Topic 7. Networks and Pathways

Network representation and analysis: strategies and limitations. Molecular interaction networks: IntAct and other databases. Reactome & KEGG pathways. Visualization of networks and pathways in Cytoscape.

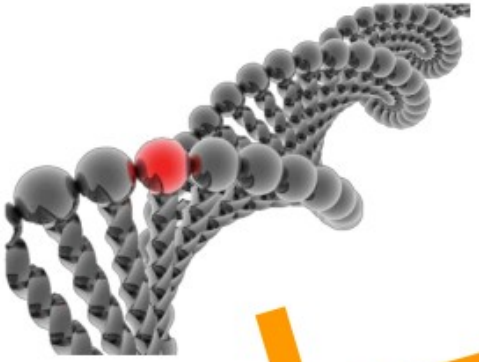
The Genotype to Phenotype challenge

Topic 5 – Genes & Genomes

Topic 6 – Functional genomics

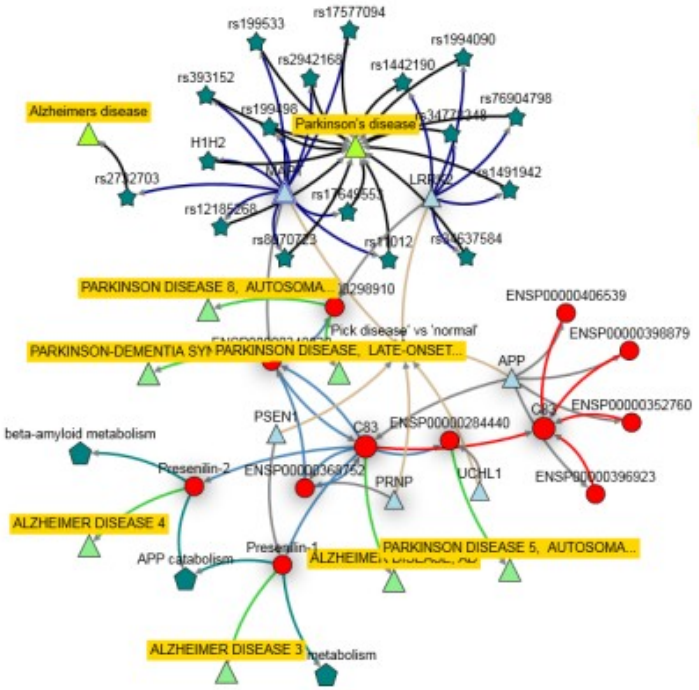
Topic 7 – Networks & Pathways

Topic 8 – Phenotypes & Diseases



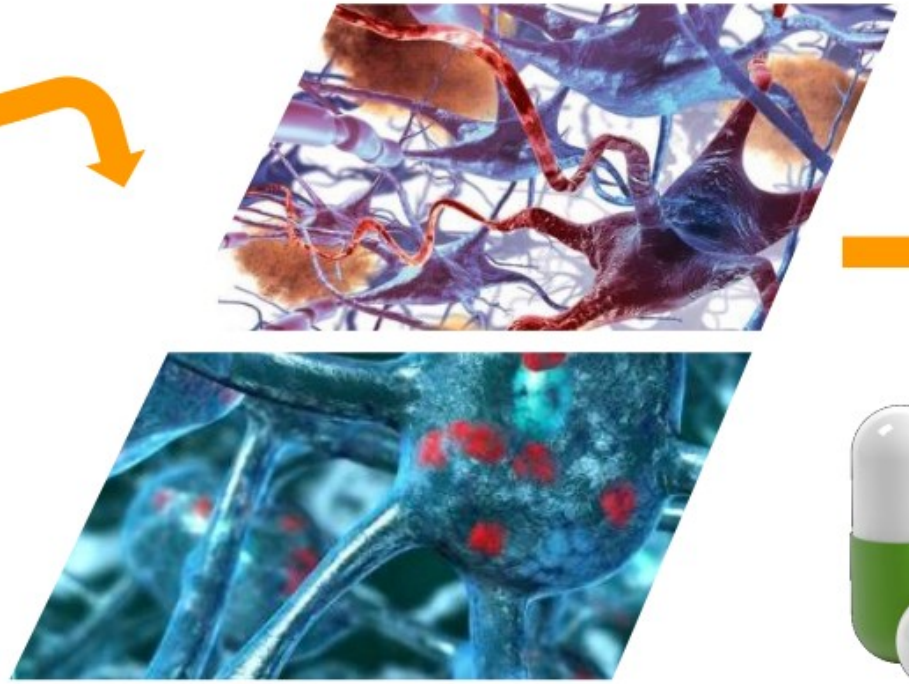
Genotype

GWAS, QTL



Biological Knowledge Discovery

Data selection, processing, transformation, integration, interpretation



Phenotype

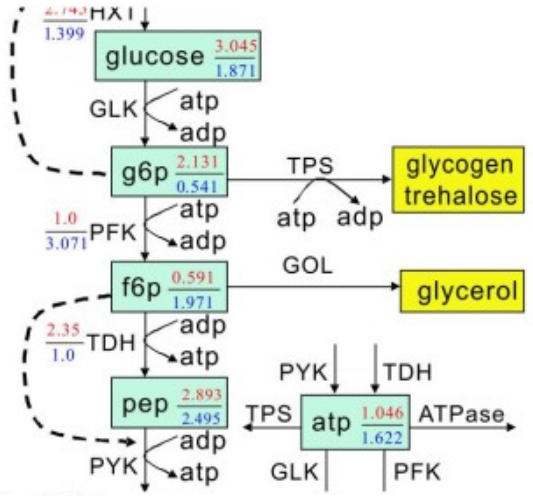
Alzheimer, Parkinson

Drugs

Precision medicine

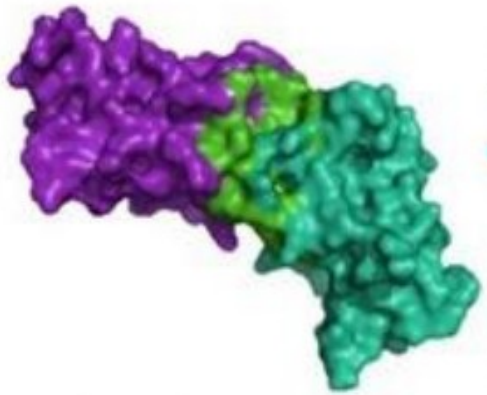
Network analysis in biology

Ecological network

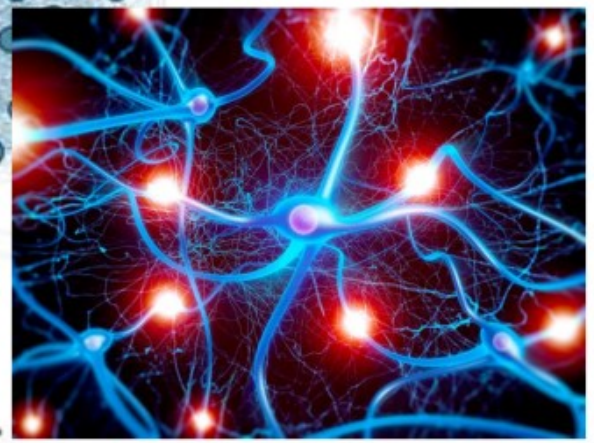


Metabolic network

Systems Biology



Molecular interaction network



Neurological network

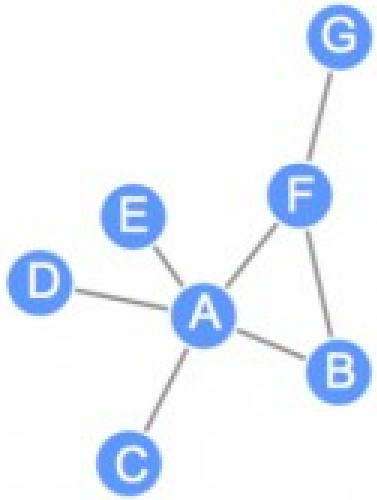
Introduction to graph theory

Graph theory is *"[…] the study of graphs, mathematical structures used to model pairwise relations between objects. A graph in this context is made up of vertices, nodes, or points which are connected by edges, arcs, or lines".*

First described by the Swiss mathematician **Leonard Euler** as applied to the problem of the seven bridges of Königsberg:

