DUBii Bioinformatique Intégrative - Université de Paris- IFB
Module 6 - Bioinformatique Intégrative

Correlation Networks and Inference of Gene Regulatory Networks

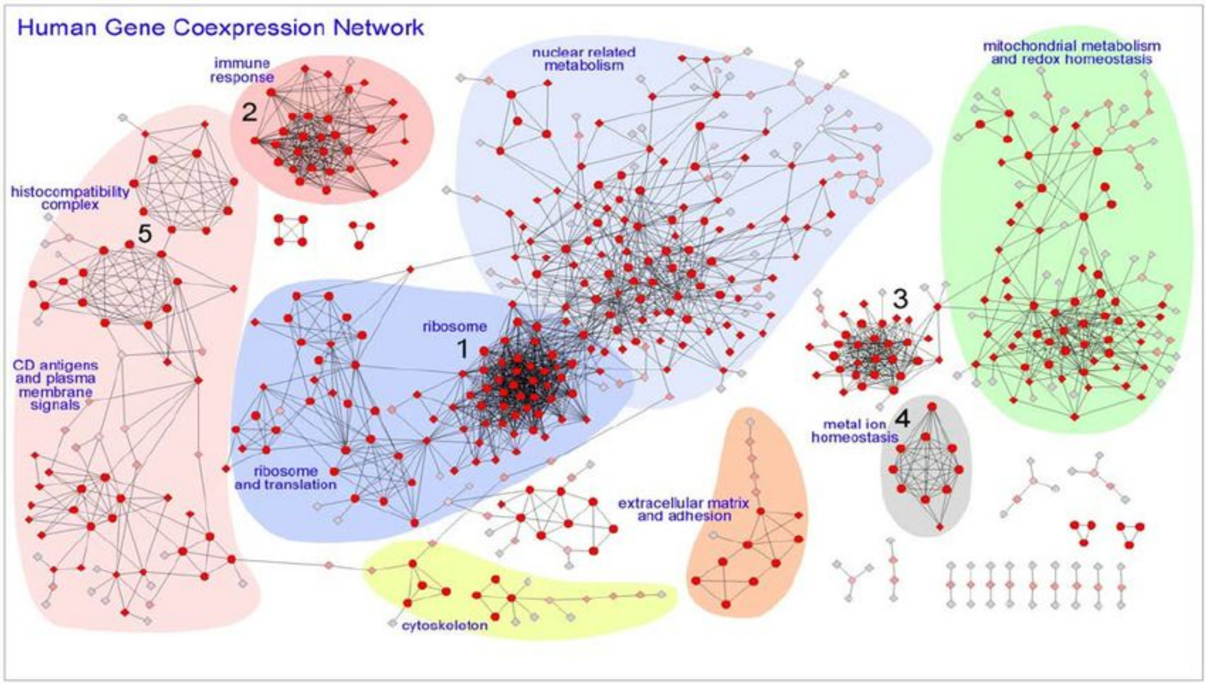
•••

Costas Bouyioukos, Université de Paris, EDC-UMR7216
Anaïs Baudot, CNRS Marseille

Outline

1.

