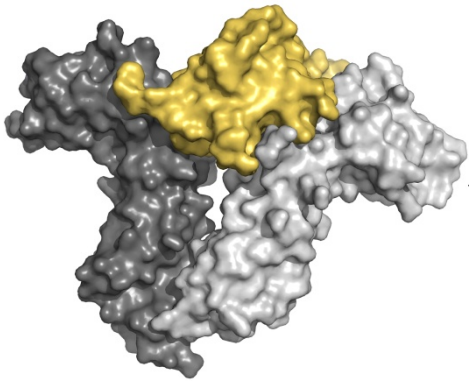
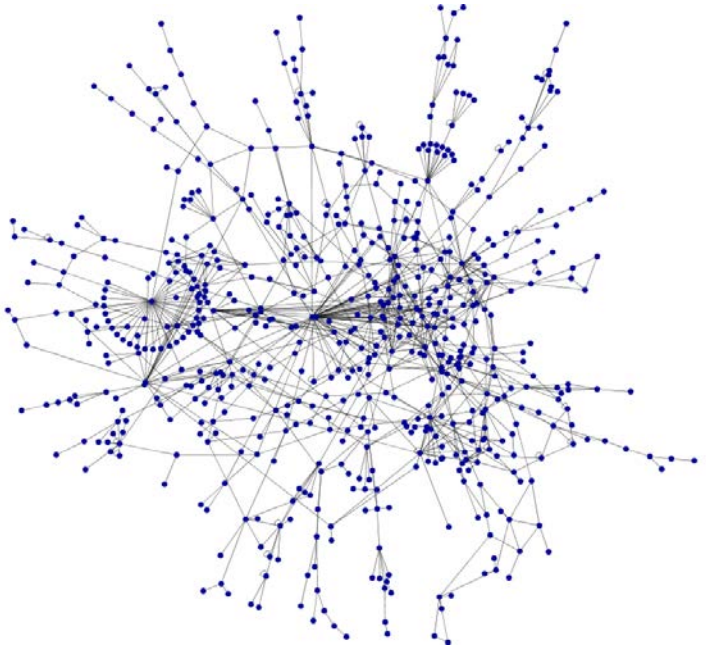
Protein-Protein-Interaction Networks

Johannes Starlinger
Extra credit: Samira Jaeger

This Lecture



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



Motivation

- Interaction: Physical binding of two or more proteins
  - E.g. signal transduction, gene regulation, metabolism, ...
  - Transient or permanent
  - Directed effect (regulates), undirected (binds), specific (activates)
- Changes in protein structure may hinder bindings and thus perturb natural cellular processes
  - Influence on all “downstream” proteins, i.e., proteins reachable through a path of interactions
- Interactome: Set of all PPIs in a cell (type, species, ...)
- Complex: Permanent binding of two or more proteins

Context-dependency

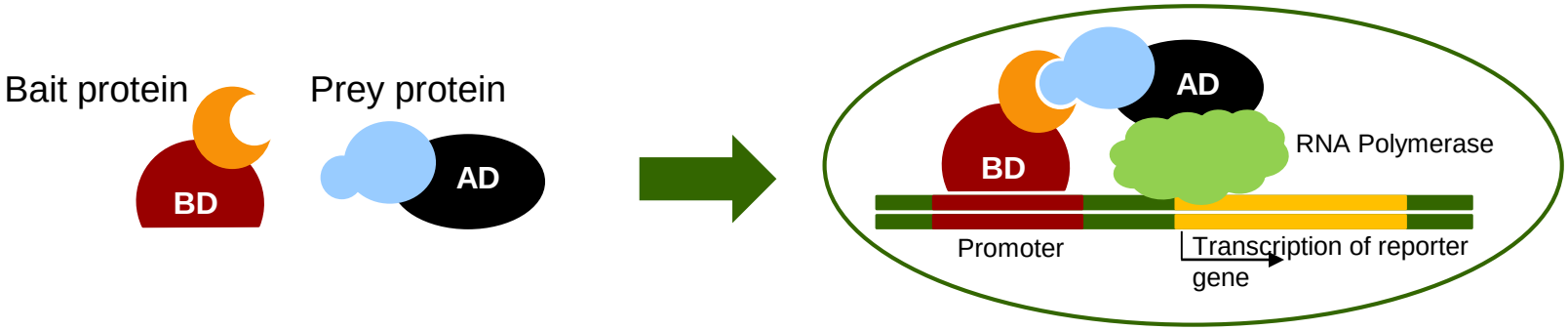
- PPI often is context-dependent
  - Cell type, cell cycle phase and state
  - Environmental conditions
  - Developmental stage
  - Protein modification
  - Presence of cofactors and other binding partners
  - ...
- Disregarded by many PPI detection methods
- Low quality of typical data sets