Graph theory: graph types and edge properties

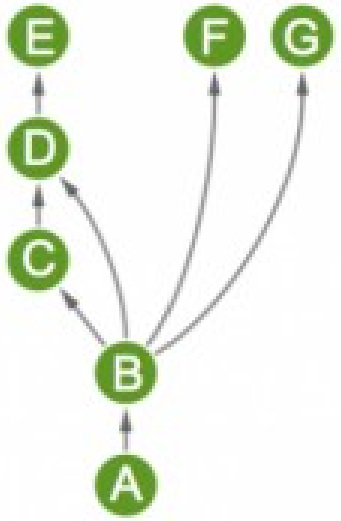
Undirected



E.g., **protein-protein interaction networks (PPINs)**

No flow implied

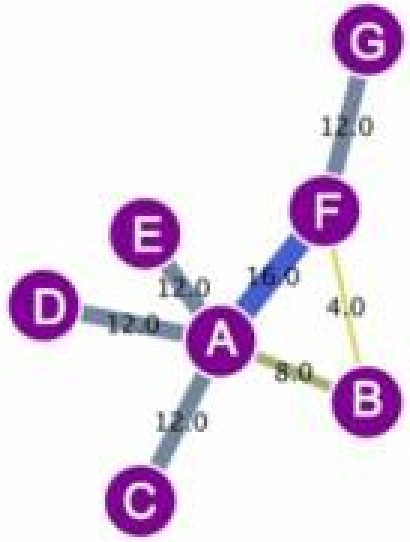
Directed



E.g., **metabolic or gene regulation networks**

Flow of signal implied, can be organized hierarchically

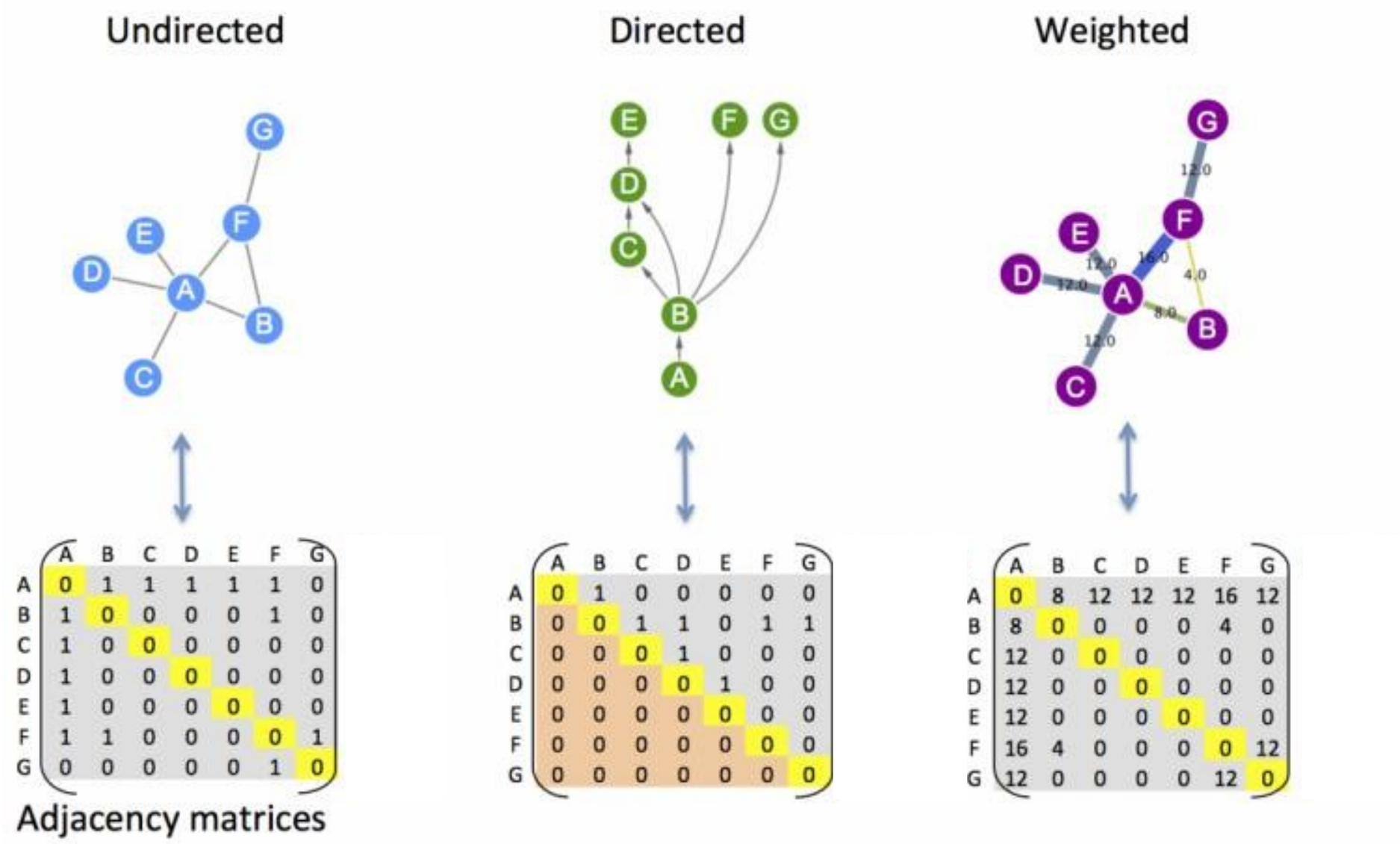
Weighted



E.g., **reliability of interaction, quantitative expression change, sequence similarity**

Quantitative value (weight) associated to directed or undirected edges

Graph theory: adjacency matrices



	A	B	C	D	E	F	G
A	0	1	1	1	1	1	0
B	1	0	0	0	0	1	0
C	1	0	0	0	0	0	0
D	1	0	0	0	0	0	0
E	1	0	0	0	0	0	0
F	1	1	0	0	0	0	1
G	0	0	0	0	0	1	0

	A	B	C	D	E	F	G
A	0	1	0	0	0	0	0
B	0	0	1	1	0	1	1
C	0	0	0	1	0	0	0
D	0	0	0	0	1	0	0
E	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0

	A	B	C	D	E	F	G
A	0	8	12	12	12	16	12
B	8	0	0	0	0	4	0
C	12	0	0	0	0	0	0
D	12	0	0	0	0	0	0
E	12	0	0	0	0	0	0
F	16	4	0	0	0	0	12
G	12	0	0	0	0	12	0

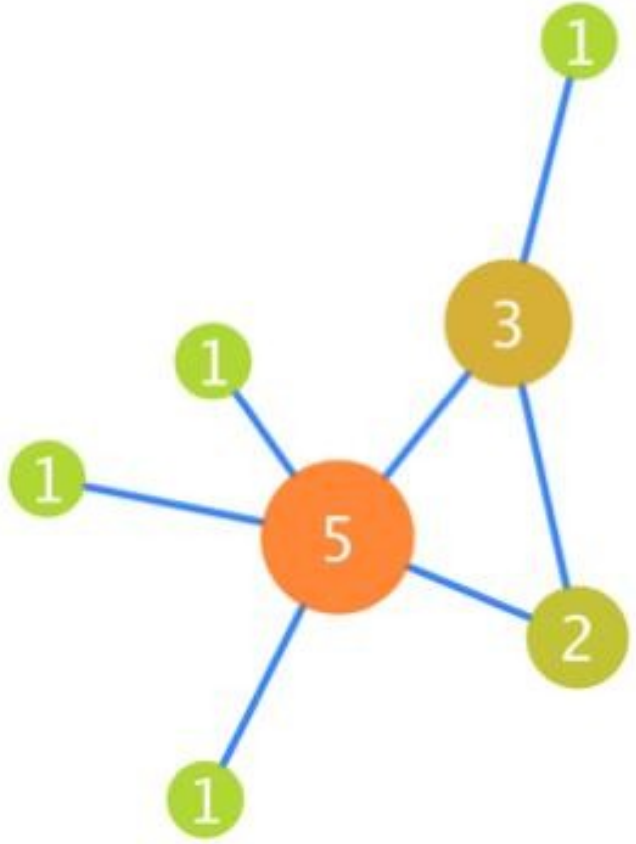
Graph theory: network topology

Topology is the way in which the nodes and edges are arranged within a network. Topological properties can apply to the network as a whole or to individual nodes and edges.

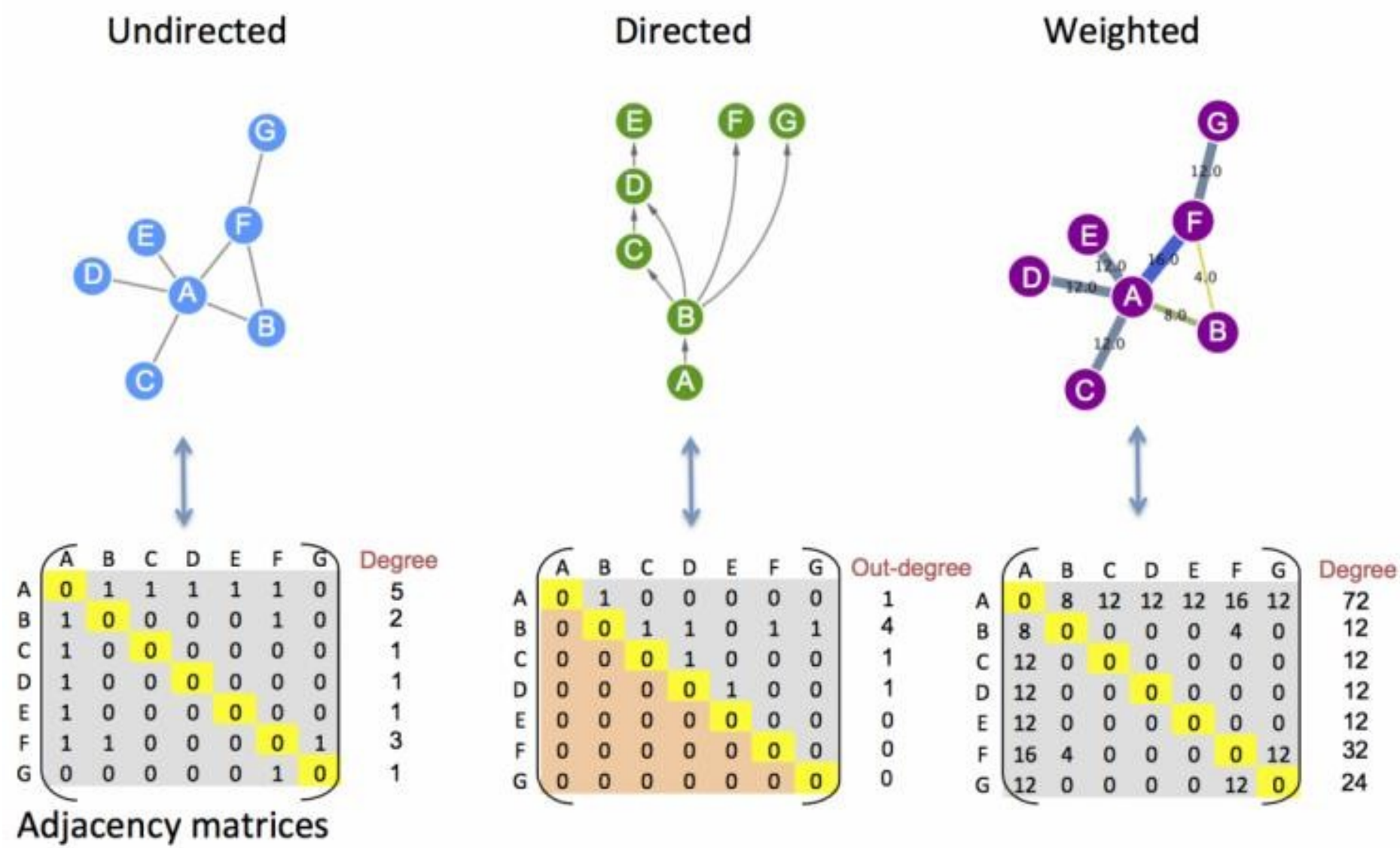
Degree

Number of edges that connect to a node.

Directed network nodes have two values for degree: **out-degree** *(number of edges coming out of the node)* and **in-degree** *(number of edges coming into the node)*.



Graph theory: adjacency matrices



	A	B	C	D	E	F	G	Degree
A	0	1	1	1	1	1	0	5
B	1	0	0	0	0	1	0	2
C	1	0	0	0	0	0	0	1
D	1	0	0	0	0	0	0	1
E	1	0	0	0	0	0	0	1
F	1	1	0	0	0	0	1	3
G	0	0	0	0	0	1	0	1

	A	B	C	D	E	F	G	Out-degree
A	0	1	0	0	0	0	0	1
B	0	0	1	1	0	1	1	4
C	0	0	0	1	0	0	0	1
D	0	0	0	0	1	0	0	1
E	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0

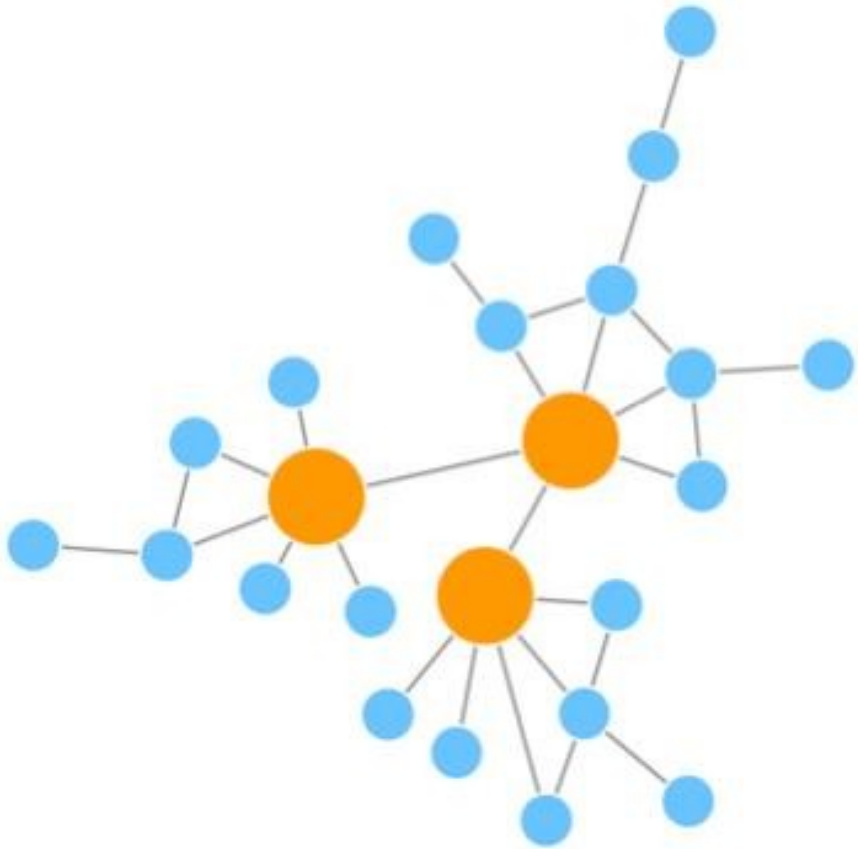
	A	B	C	D	E	F	G	Degree
A	0	8	12	12	12	16	12	72
B	8	0	0	0	0	4	0	12
C	12	0	0	0	0	0	0	12
D	12	0	0	0	0	0	0	12
E	12	0	0	0	0	0	0	12
F	16	4	0	0	0	0	12	32
G	12	0	0	0	0	12	0	24

Graph theory: network topology

Topology is the way in which the nodes and edges are arranged within a network. Topological properties can apply to the network as a whole or to individual nodes and edges.

