Community analysis

Rui Benfeitas
NBIS - National Bioinformatics Infrastructure Sweden
Science for Life Laboratory, Stockholm
Stockholm University

rui.benfeitas@scilifelab.se

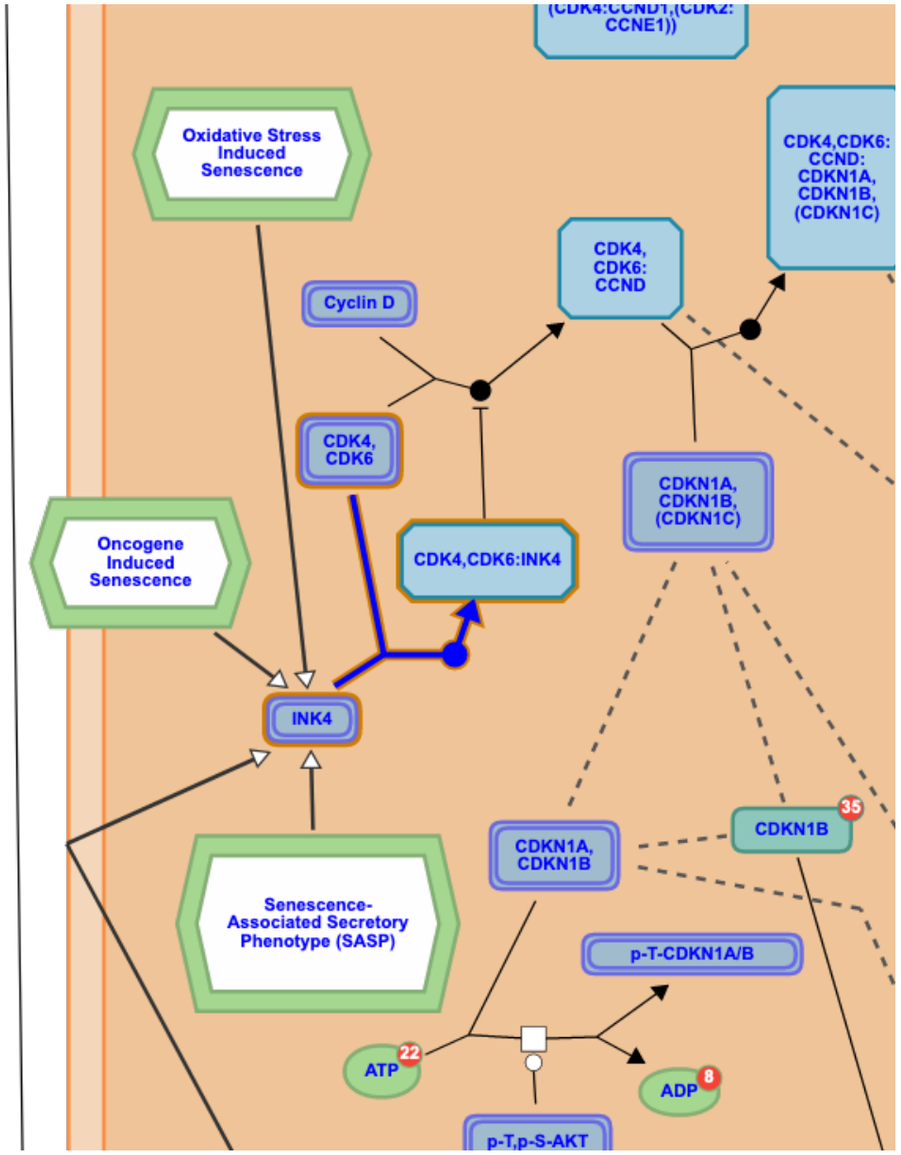
Overview

1. Introduction to network analysis
2. Terminology
3. Network inference
4. Key network properties
5. Community analysis

What are modules?

Modules are physically or functionally associated nodes that work together to achieve a distinct function