Experimental detection methods

- PPIs have been studied extensively using different experimental methods
- Many are small-scale: Two given proteins in a given condition
- High-throughput methods
  - Yeast two-hybrid assays (Y2H)
  - Tandem affinity purification and mass spectrometry (TAP-MS)

Yeast two-hybrid screens

- Test if protein A (bait) is interacting with B (prey)
  - Choose a transcription factor T and reporter gene G such that
    - T can be split in two domains: DNA binding and activation
    - If activated T binds to promoter of G, G is expressed
    - Expression of G can be measured
  - Bait is fused to DNA binding domain of T
  - Prey is fused to activating domain of T
  - Both are expressed in genetically engineered yeast cells
  - If A binds to B, T is assembled and G is expressed

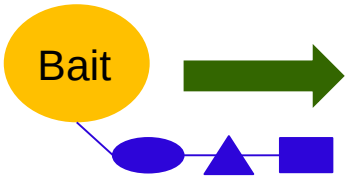


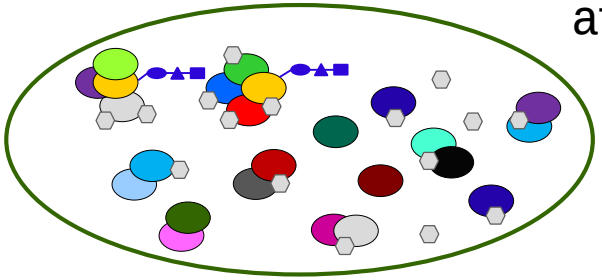
Properties

- Advantages
  - Throughput: Many preys can be tested with same bait (and vice versa)
  - Can be automized – high coverage of interactome
  - Readout can be very sensitive
- Problems
  - High rate of false positives (up to 50%)
    - Artificial environment: Yeast cells
    - No post-translational modifications
    - No protein transport
    - Unclear if proteins in vivo are ever expressed at the same time
    - ...
  - Fusion influences binding behavior – false negatives

Tandem affinity purification and mass spectrometry

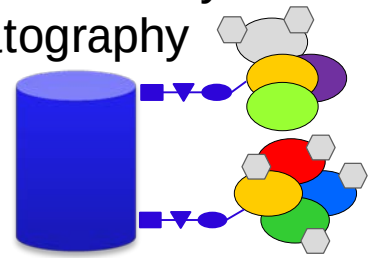
1. Tag the protein of interest



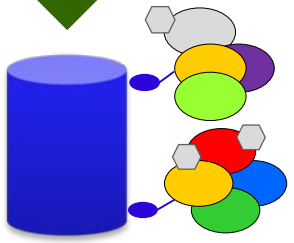


2. Protein binds in its natural environment

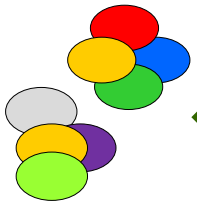
3. Complexes are fished by affinity chromatography