Scale-free networks

In scale-free networks most of the nodes are connected to a low number of neighbours (*blue*) and there are a small number of high-degree nodes (**hubs**; *orange*) that provide high connectivity to the network.

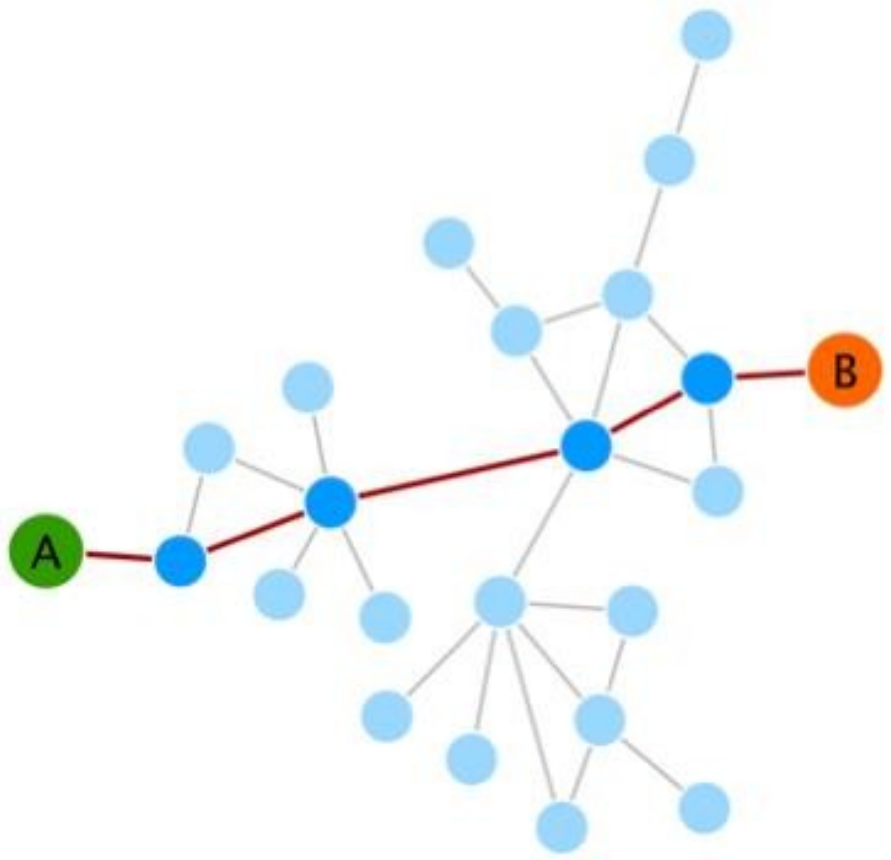


Graph theory: network topology

Topology is the way in which the nodes and edges are arranged within a network. Topological properties can apply to the network as a whole or to individual nodes and edges.

Shortest paths

Shortest distance between any two nodes.



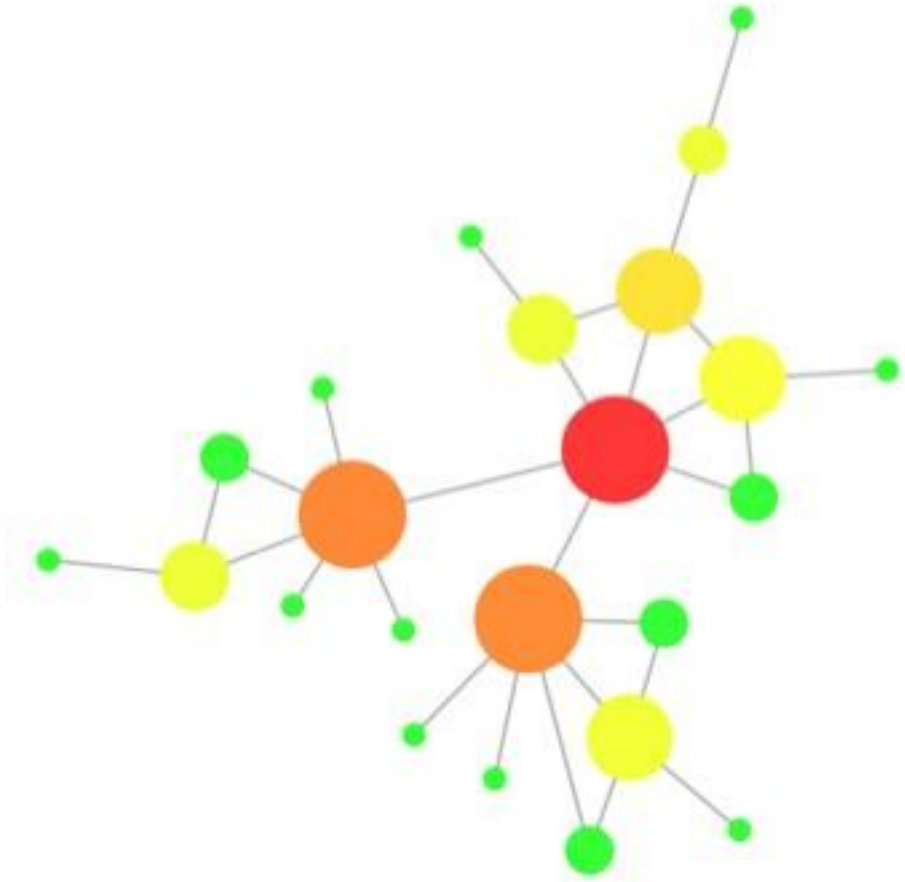
Graph theory: network topology

Topology is the way in which the nodes and edges are arranged within a network. Topological properties can apply to the network as a whole or to individual nodes and edges.

Centralities

Can be measured for nodes and for edges and gives an estimation on how important that node/edge is for the connectivity or the information flow of the network:

- **Degree centrality** *(node size)*
- **Betweenness centrality** *(warm colors)*



Graph theory: network topology

Centrality analysis

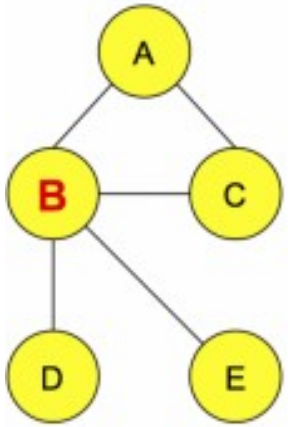
How important a node or edge is for the connectivity or information flow of the network

- Local measures:
  - **Degree of the nodes**
- Global measures:
  - **Closeness centrality**
  - **Betweenness centrality**
- Other measures:
  - **Random walks**

Measures how short the **shortest paths** are from node *i* to all other nodes. Expressed as the normalized inverse of the sum of the topological distances in the graph.

$$CC(i) = \frac{N-1}{\sum_j d(i,j)}$$

where
$i \neq j$,
d_{ij} is the length of the shortest path between nodes i and j in the network,
N is the number of nodes.



	A	B	C	D	E	farness $\sum_{j=1}^{n} d(i,j)$	$CC(i) = \frac{N-1}{\sum_j d(i,j)}$
A	0	1	1	2	2	6	(5-1)/6= 0.67
B	1	0	1	1	1	4	1.00
C	1	1	0	2	2	6	0.67
D	2	1	2	0	2	7	0.57
E	2	1	2	2	0	7	0.57

N = 5 (# of nodes)

Graph theory: network topology

Centrality analysis

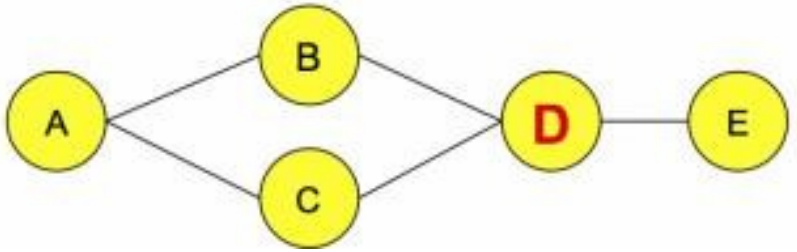
How important a node or edge is for the connectivity or information flow of the network

- Local measures:
  - **Degree of the nodes**
- Global measures:
  - **Closeness centrality**
  - **Betweenness centrality**
- Other measures:
  - **Random walks**

Measures how often a node occurs on all **shortest paths** between two nodes. Defined as the number of shortest paths in the graph that pass through the node divided by the total number of shortest paths. Calculated considering couples of nodes.

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i) / g_{jk}$$

Where g_{jk} = the number of geodesics (shortest paths) connecting *jk*, and $g_{jk}(n_i)$ = the number that node *i* is on.

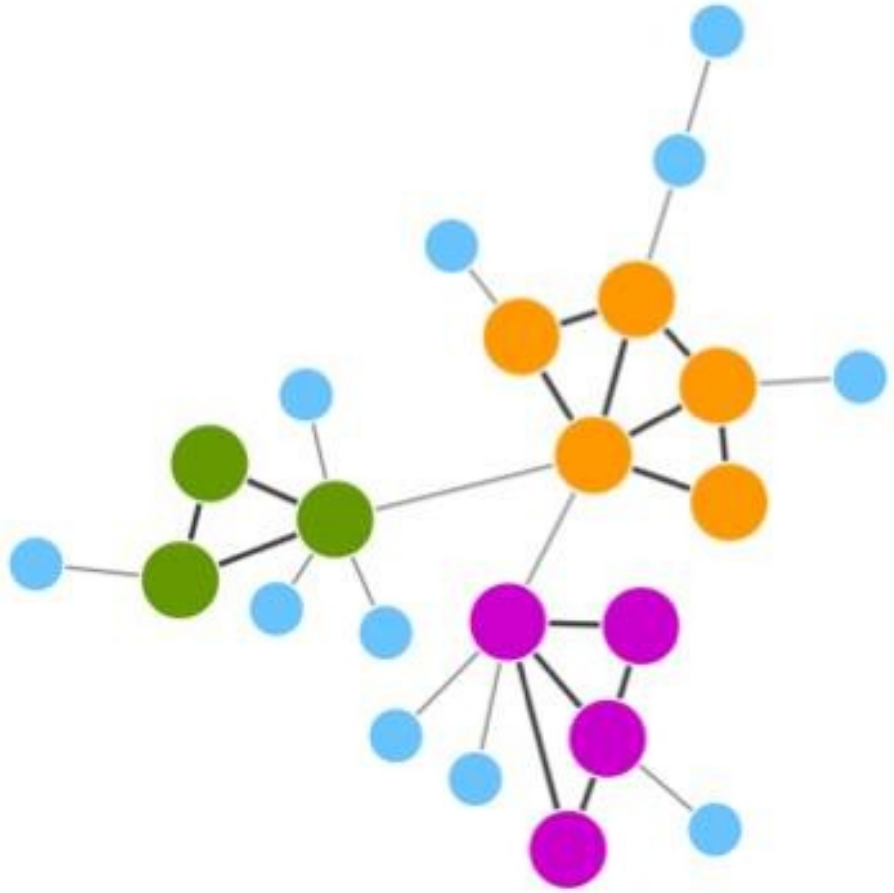


Graph theory: network topology

Topology is the way in which the nodes and edges are arranged within a network. Topological properties can apply to the network as a whole or to individual nodes and edges.

Transitivity

Presence of tightly interconnected nodes in the network called **topological clusters** or **communities**, which are more internally connected than they are with the rest of the network.