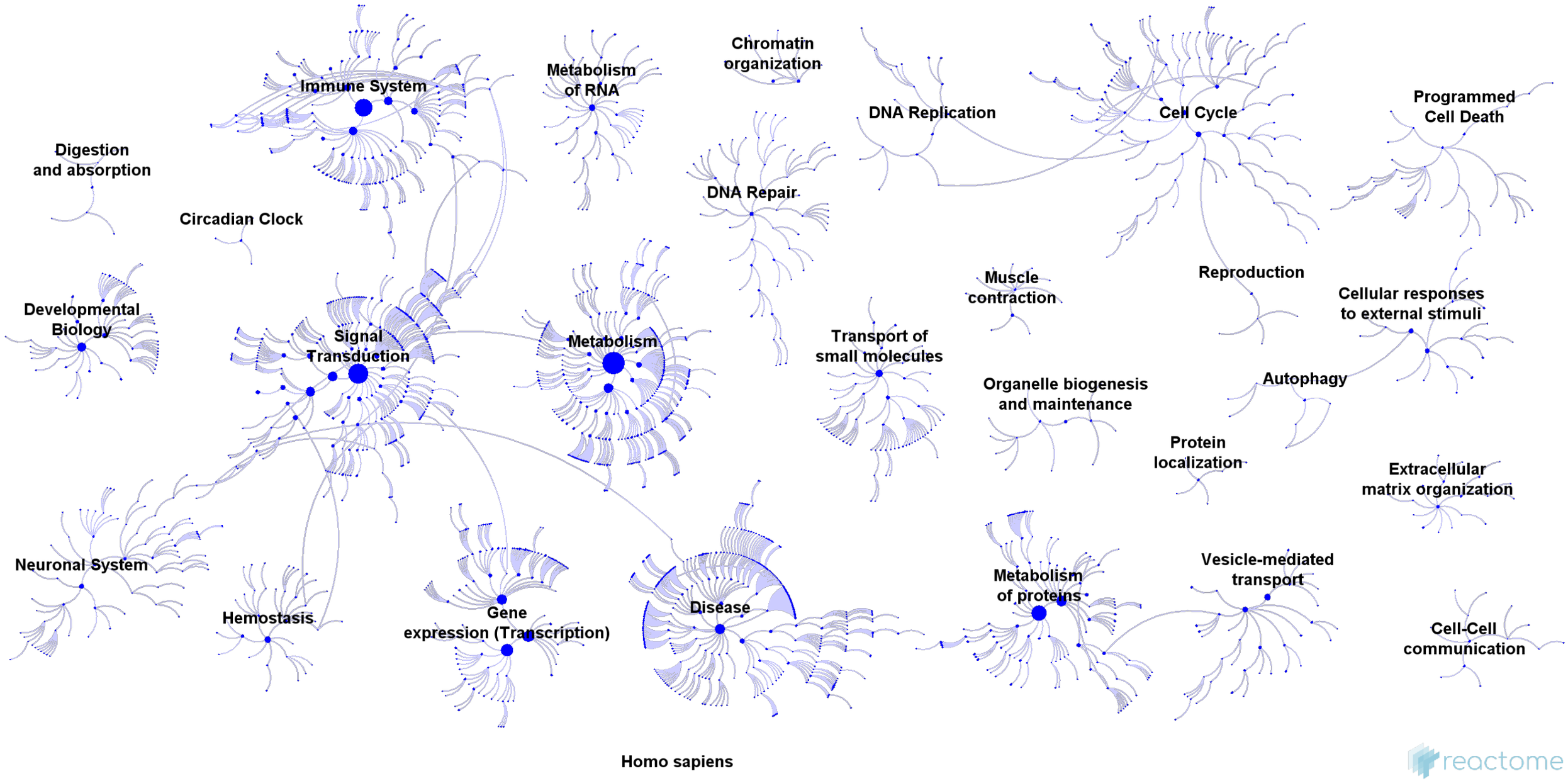
Protein complexes are physical modules



What are modules?

Pathway-associated proteins ***may*** represent functional modules

Gene Ontology



What are modules?

In addition to physical or functional modules, one may identify other types of modules

Topological: derived from their high within-module degree

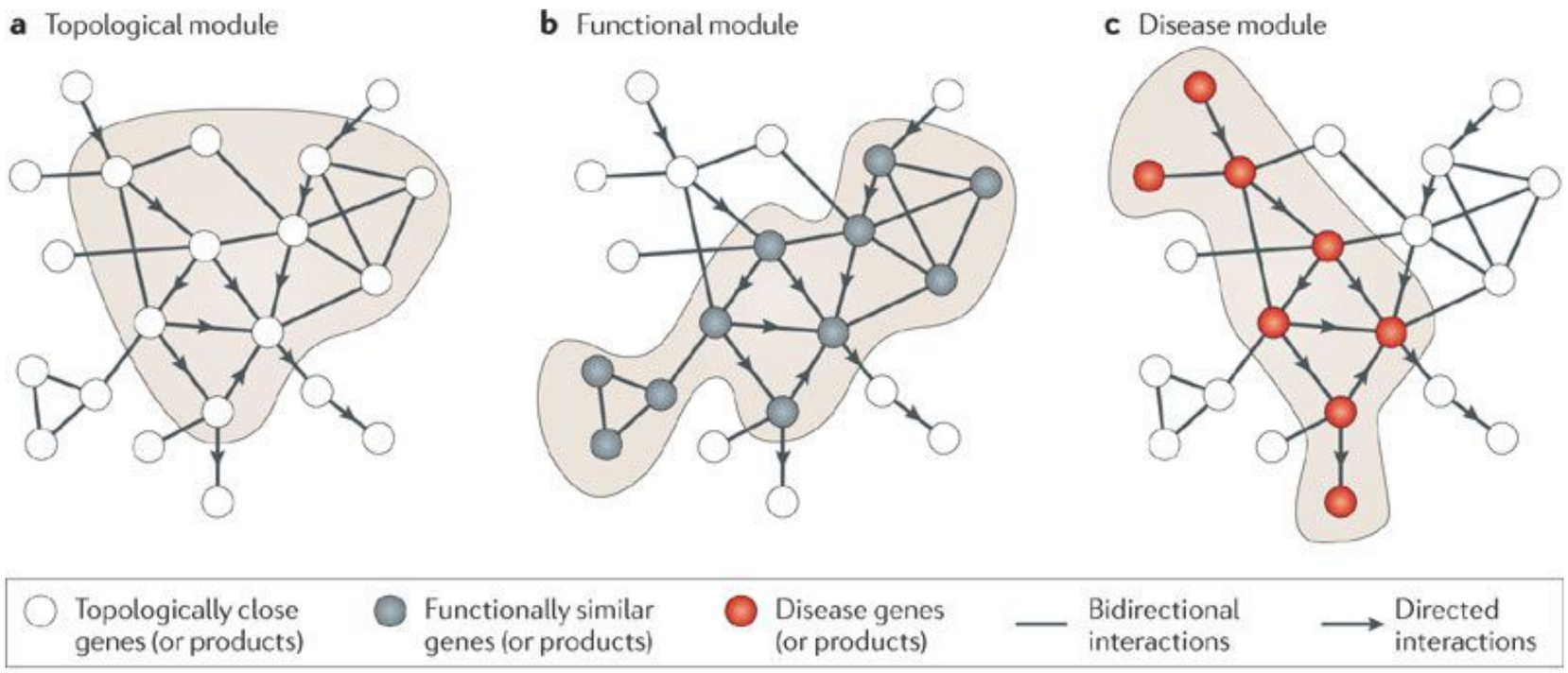
Disease: highly interconnected nodes associated with a disease response

