Unsigned network, absolute value

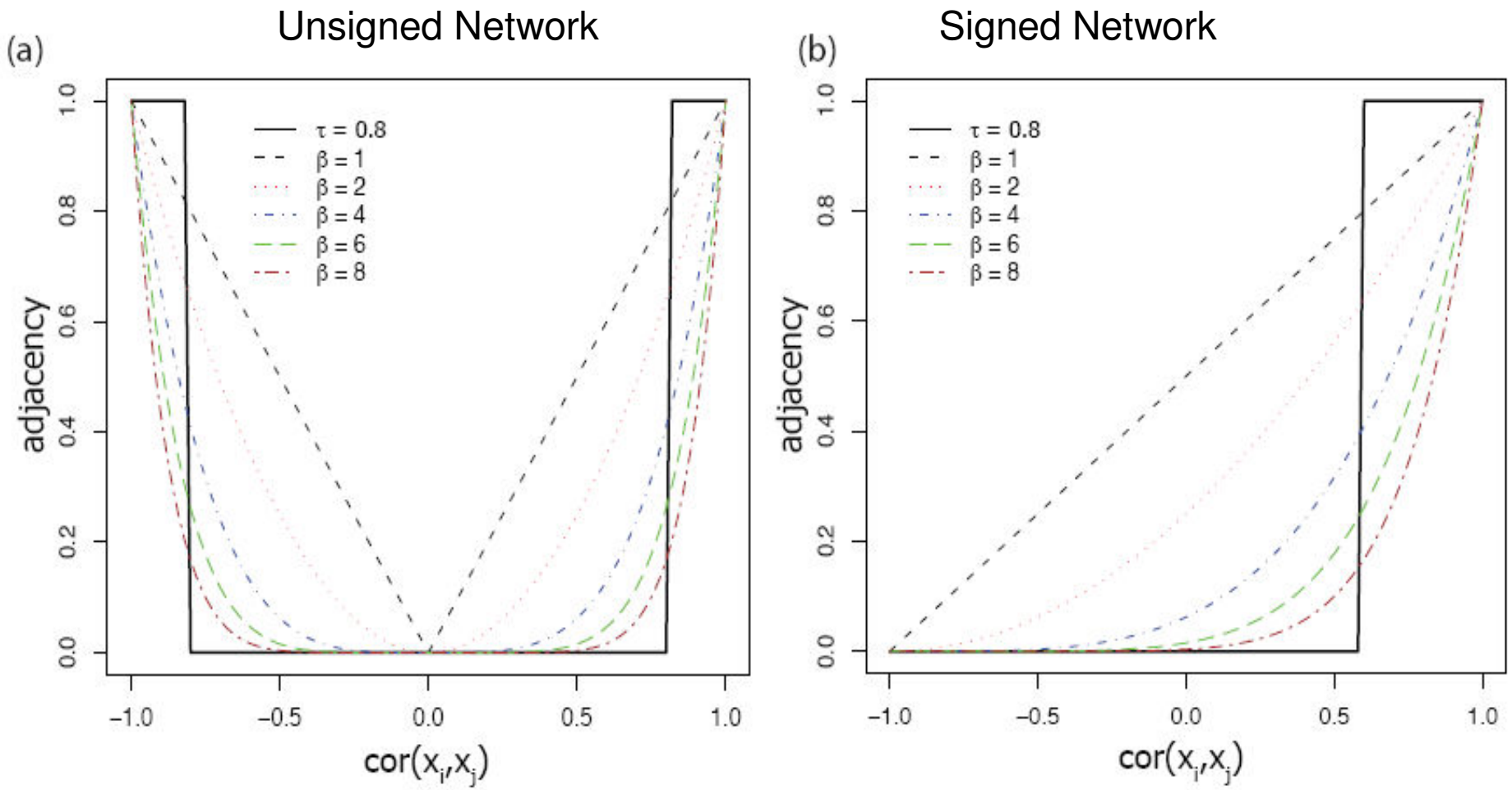
$$a_{ij} = | cor(x_i, x_j) |^{\beta}$$

Signed network preserves sign info

$$a_{ij} = | 0.5 + 0.5 \times cor(x_i, x_j) |^{\beta}$$

Default values: beta=6 for unsigned and beta=12 for signed networks.