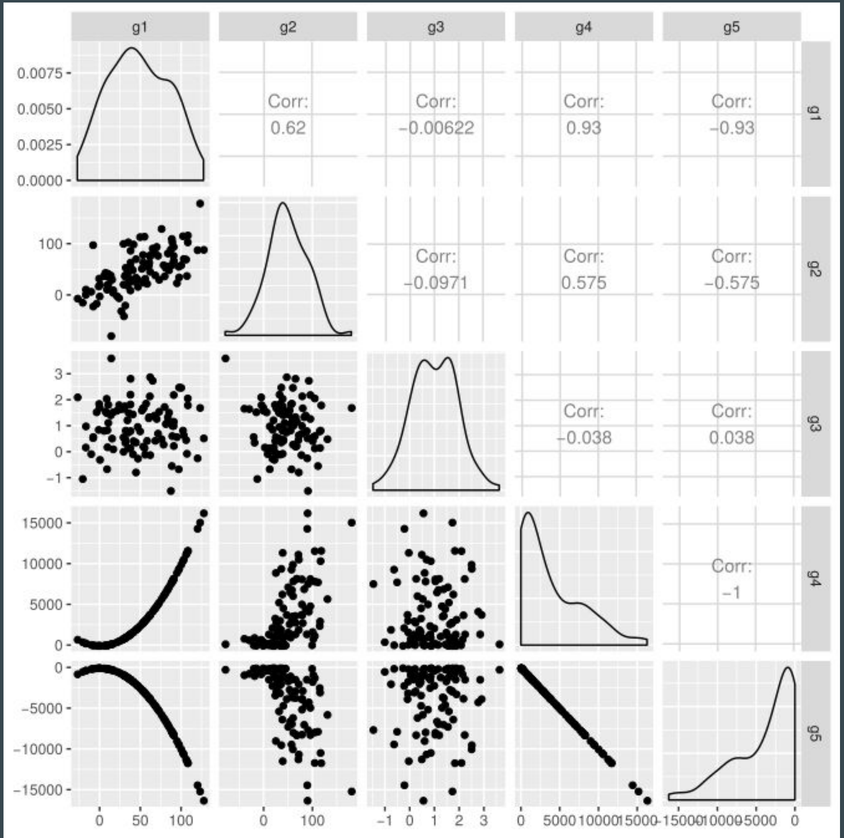
Biological motivation


Human Gene Coexpression Network
immune response
2
nuclear related metabolism
mitochondrial metabolism and redox homeostasis
histocompatibility complex
5
ribosome
1
3
CD antigens and plasma membrane signals
metal ion homeostasis
4
ribosome and translation
extracellular matrix and adhesion
cytoskeleton

- Find co-expressed genes (by using the correlation of their gene expression profiles) this is a proxy for possible co-regulation.
- Find co-functional modules (clusters) this is a proxy for common function regarding a given phenotype.

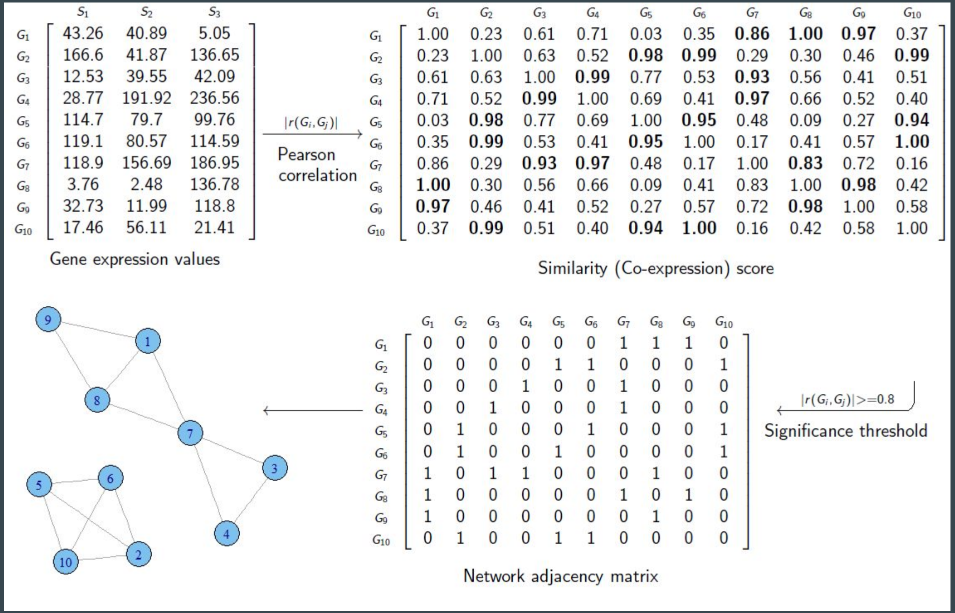
Gene Correlation matrices

- Correlation -> Is a measure of similarity between gene expression profiles.
- For a gene expression matrix, if we compute the correlation coefficient of all vs. all gene expression profiles we obtain the expression correlation matrix.
- Correlation is only ONE measure of similarity, we can use different coefficients (Pearson, Spearman etc.) and different measures (MI etc.).