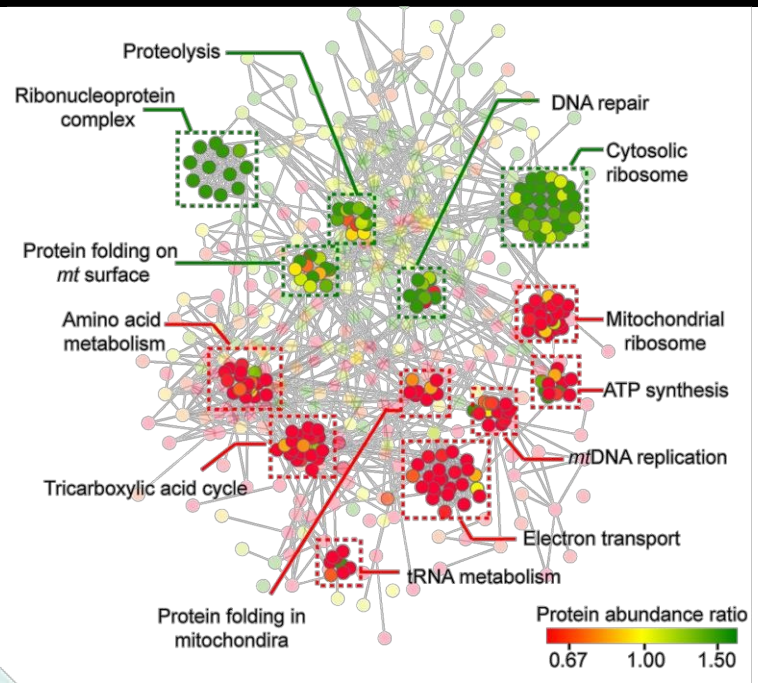
Graph theory: network topology

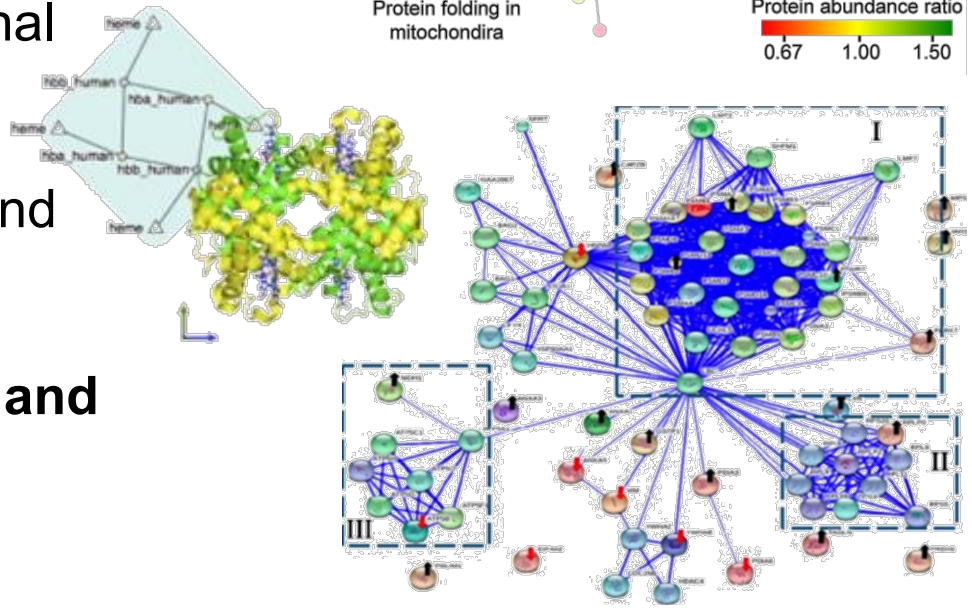
Clustering analysis

The **transitivity**, **modularity**, or **clustering coefficient** of a network is a measure of the tendency of the nodes to cluster together. High transitivity means that the network contains communities or groups of nodes that are densely connected internally. They usually reflect:

- **Functional modules:** self-contained, exchangeable functional units; nodes do not necessarily interact in time and space.
- **Protein complexes:** group of proteins that interact in time and space; multi-protein machineries with specific functions.

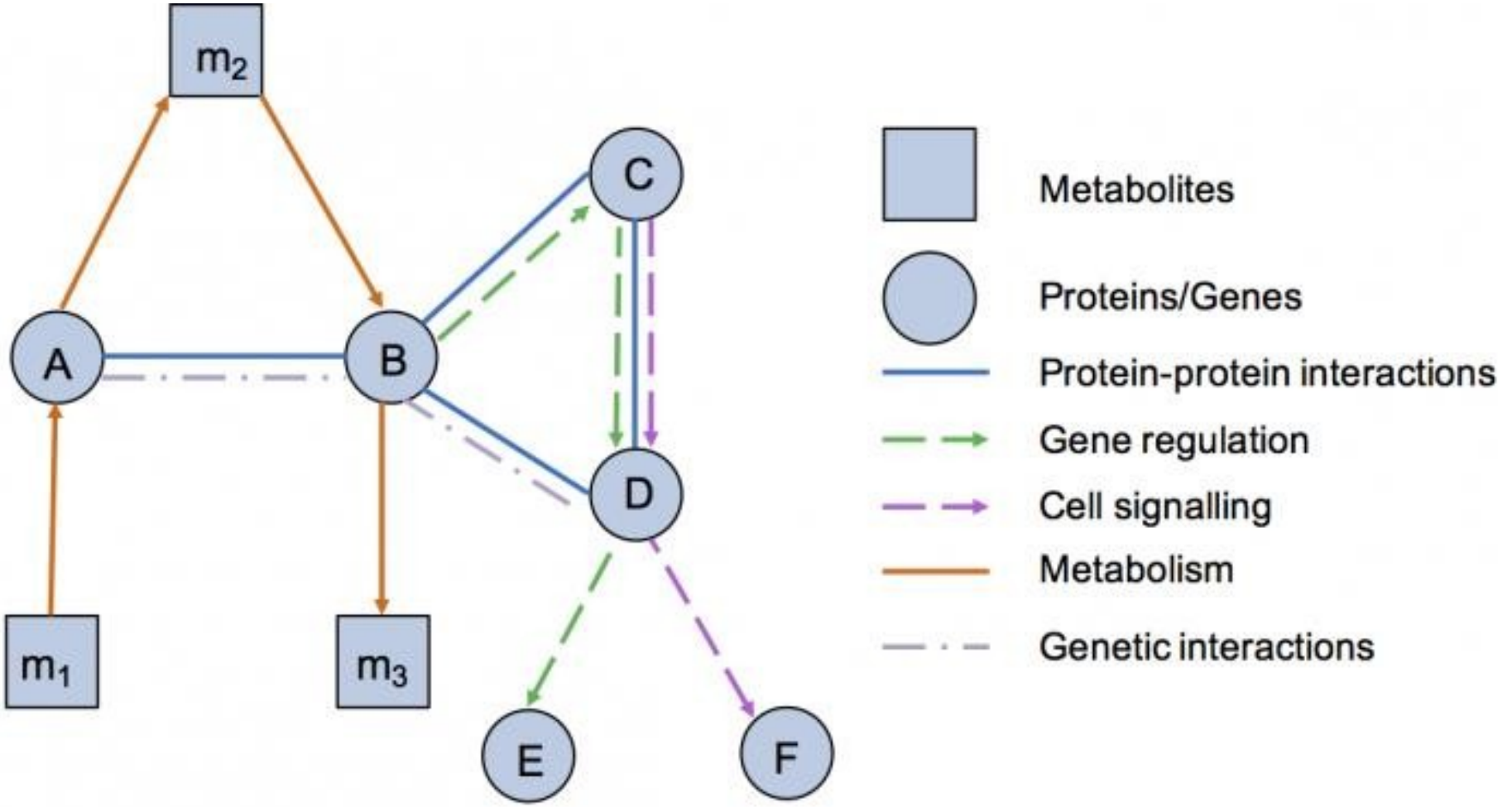
They can be switched on or off by **intermodular interactions and proteins** acting as high-level modulators.





Types of biological networks

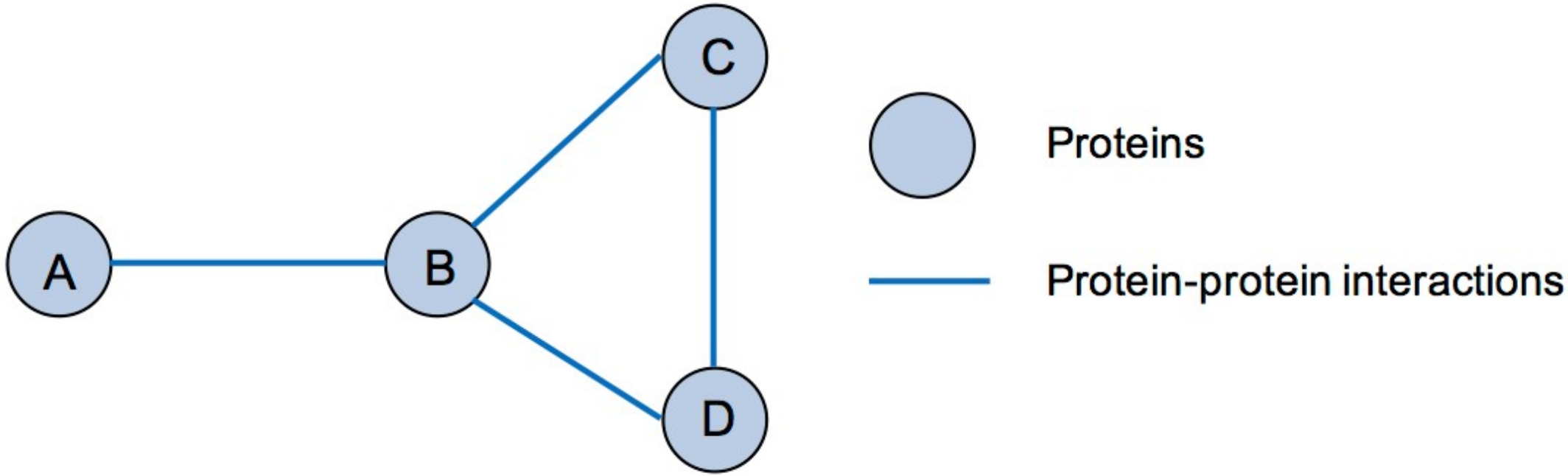
Different types of information can be represented in the shape of networks in order to model the cell:



Types of biological networks: Protein-protein interaction networks

Protein-Protein Interaction Networks (PPINs)

