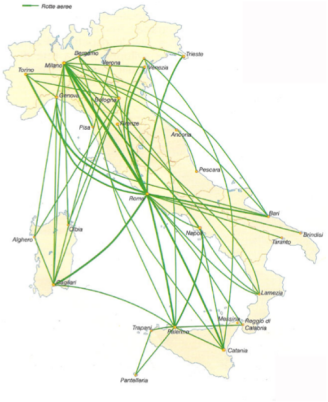
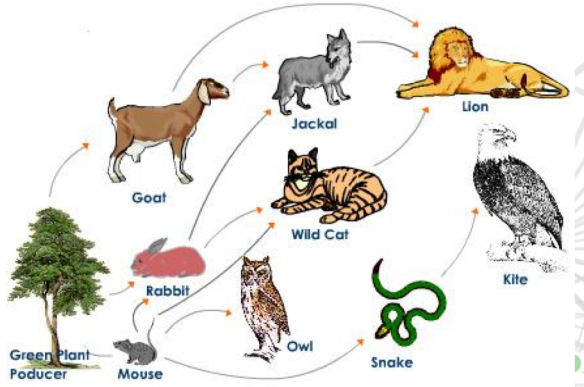
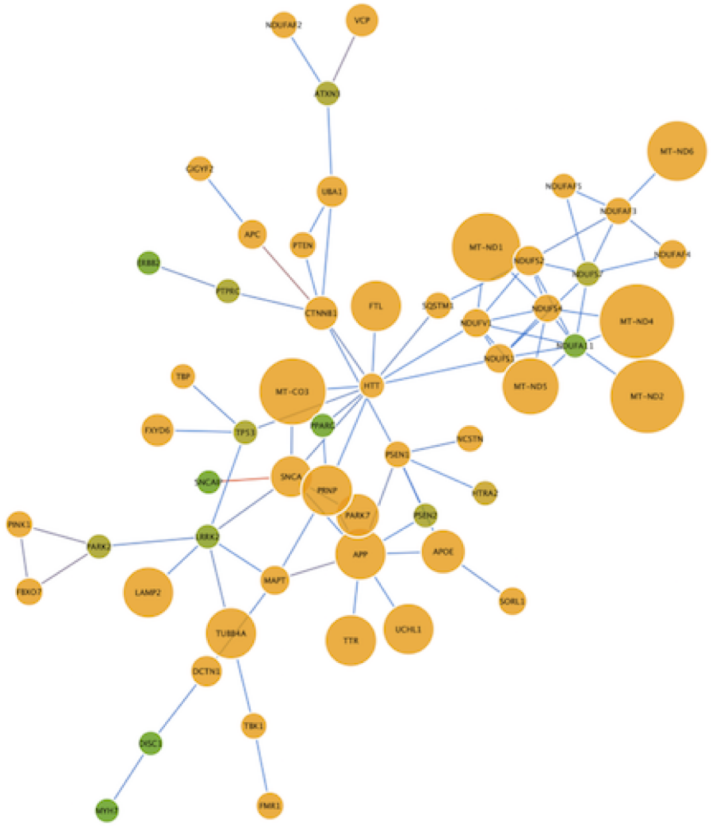
Networks are Everywhere…



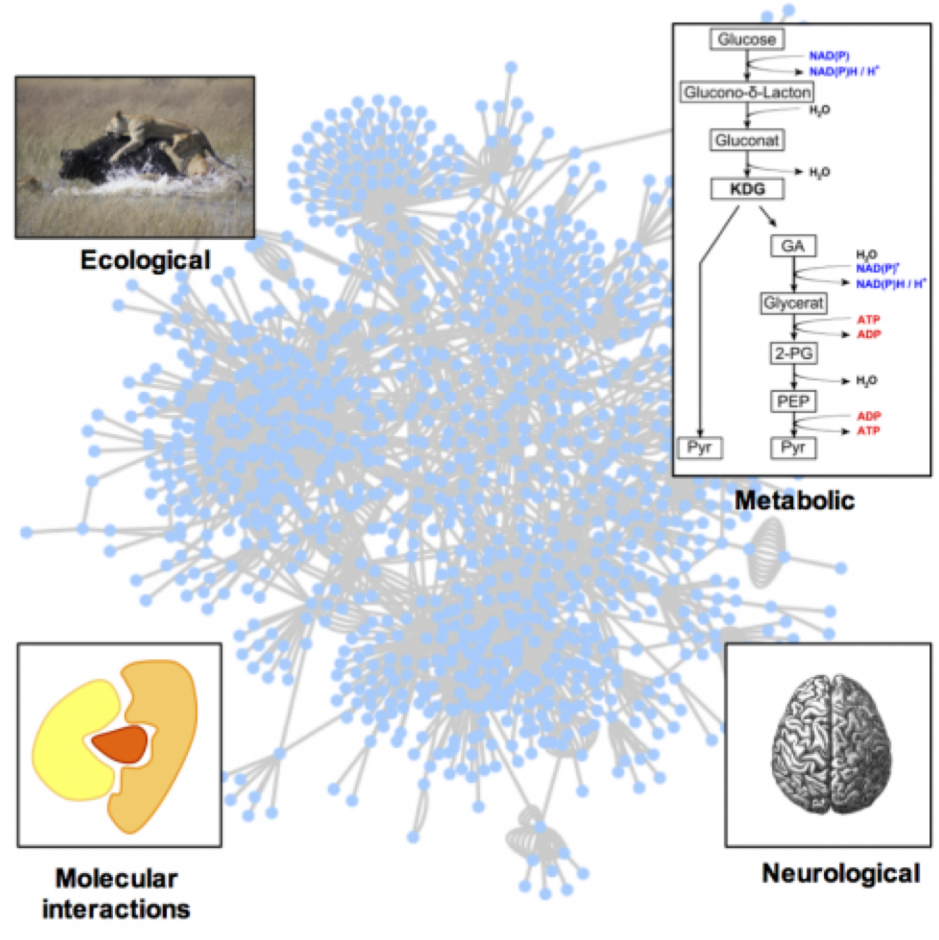


Network analysis of protein interaction data: an introduction



The content of these slides was extracted from the EBI Train Online at:
https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction
Author: Pablo Porras Millan

Network analysis in biology



- Systems biology
  - biological entities at the systemic level
  - not as individual components
  - interacting systems
- Network biology
  - Graph theory

Introduction to graph theory

- social network analysis
- application of graph theory to the social sciences

"[…] 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".
(Wikipedia)

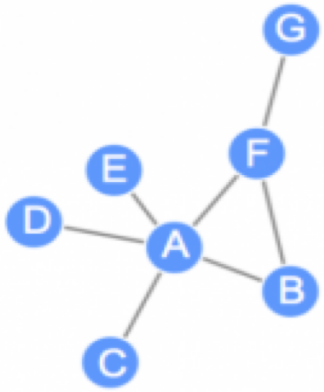
The seven bridges of Königsberg

Euler used this graph and its topological features to prove that the path did not exis