From matrices to correlation Networks

- A correlation matrix represents a (fully connected) weighted network
- Correlation sign (+/-) represents the type of association between biological entities.
- However a fully connected network is never useful. Solution:
    - Apply a threshold, dichotomise (discretise) the matrix.
- Then construct the adjacency matrix (which is the network)



$$
\begin{array}{c|ccc}
 & S_1 & S_2 & S_3 \\
\hline
G_1 & 43.26 & 40.89 & 5.05 \\
G_2 & 166.6 & 41.87 & 136.65 \\
G_3 & 12.53 & 39.55 & 42.09 \\
G_4 & 28.77 & 191.92 & 236.56 \\
G_5 & 114.7 & 79.7 & 99.76 \\
G_6 & 119.1 & 80.57 & 114.59 \\
G_7 & 118.9 & 156.69 & 186.95 \\
G_8 & 3.76 & 2.48 & 136.78 \\
G_9 & 32.73 & 11.99 & 118.8 \\
G_{10} & 17.46 & 56.11 & 21.41
\end{array}
$$

Gene expression values

$$\xrightarrow[\text{Pearson correlation}]{|r(G_i, G_j)|}$$

$$
\begin{array}{c|cccccccccc}
 & G_1 & G_2 & G_3 & G_4 & G_5 & G_6 & G_7 & G_8 & G_9 & G_{10} \\
\hline
G_1 & 1.00 & 0.23 & 0.61 & 0.71 & 0.03 & 0.35 & \mathbf{0.86} & \mathbf{1.00} & \mathbf{0.97} & 0.37 \\
G_2 & 0.23 & 1.00 & 0.63 & 0.52 & \mathbf{0.98} & \mathbf{0.99} & 0.29 & 0.30 & 0.46 & \mathbf{0.99} \\
G_3 & 0.61 & 0.63 & 1.00 & \mathbf{0.99} & 0.77 & 0.53 & \mathbf{0.93} & 0.56 & 0.41 & 0.51 \\
G_4 & 0.71 & 0.52 & \mathbf{0.99} & 1.00 & 0.69 & 0.41 & \mathbf{0.97} & 0.66 & 0.52 & 0.40 \\
G_5 & 0.03 & \mathbf{0.98} & 0.77 & 0.69 & 1.00 & \mathbf{0.95} & 0.48 & 0.09 & 0.27 & \mathbf{0.94} \\
G_6 & 0.35 & \mathbf{0.99} & 0.53 & 0.41 & \mathbf{0.95} & 1.00 & 0.17 & 0.41 & 0.57 & \mathbf{1.00} \\
G_7 & 0.86 & 0.29 & \mathbf{0.93} & \mathbf{0.97} & 0.48 & 0.17 & 1.00 & \mathbf{0.83} & 0.72 & 0.16 \\
G_8 & \mathbf{1.00} & 0.30 & 0.56 & 0.66 & 0.09 & 0.41 & 0.83 & 1.00 & \mathbf{0.98} & 0.42 \\
G_9 & \mathbf{0.97} & 0.46 & 0.41 & 0.52 & 0.27 & 0.57 & 0.72 & \mathbf{0.98} & 1.00 & 0.58 \\
G_{10} & 0.37 & \mathbf{0.99} & 0.51 & 0.40 & \mathbf{0.94} & \mathbf{1.00} & 0.16 & 0.42 & 0.58 & 1.00
\end{array}
$$