Comparing adjacency functions for transforming the correlation into a measure of connection strength



Why construct a co-expression network based on the correlation coefficient ?

1. Intuitive
2. Measuring linear relationships avoids the pitfall of overfitting
3. Because many studies have limited numbers of arrays→ hard to estimate non-linear relationships
4. Works well in practice
5. Computationally fast
6. Leads to reproducible research

Why soft thresholding as opposed to hard thresholding?

1. Preserves the continuous information of the co-expression information
2. Results tend to be more robust with regard to different threshold choices

But hard thresholding has its own advantages:
In particular, graph theoretic algorithms from the computer science community can be applied to the resulting networks

Questions:
How should we choose the power beta or a hard threshold?
Or more generally the parameters of an adjacency function?

IDEA: use properties of the connectivity distribution

Connectivity (degree) based on the entire network

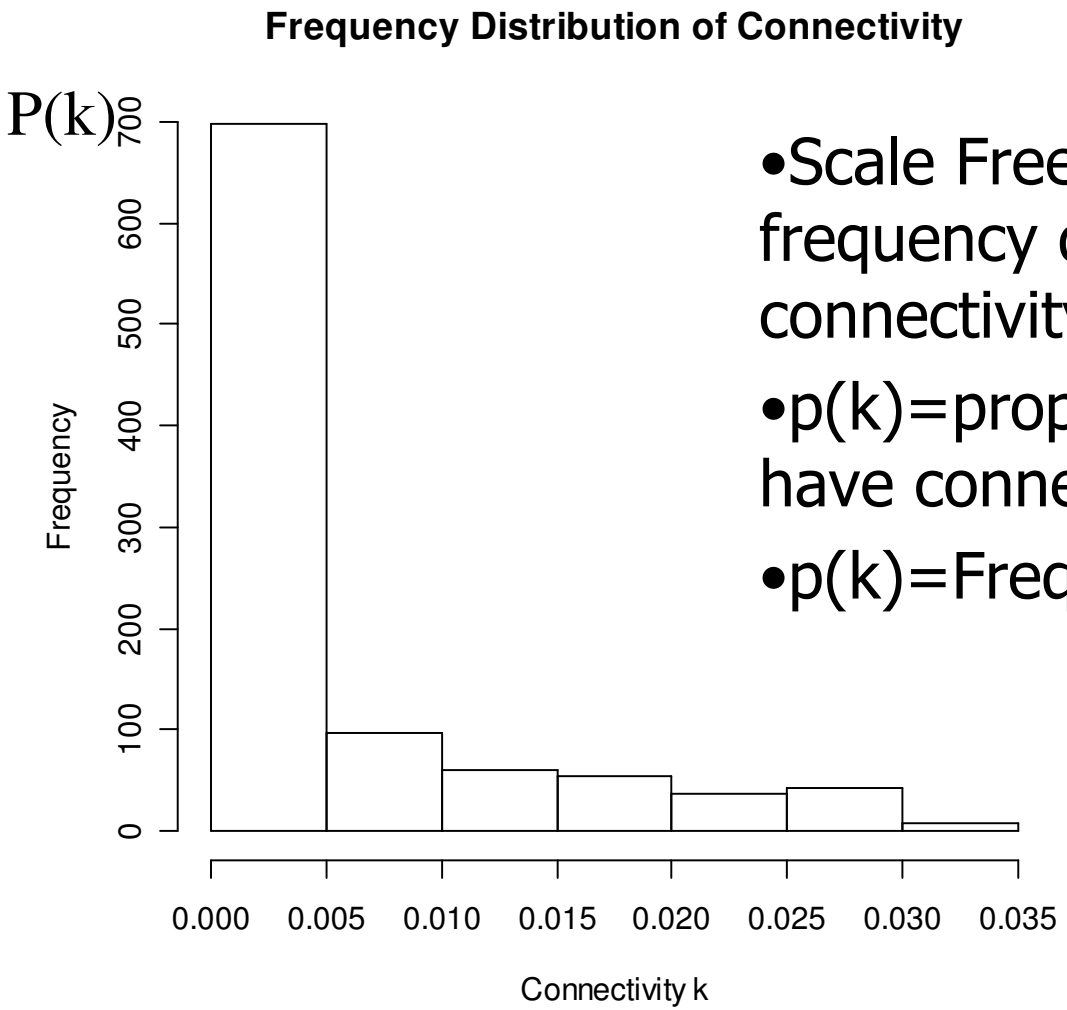
- Gene connectivity = row sum of the adjacency matrix
  - unweighted networks
    - number of direct neighbors
    - “number of friends”
  - weighted networks
    - sum of connection strengths to other nodes

$$k_i = \sum_j a_{ij}$$

Approximate scale free topology is a fundamental property of such networks (Barabasi et al)

