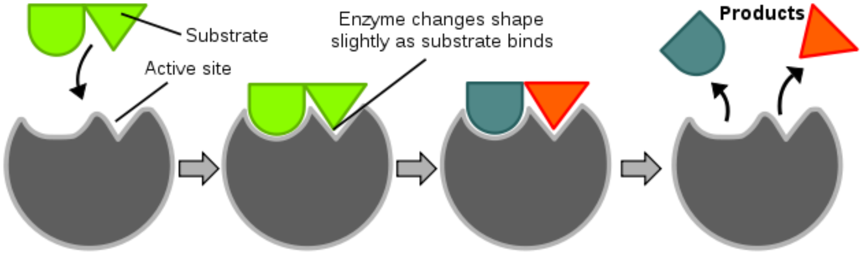
Protein interaction networks

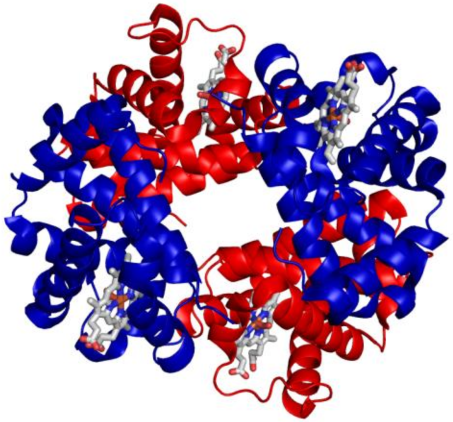
Ulf Leser and Samira Jaeger

Molecular interactions – Motivation

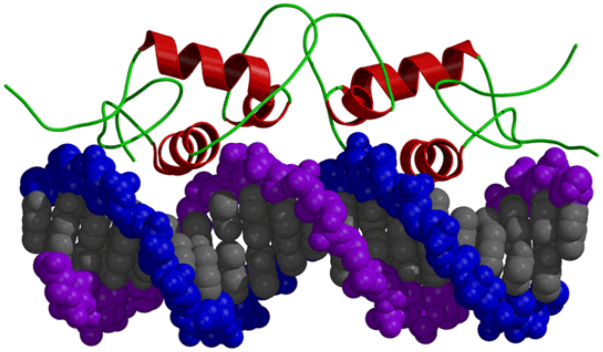
- Proteins mediate their function in complex interplay with other molecules through molecular interactions



Enzymes bind substrates to catalyzes biochemical reactions



α and β-hemoglobin chains assemble into hetero-tetramers for transporting oxygen from lungs to tissues



Transcription factors bind the DNA to induce transcription

Protein-protein interactions – Motivation

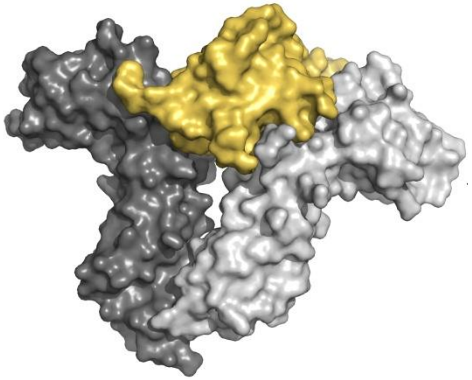
- Important class of biomolecular interactions are protein-protein interactions

- Virtually all cellular mechanisms rely on the physical binding of two or more proteins to accomplish a particular task:
  - Critical role in cellular processes, e.g. signal transduction, gene regulation, cell cycle control and metabolism
  - Alterations in protein interactions perturb natural cellular processes and contribute to many diseases, such as cancer and AIDS

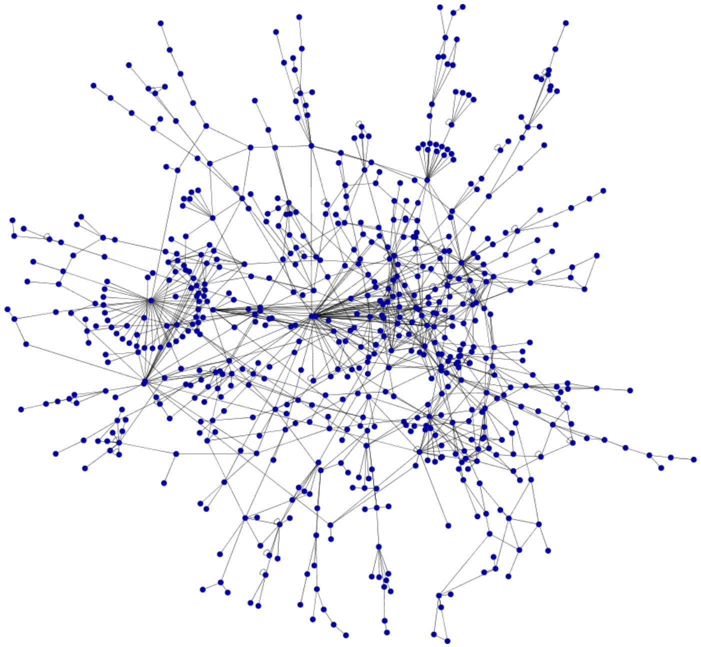
- Identifying all physical interactions within an organisms – the interactome – essential towards understanding the complex molecular relationships in living systems

This Lecture

- Protein-protein interactions
  - Characteristics
  - Experimental detection methods
  - Databases

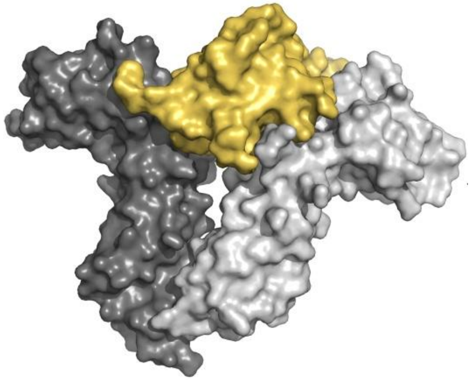


- Protein-protein interaction networks
  - Characteristics
  - Applications
    - Protein function prediction



Protein-protein interactions – Characteristics

- Protein interaction defined as physical contact with molecular docking
  - Non-covalent contacts between side chains driven by hydrophobic effects, hydrogen bonds and electrostatic interactions
- Any two proteins can interact – but on what conditions ?
- Important aspect is the biological context:
  - Cell type, cell cylce phase and state
  - Environmental conditions
  - Developmental stage
  - Protein modification
  - Presence of cofactors and other binding partners



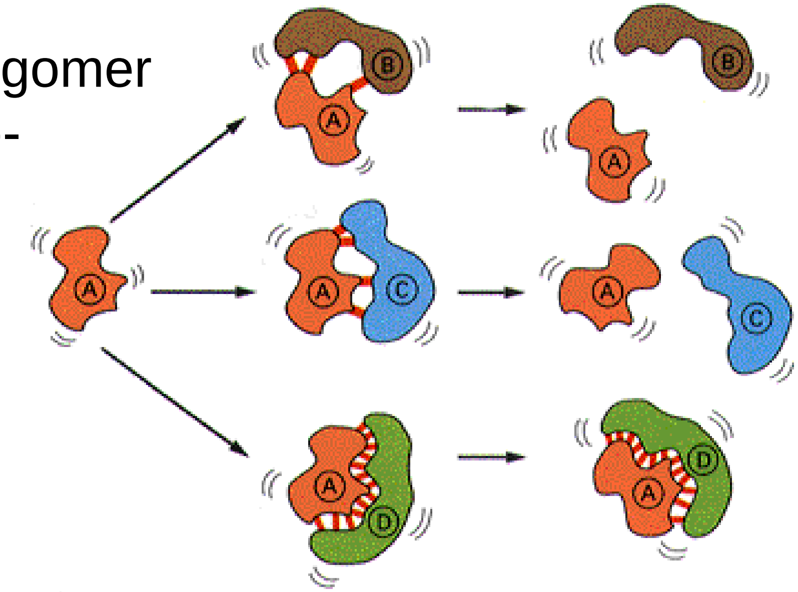
Protein-protein interactions – Characteristics

- Protein interactions differ in diverse structural and functional characteristics, e.g. composition, affinity and life time of the association

- Strength depicts whether an interaction is permanent or transient

- Specificity refers to the selective binding of interaction partners

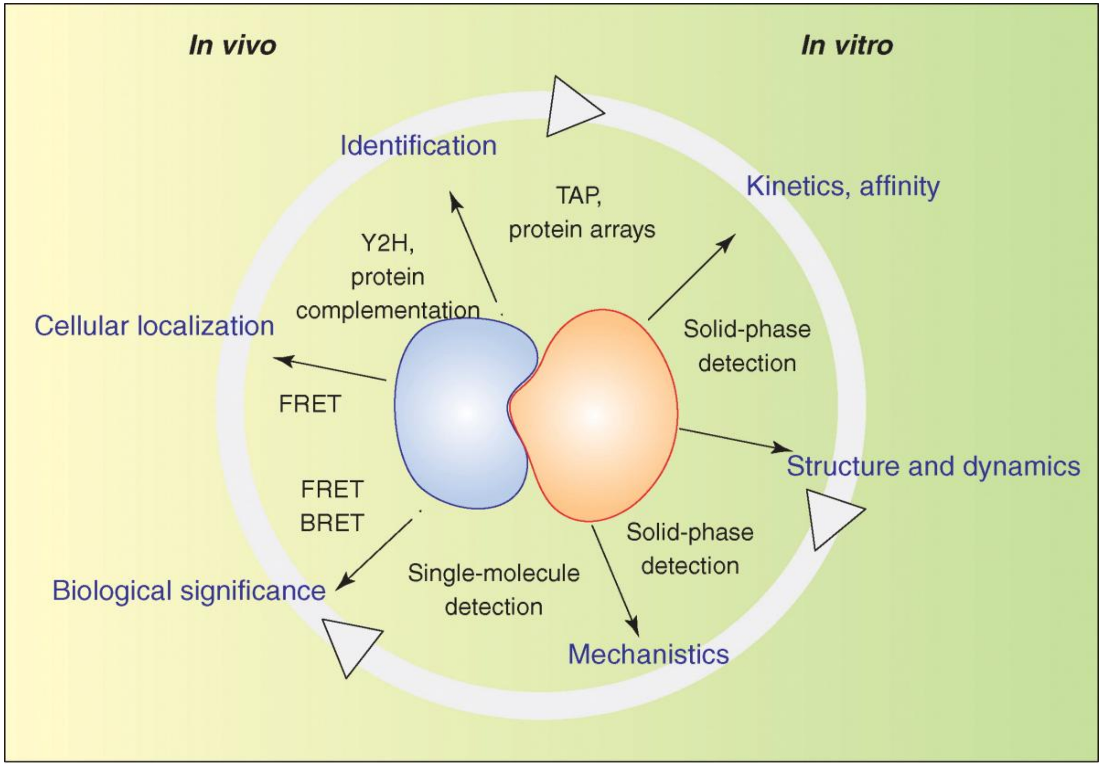
- Type of interacting subunits specifies whether an interaction forms hetero-oligomer with several different subunits or homo-oligomer with only one type of protein subunit



From The Art of MBoC³ © 1995 Garland Publishing, Inc.