Similarity (Co-expression) score

$|r(G_i, G_j)| >= 0.8$

Significance threshold

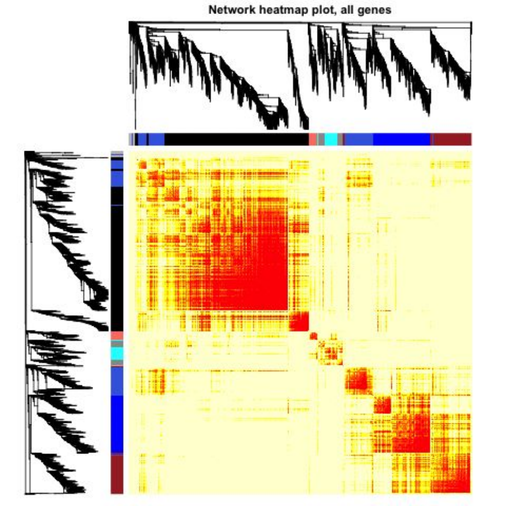
$$
\begin{array}{c|cccccccccc}
 & G_1 & G_2 & G_3 & G_4 & G_5 & G_6 & G_7 & G_8 & G_9 & G_{10} \\
\hline
G_1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
G_2 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\
G_3 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
G_4 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
G_5 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
G_6 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
G_7 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
G_8 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
G_9 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
G_{10} & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0
\end{array}
$$

Network adjacency matrix

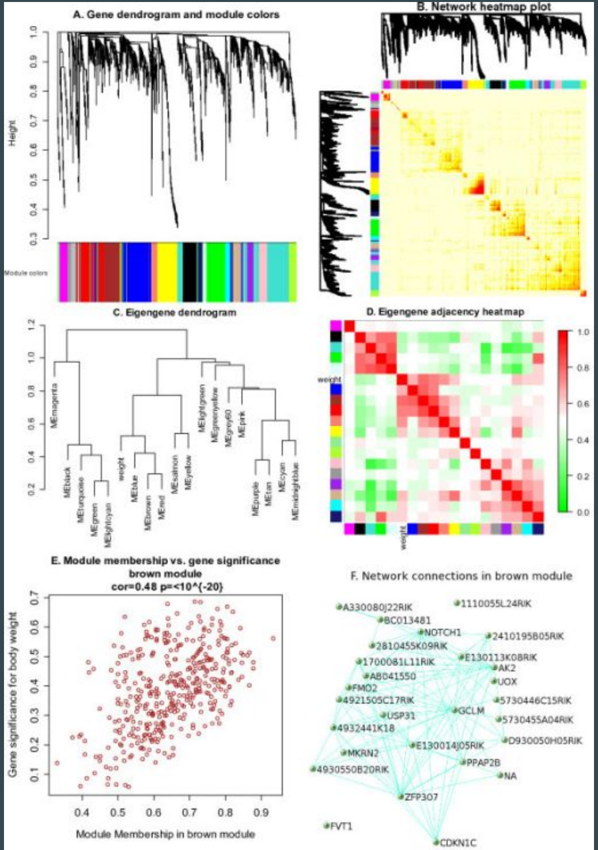
Network inference

- There are numerous algorithms to construct co-expression networks from “similarity” matrices.
- The Weighted Correlation Networks Analysis (WGCNA)
  - Power adjacency, Topological overlap similarity.
  - Identify clusters (modules) of highly correlated genes.
  - Summarise these modules by finding the “eigengene” (1st Principal Component)
  - Relate modules to external biological traits and to each other via the eigen gene.