Graph theory: graph types and edge properties

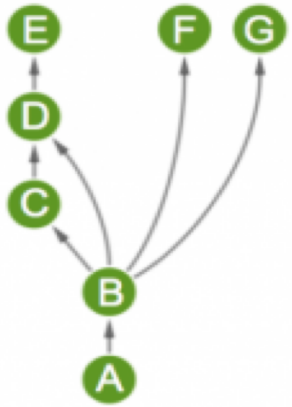
Types of network edges

Undirected



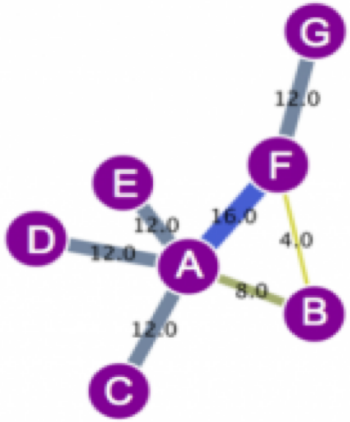
- PPI networks

Directed



- Metabolic networks
- Gene regulation networks

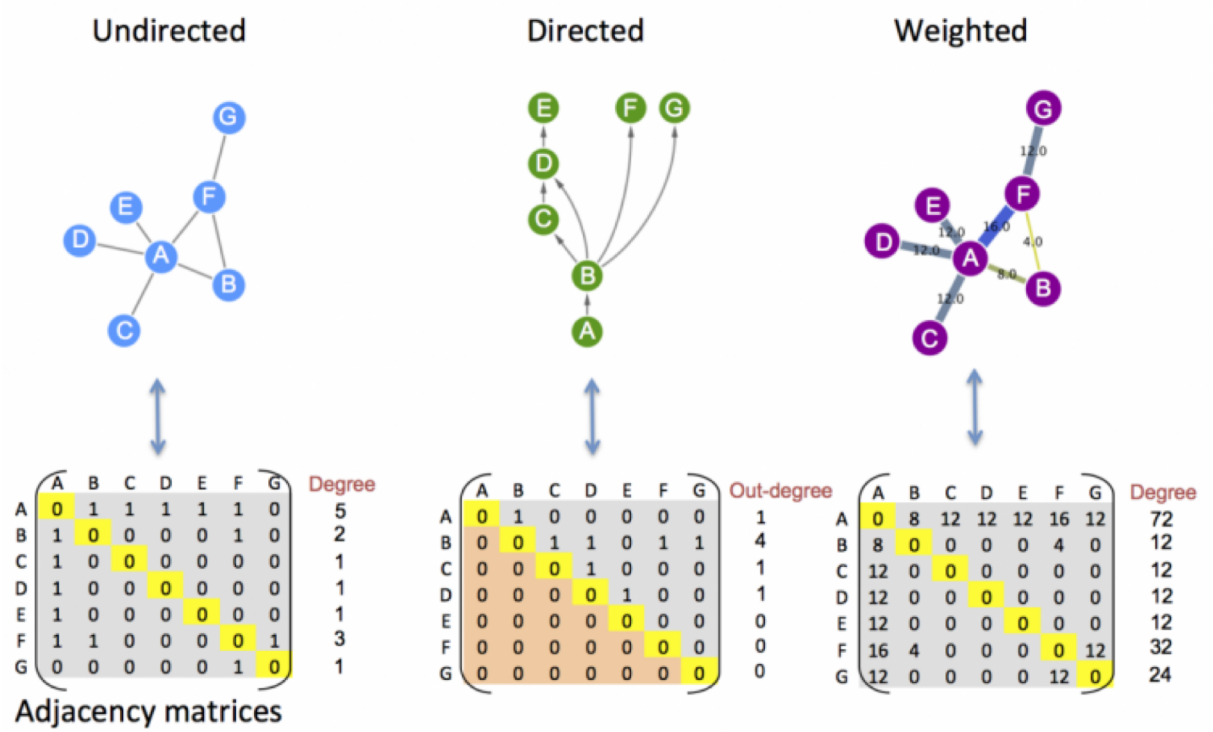
Weighted



- Reliability of an interaction
- Quantitative expression change that a gene induces over another
- Relatedness of two genes in terms of seq similarity

Graph theory: adjacency matrices

Rows and columns are assigned to the nodes in the network and the presence of an edge is symbolised by a numerical value

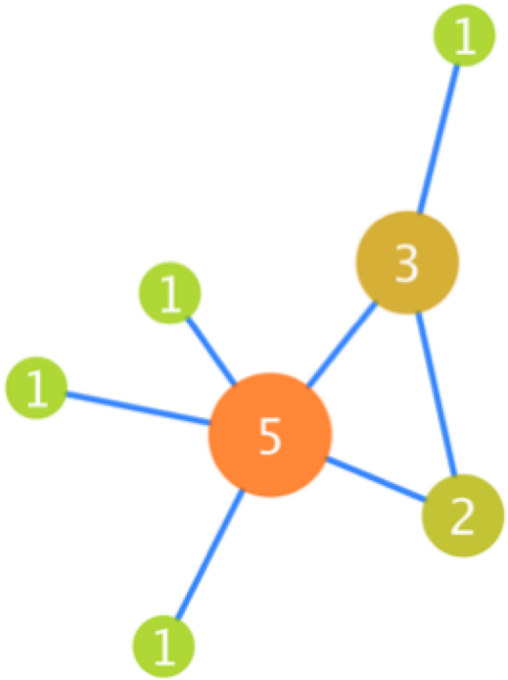


Sign of the values can be used to indicate stimulation or inhibition

Graph theory: network topology

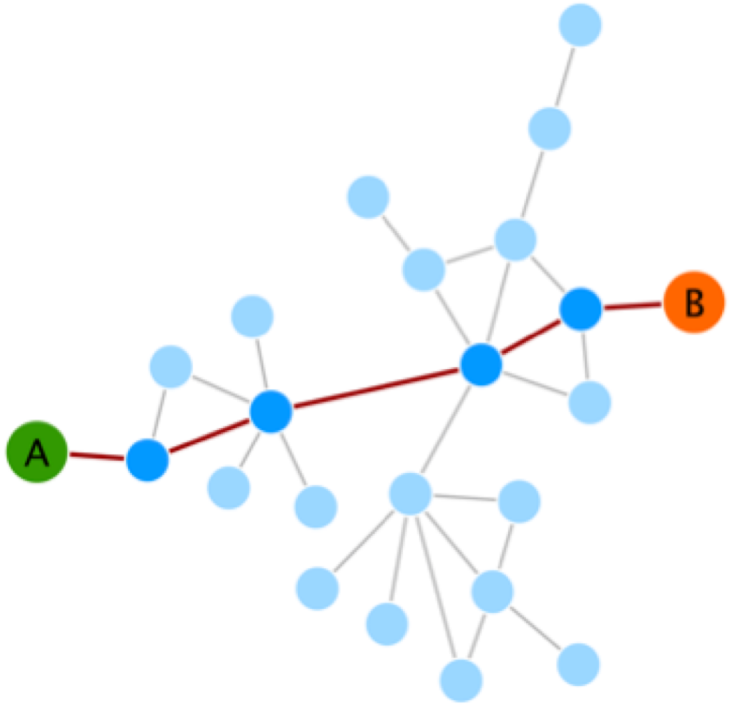
topological properties help identify relevant sub-structures within a network