Experimental detection methods

- Protein-protein interactions have been studied extensively by different experimental methods



Piehler, 2005

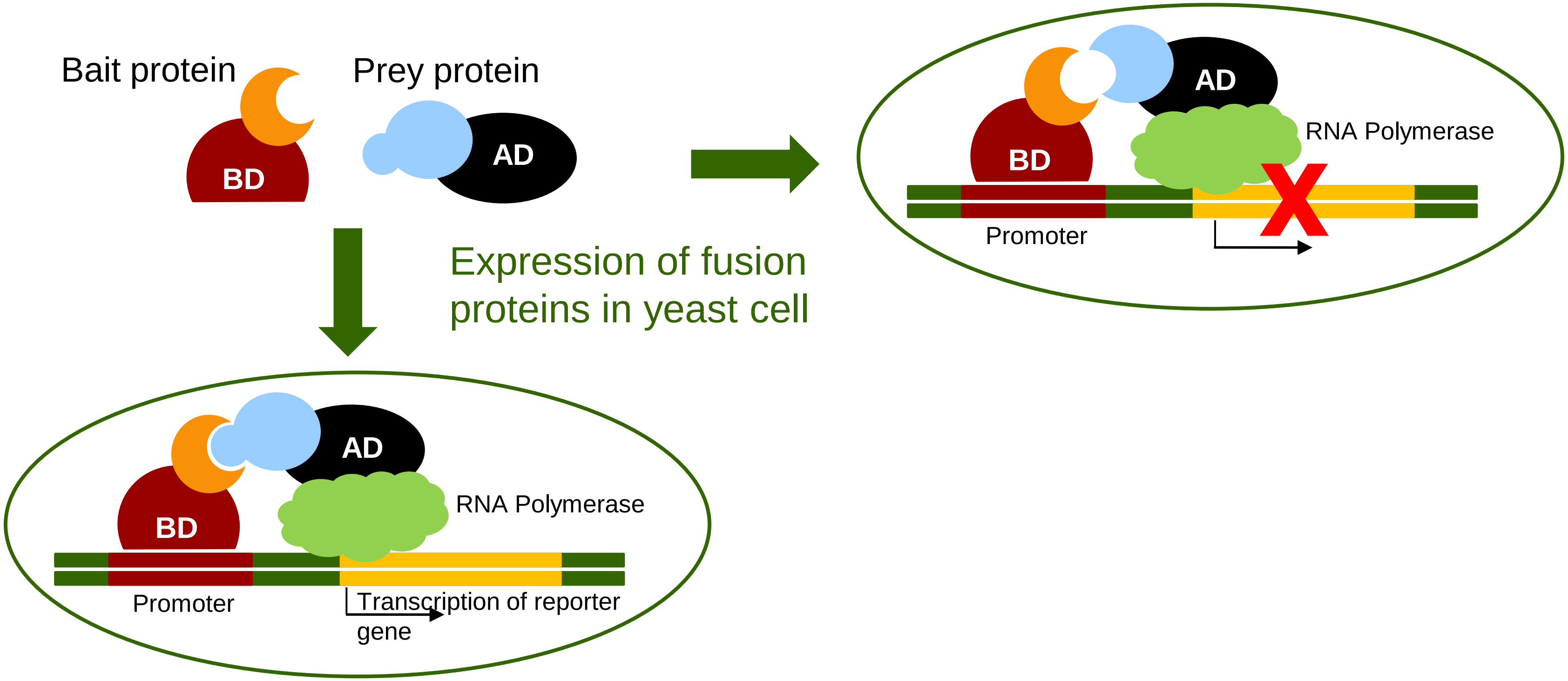
Current Opinion in Structural Biology

- Small-scale techniques
- Large-scale techniques
  - Yeast two-hybrid assays (Y2H)
  - Tandem affinity purification and mass spectrometry (TAP-MS)

- Cell assay
  - *in vivo* vs. *in vitro*
- Type of interaction
  - binary vs. complex
- Type of characterization

Yeast two-hybrid screens

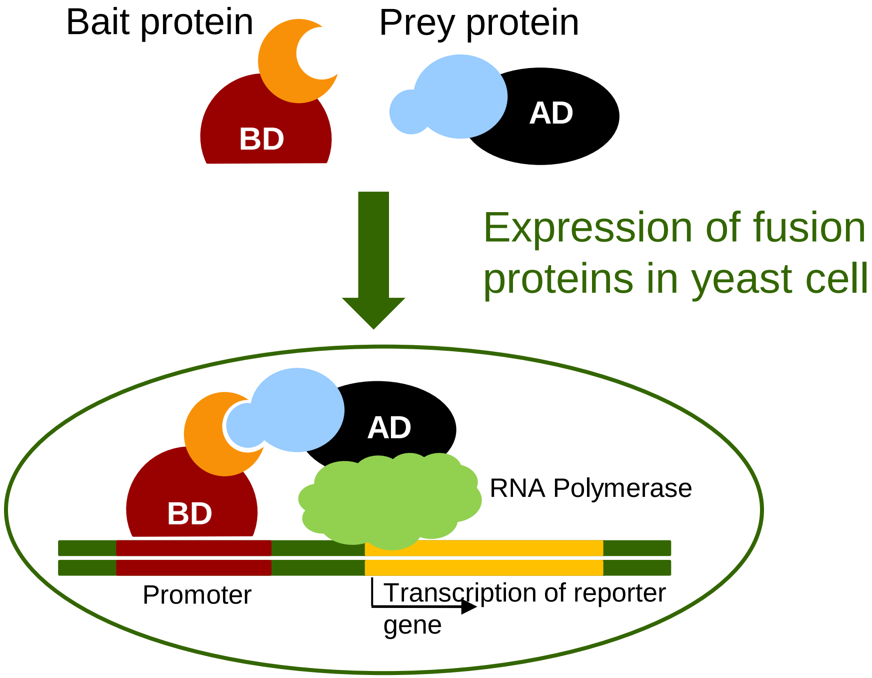
- Y2H is a molecular genetic tool, in which an interaction reconstitutes a transcription factor that activates expression of reporter genes
- Transcription factors require two domains: DNA binding domain (BD) and an activation domain (AD)

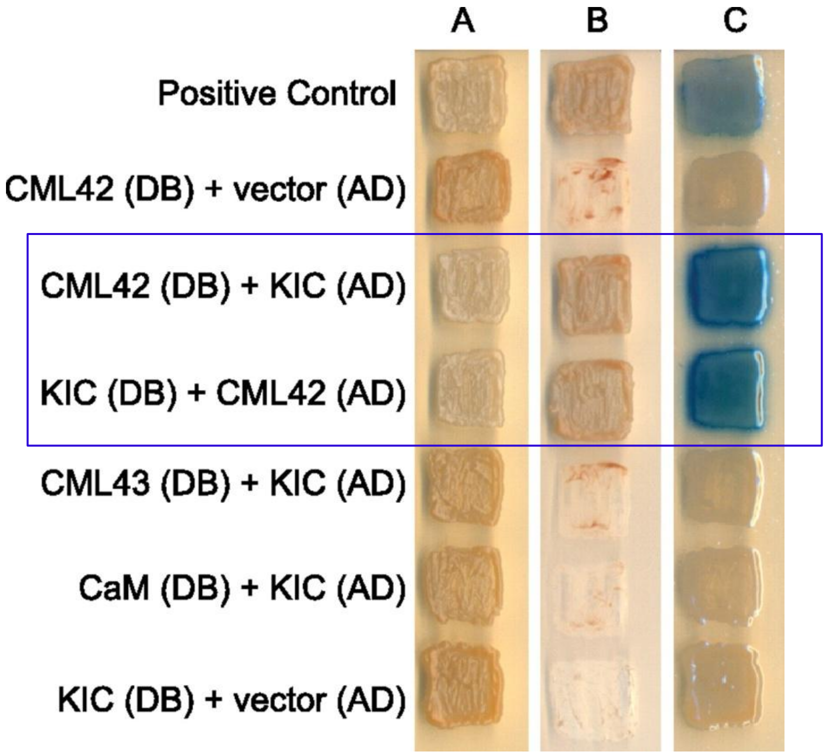


Expression of fusion proteins in yeast cell

Yeast two-hybrid screens

- Y2H is a molecular genetic tool, in which an interaction reconstitutes a transcription factor that activates expression of reporter genes
- Transcription factors require two domains: DNA binding domain (BD) and an activation domain (AD)


Bait protein
BD
Prey protein
AD
Expression of fusion proteins in yeast cell
AD
BD
RNA Polymerase
Promoter
Transcription of reporter gene



Dobney S et al. J. Biol. Chem.