Drug: highly interconnected nodes associated with a drug response

Subgroup: highly interconnected nodes associated with a sample subgroup (e.g. cancer subtype)

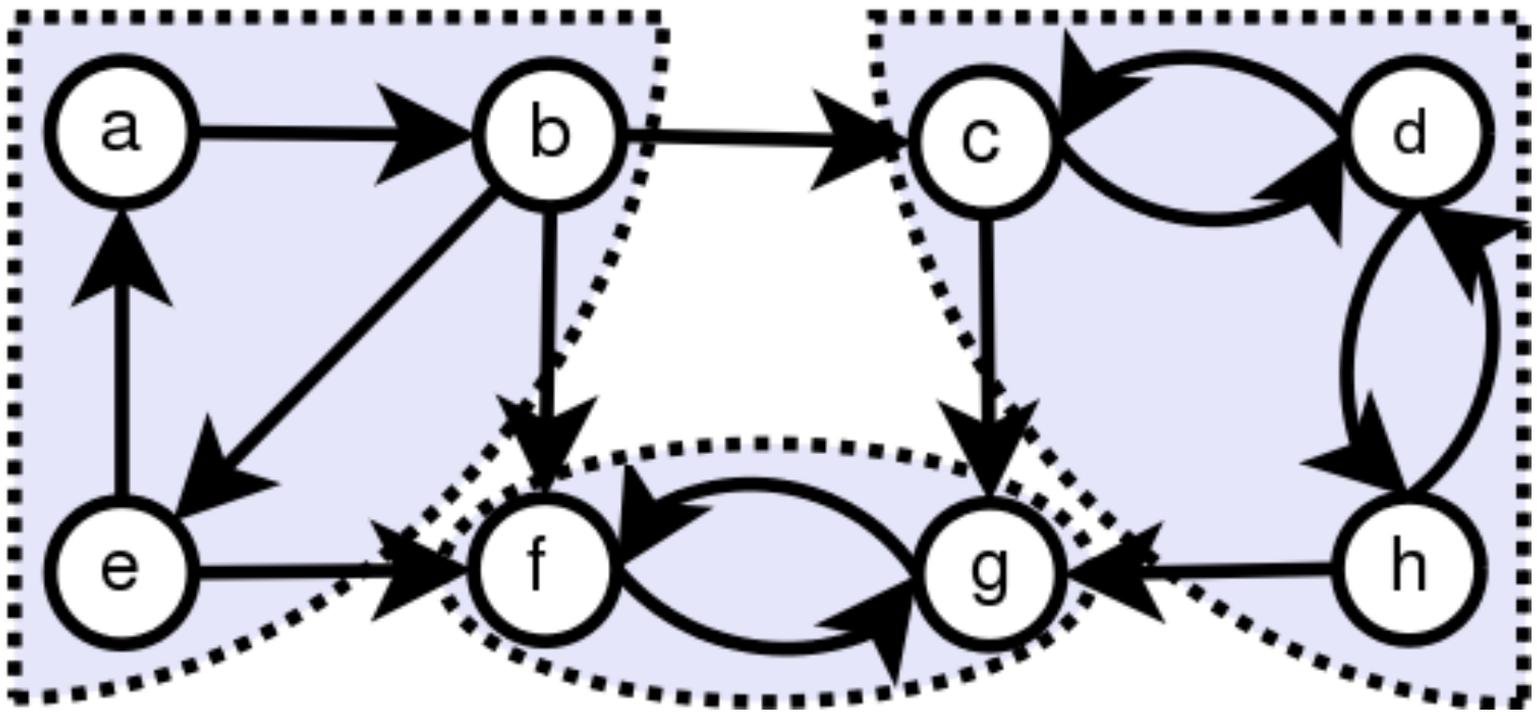
Tissue-, cell-type-specific: highly interconnected nodes associated with a specific tissue or cell type