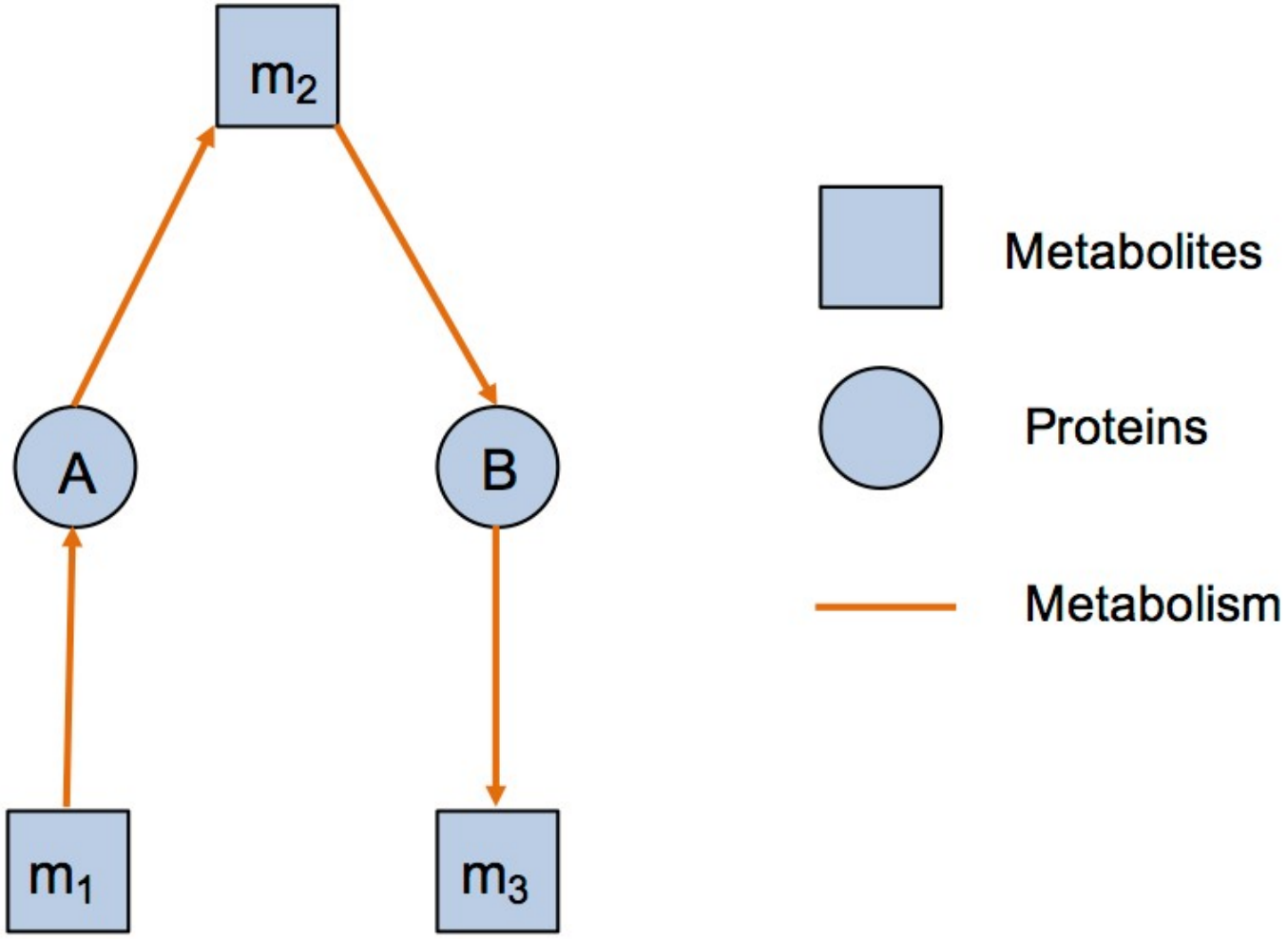
Represent the physical relationships between proteins. Proteins are represented as nodes that are linked by undirected edges.



Types of biological networks: Metabolic networks

Metabolic networks

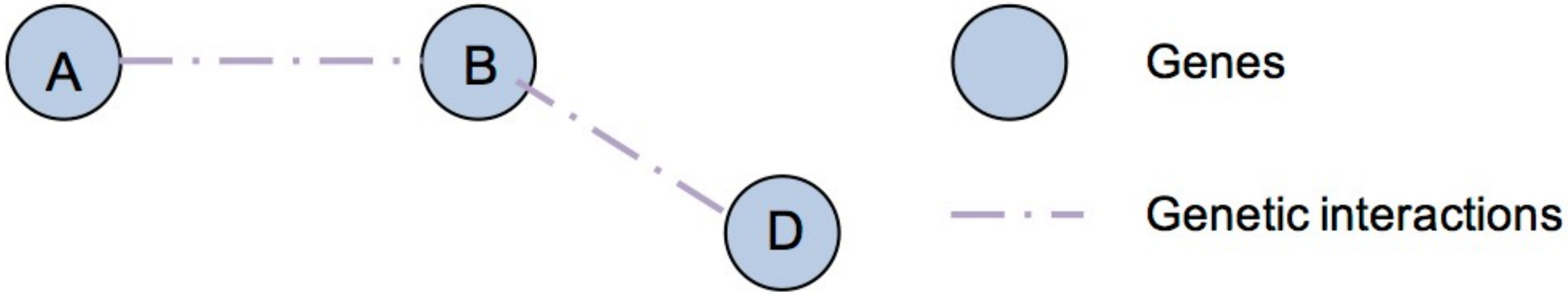
Represent the biochemical reactions that allow an organism to grow, reproduce, respond to the environment and maintain its structure. Metabolites and enzymes are represented as nodes and the reactions describing their transformations are represented as directed edges.



Types of biological networks: Genetic interaction networks

Genetic interaction networks

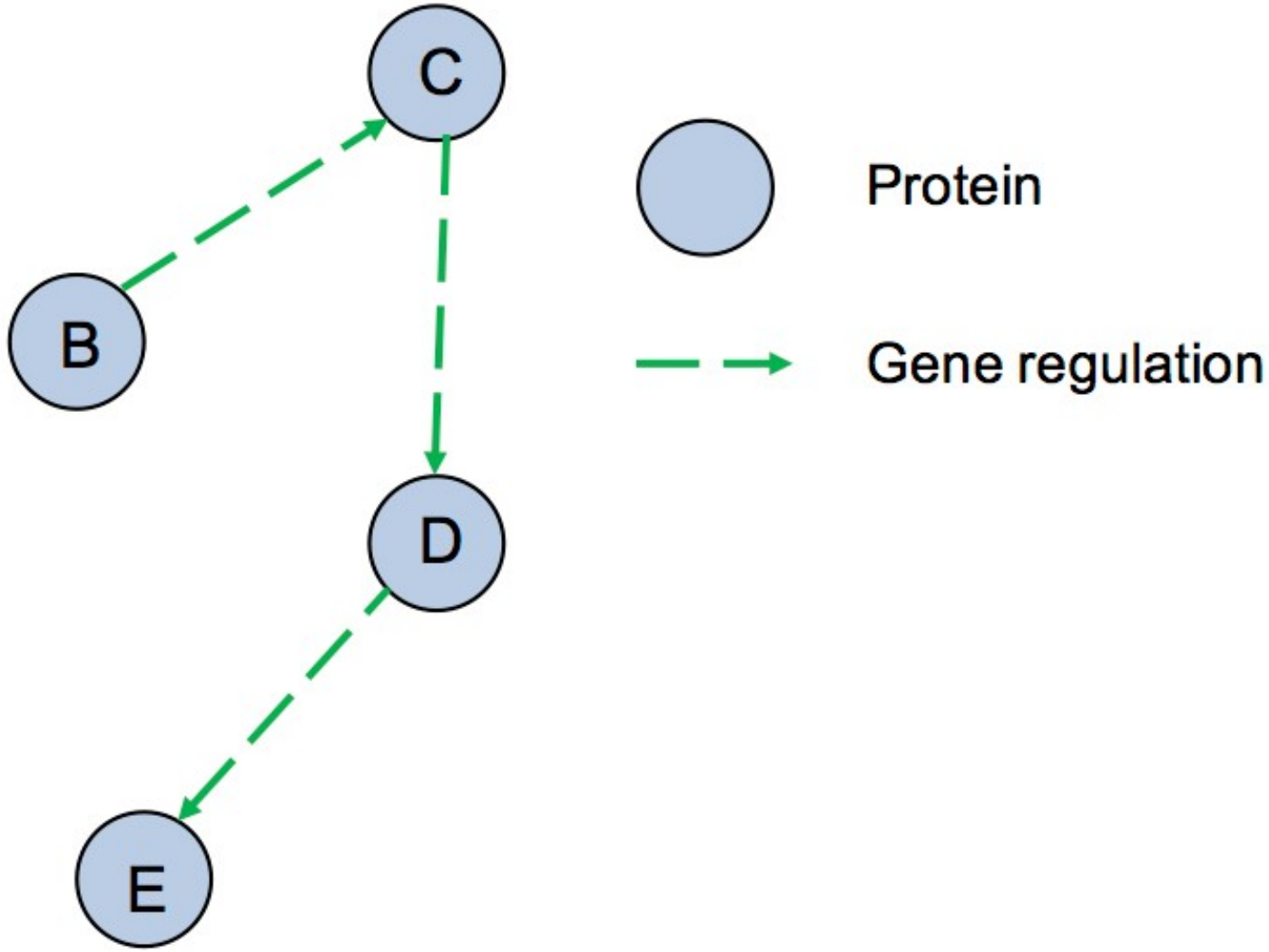
Genetic interaction is the synergistic phenomenon where the phenotype resulting from simultaneous mutations in two or more genes is significantly different from the phenotype that would result from adding the effects of the individual mutations. Genes are represented as nodes and their relationships as edges.



Gene/transcriptional regulatory networks

Gene/transcriptional regulatory networks

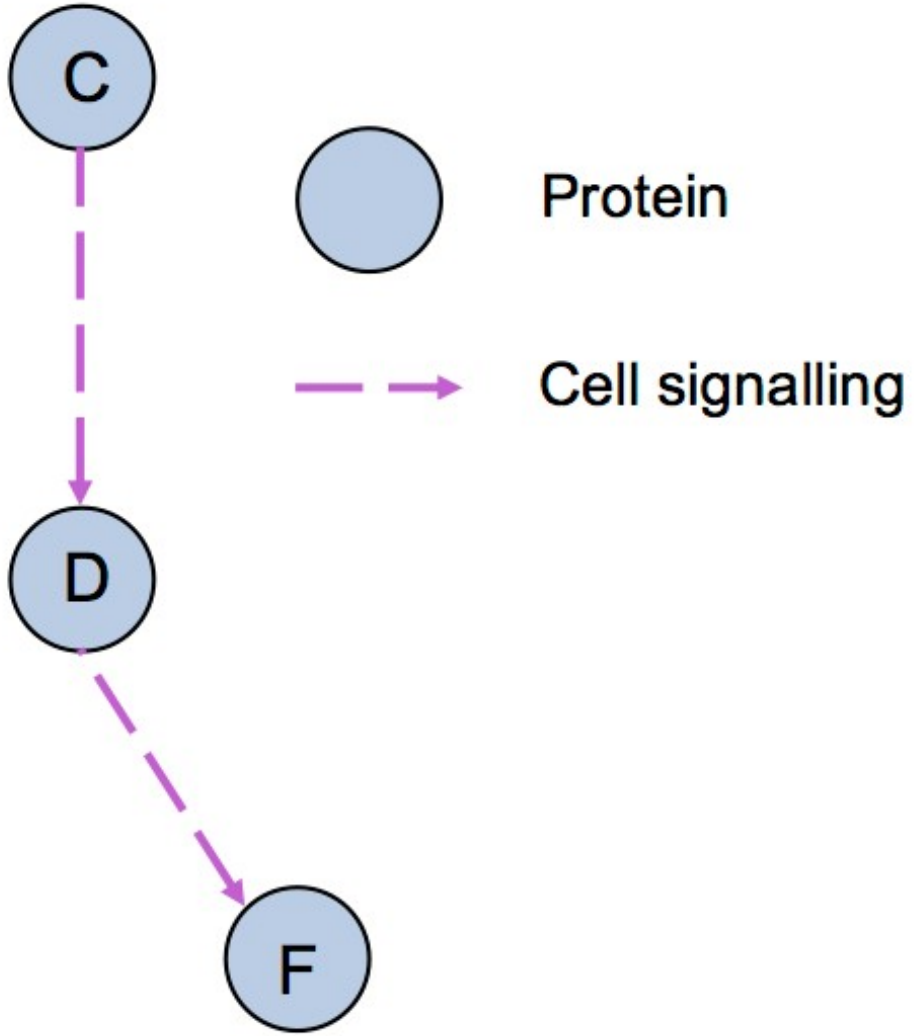
Represent how gene expression is controlled. Genes and transcription factors are represented as nodes, while the relationship between them is depicted by different types of directional edges.