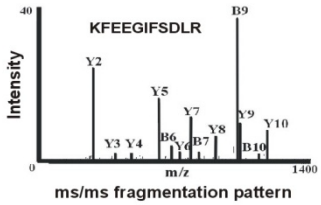


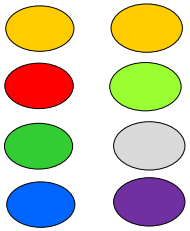
4. Purification



5. Purified protein complexes



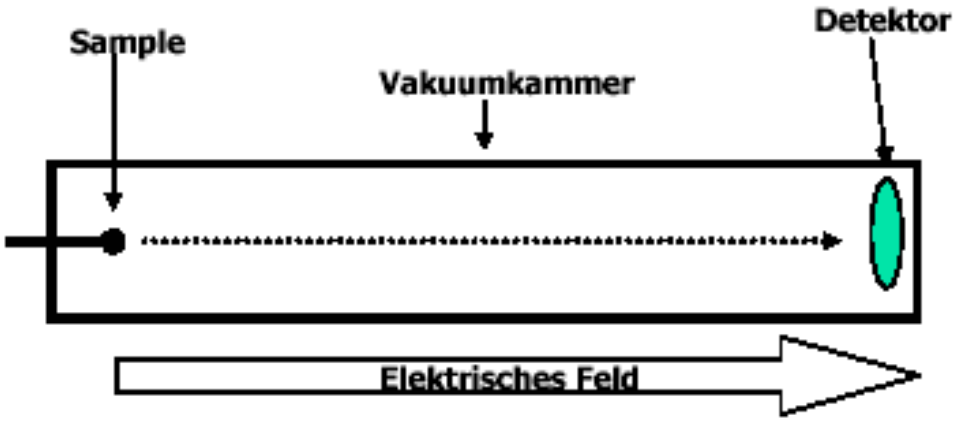




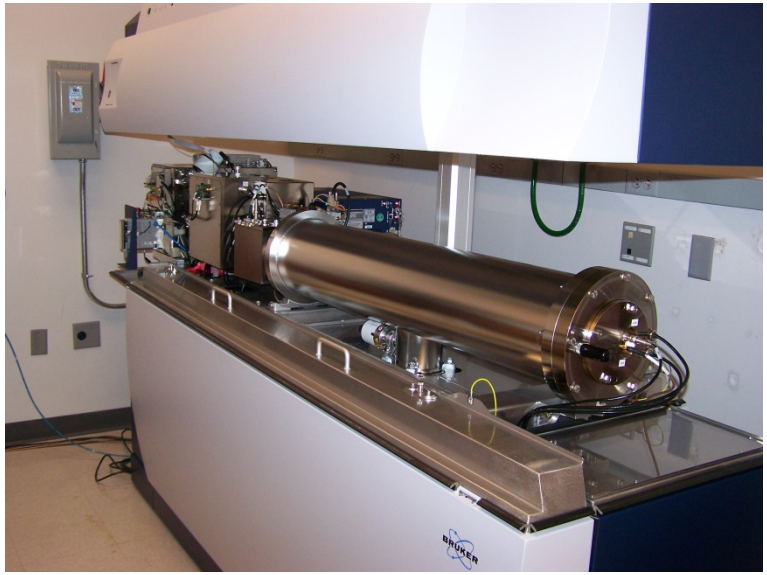
6. Identification of associated proteins by mass spectrometry

Mass Spectrometry

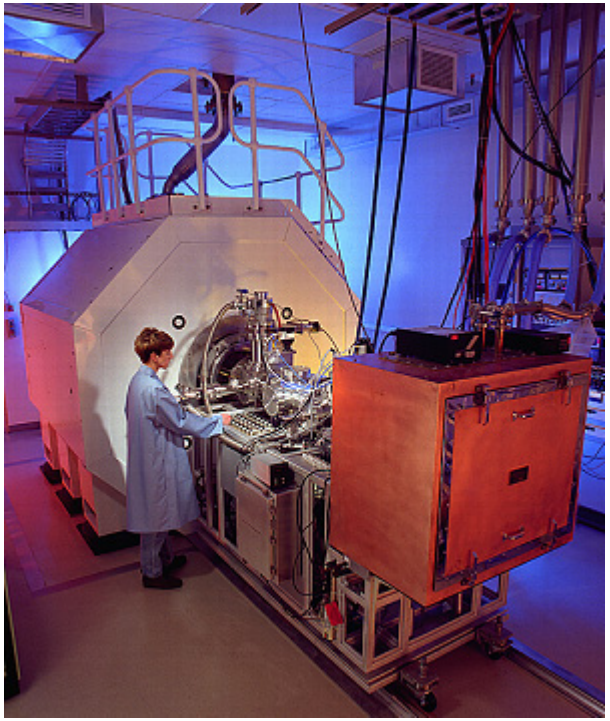
- Accelerate particles (must be charged) in an electric field
- Detector measures hits at back wall
- Time of flight (ToF) proportional to mass
  - Other techniques exist (magnetic drift, ...)
- Spectrum of mass peaks is used to identify particle