Highly interlinked local regions of a network



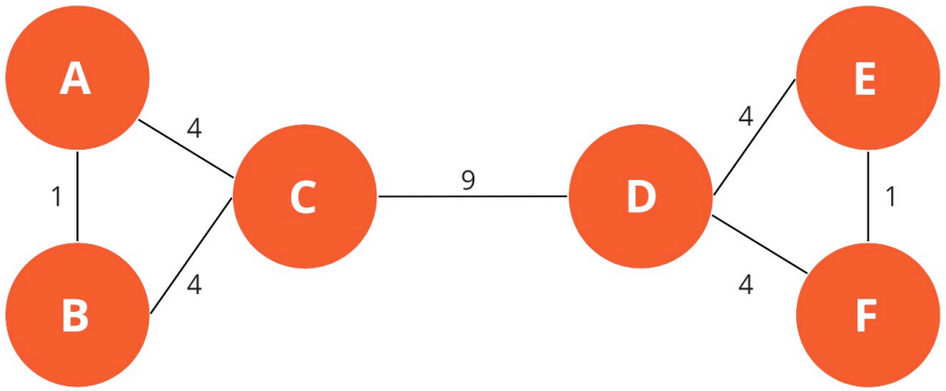
Barabasi 2011

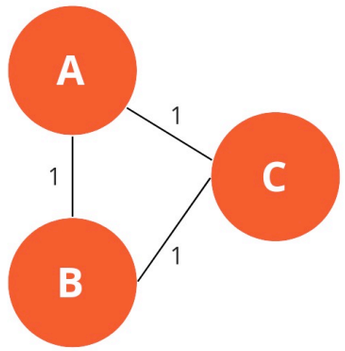
Module detection: Connected components

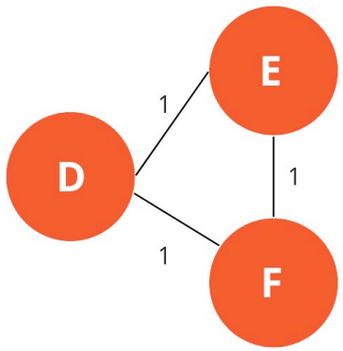


Girvan-Newman algorithm

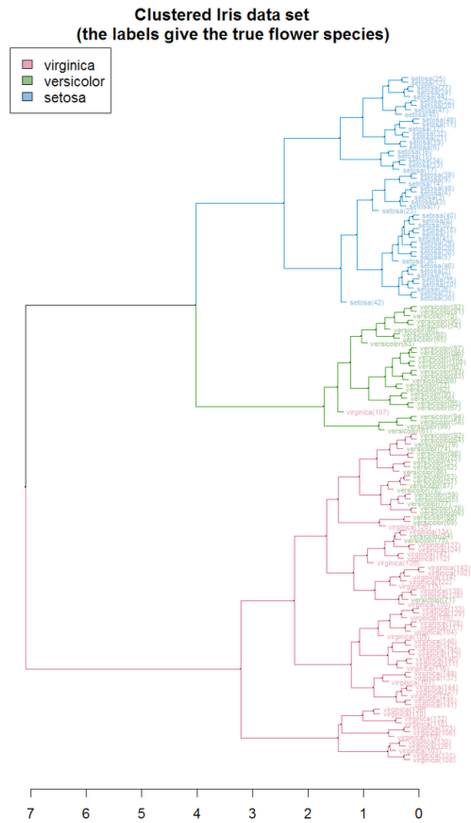
Recall: **Betweenness** considers the number of shortest paths passing through each edge







1. Calculate edge betweenness for each edge;
2. Remove edge with highest betweenness;
3. Recalculate betweenness centrality for all edges;
4. Repeat until no edges left



Modularity

Modularity is a property of the network

Modularity (Q) measures the tendency of a graph to be organised into modules

Modules computed by comparing probability that an edge is in a module vs what would be expected in a random network

$$Q \propto \sum_{s \in S} [(e_s) - (\text{expected } e_s)]$$

edges in group s

Random network with same number of nodes, edges and degree per node

$-1 < Q < 1$

Q = 1: much higher number of edges than expected by chance

Q = -1: lower number of edges than expected by chance

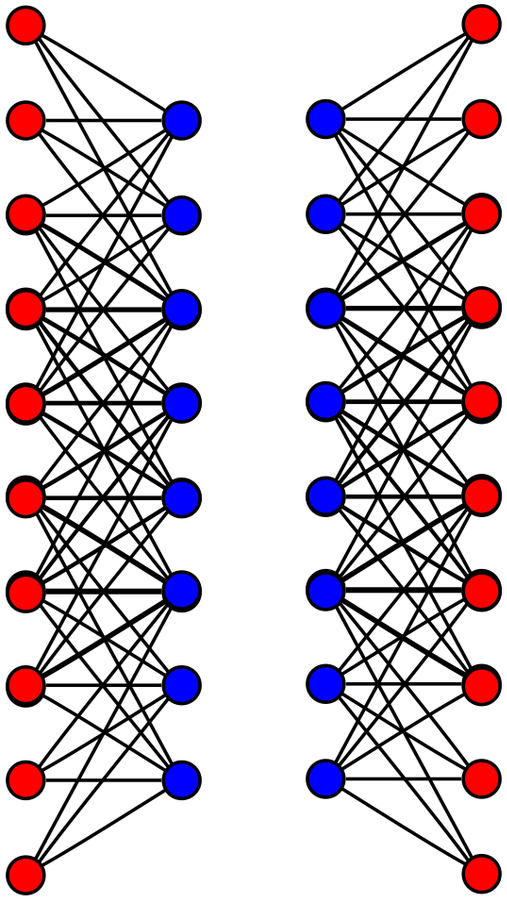
Q > 0.3 - 0.7 means significant community structure

Modularity

Modularity is different than **clustering coefficient**:

Graph composed of two bipartite complete subgraphs:

high Q but low connectivity (C)



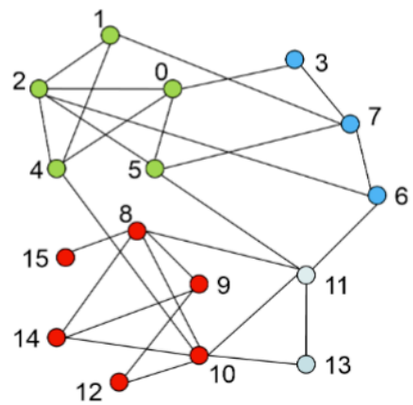
Module detection: Louvain algorithm

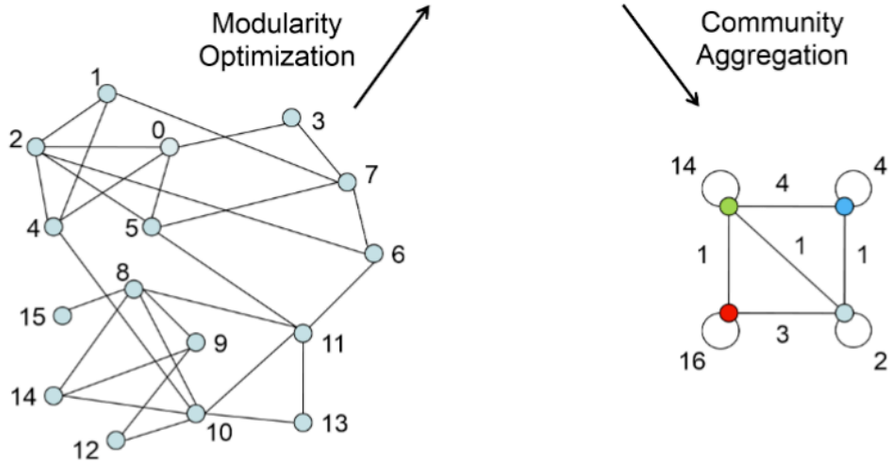
Phase 1: greedy modularity optimisation

1. Start with 1n/community
2. Compute Q by moving *i* to the community of *j*
3. If ΔQ>1, node is placed in community
4. Repeat 1-3 until no improvement is found. Ties solved arbitrarily