Degree



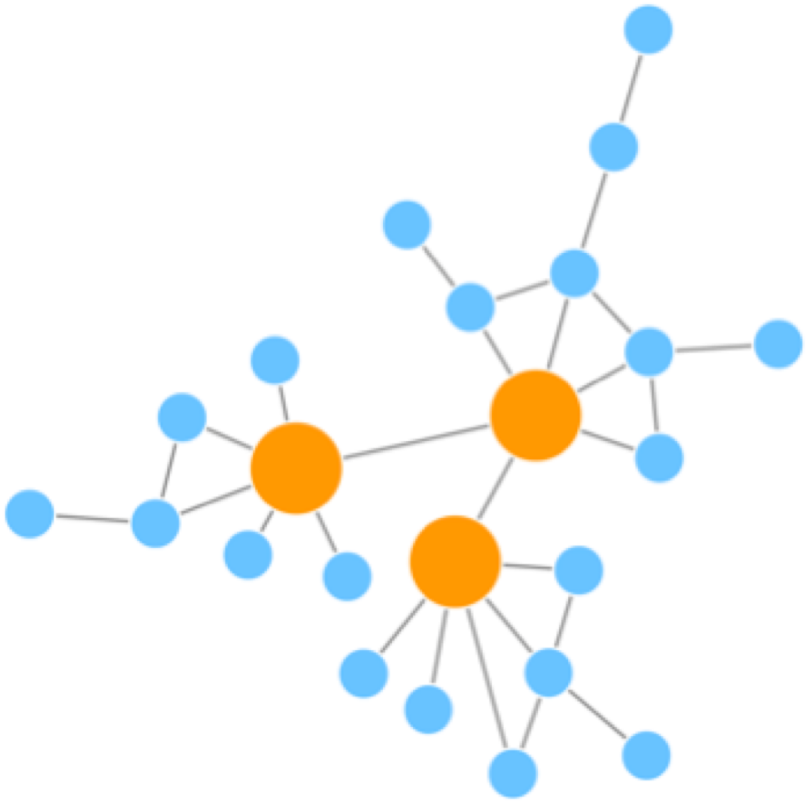
Graph theory: network topology

Shortest paths



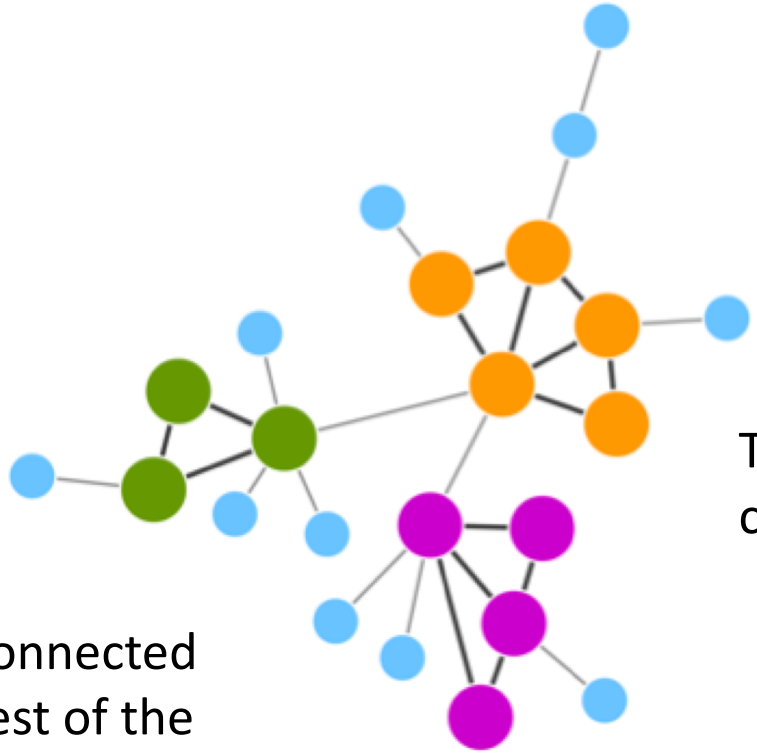
Graph theory: network topology

Scale-free



Graph theory: network topology

Transitivity

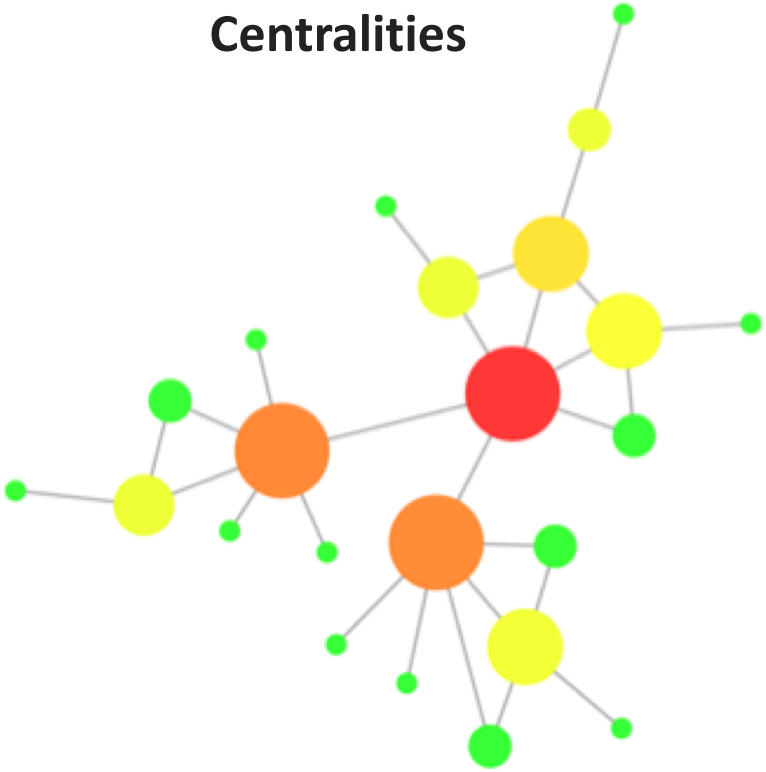


Topological clusters or communities

Nodes more internally connected than they are with the rest of the network

Graph theory: network topology

Centralities



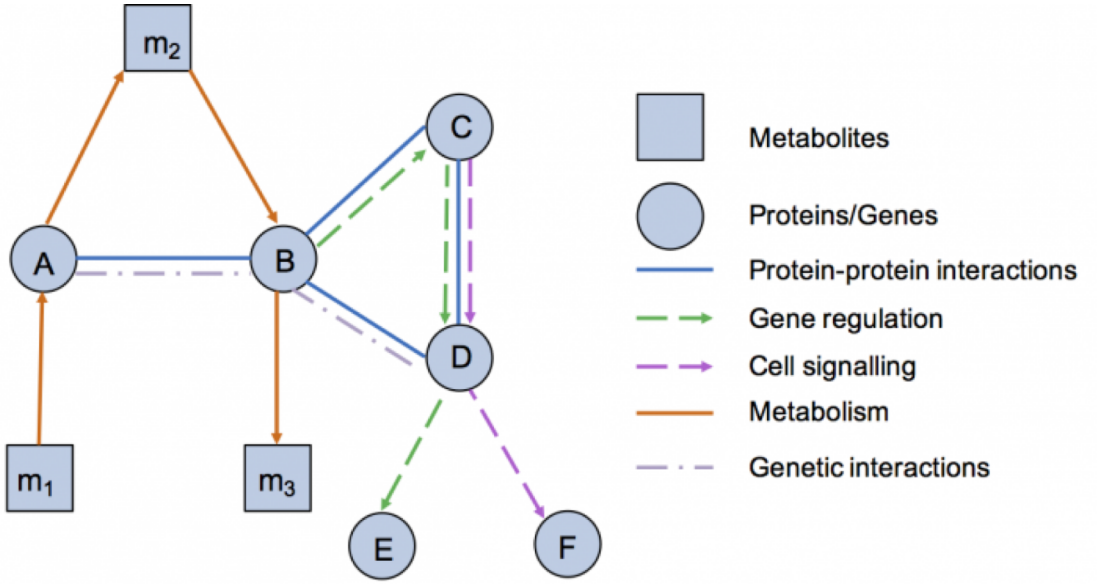
Can be measured for nodes and for edges

Estimation on of how important the node/edge is for the connectivity or the information flow of the network

Examples:

- Degree centrality (node size)
- Betweenness centrality (warm colours)

Types of biological networks

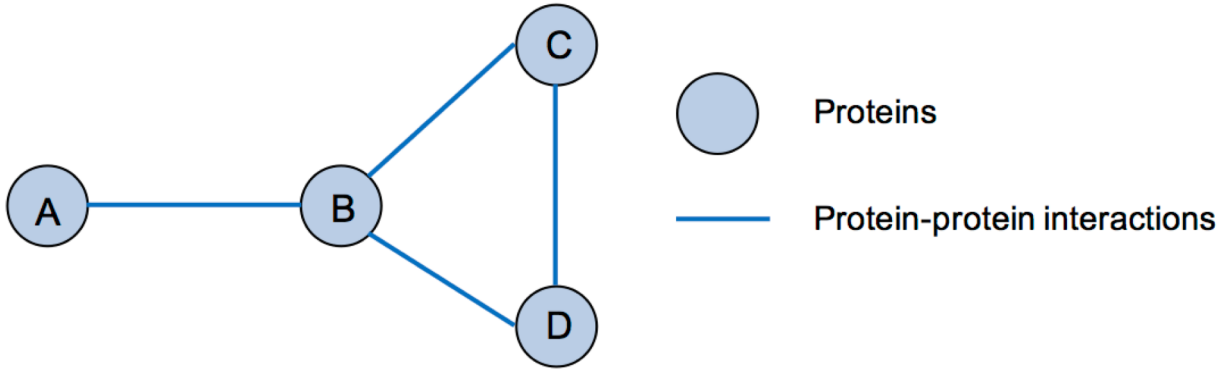


Some of the most common types of biological networks are:

1. Protein-protein interaction networks
2. Metabolic networks
3. Genetic interaction networks
4. Gene / transcriptional regulatory networks
5. Cell signalling networks