Mass Spectrometry



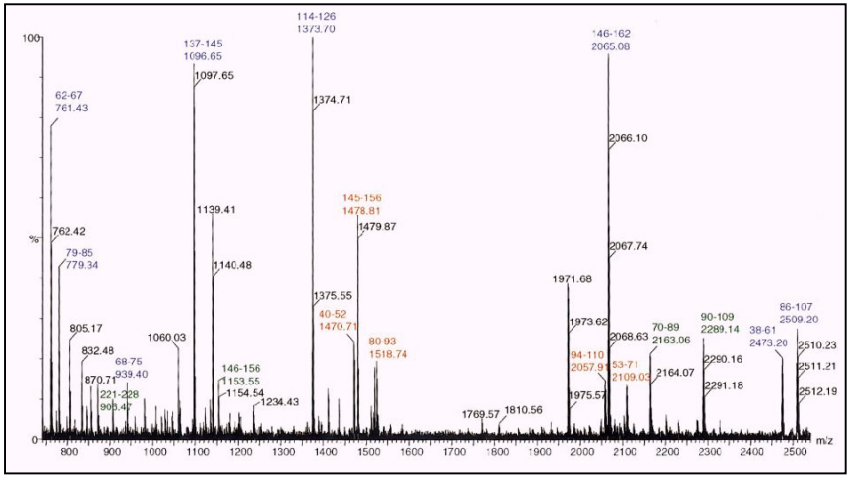
Source: http://imr.osu.edu



Source: http://www.sysbio.org

Mass Spectrometry for Protein Identification

- Problem: Proteins are fragile and break during acceleration
- Solution
  - Break proteins at defined points before acceleration (digestion)
  - Measure peptides (each peptide one signal – time of flight)
  - Identify protein based on spectrum of peptide hits
- In theory, every protein has an almost unique spectrum
  - Using modern MS/MS, even different forms of the same protein are separable

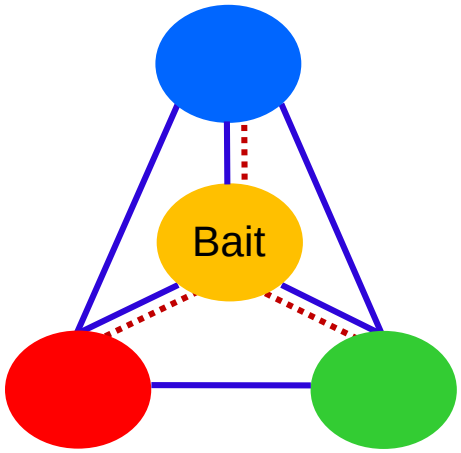


Properties of TAP-MS

- Advantages
  - Can capture PPI in (almost – the tag) natural conditions
  - Single bait can detect many interactions in one experiment
  - Few false positives
- Disadvantages
  - Tag may hinder PPI – false negatives
  - Purification and MS are delicate processes
  - Difficult MS since the input is a mixture of different proteins
  - Individual complexes are not identified
  - Internal structure of complex is not resolved

Matrix / Spokes Model

- Direct interactions can not be distinguished from interactions mediated by other proteins in a complex
- 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