- 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 but are sensitive to the targeted attack on hub nodes
- It has been demonstrated that metabolic networks exhibit scale free topology at least approximately.

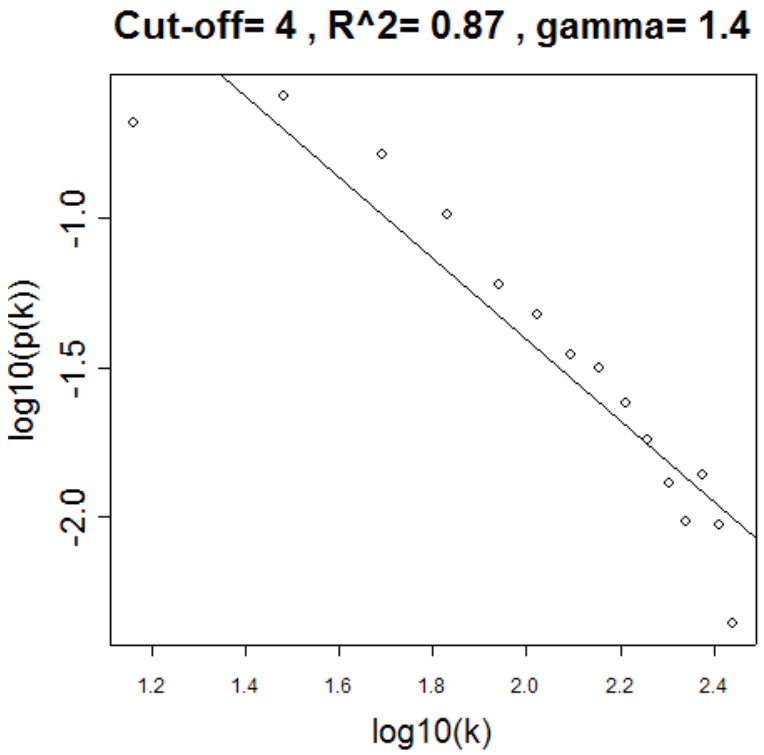
P(k) vs k in scale free networks



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

How to check Scale Free Topology?

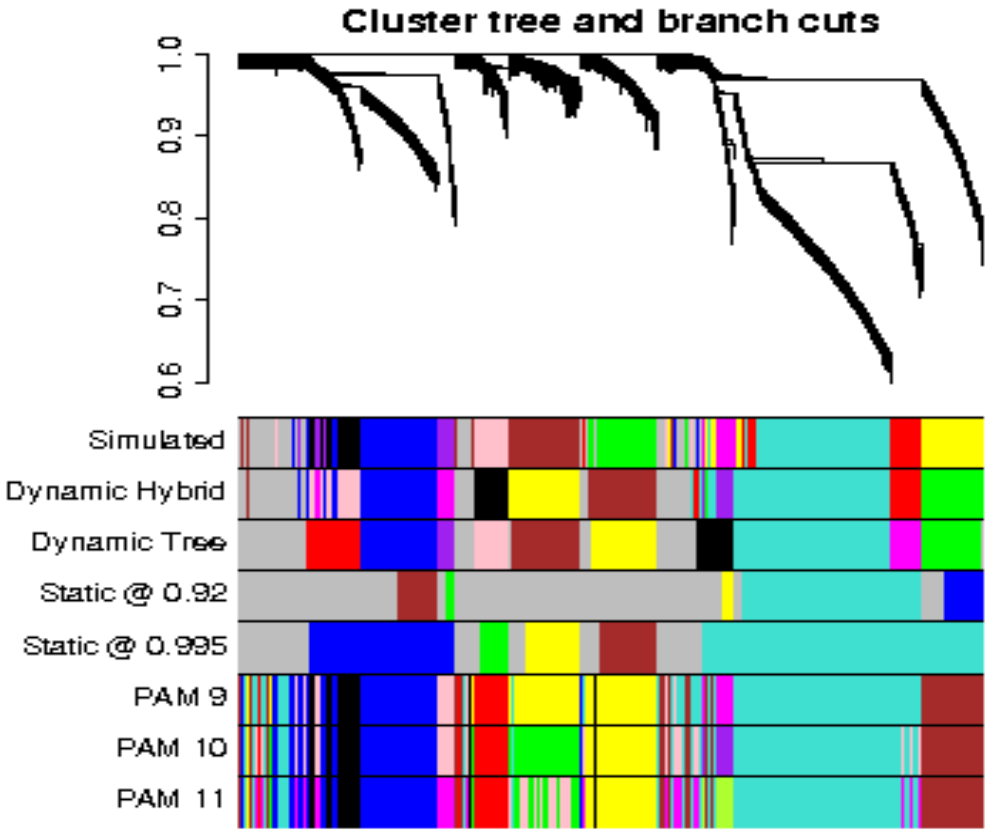
Idea: Log transformation p(k) and k and look at scatter plots