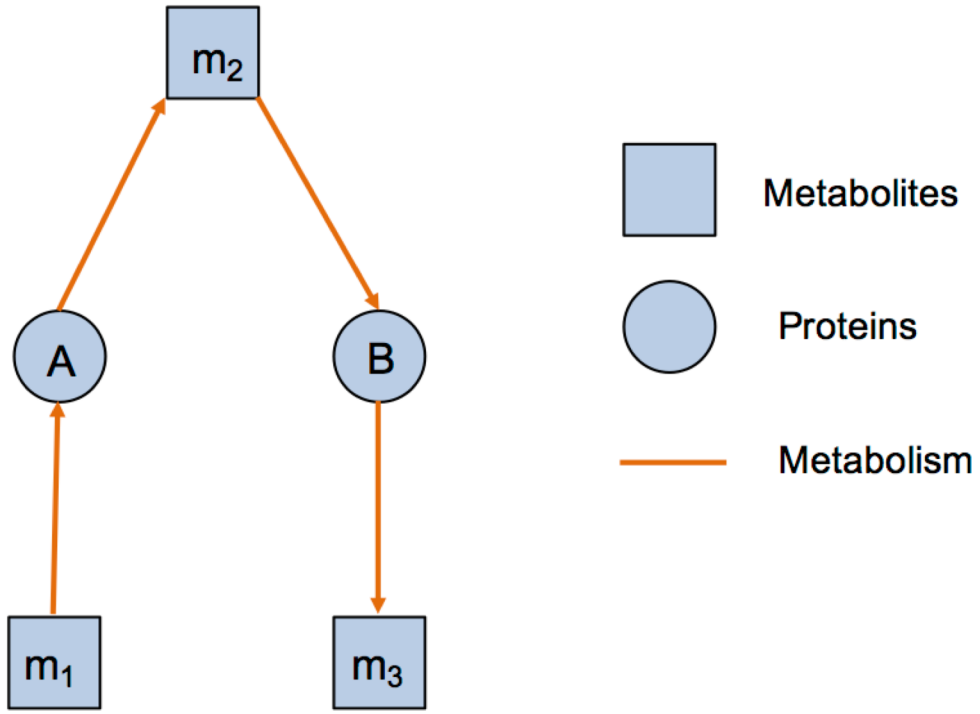
Types of biological networks

Protein-protein interaction networks



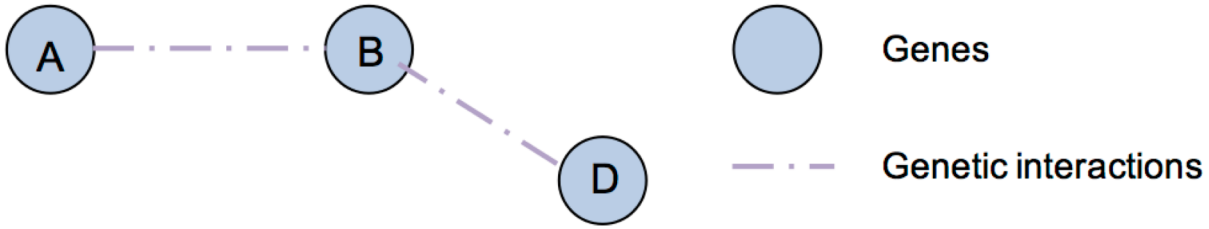
Types of biological networks

Metabolic networks



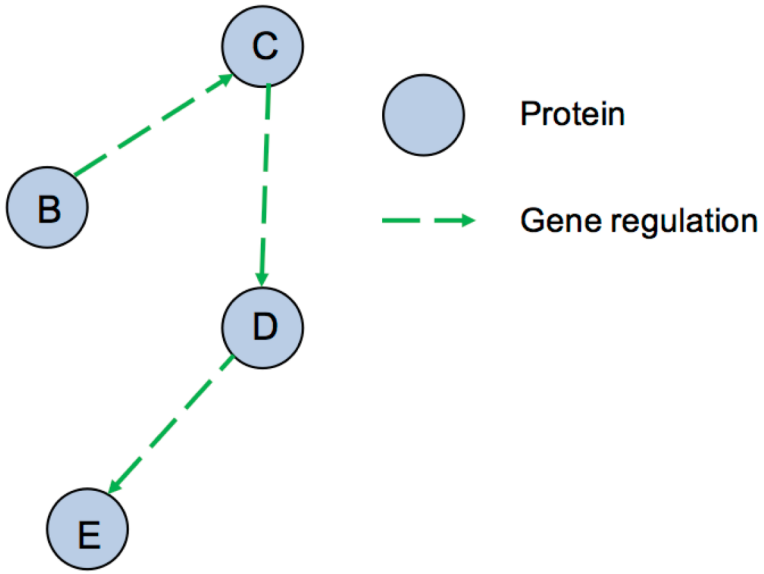
Types of biological networks

Genetic interaction networks



Types of biological networks

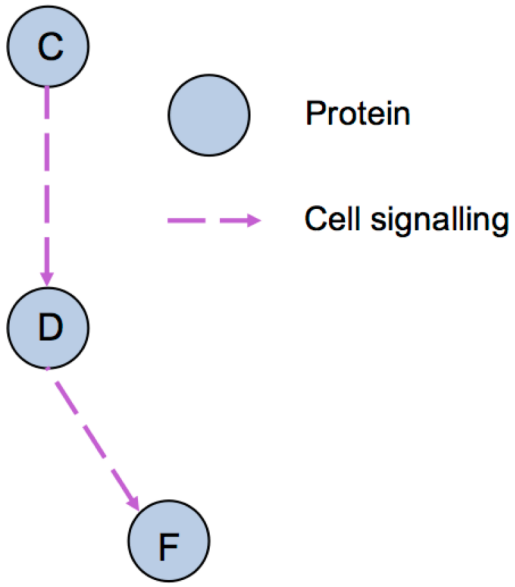
Gene / transcriptional regulatory networks



Sources: databases representing consensus knowledge of gene regulation (e.g. Reactome or KEGG)

Types of biological networks

Cell signaling networks



Sources: Pathway databases, Reaction network databases

The sources of data underlying biological networks

Biological datasets are noisy and incomplete

- **Manual curation of scientific literature**
- **High-throughput datasets**
- **Computational predictions**
- **Literature text-mining**

Protein-protein interaction networks

PPI networks are mathematical representation of the physical contacts between proteins in the cell

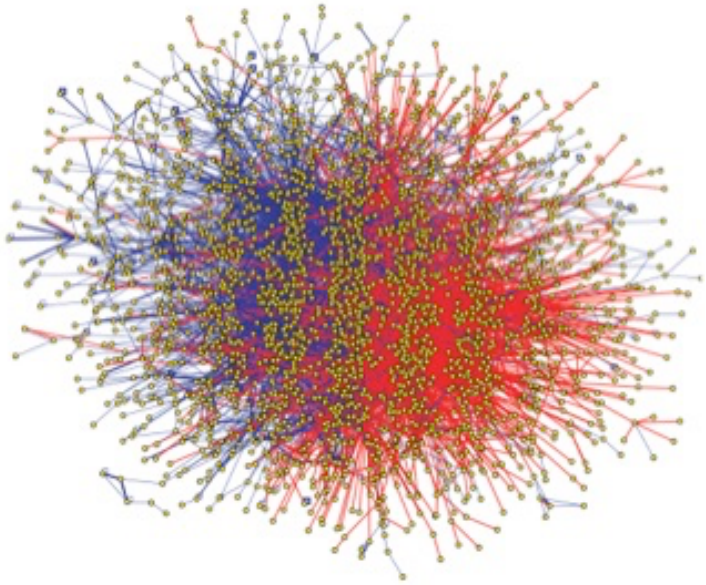
Knowledge of PPIs can be used to:

- assign putative roles to uncharacterised proteins;
- add fine-grained detail about the steps within a signalling pathway; or
- characterise the relationships between proteins that form multi-molecular complexes such as the proteasome.

Protein-protein interaction networks

The **interactome** is the totality of PPIs that happen in a cell, an organism ot a specific biological context.

Yeast

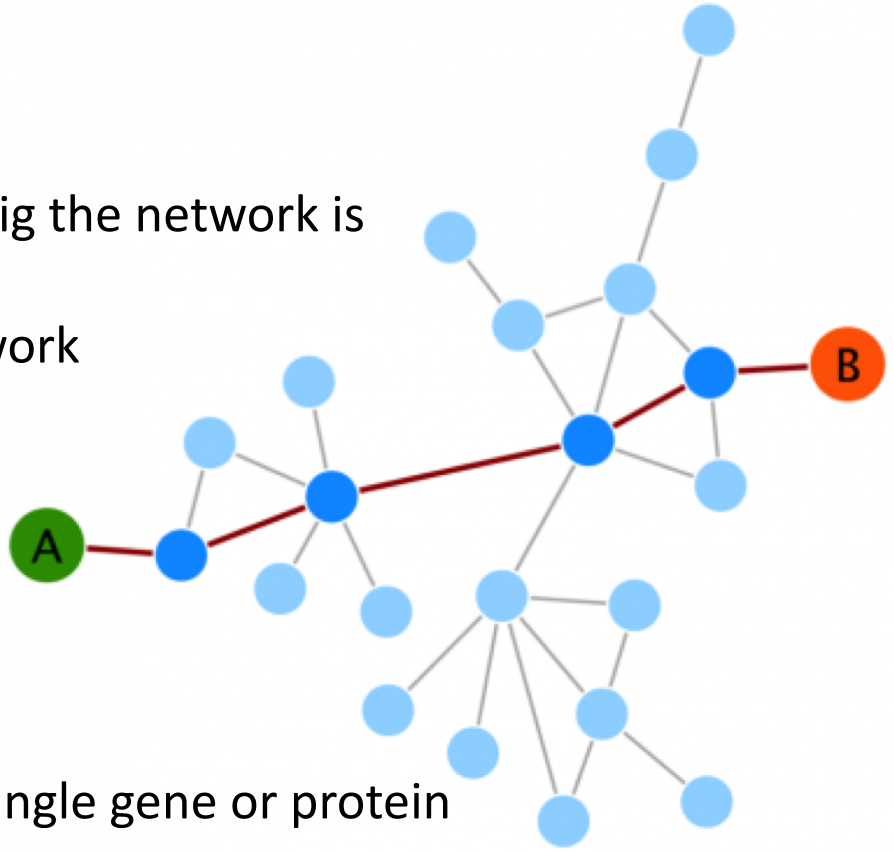


Human

Our current knowledge of the interactome is both **incomplete** and **noisy**.