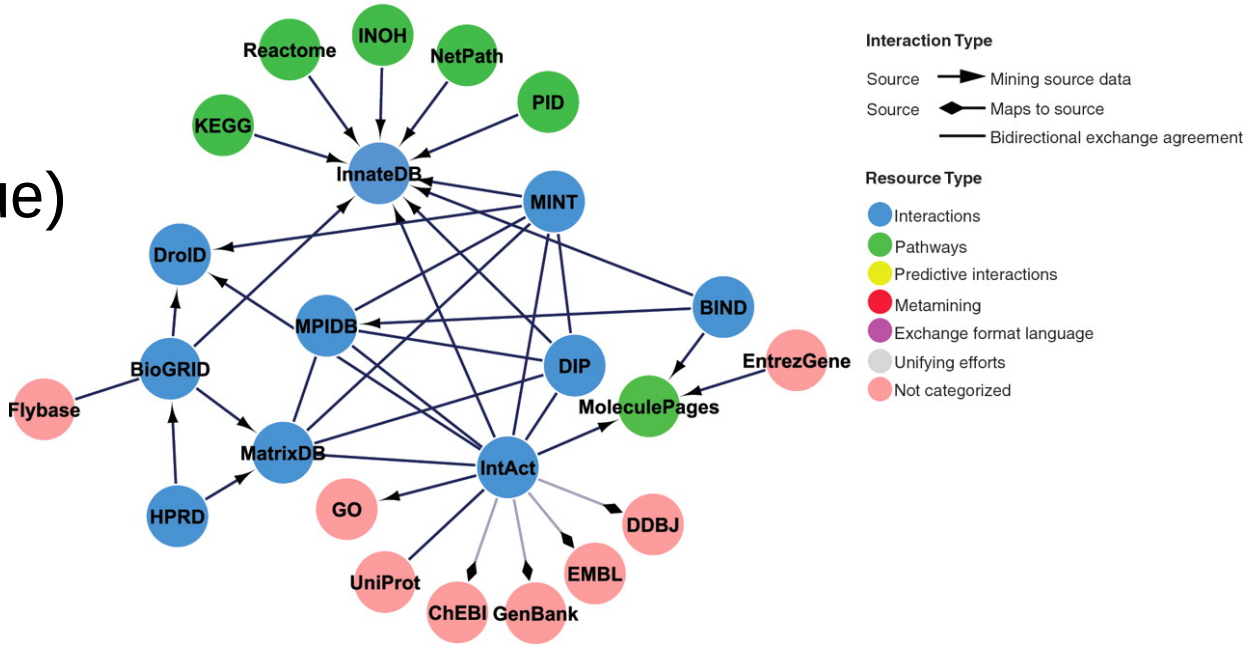


# Proteins	Matrix	Spokes
4	6	3
10	45	9
80	3540	79

PPI Databases [KP10]

- There are >300 DBs related to PPI and pathways
  - See http://www.pathguide.org

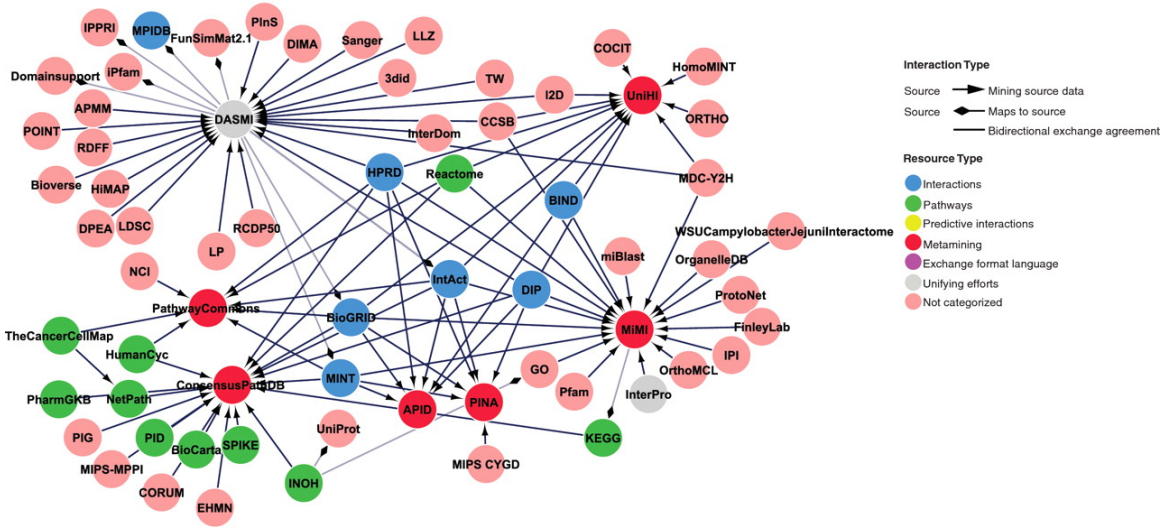
- Manually curated "source" DBs (blue)
  - Gather data from low-throughput methods



PPI Databases

- There are >300 DBs related to PPI and pathways
  - See http://www.pathguide.org