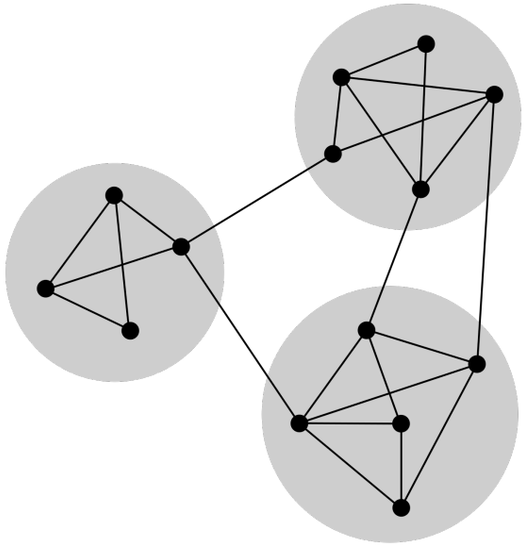
Phase 2: coarse grained community aggregation

5. Link nodes in a community into single node.
6. Self loops show intra-community associations
7. Inter-community weights kept
8. Repeat phase 1 on new network





Other methods:
Walktrap
Label propagation
…
(benchmarking)



Community characterisation

Clustering coefficient and degree distribution

Enrichment analysis

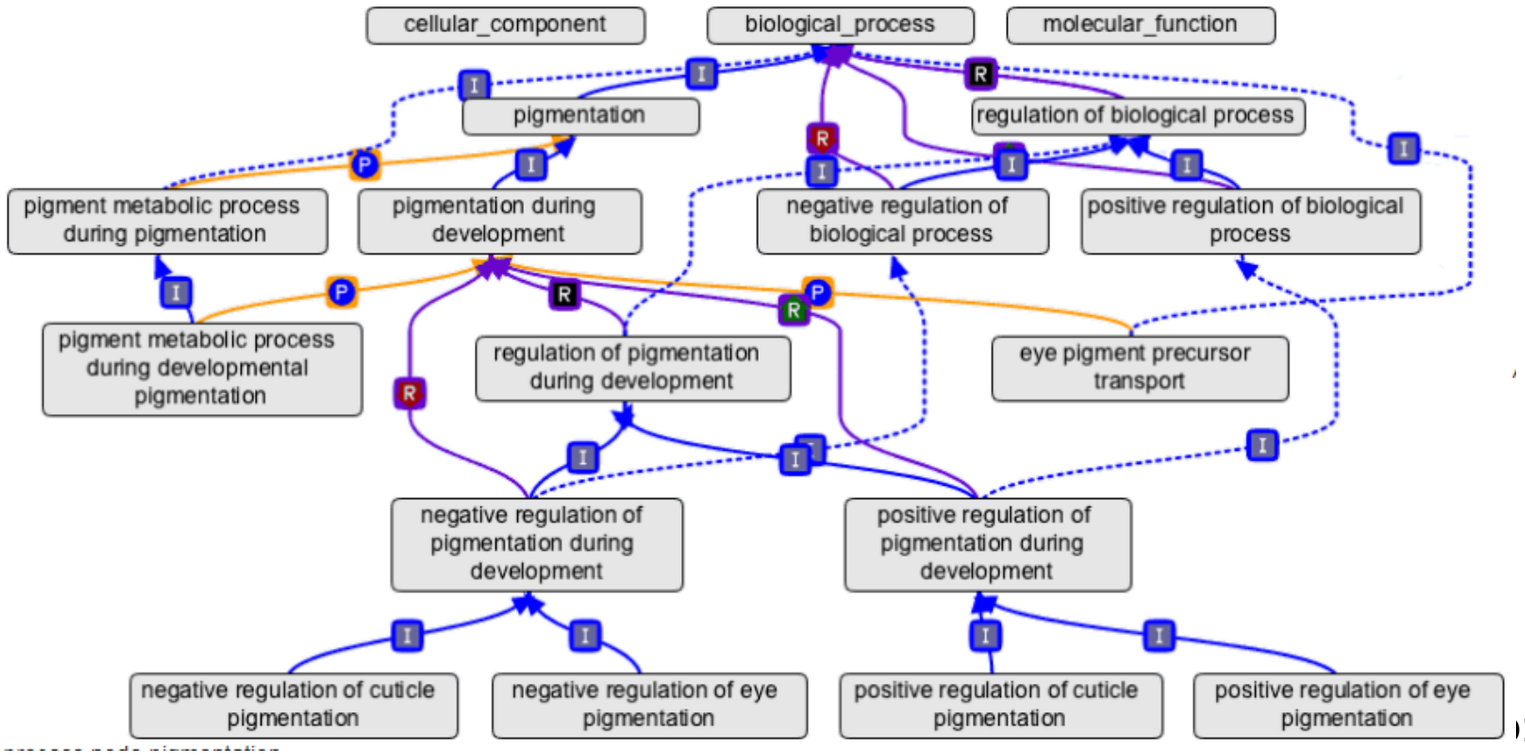
Hypothesis: community-associated features show coordinated changes associated with common biological processes