Properties of PPINs: small world effect

- great connectivity between proteins
- the network's diameter is small, no matter how big the network is
- “six degrees of separation”
- efficient and quick flow of signals within the network



- Biological systems are extremely robust
- no dramatic consequences of perturbations in a single gene or protein

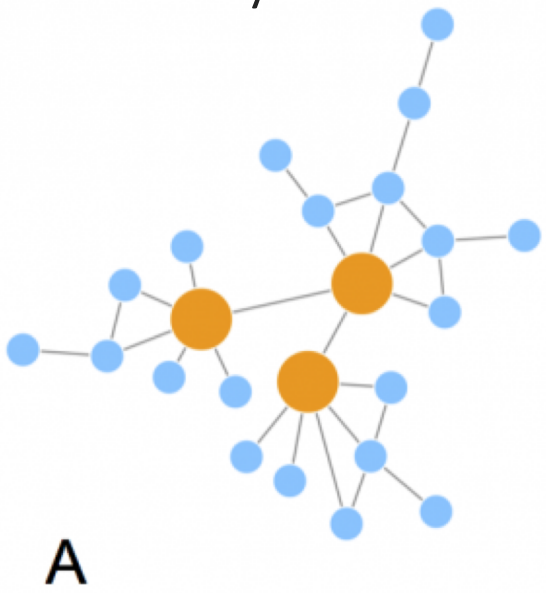
Diameter: the maximum number of steps separating any two nodes

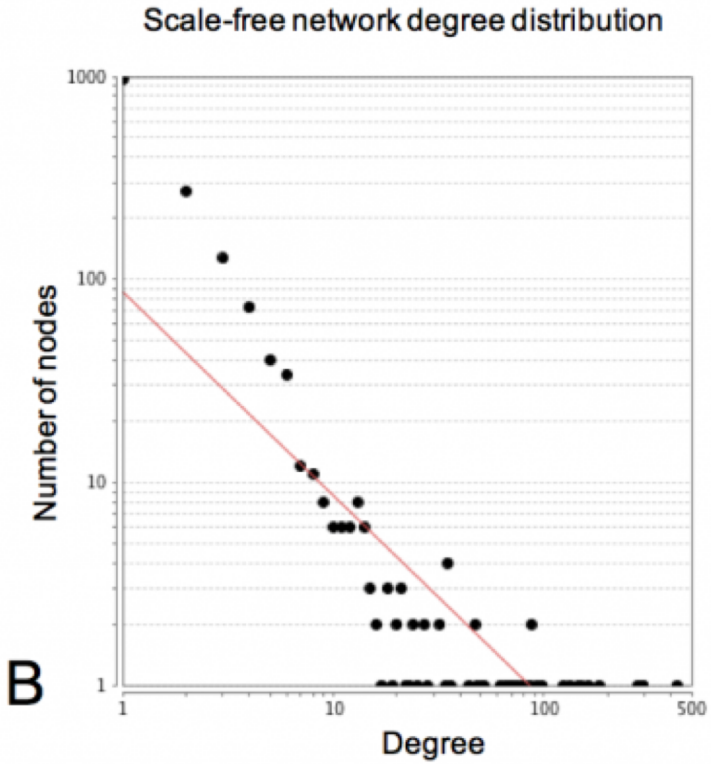
Properties of PPINs: scale-free networks

- The majority of nodes (proteins) have only a few connections to other nodes
- some nodes (hubs) are connected to many other nodes in the network.

Preferential attachment model: edges are preferentially attached to those nodes with a highest degree

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





Properties of PPINs: transitivity

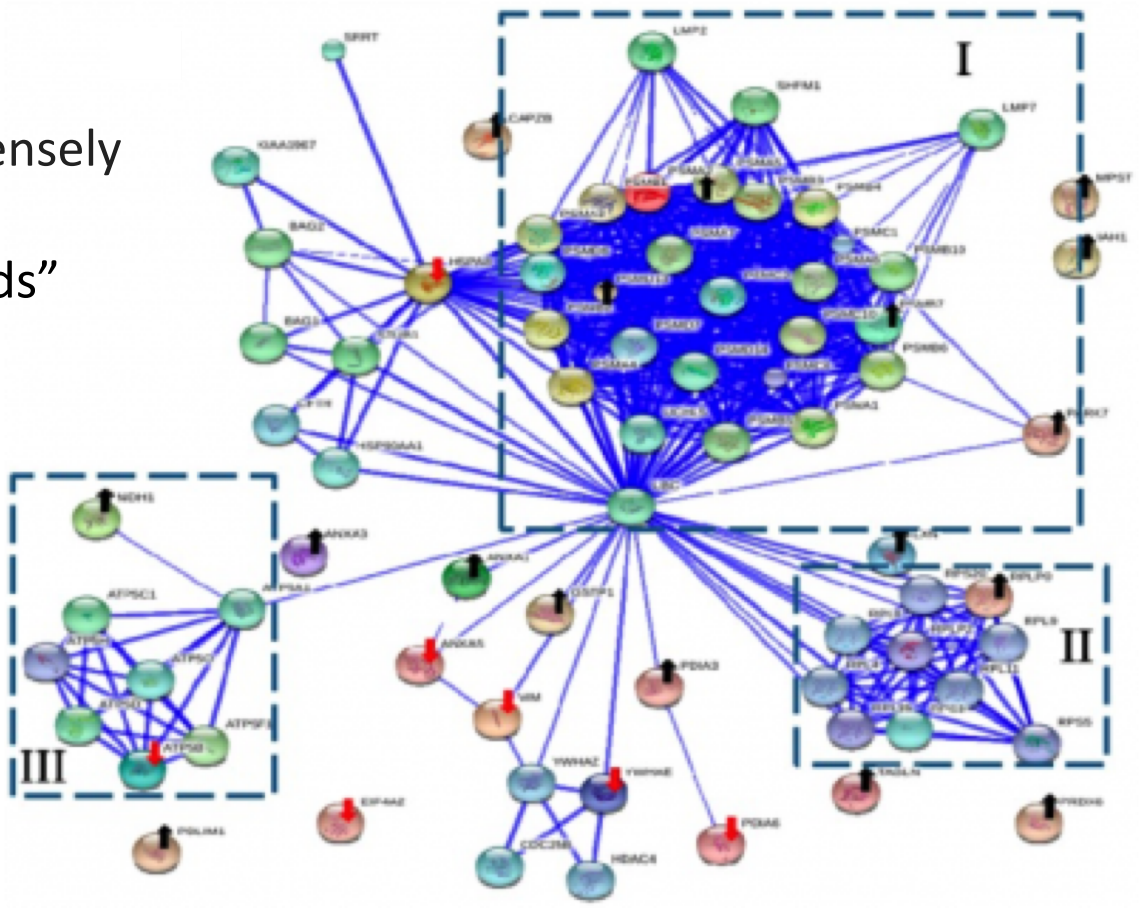
The **transitivity** or **clustering coefficient** of a network: measure of the tendency of the nodes to cluster together

High transitivity:

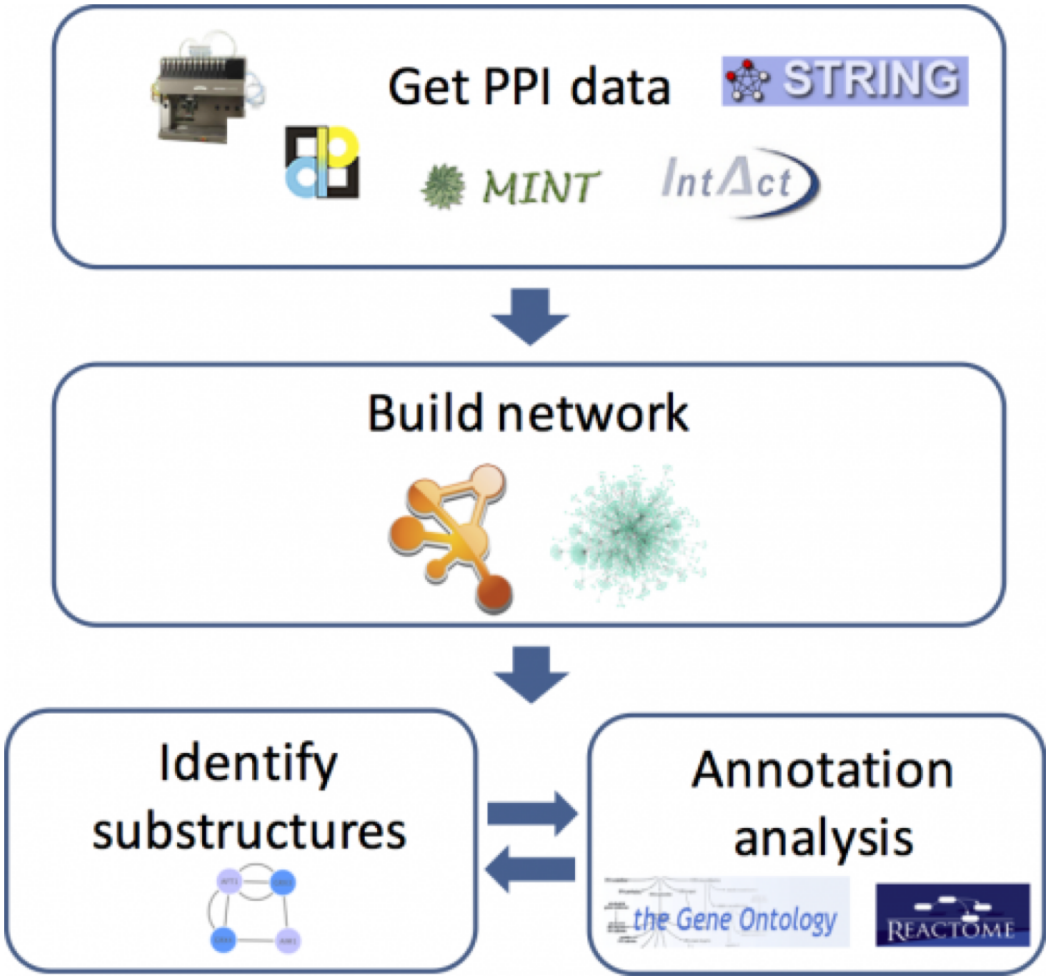
- communities or groups of nodes are densely connected internally
- “the friends of my friends are my friends”