WGCNA Module - network Detection

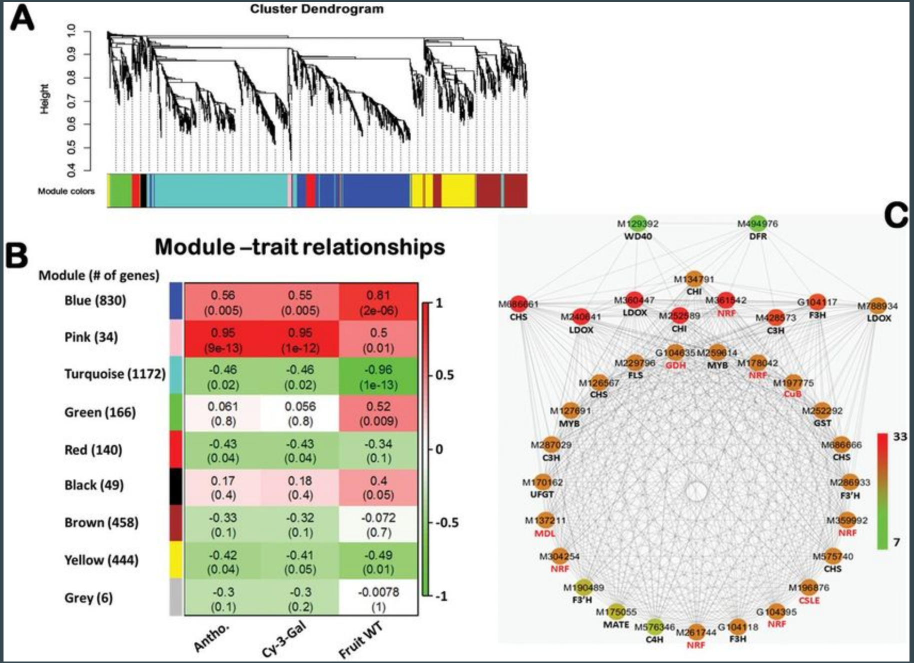


- Start by the similarity/correlation matrix.
- Apply the a “soft thresholding” approach.
- Hierarchical clustering of the “similarity” (correlation) matrix.
- DeepTreeCut to Identify Topological Overlapping Modules to separate the modules



Eigen-genes and phenotypes

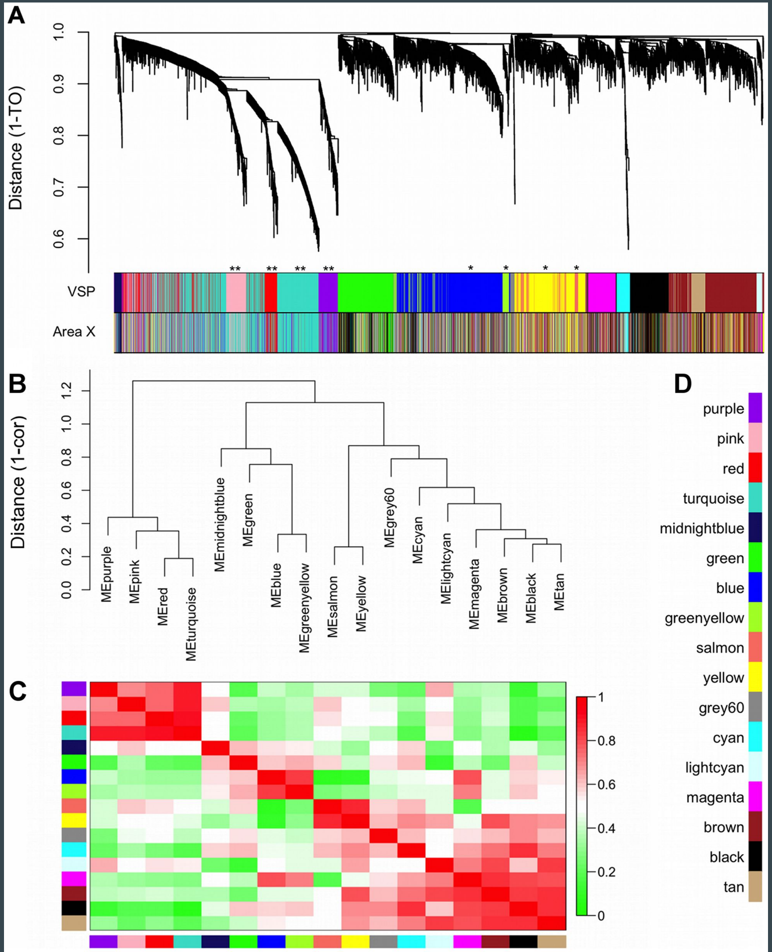
- The concept of “Eigengene”
  - A “virtual” gene with a gene expression profile such that it represents better than anything else all the genes in the module.
  - Can be visualised as the “barycentre” of a module.
  - Mathematically is the first eigenvector (principal component) of a eigen decomposition of the correlation matrix.
- We use this module signature to relate modules between them ad between modules and external biological traits (phenotypes, clinical data, features). By calculating their correlation coefficient.