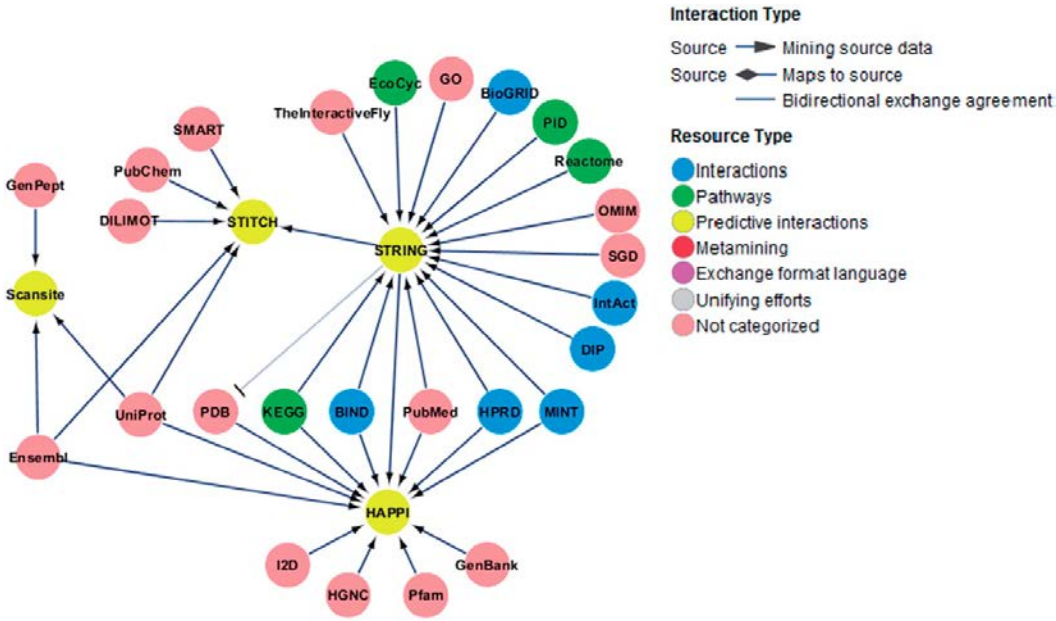
- Manually curated "source" DBs
- DBs integrating other DBs and HT data sets (red)



PPI Databases

- There are >300 DBs related to PPI and pathways
  - See http://www.pathguide.org

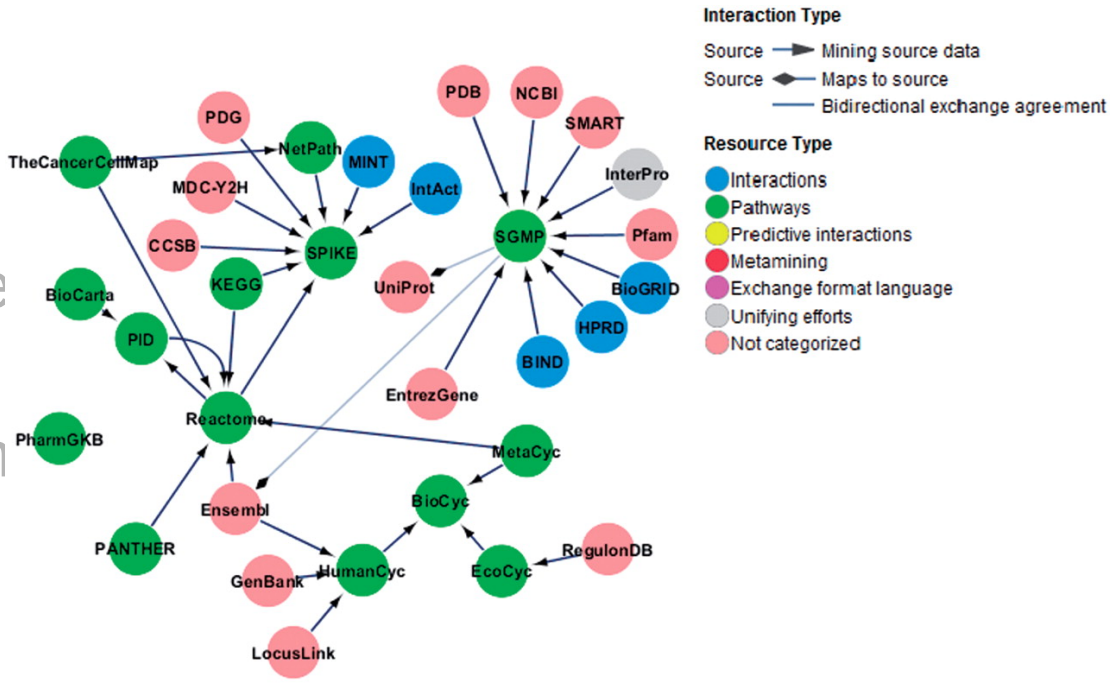
- Manually curated "source" DBs
- DBs integrating others and HT data sets
- Predicted interactions (yellow)



PPI Databases

- There are >300 DBs related to PPI and pathways
  - See http://www.pathguide.org

- Manually curated "source" DBs
- DBs integrating othe and HT data sets
- Predicted interaction
- Pathway DBs (green)



A Mess [KP10]

- Different definitions of a PPI
  - Binary, physical interaction
  - Complexes, pairs, pathways
  - Transient, functional association
- Consistency: Some integrated DBs have “imported” more data than there is in the sources
- Databases overlap to varying degrees
- Different reliability of content
- Literature-curated DBs do not guarantee higher quality than high-throughout experiments [CYS08]
  - Re-annotation reveals inconsistencies, subjective judgments, errors in gene name assignment, ...