Linear model fitting R^2 index can be used to quantify goodness of fit

How to detect network modules (clusters) ?

How to cut branches off a tree?



Module=branch of a cluster tree

Dynamic hybrid branch cutting method combines advantages of hierarchical clustering and pam clustering

Module Definition

- Numerous methods have been developed
- Average linkage hierarchical clustering coupled with the topological overlap dissimilarity measure has proven to be useful.
- Once a dendrogram is obtained from a hierarchical clustering method, choose a height cutoff to arrive at a clustering.
- Modules correspond to branches of the dendrogram

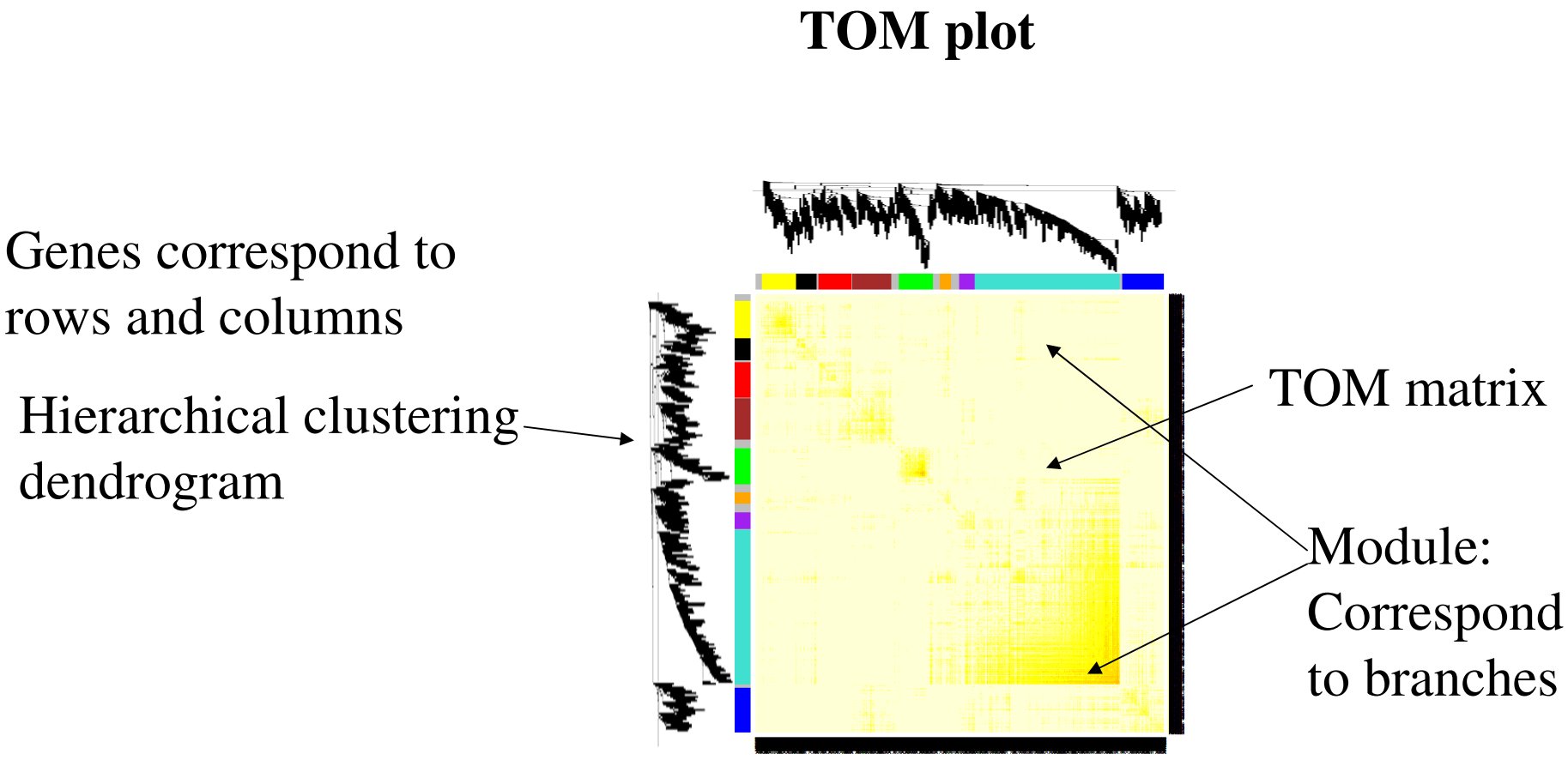
The topological overlap dissimilarity is used as input of hierarchical clustering