Yeast two-hybrid screens

Benefits

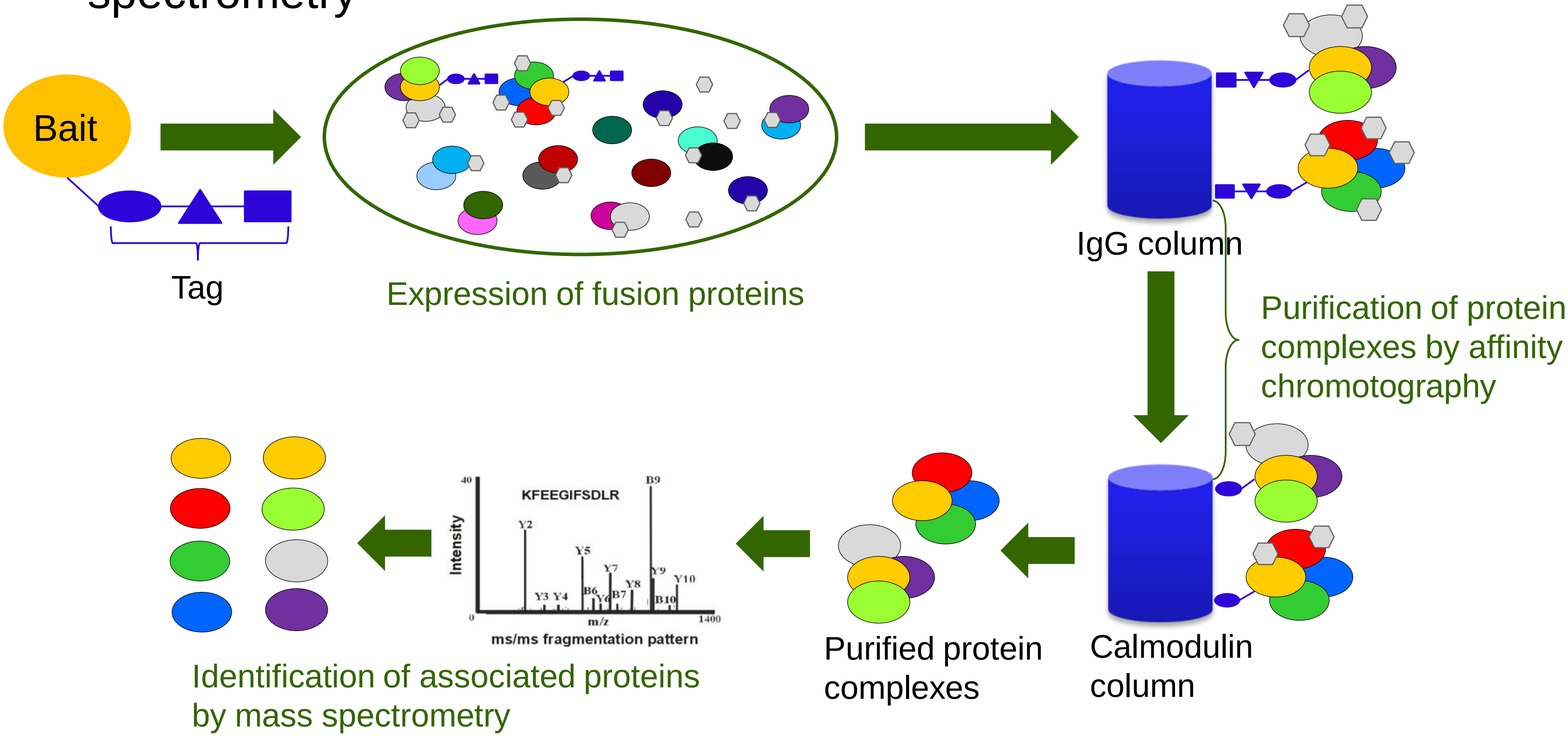
- Large-scale analysis
- Sensitive *in vivo* technique
- Identification of direct, transient and unstable interactions
- Genetic code of any fusion protein may be introduced into yeast cells

Drawbacks

- Poor reliability
  - High false positive rate up to 50% (!)
  - High false negative rate
- Analysis of proteins in nucleus rather than in their native compartement
- Stable expression of fusion protein might be a problem
- Essential post-translational modification of non-yeast proteins may not be carried out

Tandem affinity purification and mass spectrometry

- TAP-MS involves biochemical isolation of protein complexes and subsequent identification of their constituting proteins using mass spectrometry



Expression of fusion proteins

IgG column

Purification of protein complexes by affinity chromotography

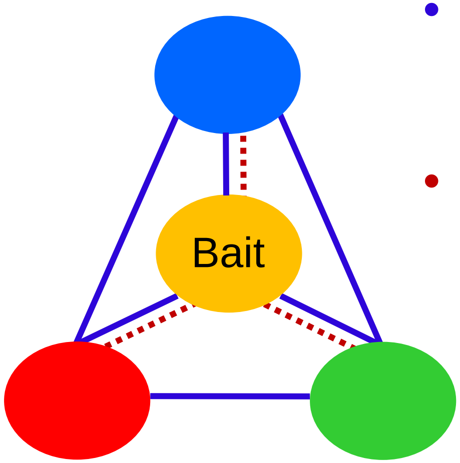
Calmodulin column

Purified protein complexes

Identification of associated proteins by mass spectrometry

Matrix and spoke model

- Direct interactions can not be distinguished from interactions mediated by other proteins in a complex
- How many interactions are detected from a TAP-MS purification ?



- Matrix model: infers interactions between all proteins of a purified complex → **(N*N -1)/2**
- Spokes model: infers only interactions between the bait and the co-purified proteins → **N – 1**

# Proteins	Matrix	Spoke
4	6	3
10	45	9
80	3540	79

Tandem affinity purification and mass spectrometry

Benefits

- Large-scale analysis
- Detection of protein complexes/interactions in correct cellular enviroment and detect several members of a complex

Drawbacks

- No direct translatation into binary interactions
- Protein complexes not present under given conditions are missed
- Loosely associated proteins of a complex might be washed of during purification
- Protein targeting might interfere with complex formation

Protein-protein interaction databases

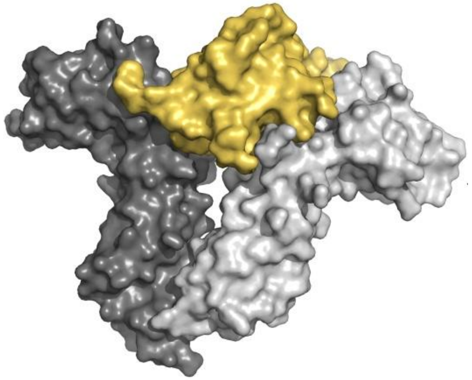
Database	Species	Proteins	Interactions
IntAct	No restriction	53.276	271.764
BioGrid	No restriction	30.712	131.638
DIP	No restriction (372)	23.201	71.276
MINT	No restriction	31.797	90.505
HPRD	Human only	30.047	39.194
MMPPI	Mammals		
⋮			
STRING	No restriction (630)	2.590.259	
UniHI	Human only	22.307	200.473
OPID	Human only		
⋮			

Experimentally verified protein interactions

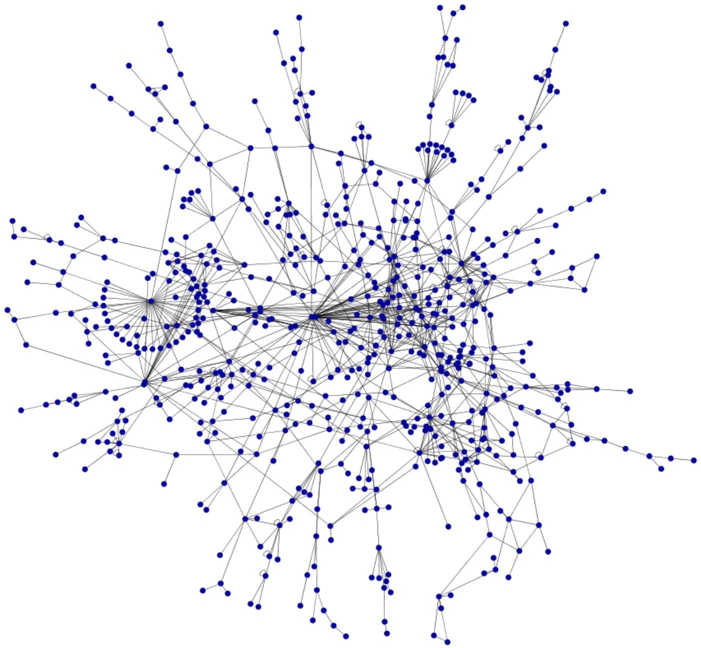
Experimentally verified and predicted protein interactions

This Lecture

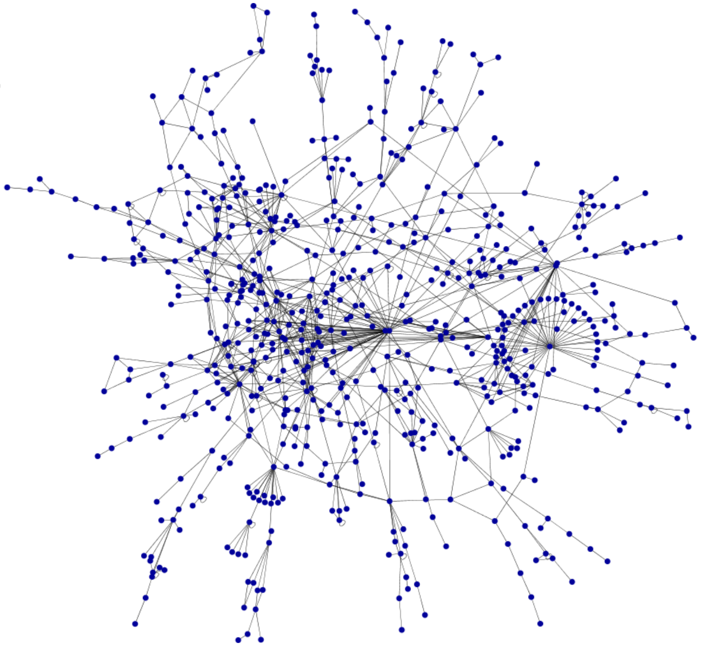
- Protein-protein interactions
  - Characteristics
  - Experimental detection methods
  - Databases



- Protein-protein interaction networks
  - Characteristics
  - Applications
    - Protein function prediction



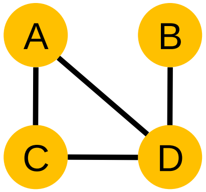
Protein-protein interaction networks



- Binary interaction data can be assembled to protein-protein interaction networks

- Networks are represented as graphs
- Definition of a graph: *G = (V,E)*
  - *V* is the set of nodes (proteins)
  - *E* is the set of edges (interactions)

- Computational representation of graphs:



Adjacency lists:
(ordered) pairs of nodes

{ (A,C), (A,D), (B,D), (C,A), (C,D) (D,B) , (D,C), (D,A) }

Adjacency matrix

	A	B	C	D
A	0	0	1	1
B	0	0	0	1
C	1	0	0	1
D	1	1	1	0

Graph-theoretic concepts

- A graph is defined by *G = (V,E)*, where *V* is the set of nodes and *E* is the set of edges connecting pairs of nodes

- The distance between two nodes is the number of edges on the shortest path between them
  - Diameter is the maximum distance between any two node

- The neighborhood of node is the set of nodes connected to it
  - *n*-neighborhood of a node is the set of nodes with distance *n*