Enrichment analysis

GO-terms, pathways, subcellular location, TF-targets, disease, drugs

Tests for significant overlap between groups

Some biological processes may have no biological meaning in your analysis



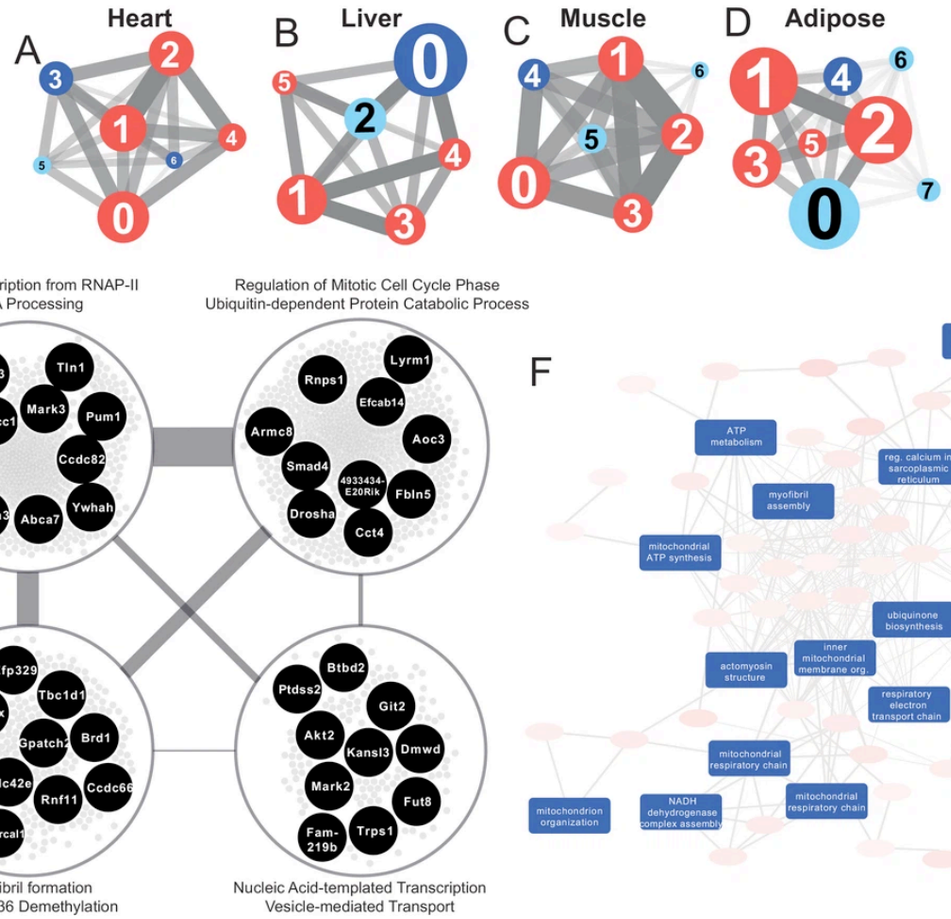
://geneontology.org/

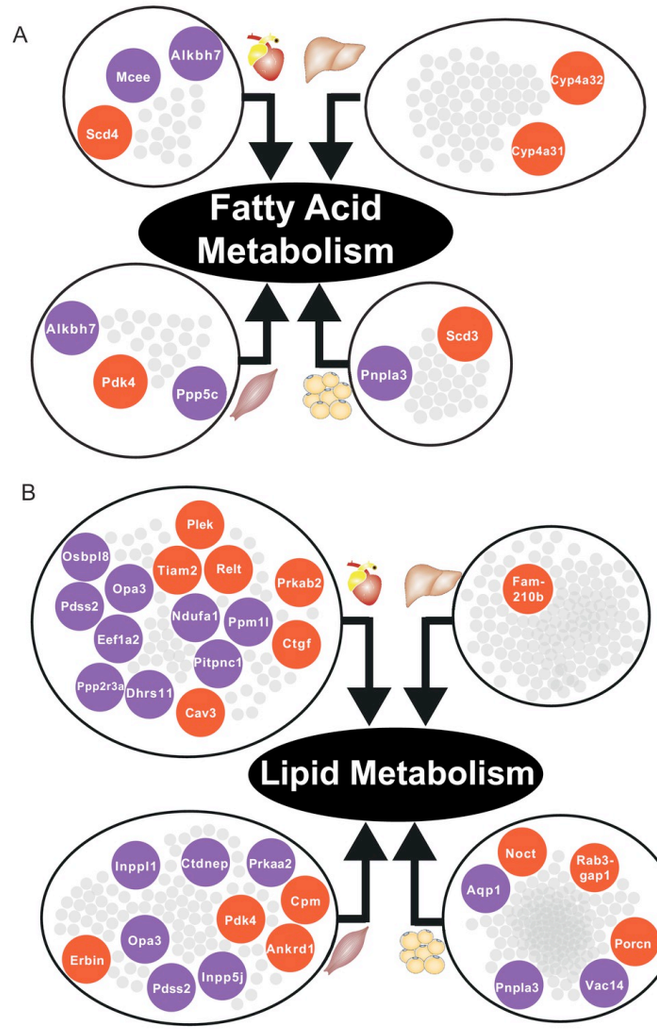
Multi-tissue network analysis

Graph analysis

Community characterization

GEMs

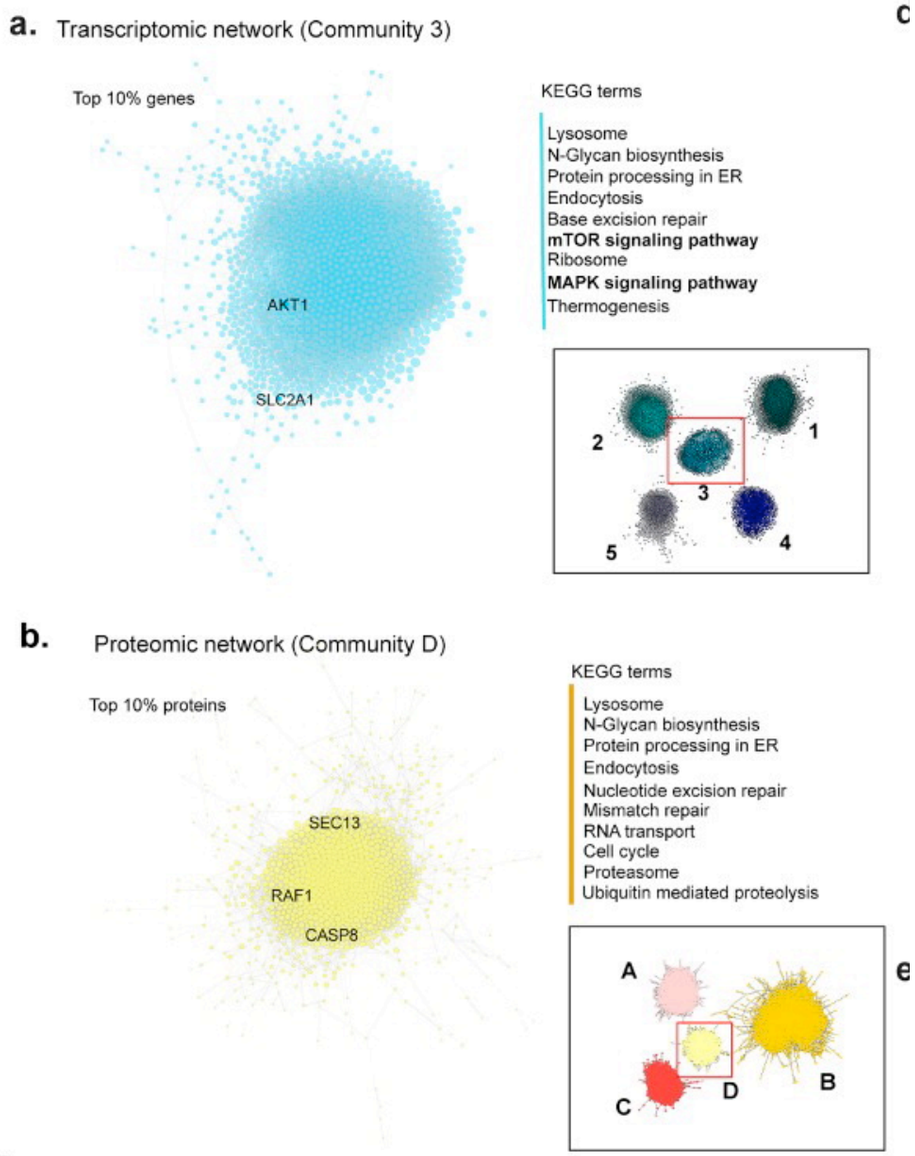


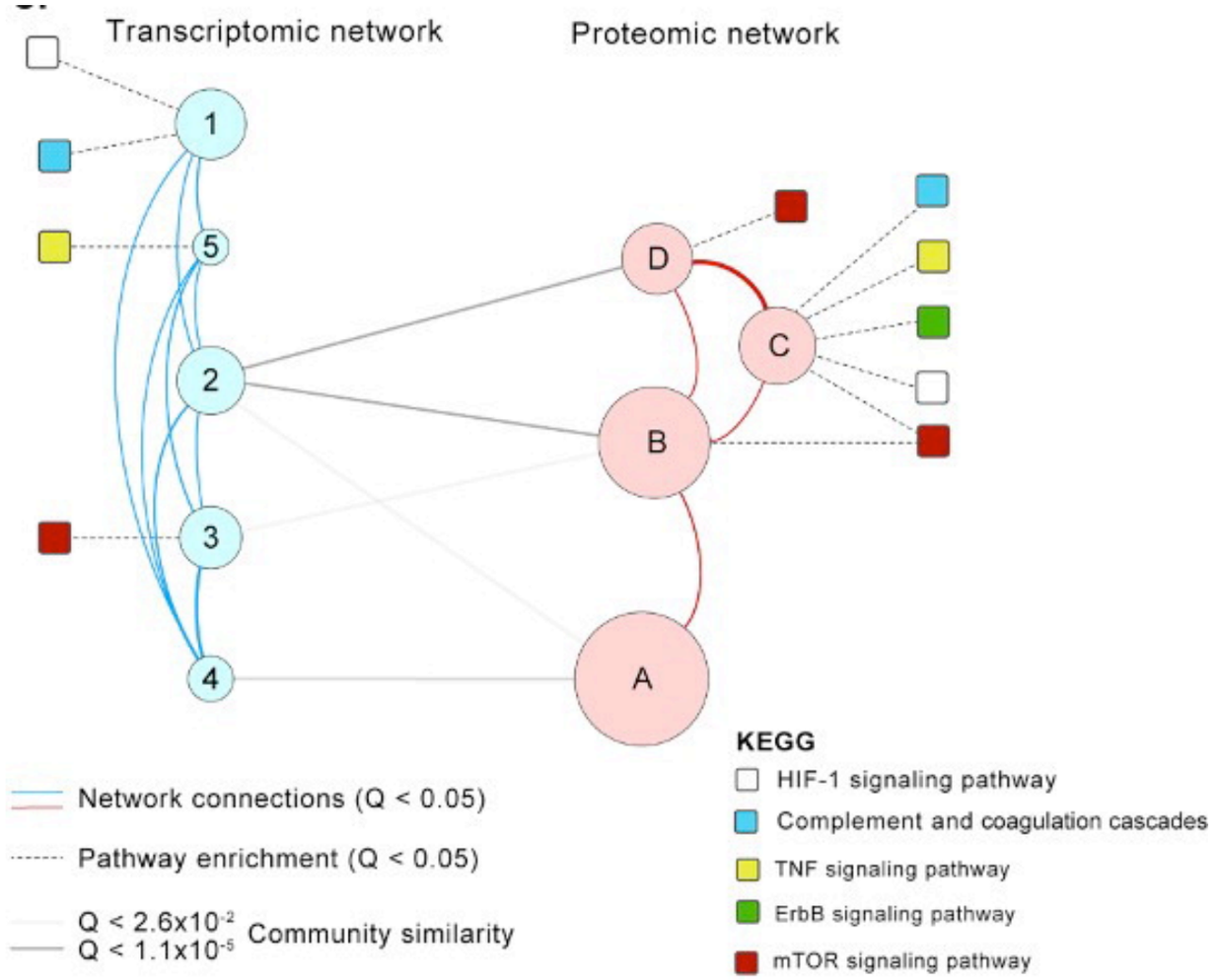


Multi-tissue network analysis

Graph analysis

Community characterization

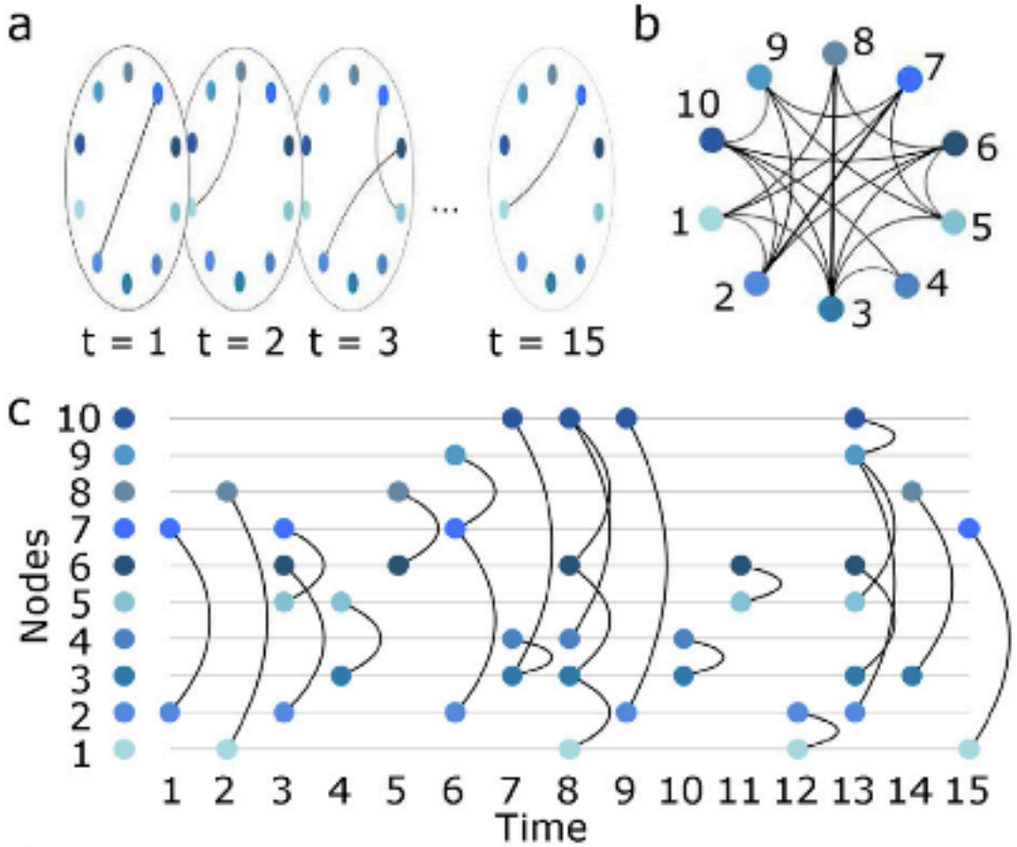




Dynamic network analysis

Dynamic encoding

Static representation



Expands many of the concepts stated:

- Paths, edge weight
- Latency (shortest time to move between nodes, i.e. *fastest path*)
- Dynamic clustering coefficient
- Dynamic closeness centrality
- Temporal *small-worldness*