Module (# of genes)	Antho.	Cy-3-Gal	Fruit WT
Blue (830)	0.56 (0.005)	0.55 (0.005)	0.81 (2e-06)
Pink (34)	0.95 (9e-13)	0.95 (1e-12)	0.5 (0.01)
Turquoise (1172)	-0.46 (0.02)	-0.46 (0.02)	-0.96 (1e-13)
Green (166)	0.061 (0.8)	0.056 (0.8)	0.52 (0.009)
Red (140)	-0.43 (0.04)	-0.43 (0.04)	-0.34 (0.1)
Black (49)	0.17 (0.4)	0.18 (0.4)	0.4 (0.05)
Brown (458)	-0.33 (0.1)	-0.32 (0.1)	-0.072 (0.7)
Yellow (444)	-0.42 (0.04)	-0.41 (0.05)	-0.49 (0.01)
Grey (6)	-0.3 (0.1)	-0.3 (0.2)	-0.0078 (1)

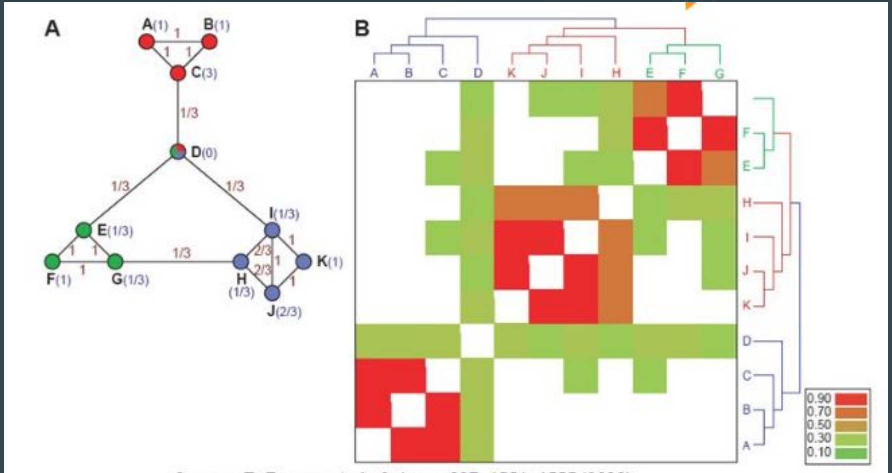
Module detection

- For the calculation of this soft threshold many algorithms have been developed. WGCNA is using DynamicTreeCut (thresholding differently for different modules).
- Clustering of the eigengenes also reveals relations between module behaviours.