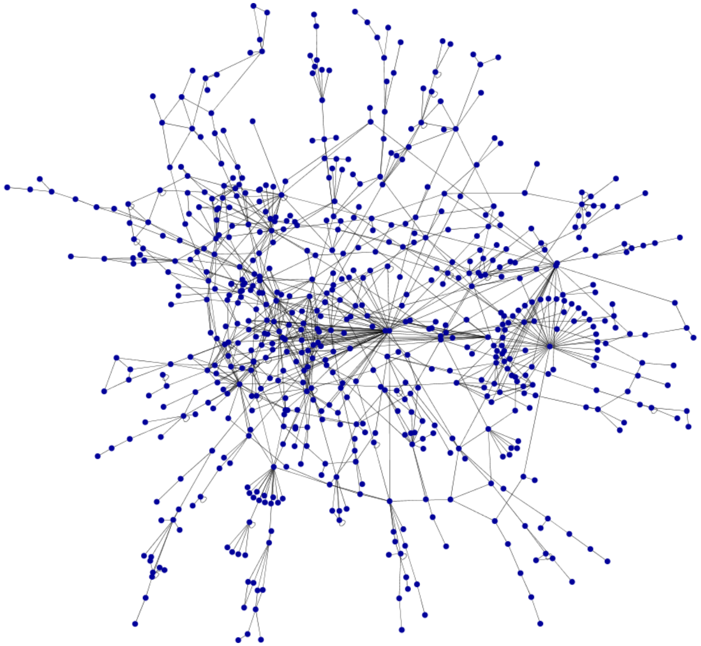
- A clique is a fully connected subgraph, a subgraph in which every two nodes are connected by an edge
  - *k*-core is a subgraph where each node has at least *k* interactions

- The density is the fraction of edges a graph has given all possible pairs of nodes

$$D_G = \frac{2|E|}{|V|(|V|-1)}$$

Protein-protein interaction networks

- Why study protein interaction networks ?
  - Elucidate the relationship between network structure and biological function
  - Discover novel protein function
  - Identify functional modules and conserved interaction patterns
  - Associate proteins with phenotypes or disease
  - Study pharmacological drug-target relationships



Topological network properties

- Topology of a network reveals its organization on different levels
  - Local and global characteristics provide insights in network evolution, stability and dynamics

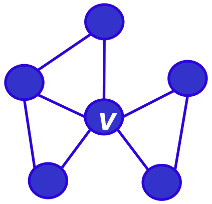
- Common properties of biological networks
  - Small world property
  - Clustering coefficient
  - Degree distribution
  - Network centrality

- Modular network organization

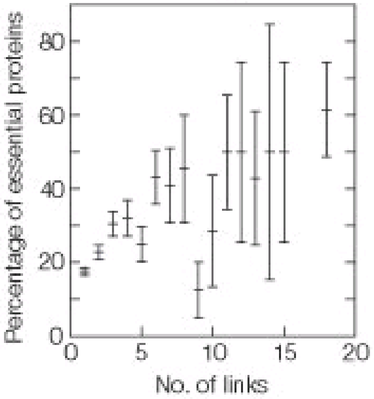
Network centrality

- Network centrality analysis identifies interesting elements/proteins within a network
- Quantitative measure to determine a proteins relative position in a network
- Example – Degree centrality:
  - Degree of a node = number of edges to other nodes

$$C_D(v) = \frac{\deg(v)}{|V|-1}$$



- High centrality in interaction networks correlates with:
  - Gene essentiality
  - Evolutionary importance
  - Conservation rate
  - Likelihood to cause human disease

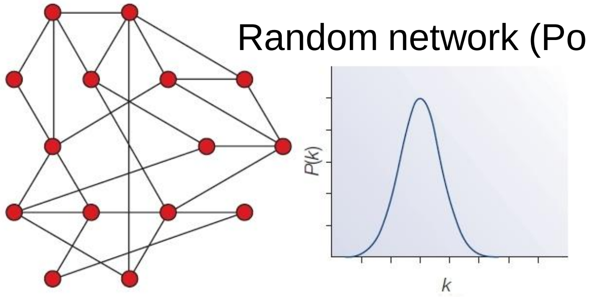


Degree distribution

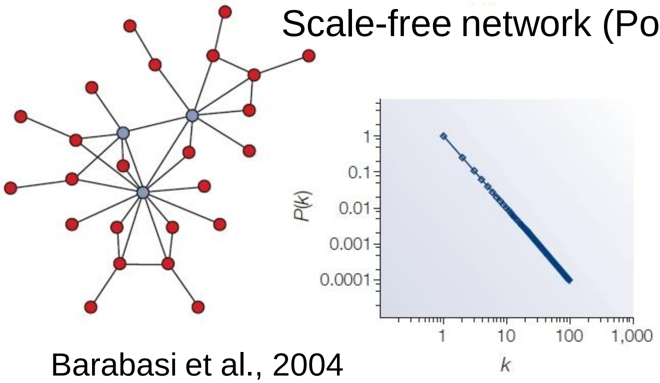
- Degree distribution P(k): probability that a node has exactly *k* links
  - Count the number of nodes N(k) with *k* = 1, 2, … links and divide by N
- Allows to distinguish between different network classes

- Common network distributions
  - Poisson: $P(k) = \frac{e^{-d} d^{k}}{k!}$
  - Exponential: $P(k) \sim e^{-k/d}$
  - Power-law: $P(k) \sim k^{-\gamma}$

Random network (Poisson)



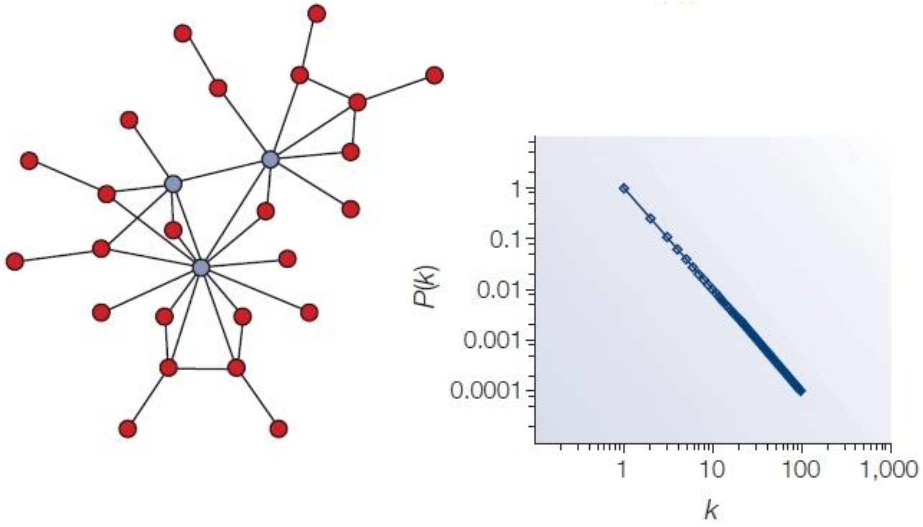
Scale-free network (Power law)



Barabasi et al., 2004

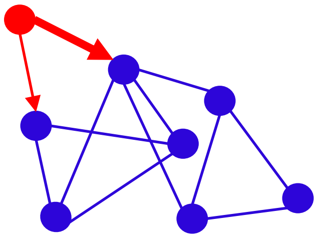
Scale-free networks

- Scale-free networks, e.g. protein interaction networks, follow a power law distribution: $P(k) \sim k^{-\gamma}$, with degree exponents 2<γ<3
  - Characterized by a small number of highly connected nodes known as hubs
- Scale-free topology typical feature of interaction networks
  - Most proteins participate in few interactions and few proteins in dozens
- Resistent to random failure, but prone to vulnerable attacks especially against hubs

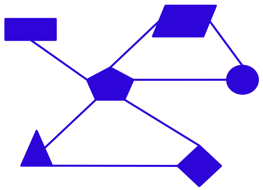


Scale-free networks

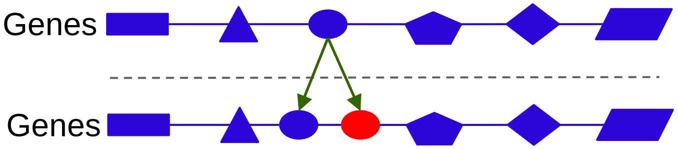
- Evolutionary origin of scale free networks
  - Growth: networks emerge through addition of new nodes
  - Preferential attachment: new nodes prefer to link to more connected nodes



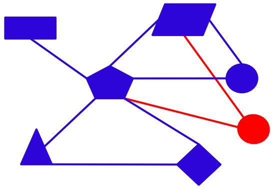
  - In interaction networks: scale-free property is thought to originate from gene duplications



Network before duplication



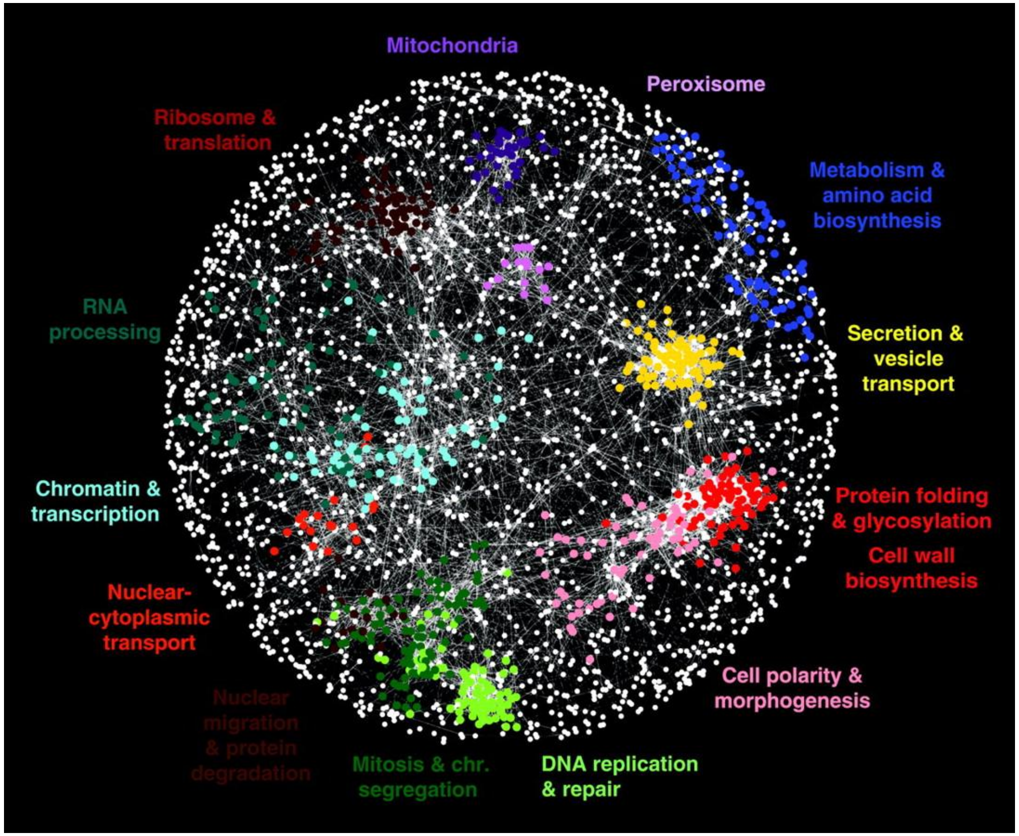
Gene duplication



Network after duplication

Modular network organization

- Complex cellular function is carried out in a highly modular manner
- Modular organization is reflected in a modular network structure