Cell signaling networks

Cell signaling is the communication system that controls cellular activities. Signaling pathways represent the ordered sequences of events and model the information flow within the cell. Elements in the pathway (*e.g.* proteins, nucleic acids, metabolites) are represented as nodes and the flow of information is represented by directed edges. Two types of resources:

- **Pathway databases** *(e.g., Reactome, KEGG)*
- **Reaction network databases** *(e.g., Signor, SignaLink, SPIKE)*



Data sources underlying biological networks

Biological datasets are inherently noisy and incomplete. Sources:

- **Manual curation of scientific literature**: High-quality, but expensive and time consuming.
- **High-throughput datasets**: Large, systematically produced datasets, but biased by the chosen technique. *E.g. yeast two-hybrid, or affinity purification followed by mass spectrometry.*
- **Computational predictions**: Experimental evidence as the basis to predict unexplored relationships between biological entities. Datasets are noisy. *E.g. extrapolation from one species to another.*
- **Literature text-mining**: Computationally extract systematically represented relationships from the published literature. Datasets are noisy.

Protein-protein interation networks

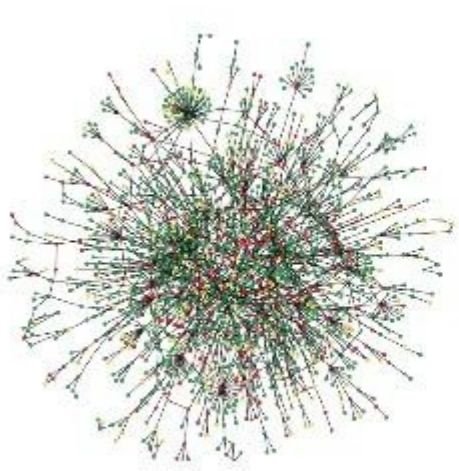
Protein-protein interactions (PPIs)

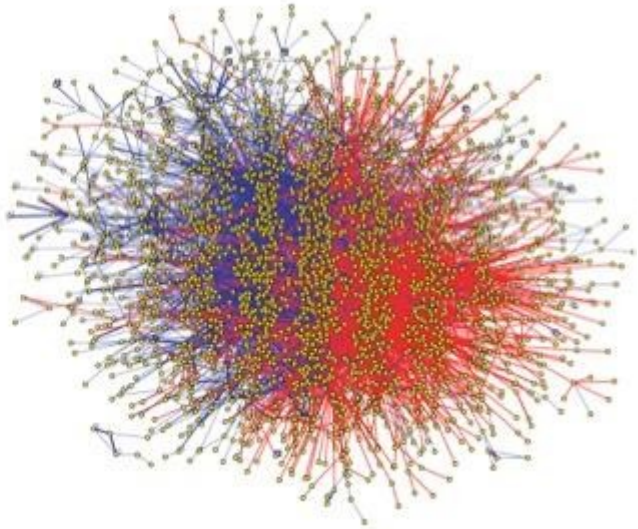
Physical contacts between proteins in the cell:

- Are specific
- Occur between defined binding regions
- Serve a specific function
- Can be stable (protein complexes) or transient (dynamic)

Interactome

The totality of PPIs that happen in a cell, an organism or a specific biological context.





Techniques: *high-throughput affinity purification + mass-spectrometry, yeast two-hybrid assay, bioinformatics prediction algorithms.* **Databases:***IntAct.*

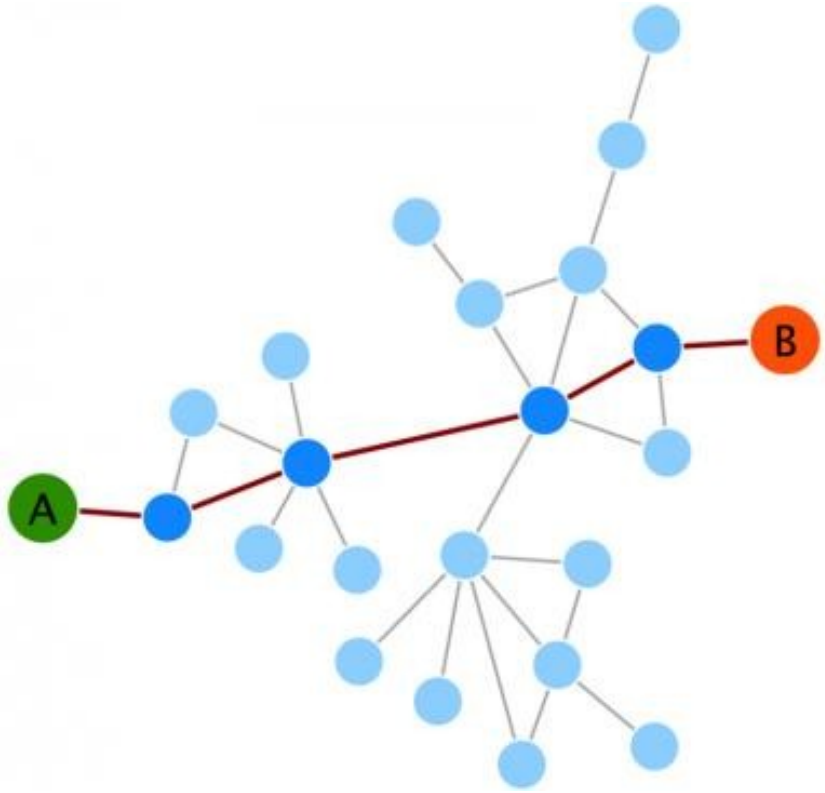
Properties of PPINs: small world effect

Small world effect

Great connectivity between proteins: the network's diameter (the maximum number of steps separating any two nodes) is small (less than six steps), no matter how big the network is.

Biological consequence: Efficient and quick flow of signals within the network

Question: Biological systems are extremely robust. If the network is so tightly connected, why don't perturbations in a single gene or protein have dramatic consequences for the network?



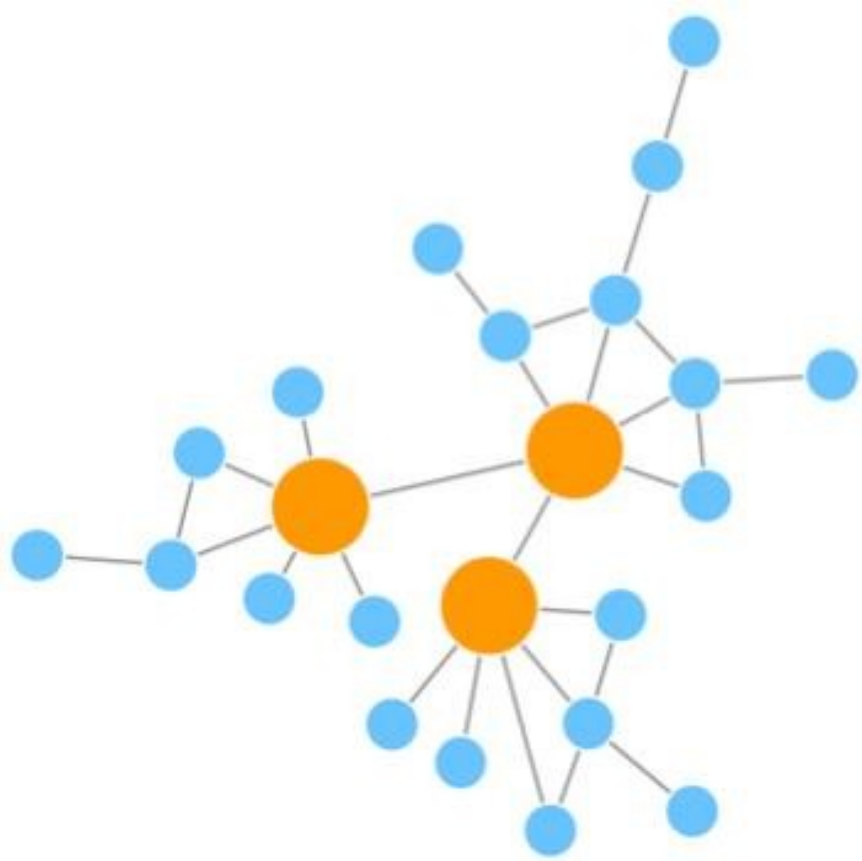
Properties of PPINs: scale-free networks

Scale-free networks

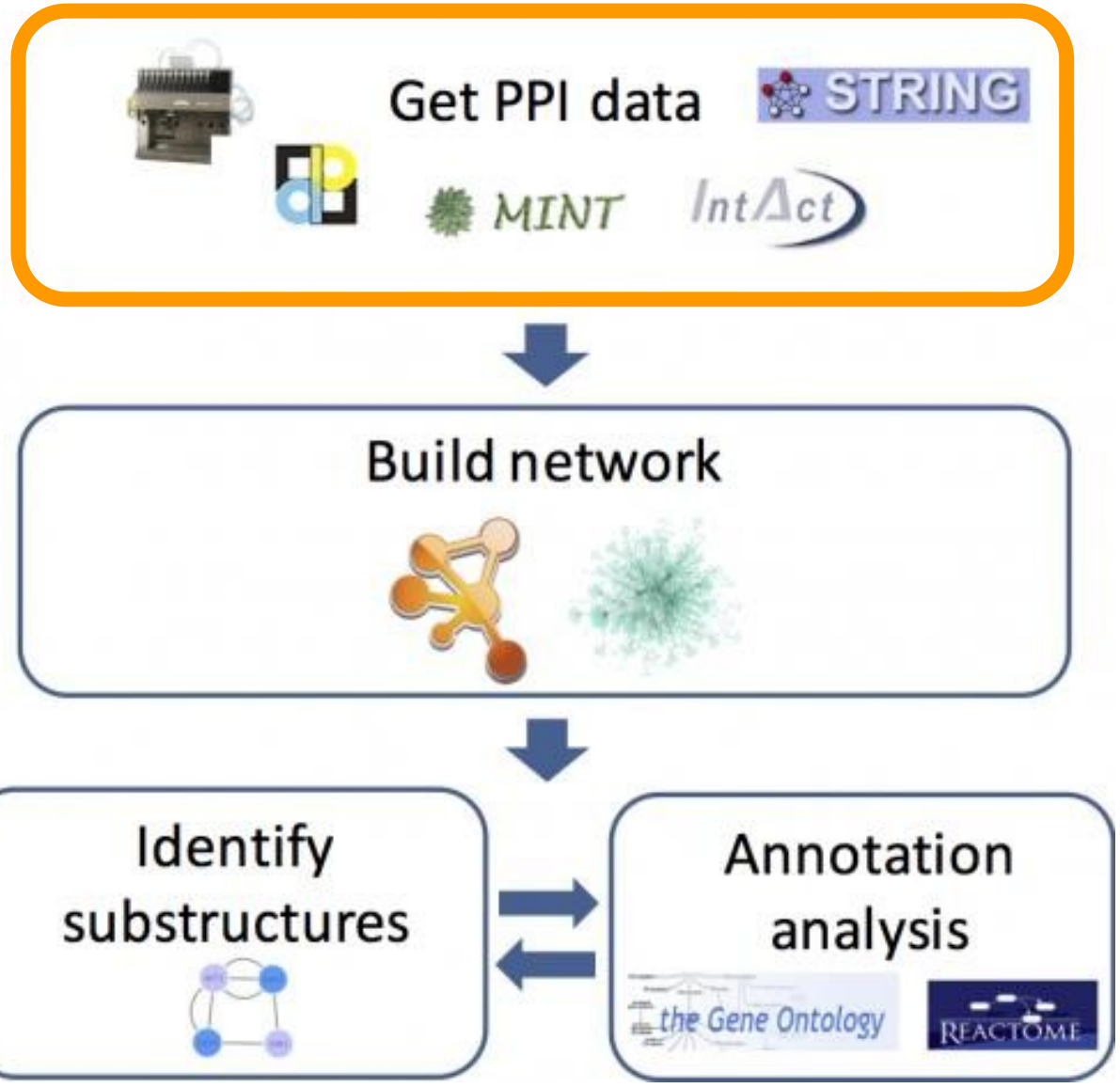
The majority of nodes (proteins) have only a few connections to other nodes *(small degree; blue)*, whereas some nodes (hubs) are connected to many other nodes in the network *(high degree; orange)*.