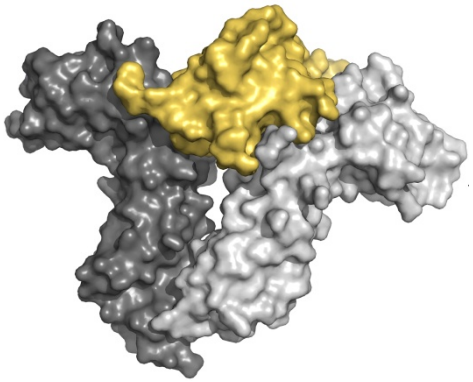
Concrete Examples (numbers slighty outdated)

Database	Species	Proteins	Interactions
IntAct	No restriction	53.276	271.764
BioGrid	No restriction	30.712	131.638
DIP	No restriction	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		
OPID	Human only		
⋮			

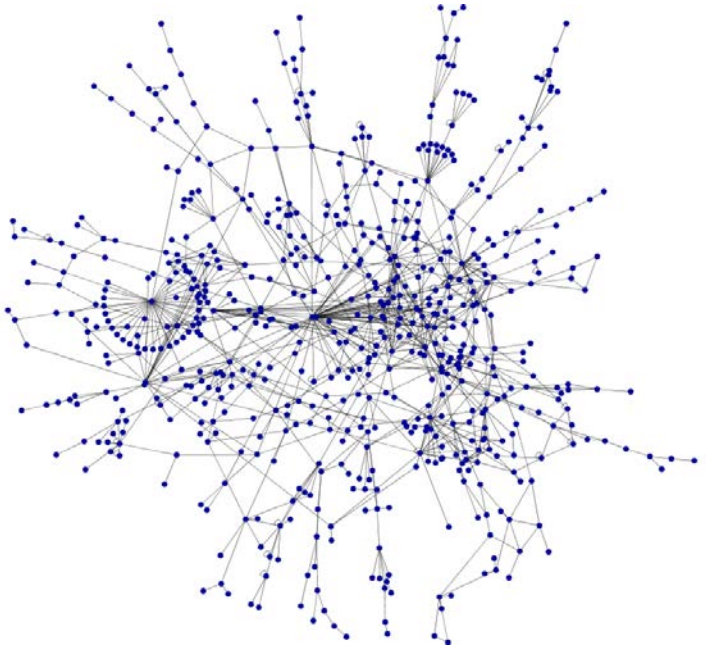
Experimentally verified

Experimentally verified and / or predicted

This Lecture

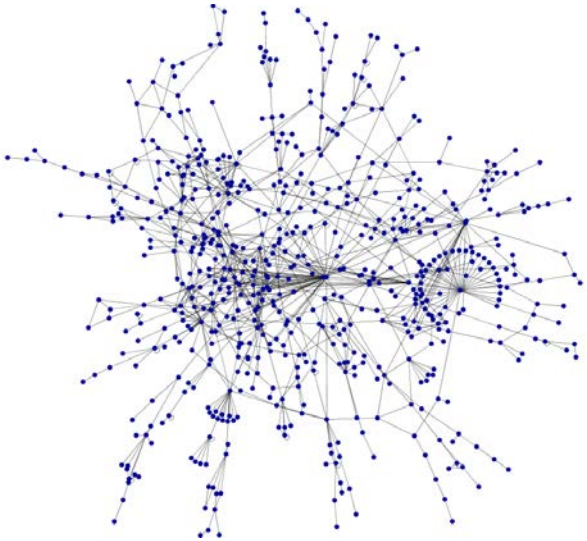


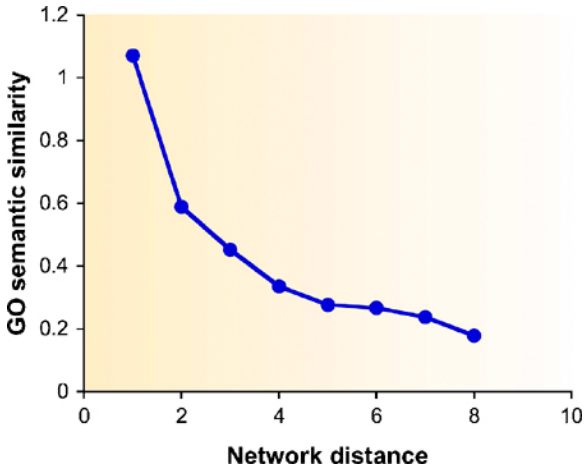
- Protein-protein interactions
- Biological networks
  - Scale-free graphs
  - Cliques and dense subgraphs
  - Centrality and diseases



Some Fundamental Observations

- Proteins that are close in the network share function more frequently
- Central proteins are vital
- Complexes form dense subgraphs