Mitochondria
Peroxisome
Ribosome & translation
Metabolism & amino acid biosynthesis
RNA processing
Secretion & vesicle transport
Chromatin & transcription
Protein folding & glycosylation
Cell wall biosynthesis
Nuclear-cytoplasmic transport
Nuclear migration & protein degradation
Cell polarity & morphogenesis
Mitosis & chr. segregation
DNA replication & repair

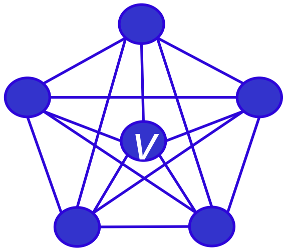
Costanzo et al., Nature, 2010

Clustering coefficient

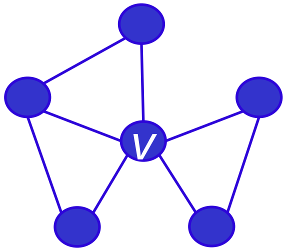
- Modules (or cluster) are densely connected groups of nodes
- Cluster coefficient *C* reflects a network's potential modularity and characterizes the tendency of nodes to cluster ('triangle density')

$$C_v = \frac{2E_v}{d_v(d_v - 1)} \longrightarrow <C> = \frac{1}{|V|}\sum_{v \in V} C_v$$

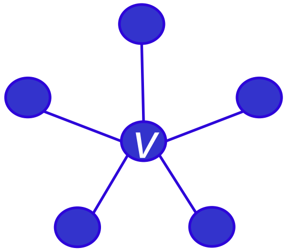
  - E_v = number of edges between neighbors of *v*
  - d_v = number of neighbors of *v*
  - $\frac{d_v(d_v-1)}{2}$ = maximum number of edges between neighbors d_v
- Example:



C_v = 10/10 = 1



C_v = 3/10 = 0.3



C_v = 0/10 = 0

Functional modules

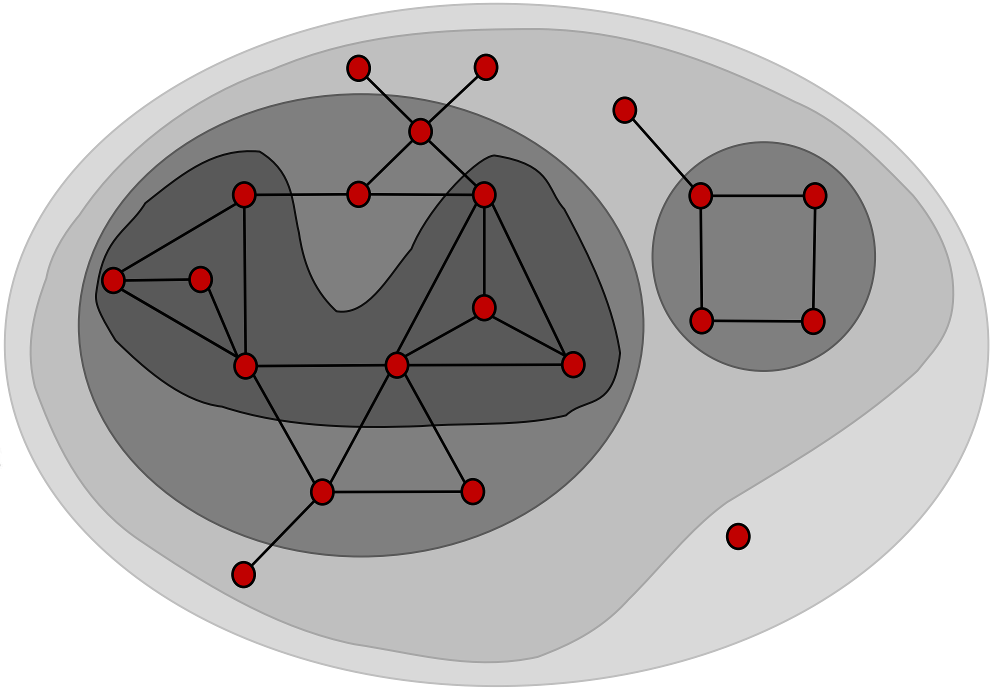
- Two types of modules are distinguished in interaction networks
  - Protein complexes: proteins that interact with each other at the same time and place
  - Functional modules: proteins that participate in the same process but interacting at different times and places

- Finding functional modules → find densely connected subgraphs:
  - Cliques / k-cores
  - Network clustering
  - Network alignment
  - Decompose networks into subnetworks according to particular topological properties

K-cores

- Identifying cliques in graphs is NP complete
- Approximation: *k*-cores
- A *k*-core of a graph *G* is defined as maximally connected subgraph of *G* in which all vertices have degree at least *k*

```
while modify_kcore == true do
  modify_kcore = false
  for all $v \in V$ do
    consider deg($v$) and the degree of N($v$)
    if deg($v$) $<k$ then
      prune $v$
      modify_kcore = true
    else
      if $v$ has $n$ neighbors with degree $<k$ then
        deg($v$) = deg($v$) - $n$
        if deg($v$) $<k$ then
          prune $v$
          modify_kcore = true
```



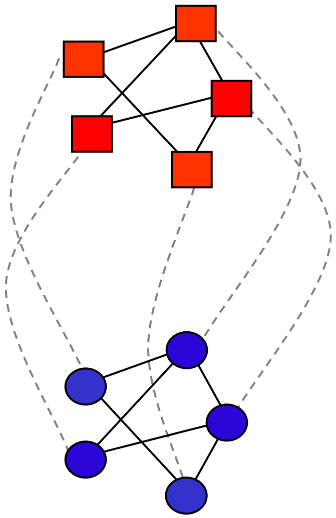
Network Alignment

1) Detection of orthologous proteins

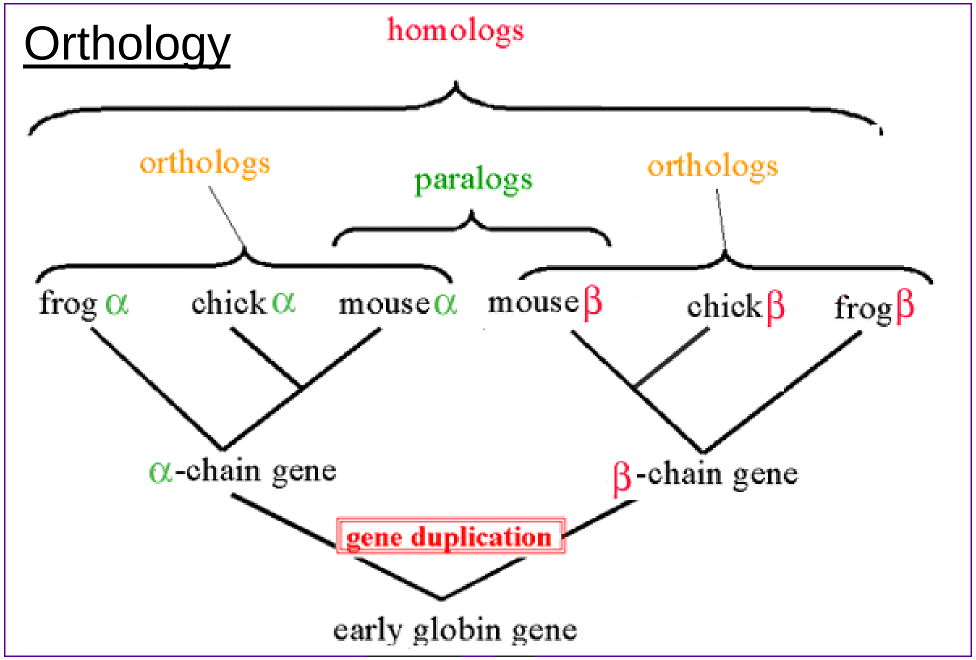
2) Identification of interologs & assembly to conserved and connected subgraphs (CCS)

ACGGT_AGATA

CGGTCAGAT



Network Alignment

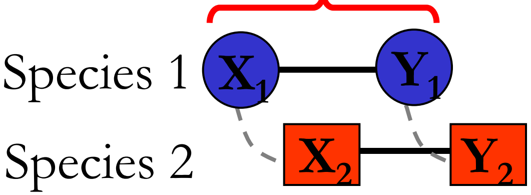


In-paralogs (‚recent‘ paralogs)

Out-paralogs (‚ancient‘ paralogs)