How the network is build

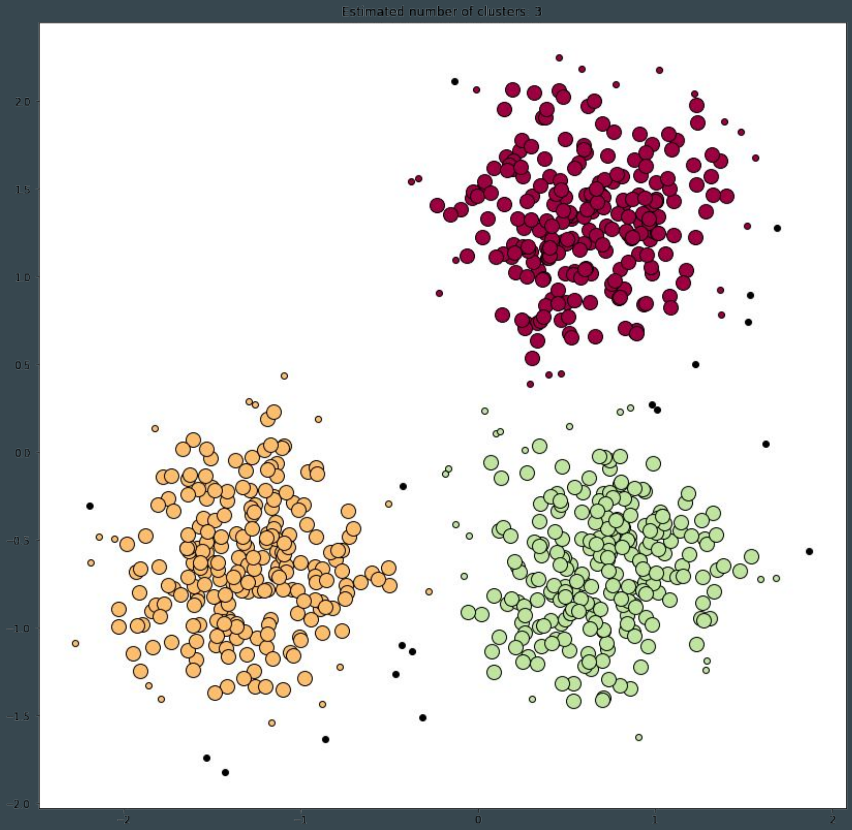
- The central algorithm for network construction is TOM (signed or unsigned)
- TOM stand for e Topological Overlap Matrix
- As any approach in correlation/coexpression networks is trying to remove spurious connections.



Source: E. Ravasz et al., Science 297, 1551 -1555 (2002)