Communities may reflect:

- Functional modules (exchangeable functional units)
- Protein complexes
- intermodular interactions and proteins



Building and analysing PPINs



Network representation and analysis tools

Cytoscape

- one of the most popular network analysis tools
- an open-source
- Java-based
- multi-platform desktop application
- widely used for network representation, integration and analysis.

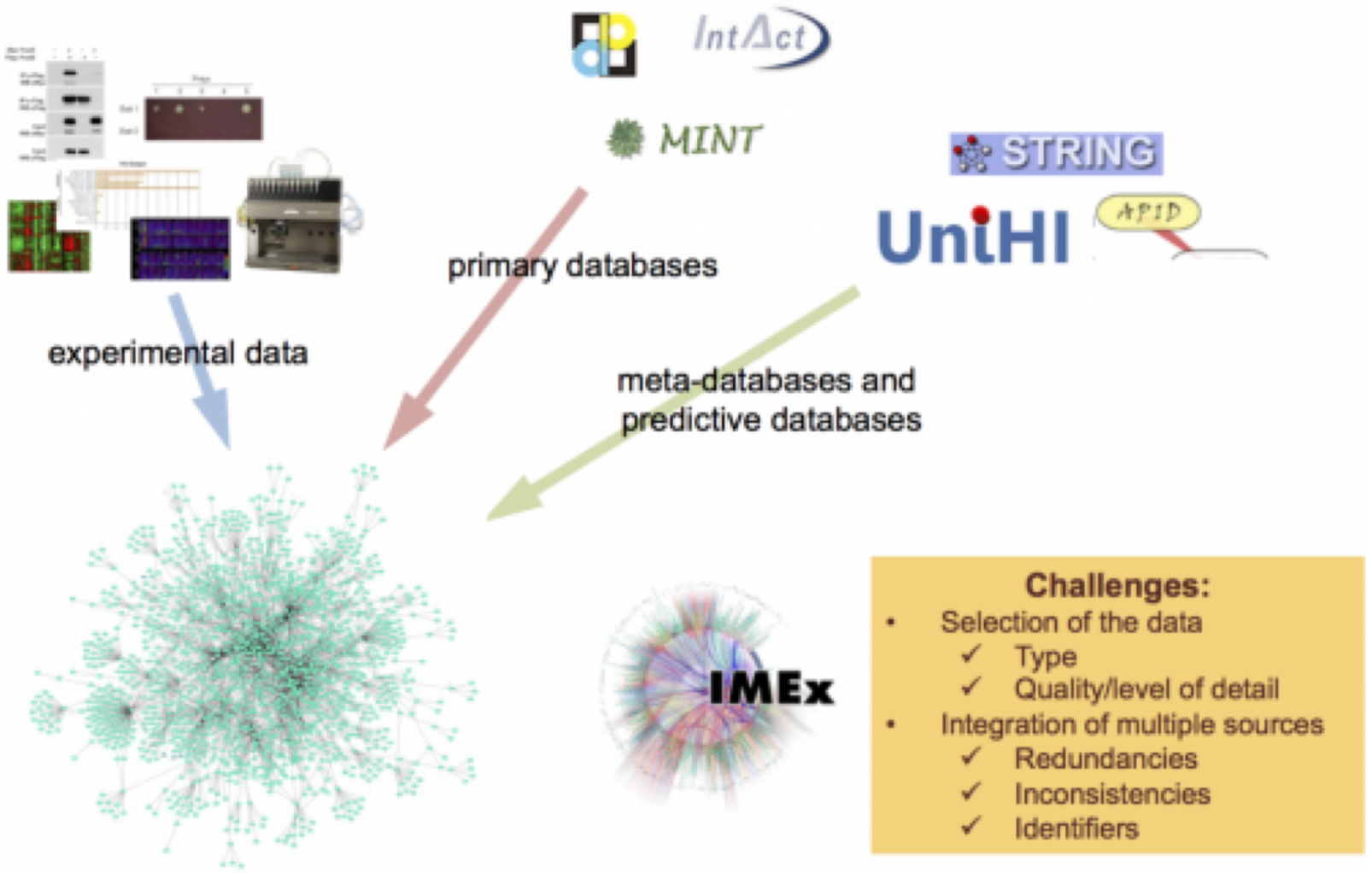
Advantages

- *Cytoscape apps*
- *Automation*

Limitations

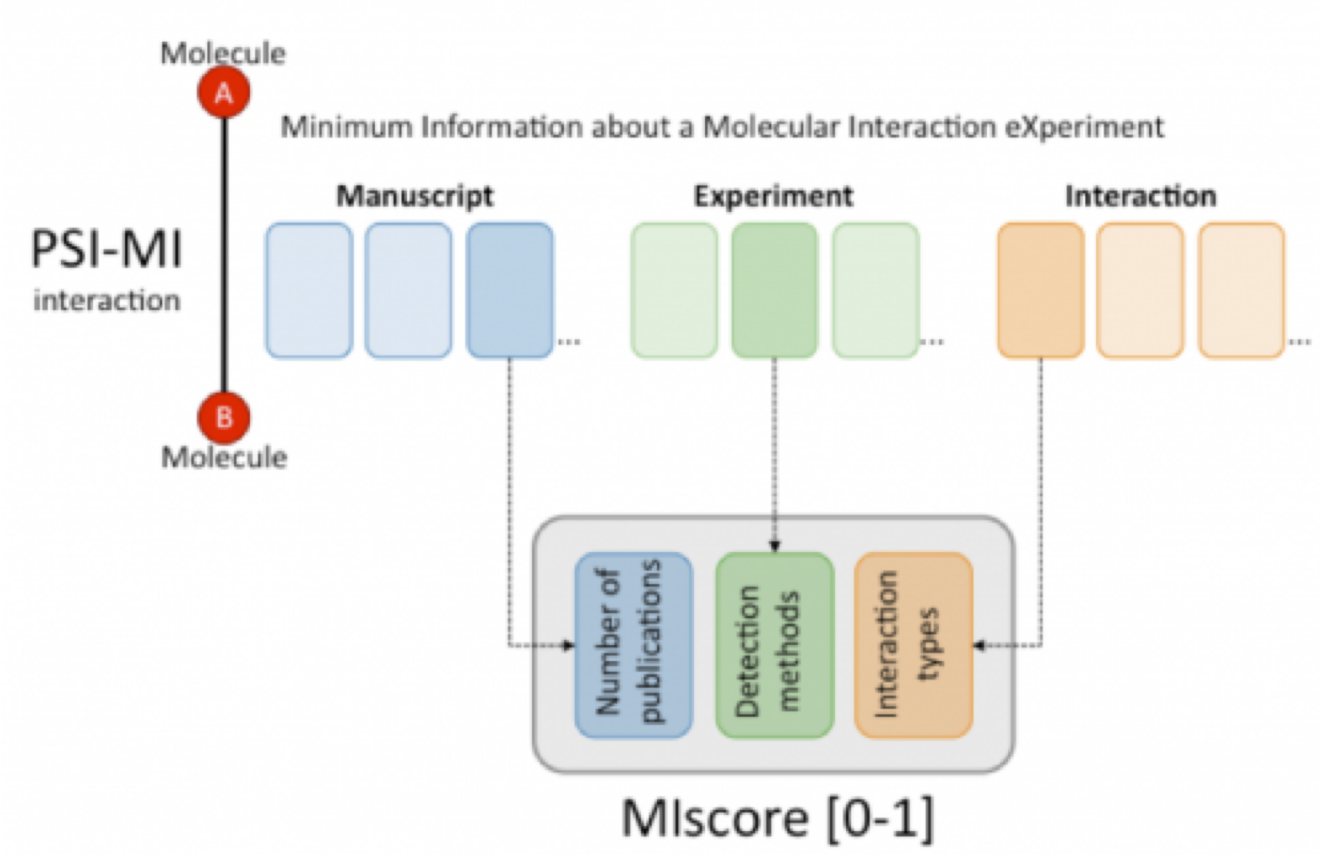
- demanding in terms of computing resources when it comes to large-scale networks