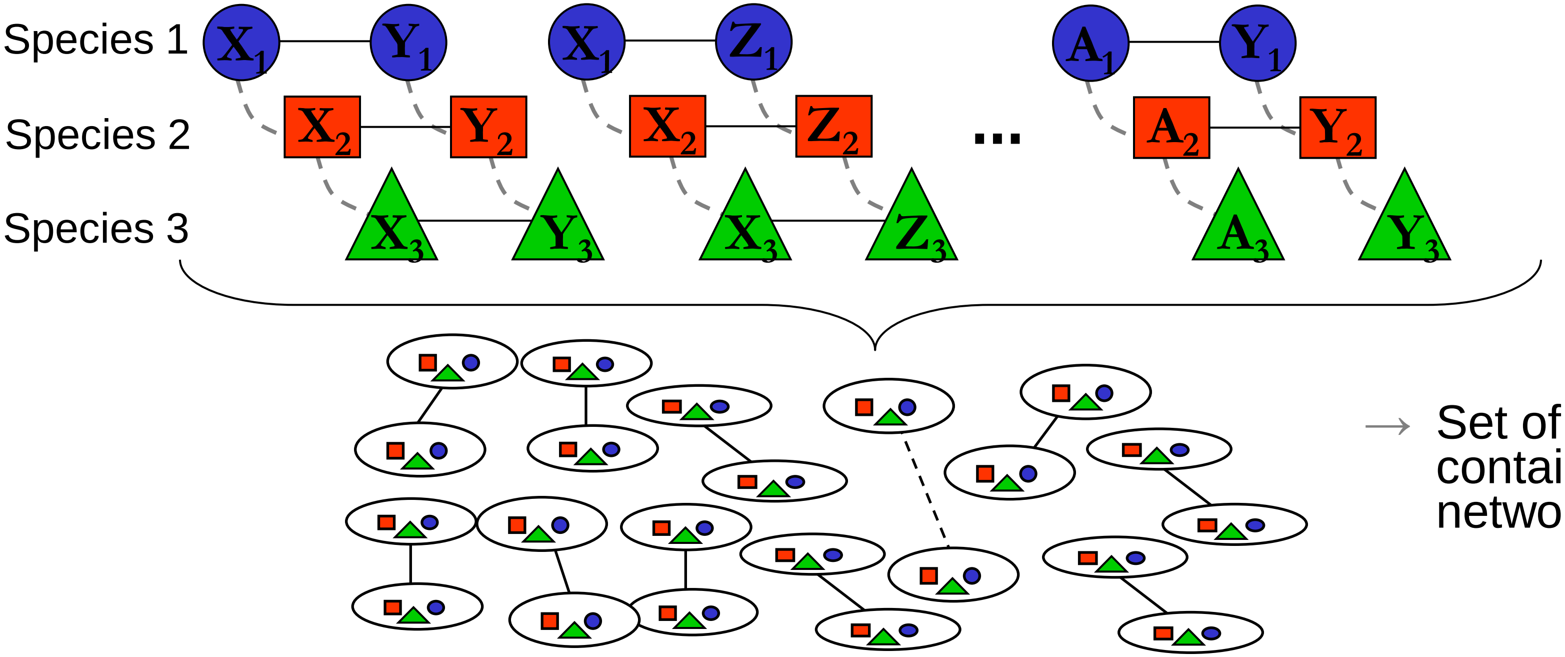
Interologs

Interolog = orthologous interaction



Network Alignment

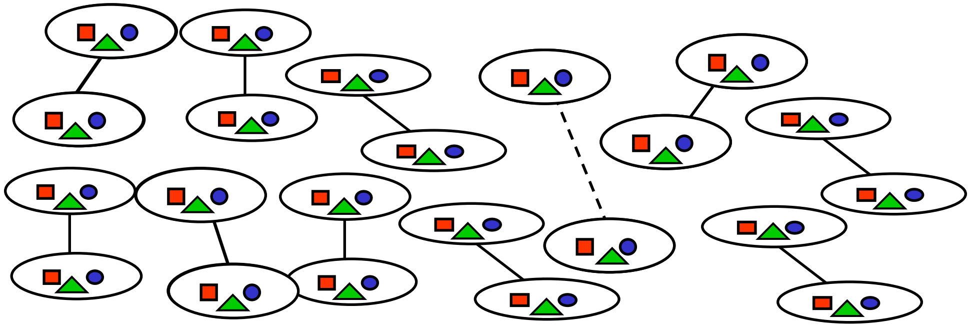
- (1) Identification of interologs and (2) assembly to subgraphs
- Modification of algorithm for frequent subgraph discovery



→ Set of interactions contained in all networks

Network Alignment

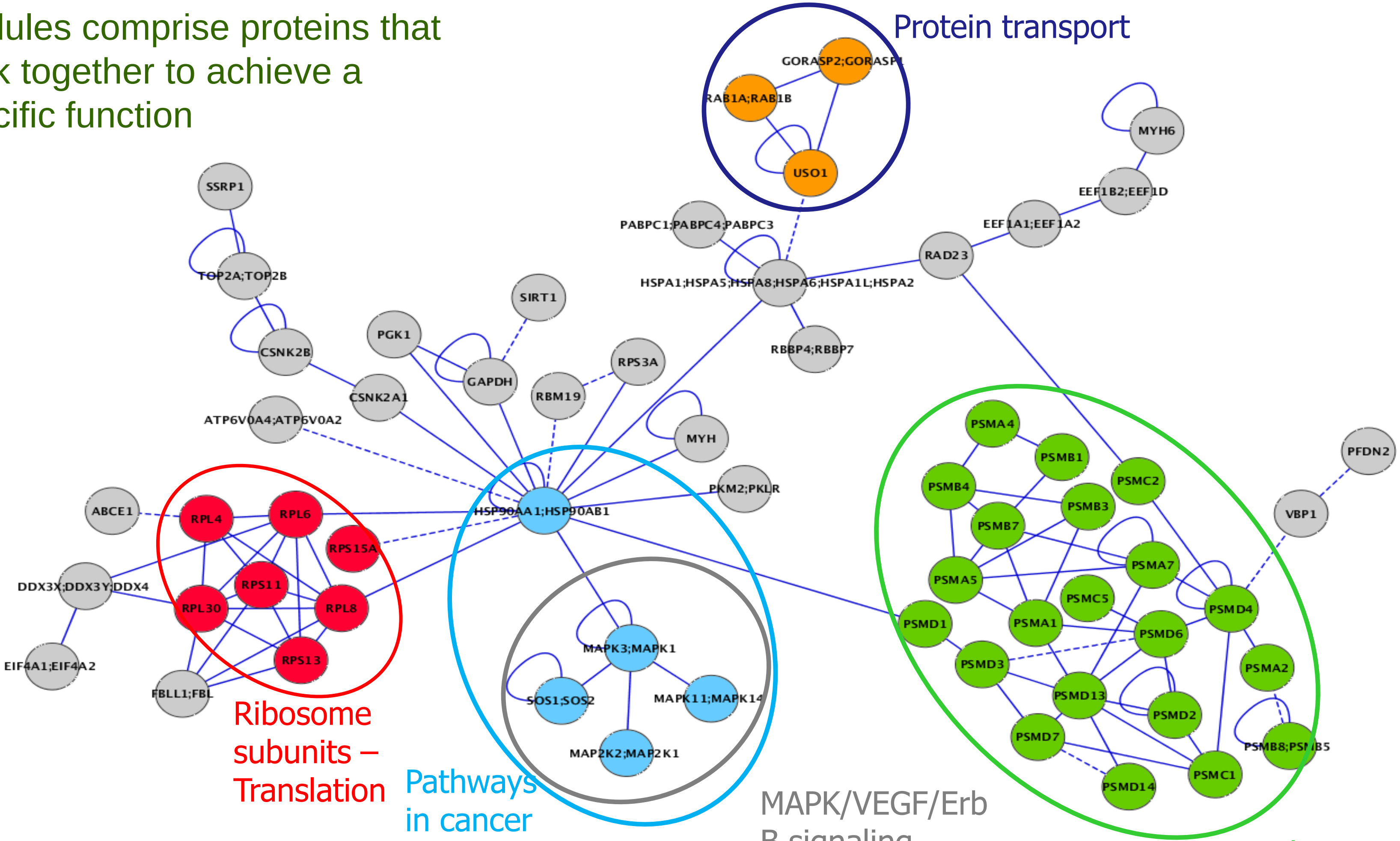
(2) Assembly of conserved PPIs to maximally connected subgraphs



→ Maximally connected and conserved subgraph → CCS

Network alignment

Modules comprise proteins that work together to achieve a specific function



Protein transport

Ribosome subunits – Translation

Pathways in cancer

MAPK/VEGF/ErbB signaling pathway

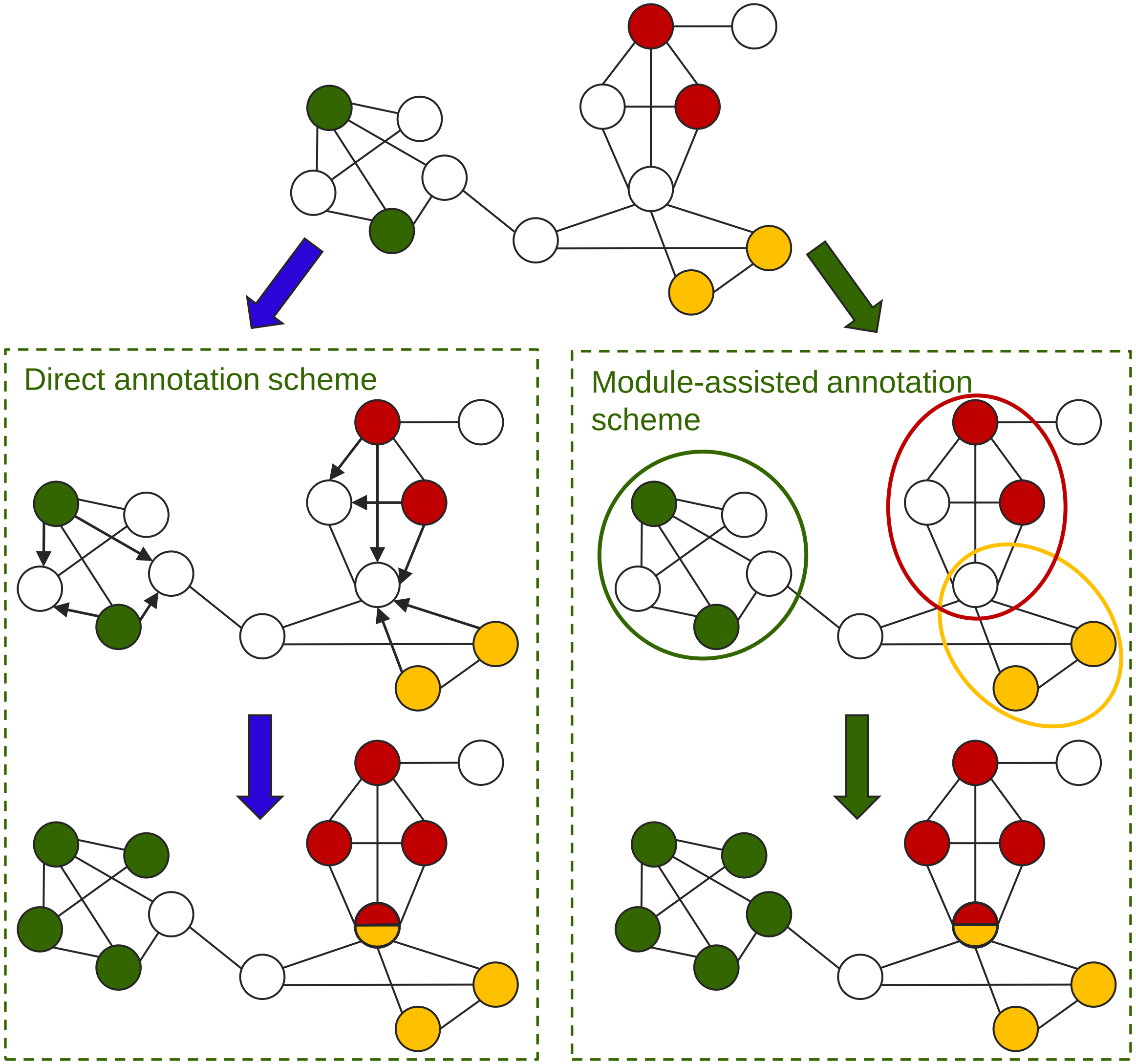
Proteasome subunits – Protein degradation

Network-based protein function prediction

- Knowing a protein's function is essential for understanding biological processes, cellular mechanisms, evolutionary changes and the onset of diseases
- Protein interactions reflect the biological role of proteins within the cells
- Neighboring proteins in a network are likely to share function (guilt-by-association)
- Function might be inferred:
  1. By transferring known functions from directly or indirectly interacting proteins.
  2. Based on the protein complexes a protein belongs to.