Biological properties:

- Stability
- Invariant to changes of scale
- Vulnerable to targeted attack



Building and analyzing PPINs: sources of PPI data



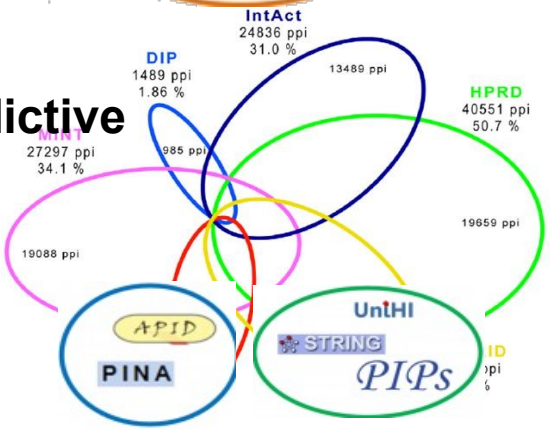
IMEx The International Molecular Exchange Consortium

Sources of PPI data

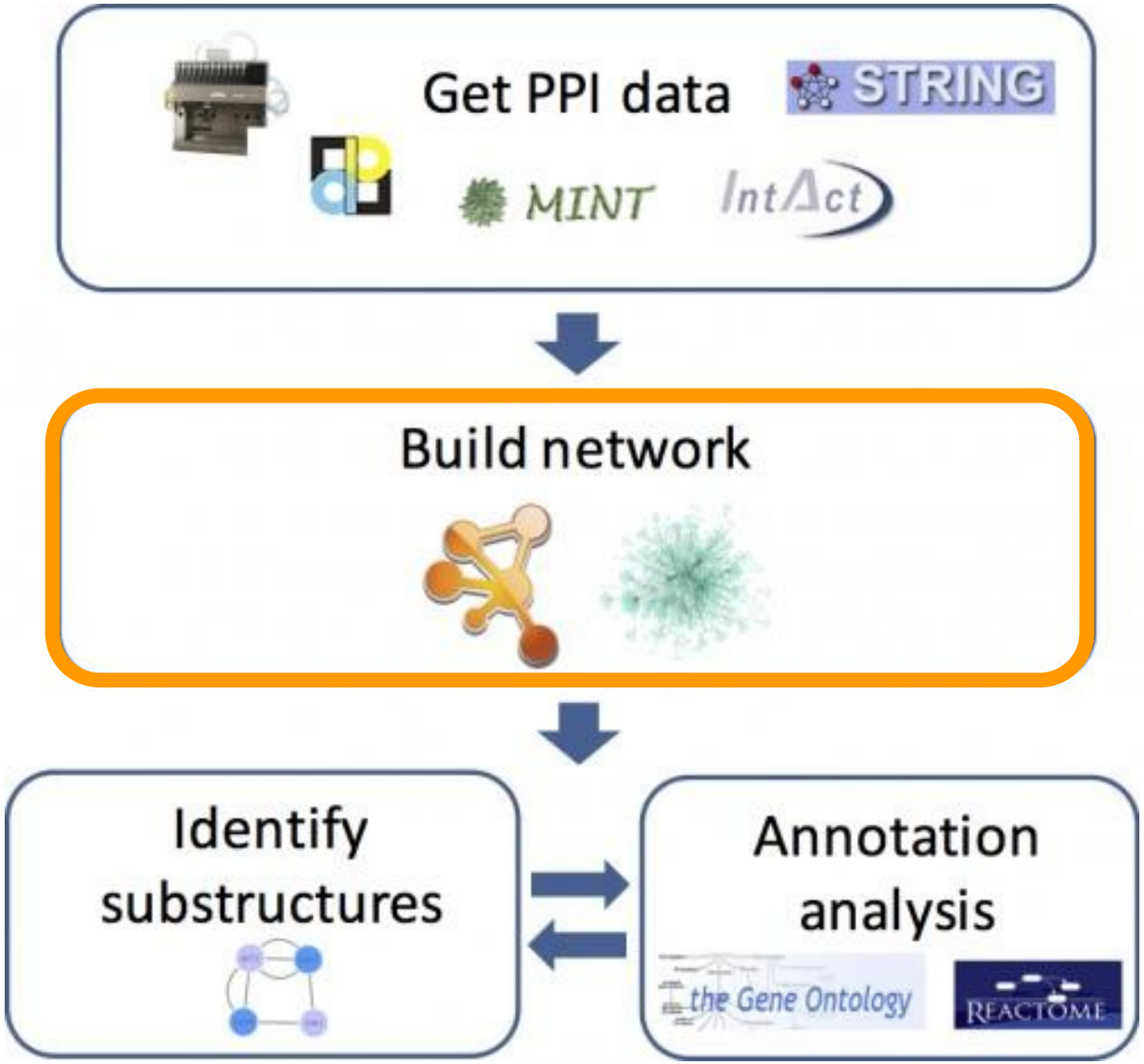
- **Experimental data**

- **Primary databases**

- **Secondary (Meta-) databases and predictive databases**

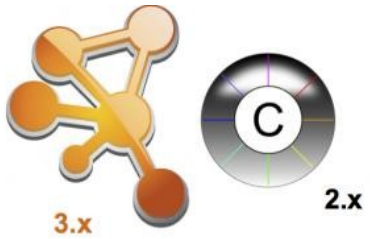


Building and analyzing PPINs: representation and analysis tools



Network representation and analysis tools

- **Cytoscape** (+apps)



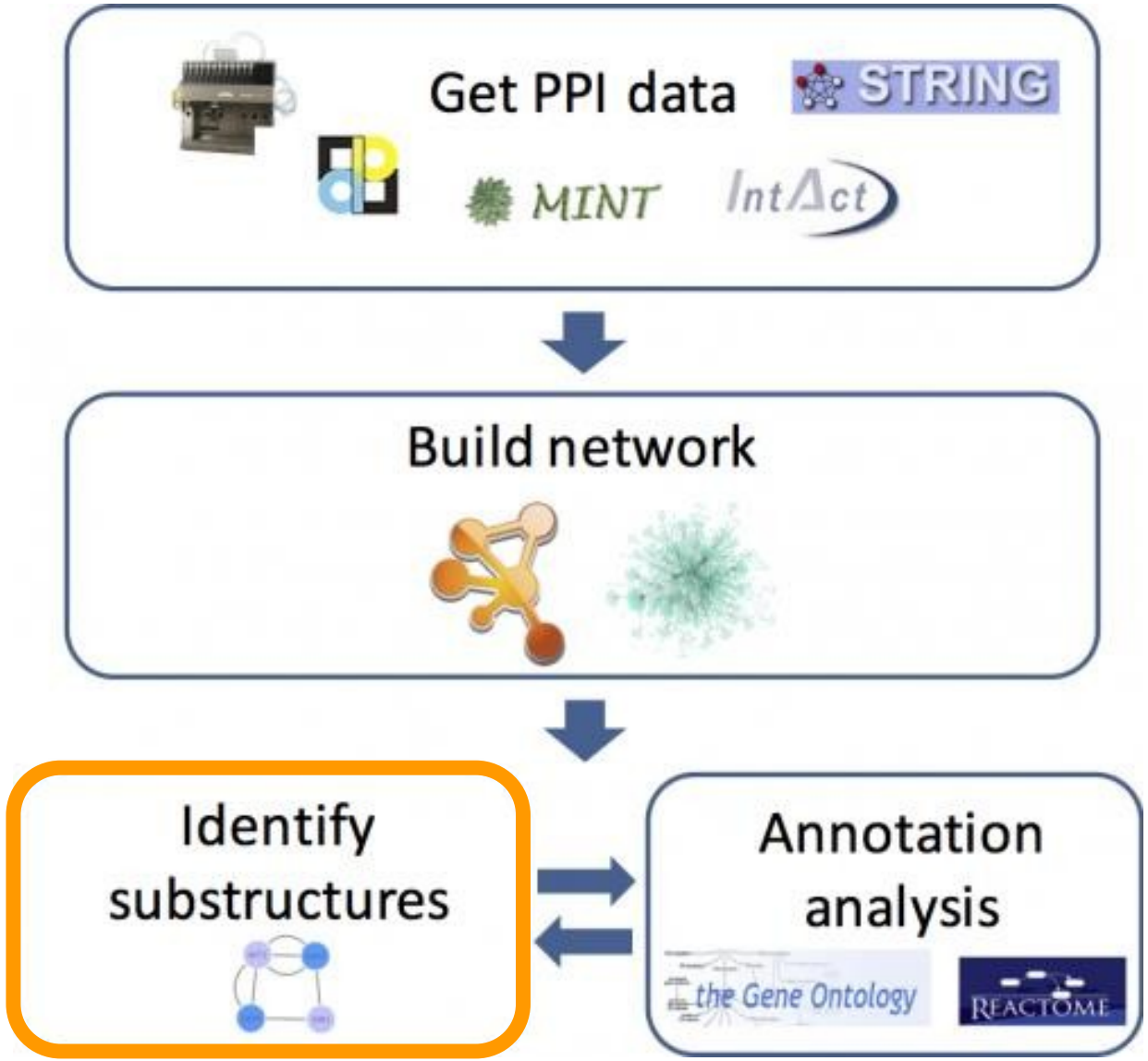
- **Gephi**

Gephi

- **Programmatic** (R, Python, C)