Sources of PPI data

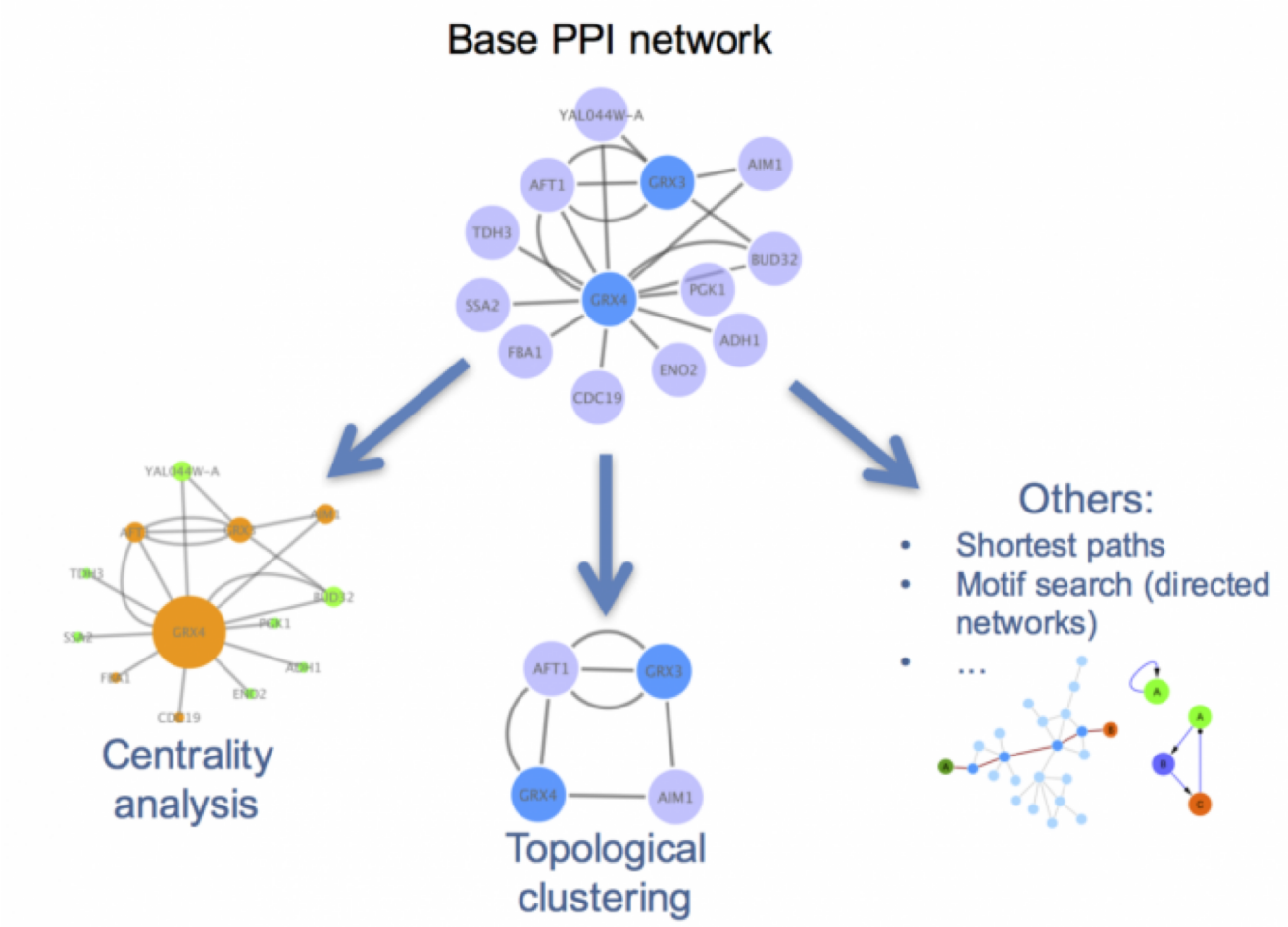


Assessing reliability and measuring confidence

- Contextual biological information
- Count how many times a given interaction has been reported in the literature (MIscore)
- Aggregated methods (INTscore)



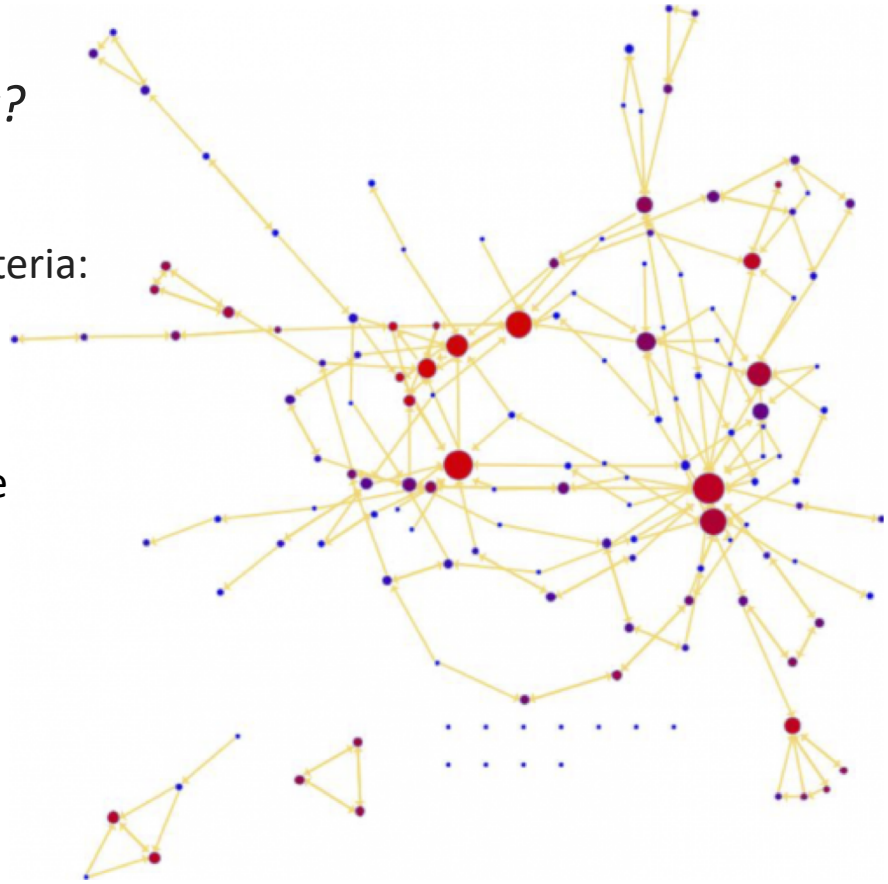
Topological PPIN analysis



Centrality analysis

Centrality: estimation on how important a node or edge is for the connectivity or the information flow of the network

Which protein is the most important and why?



Centrality can be measured using different metrics and criteria:

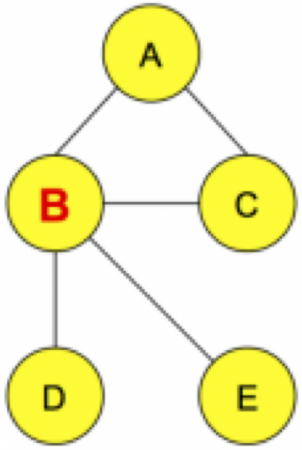
- **Degree of the nodes**
  - **local** measure
  - does not take into account the rest of the network
  - Importance depends strongly on the network's size
- **Global centrality measures**
  - take into account the whole of the network
  - independent of network size
  - **Closeness**
  - **Betweenness**
- **Other measures of centrality**
  - 'random walks'
  - Measure 'time' or 'speed' needed to reach other nodes

Closeness centrality

- Estimates how fast the flow of information would be through a given node to other nodes
- Measures how short the shortest paths are from node *i* to all nodes
- Sometimes closeness centrality is also expressed simply as the inverse the farness

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

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

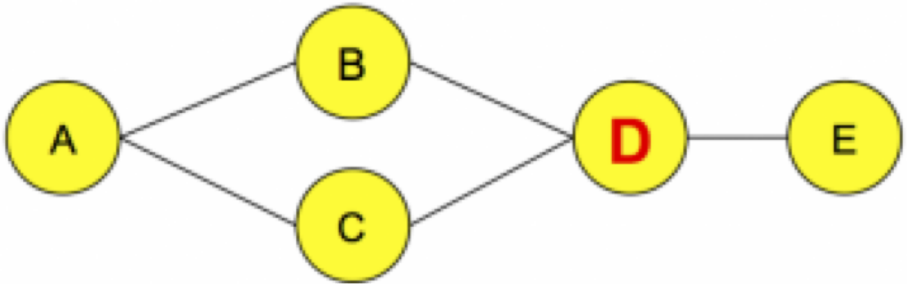
N = 5 (# of nodes)

Betweenness centrality

The betweenness of a node *N* is calculated considering couples of nodes (*v1, v2*) and counting the number of shortest paths linking those two nodes, which pass through node *N.* Next the value is related to the total number of shortest paths linking *v1* and *v2*.