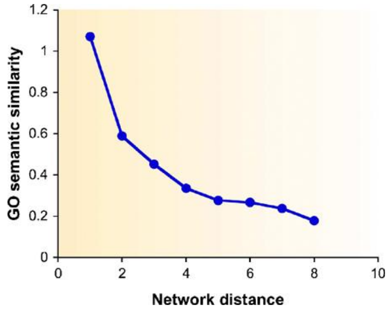
Network-based protein function prediction

- Study the set of neighbors
- Consider position of the protein within its neighborhood,
- Consider position of the protein in the entire interaction network

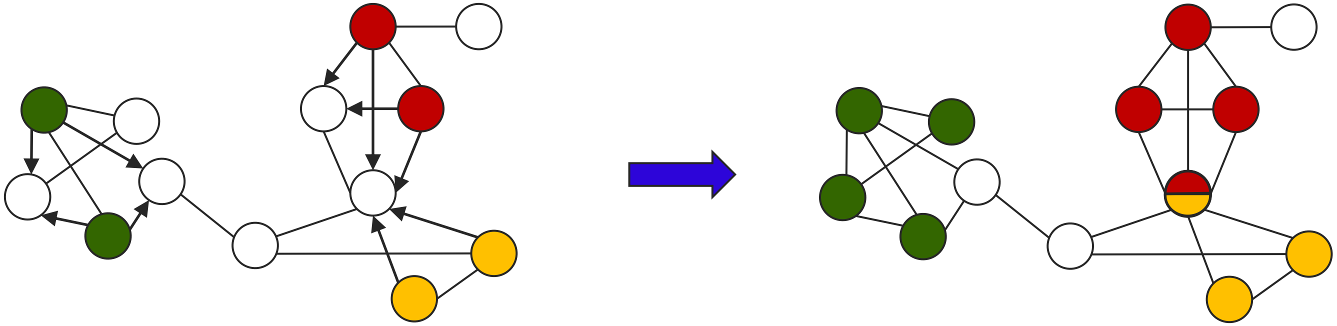


Direct annotation scheme

- Correlation between network and functional distance: the closer two proteins are in the network, the more similar are their functional annotation

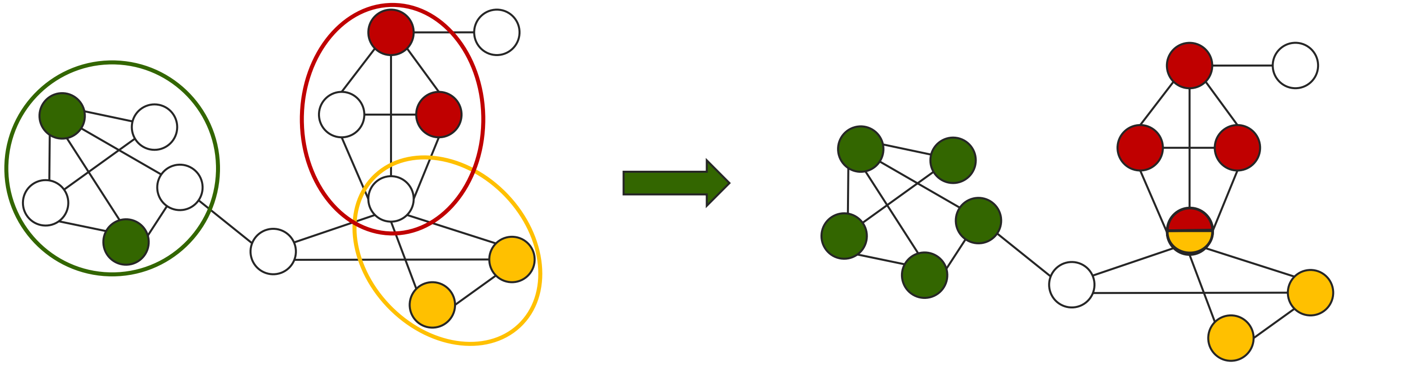


- Majority-rule based on most common function(s) annotated to the direct interaction partners of a protein – proteins are associated with the most frequent functions of its direct neighbors



Module-assisted annotation scheme

- Based on hypothesis: cellular function is organized in a highly modular manner
- Module-based function assignment:
  1. First compute clusters (or modules) within the protein network.
  2. Proteins in a cluster are associated with annotations that are enriched within the module
     - Common functional annotations shared by the majority of the module's proteins
     - Over-represented function that are enriched in a cluster according to the hypergeometric distribution



Direct vs. Module-based methods

- Direct methods
  - + Accurate predictions
  - + Provide high coverage
  - – More sensitive to high level of false positives

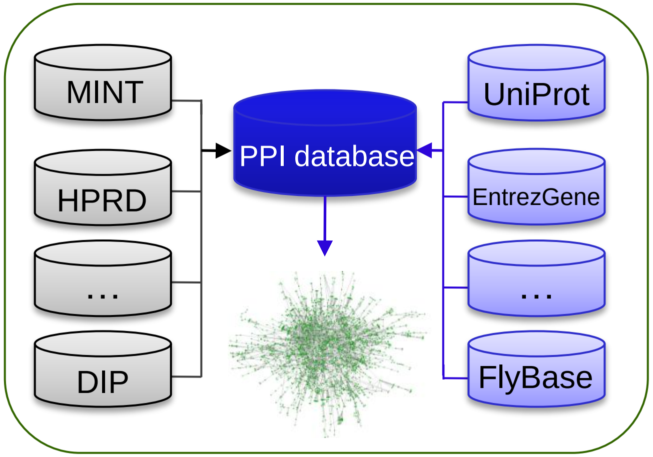
- Module-based methods
  - + More robust to missing or false interactions
  - + Performance improves in networks with less functional coverage
  - – Predict function only in dense network regions → reduced coverage
  - – Less accurate than simple direct methods

– Both methods work within a species, which disregards functional information available in evolutionary related other species

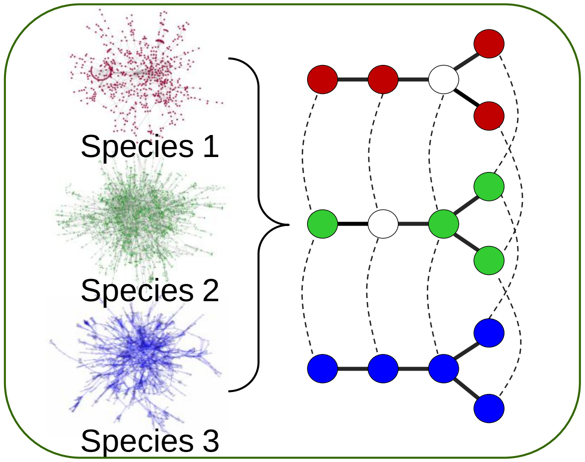
Combining modularity, conservation, and direct interactions in one method

Combine modularity, conservation, and direct interactions

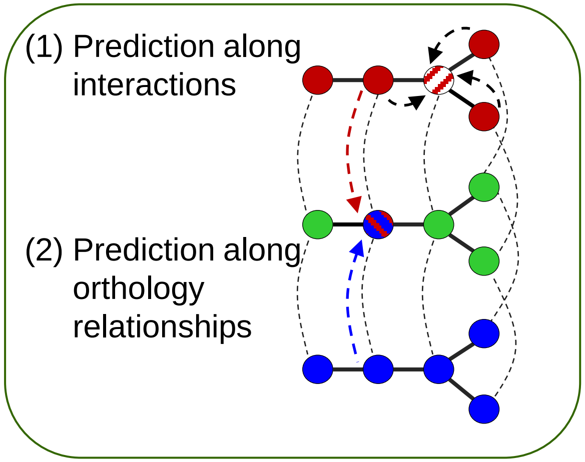
- Assumption: structural conservation in networks correlates with functional conservation in networks which can be exploited for predicting protein functions