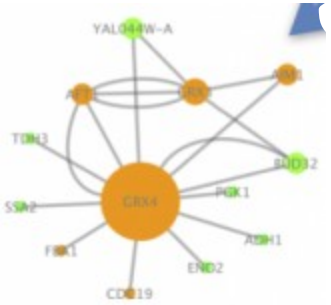
igraph NetworkX

Building and analyzing PPINs: topological PPIN analysis

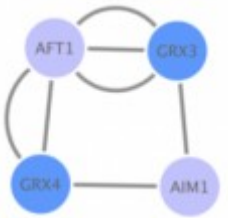


Topological PPIN analysis

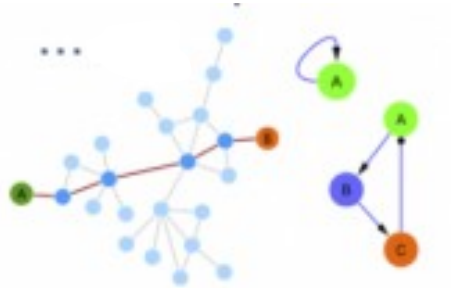
- **Centrality analysis**



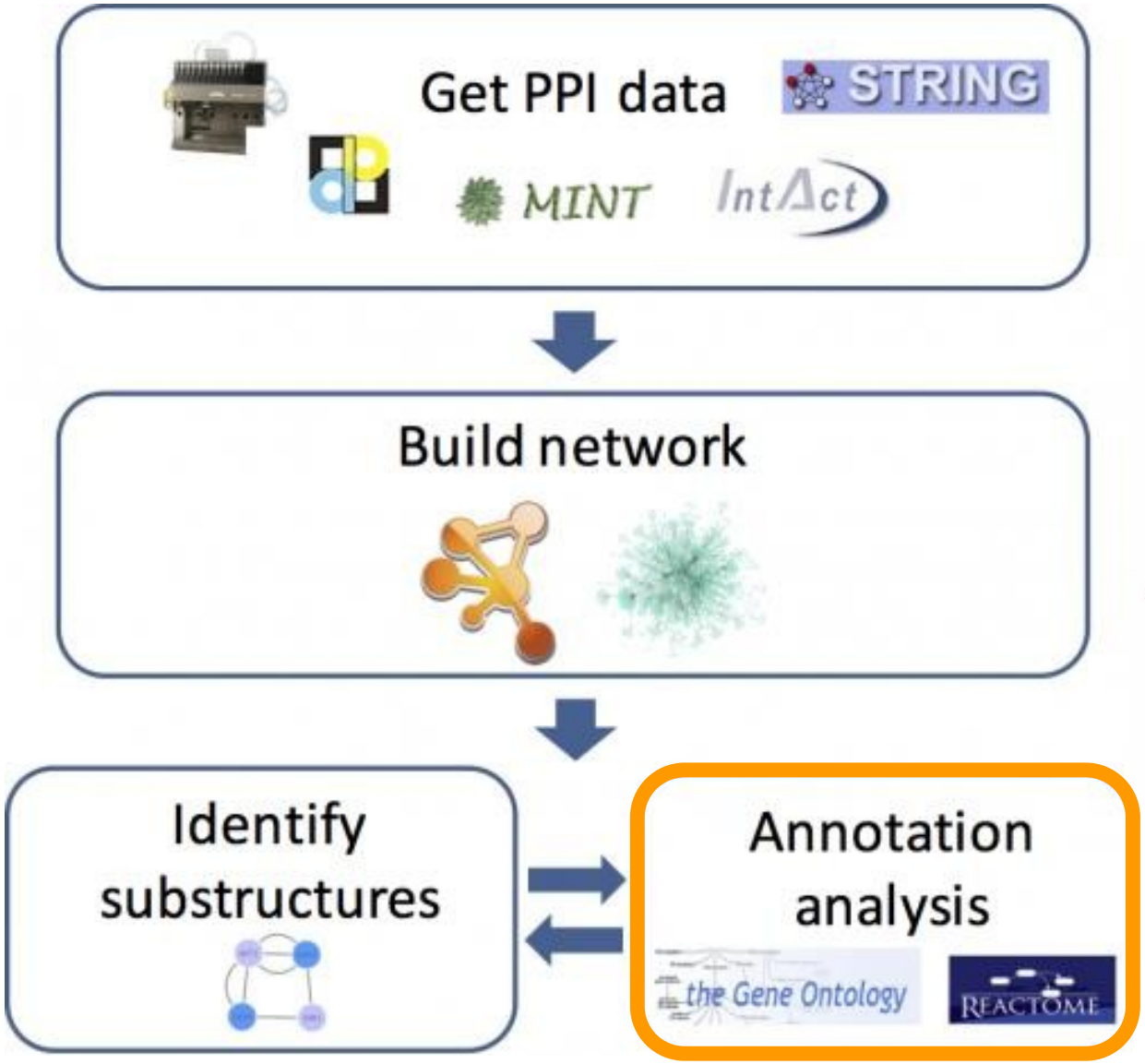
- **Topological clustering**



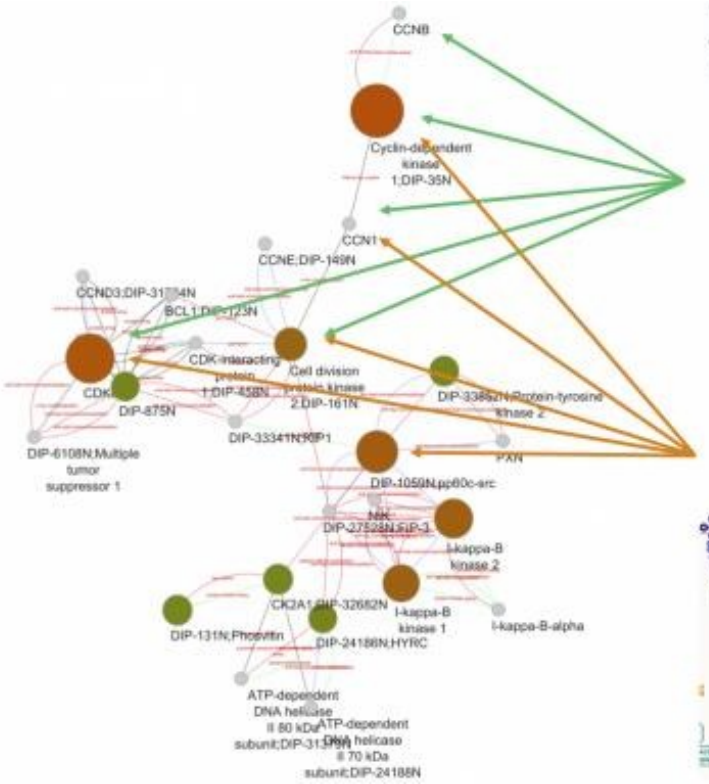
- **Shortest paths, motif search, etc.**



Building and analyzing PPINs: annotation enrichment analysis



Annotation enrichment analysis

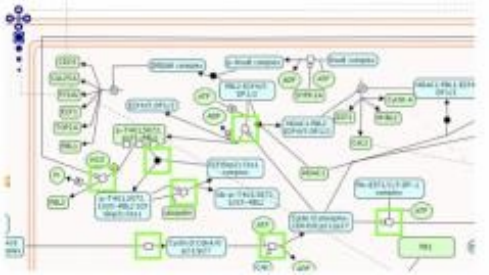


the Gene Ontology

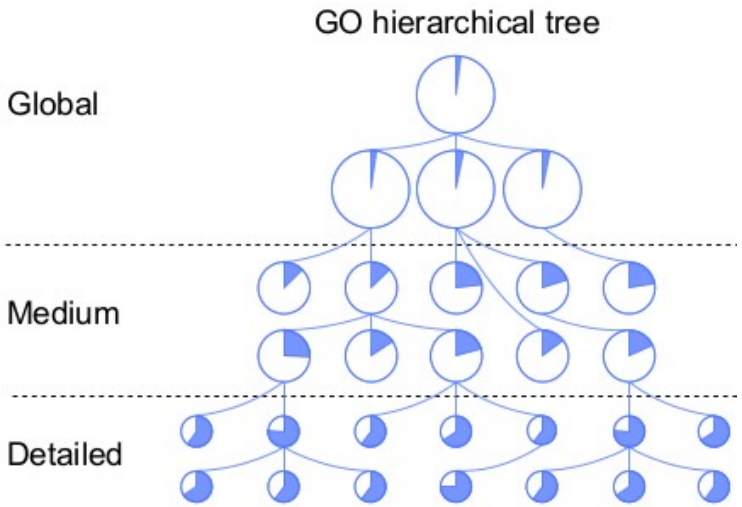
GO:0019908:
nuclear cyclin-dependent protein kinase holoenzyme complex

REACTOME

REACT_821.2:
Cyclin D associated events in G1

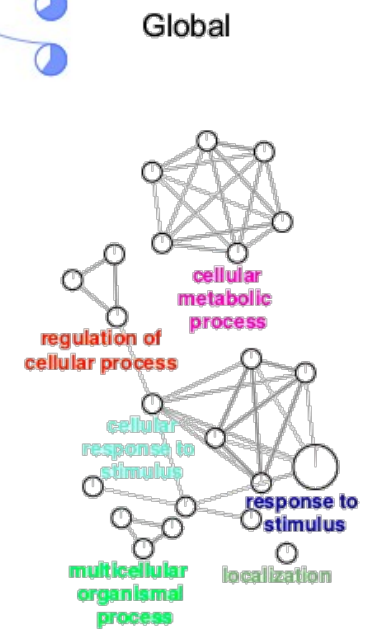


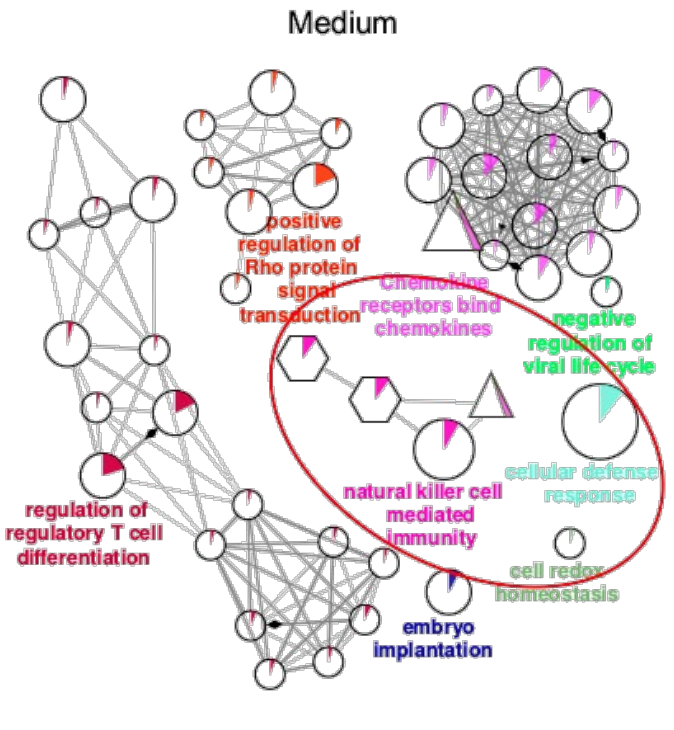
Building and analyzing PPINs: annotation enrichment analysis

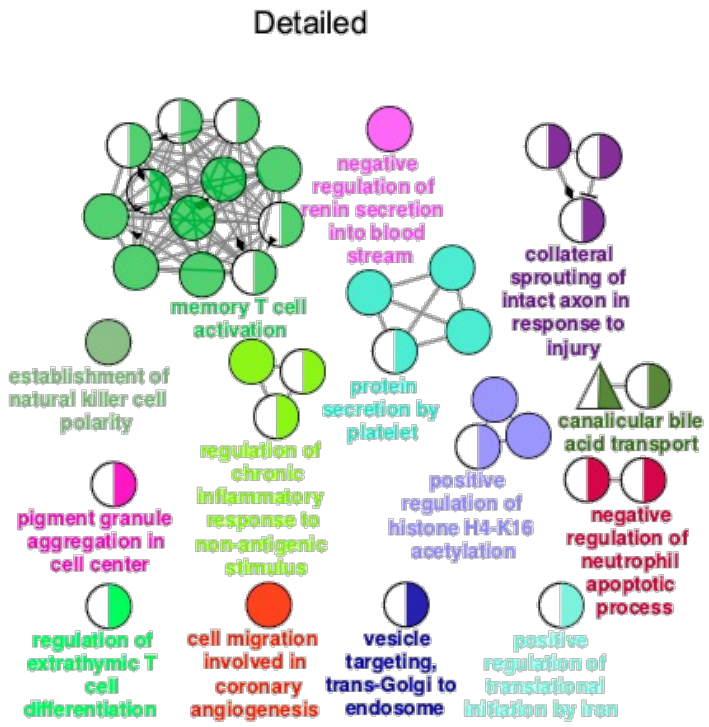


Problems:

- “Popular” proteins are better annotated (i.e. bias)
- Onthology granularity/depth







Summary

- **Biological networks**
  - Several types: genetic, metabolic, cell signaling, protein-protein interaction, etc.
  - Represented and analyzed using the tools provided by graph theory
  - Represented by nodes (entities) and edges (connections)
- **Protein-protein interaction networks**
  - Small-world effect: network diameter ~ 6 steps
  - Scale-free: few nodes (hubs) a lot more connected than average
  - Transitivity: communities/clusters
- **Analyzing PPINs**
  - Topological methods
    - Centrality analysis: degree centrality (local) vs. closeness/betweenness (global)
    - Clustering analysis: community detection; functional modules vs protein complexes
  - Annotation enrichment analysis