$$TOM_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

$$DistTOM_{ij} = 1 - TOM_{ij}$$

Using the topological overlap matrix (TOM) to cluster genes

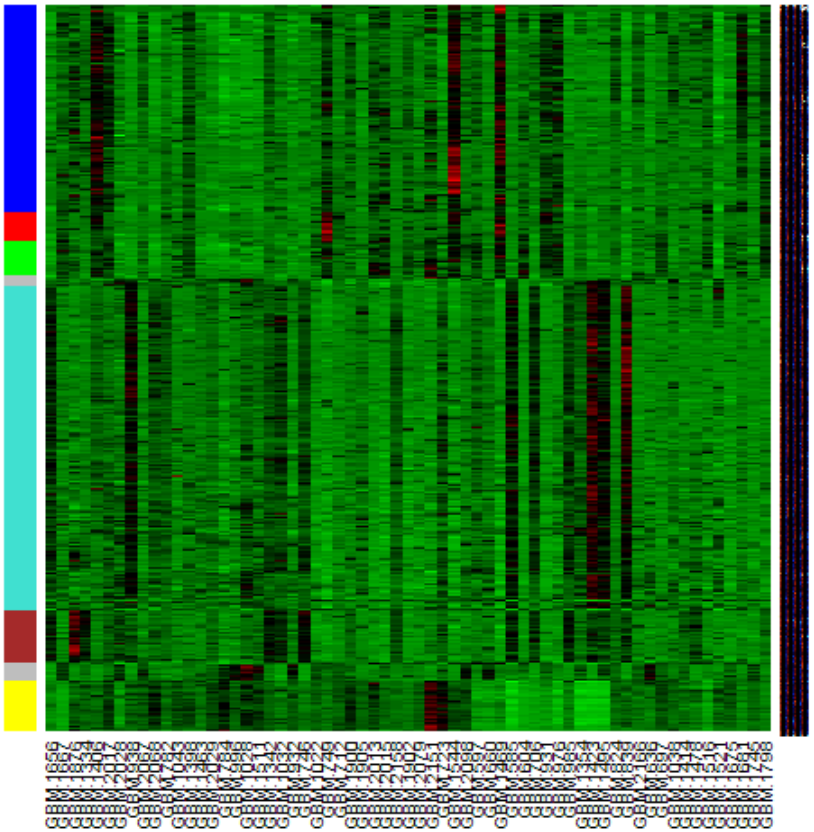
- Here modules correspond to branches of the dendrogram



Heatmap view of module

Columns= tissue samples

Rows=Genes
Color band indicates
module membership



Message: characteristic vertical bands indicate tight co-expression of module genes