Additional reading

- Network Science - Textbook on graph theory and network analysis.
- Communication dynamics in complex brain networks - Discussion about whether and how network topology may be applied to study the brain networks.
- A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models - General review and discussion on methods to use in genome-scale metabolic models.
- Analysis of Biological Networks - General introduction into biological networks, network notation, and analysis, including graph theory.
- Multi-omics approaches to disease - Introduction to how integrative approaches may be applied in disease

Additional references displayed as hyperlinks in each slide.

Additional reading

- Analysis of Biological Networks - General introduction into biological networks, network notation, and analysis, including graph theory.
- Using graph theory to analyze biological networks - overview of the usage of graph theory in biological network analysis
- Survival of the sparsest: robust gene networks are parsimonious - analysis of network complexity and robustness.
- Network biology: understanding the cell's functional organization - Overview of key concepts in biological network structure
- Graph Theory and Networks in Biology - extended perspective on how graph analysis is applied in biology
- Modularity and community structure in networks

Additional references displayed as hyperlinks in each figure.

Enrichment analysis

Important databases with gene-sets:

- MSigDB (gene)
- Enrichr (gene)
- KEGG (metabolite, gene)
- DIANA (miRNA)
- MetaboAnalyst (metabolite)
- DAVID (web)
- Reactome (web)

Creating custom sets and joint sets

Mapping your data to common IDs

- Easy for genes and proteins: use DAVID, Biomart, or MyGene (in Python or R)
- Hard for other types