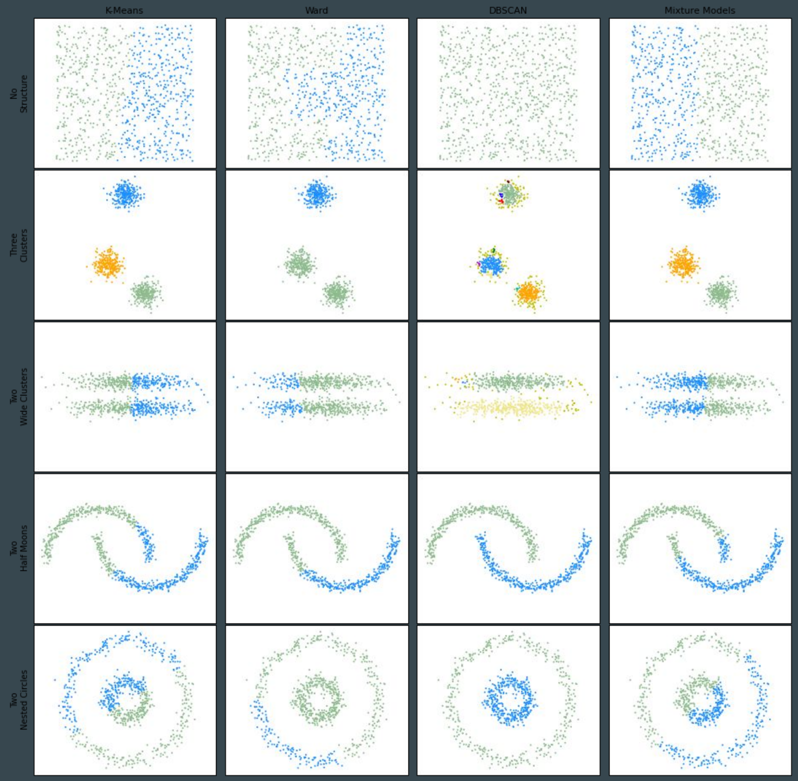
Introduction to clustering

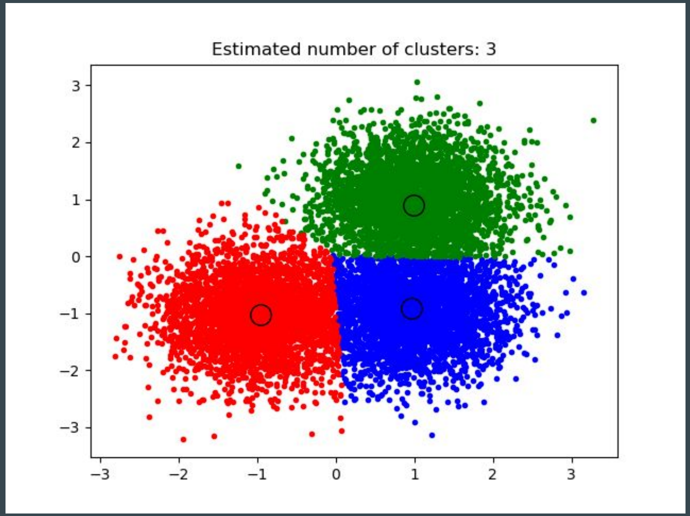
- Clustering, a general family of unsupervised methods to group entities which have similarities.
- Clustering algorithms operate on any kind of similarity measures:
  - Euclidean and other types of distance.
  - Correlation(s)
  - Covariance(s)
- Hierarchical clustering:
  - Returns a tree representing the relationships between data.



Problems of clustering

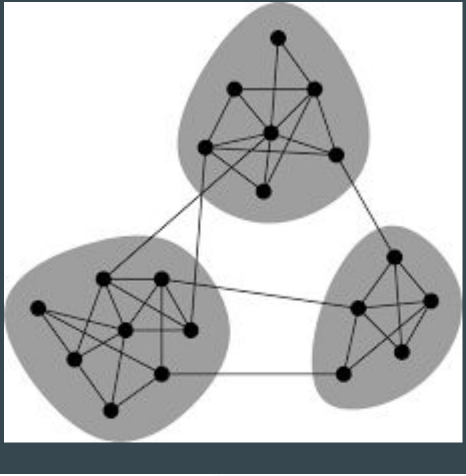
- Easy to use -thus- easy to fail methods.
- Difficult to determine a priori the “real” number of clusters.
- Many different algorithms/approaches -difficult to chose.
- Biological data is highly correlated.

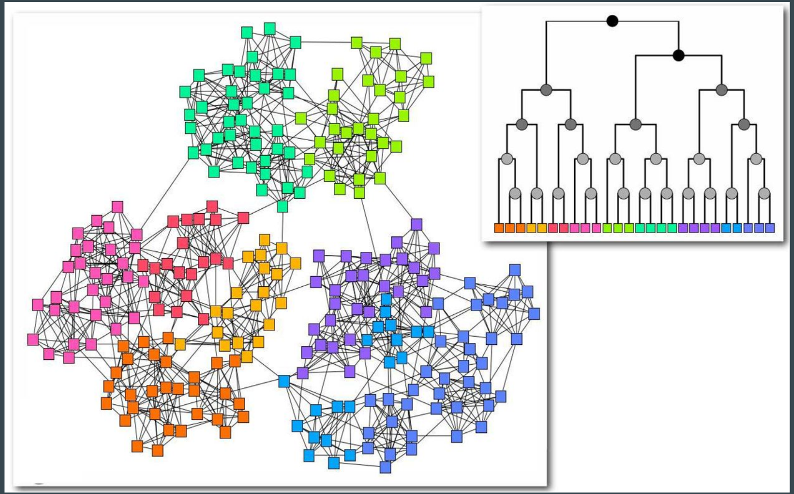




Introduction to network clustering

- Cluster areas of networks that are more connected w.r.t the rest of the network. Network clusters are also called modules.
- Helps to identify related elements and thus relate their function (guilt-by-association)
- Allows the association of modules with biological traits of interest
- Used extensively in PPIs (to identify interactions and complexes) and in GRNs to identify co-expression modules.





The WGCNA package

- We will work with an “easy to use”, “user friendly” R package that implements all the functionalities of the WGCNA method and networks.
- We will export the inferred network to Cytoscape for visualisation.