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

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

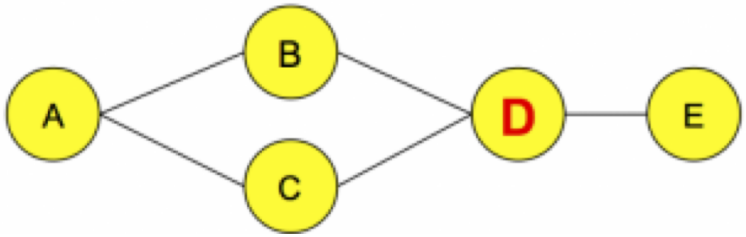


Betweenness centrality

- the number of shortest paths in the graph that pass through the node divided by the total number of shortest paths
- measures how often a node occurs on all shortest paths between two 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.



- **Betweenness centrality** is based on communication flow
- Nodes with a high betweenness centrality lie on communication paths and can control information flow
- important proteins in signalling pathways and can form targets for drug discovery

Clustering analysis

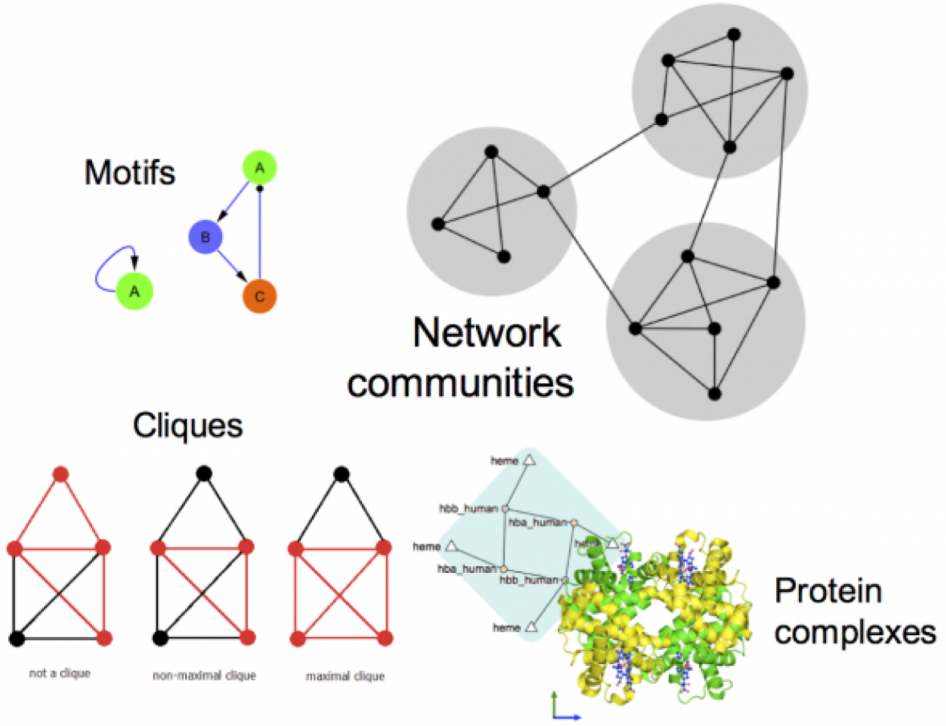
Community / Cluster
group of nodes that are more connected within themselves than with the rest of the network.

Module
exchangeable functional units in which the nodes (proteins) do not have to be interacting in the same time or space. Intrinsic functional properties do not change when it is placed in a different context

Complex
Group of proteins that interact with each other at the same time and in the same space, forming relatively stable multi-protein machinery

Clique
A subset of nodes in which every node is connected with every other member of the clique.

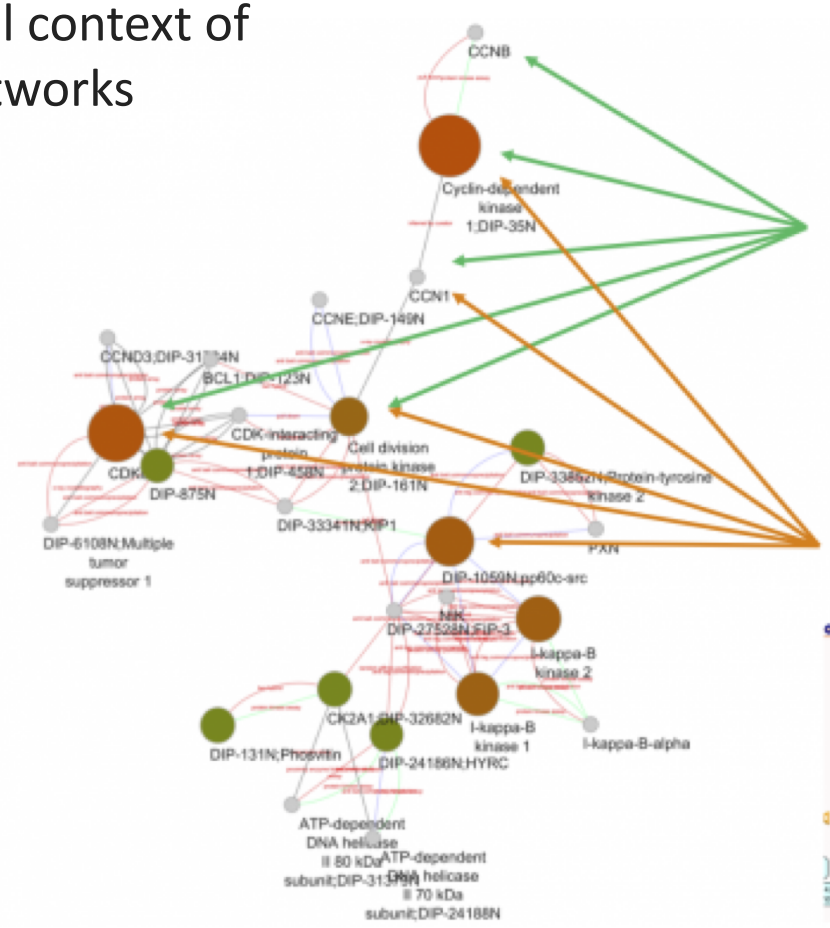
Motif
Motifs are statistically over-represented sub-graphs in a network



Annotation enrichment analysis

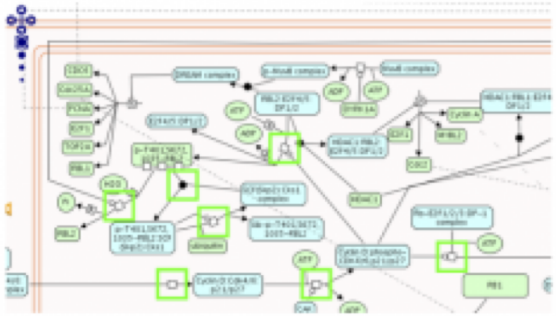
Goal: understand the biological context of protein-protein interaction networks

"When sampling X *proteins (test set) out of* N *proteins (reference set; graph or annotation), what is the probability that* x*, or more, of these proteins belong to a functional category* C *shared by* n *of the* N *proteins in the reference set."*



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

REACT_821.2:
Cyclin D associated events in G1



Limitations of annotation enrichment

- The main limitations of annotation enrichment come from the annotations themselves
- more "popular" proteins are better annotated
- This introduces a certain bias into the statistical analysis
- Another limitation of annotation enrichment is complexity and detail of annotation associated with large gene or protein sets
- resources such as Reactome and, especially, GO can be very complex and detailed in their annotation leading to the generation of overwhelmingly complicated networks of inter-related and similar terms
- use of simplified ontologies

Summary

Biological networks

- graph theory
- several types of biological networks: genetic, metabolic, cell signalling
- Networks are represented by nodes and edges

Summary

Protein-protein interaction networks

- **Small-world effect**: Network diameter is usually small (~ 6 steps), no matter how big the network is
- **Scale-free**: A small number of nodes (hubs) are lot more connected than the average node
- **Transitivity**: The networks contain communities of nodes that are more connected internally than they are to the rest of the network.

Summary

Analysing PPINs

- When building a PPIN it is important to be aware of the **type and quality of the data** used.
- Confidence scoring tools such as MIscore can help select the best characterised interactions.
- two of the most used topological methods to analyse PPINs are:
  - **Centrality analysis:** Which identifies the most important nodes in a network, using different ways to calculate centrality.
  - **Community detection:** Which aims to find heavily intinter-connected components that may represent protein complexes and machineries
- **Annotation enrichment analysis** is a complementary tool often used when analysing PPINs (GO, Reactome)