a) Data Integration

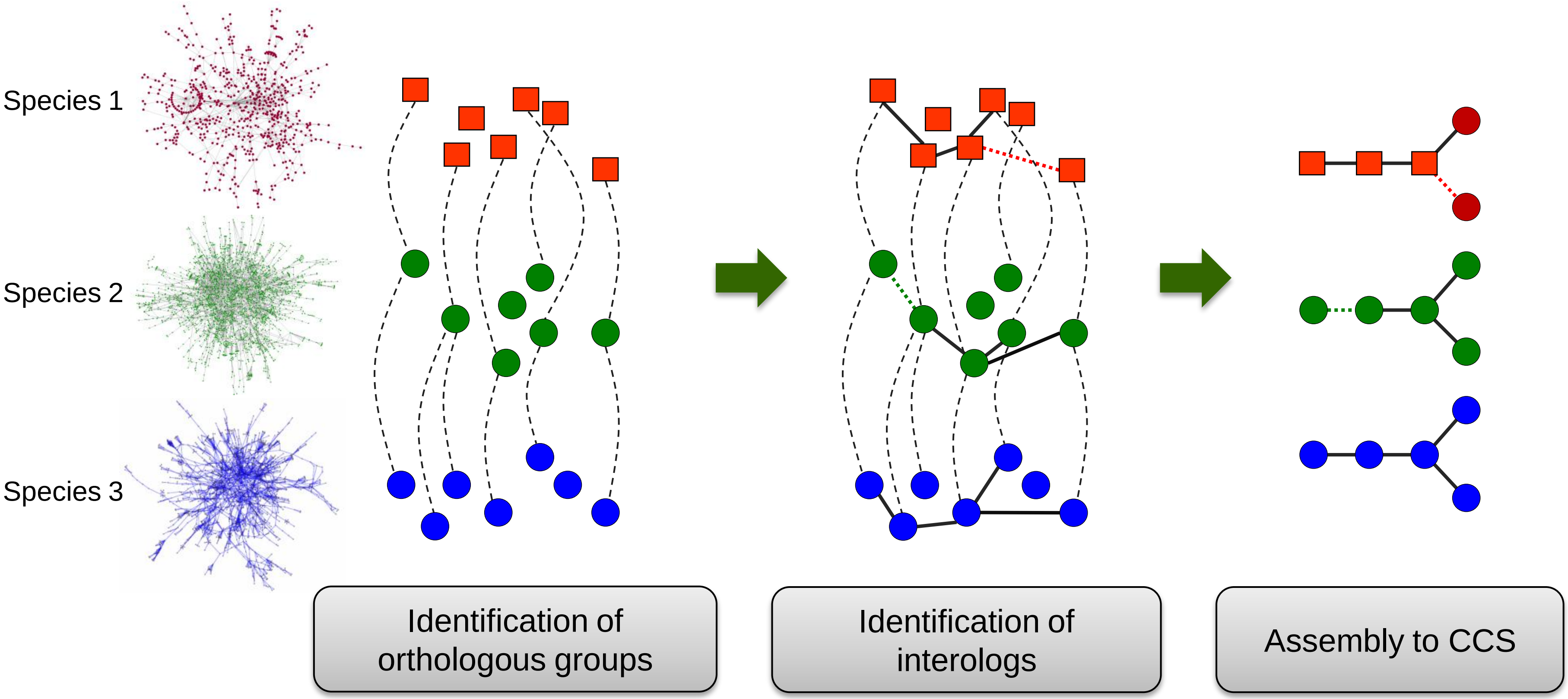


b) Network Comparison


(1) Prediction along interactions
(2) Prediction along orthology relationships

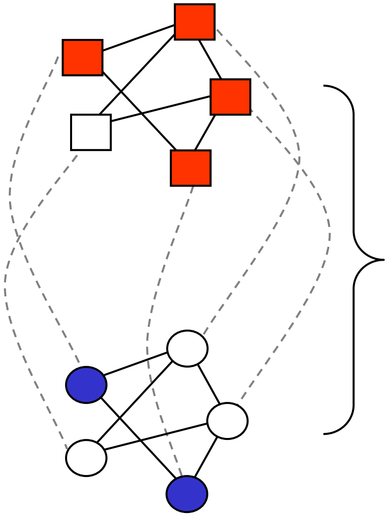
c) Protein Function Prediction

Network Comparison


Species 1
Species 2
Species 3
Identification of orthologous groups
Identification of interologs
Assembly to CCS

CCS-based function prediction

- Analyze proteins within CCS that are defined by evolutionary conserved processes

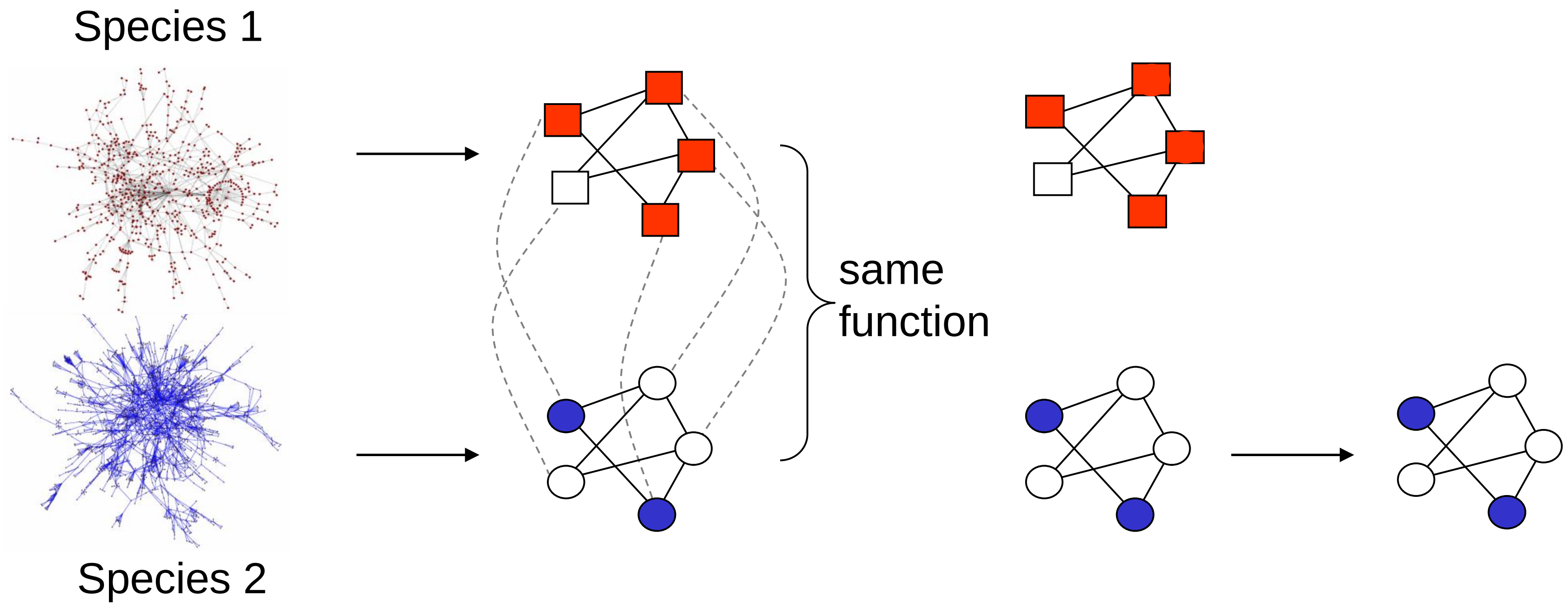


Evolutionary conserved CCS = Functionally coherent?

- Combine comparative cross-species genomics and functional linkage within species-specific networks, predict function from
  - Orthology relationships
  - Direct interaction partners

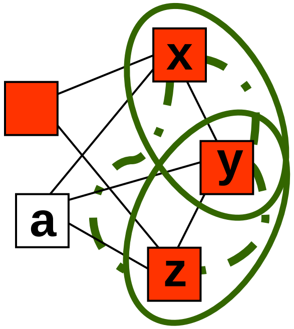
CCS-based function prediction

1) Exploiting Orthology Relationships across multiple species



CCS-based function prediction

2) Exploiting Protein Neighbours – Based on functional similarity between proteins

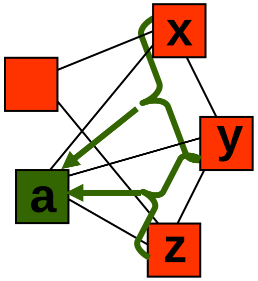


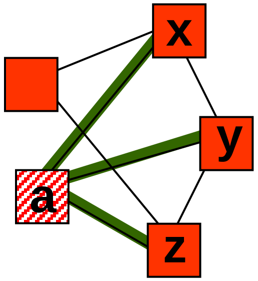
GO_Sim(x,y) > *t*

GO_Sim(x,z) < *t*

GO_Sim(y,z) > *t*

common GO annotations





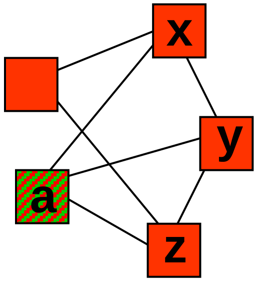
GO_Sim(a,x) > *t*

GO_Sim(a,y) > *t*

GO_Sim(a,z) > *t*

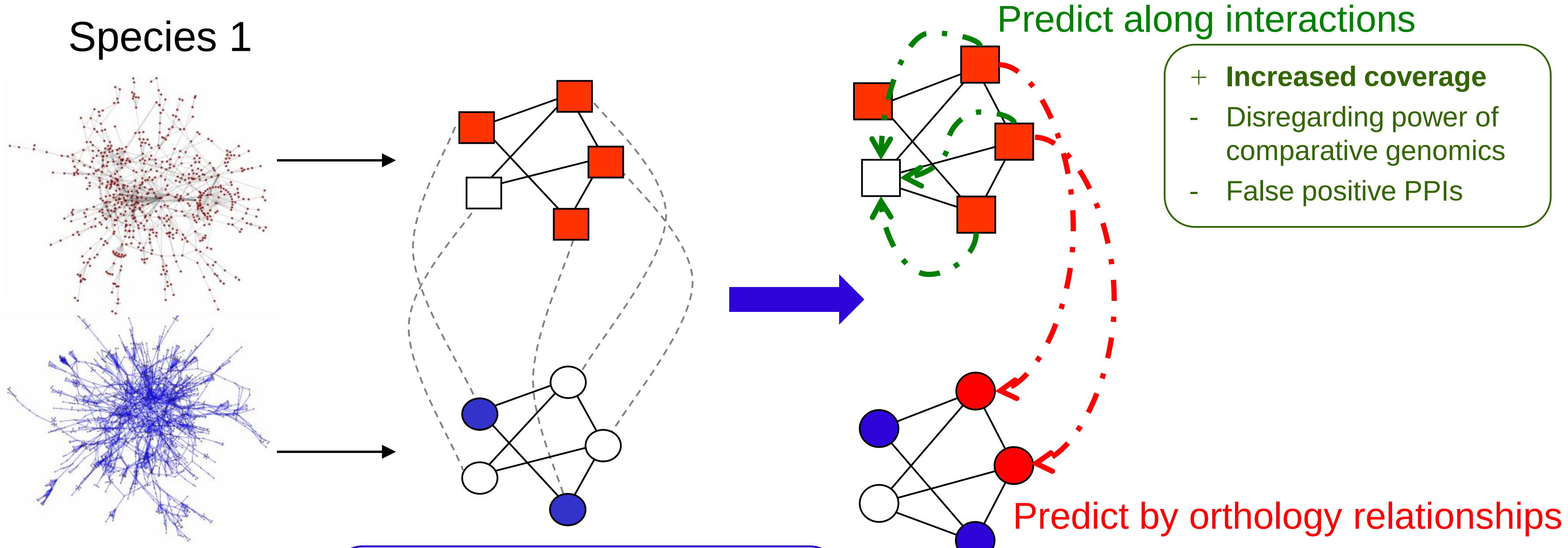
Candidate GO terms = {GO_1, GO_2, …, GO_n}

	protein	GO Sim	GO_i
candidate	a	-	? ✓
training	x	0.7	✓
	y	0.75	-
	z	0.71	✓



CCS-based function prediction

Combining modularity, conservation and interaction



Species 1

Predict along interactions

- + **Increased coverage**
- - Disregarding power of comparative genomics
- - False positive PPIs

Species 2

- + **More robust to missing or false interactions**
- + **Good performance in networks with less functional coverage**

Predict by orthology relationships

- + **Exploit knowledge of well-studied species**
- + **High precision**
- - Limited coverage
- - Restricted to proteins with characterized orthologs

Summary

- Analysis of protein interaction data and protein interaction network facilitates the understanding of cellular organization, function and processes

- Public databases provide large repositories of interaction data of varying quality and quantity

- Sucessfully used to infer novel function and disease associations from interaction partners