- Functional modules are subgraphs
- Certain subgraphs can be found significantly more often than expected by chance (why?)



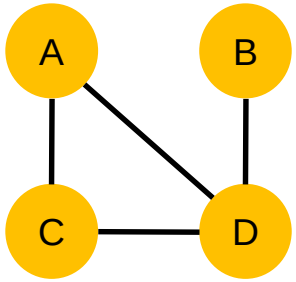


Protein-protein interaction networks

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

Adjacency lists

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

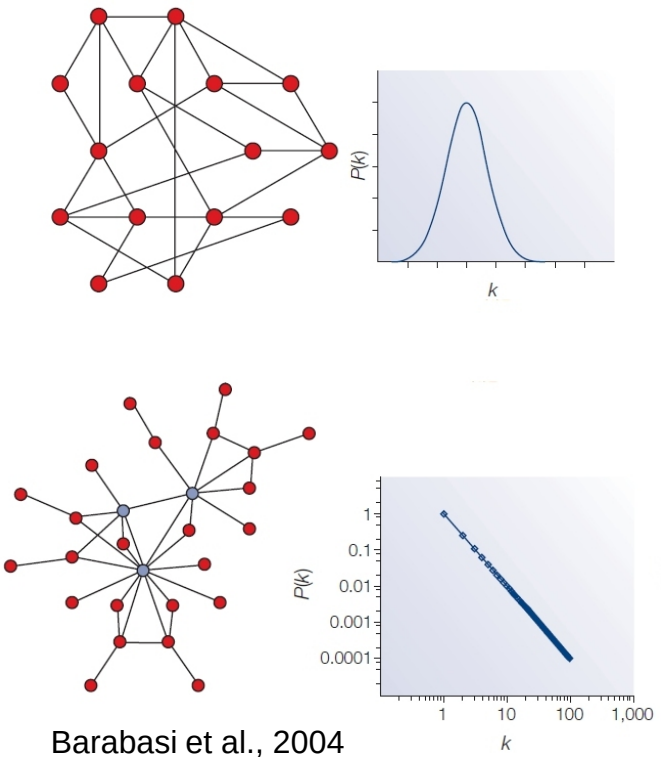


Adjacency matrix

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

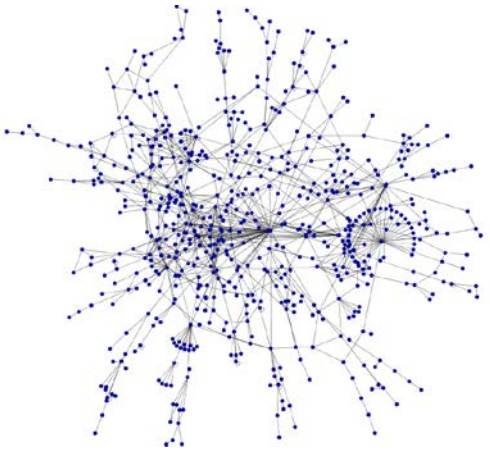
Degree distribution

- Degree distribution P(k): relative frequency of nodes with degree *k*
- Used to define different classes of networks
- Common distributions
  - Poisson $P(k) \sim \frac{\lambda^k}{k!} e^{-\lambda}$
    - Random networks
  - Power-law $P(k) \sim k^{-\gamma}$
    - Scale-free networks

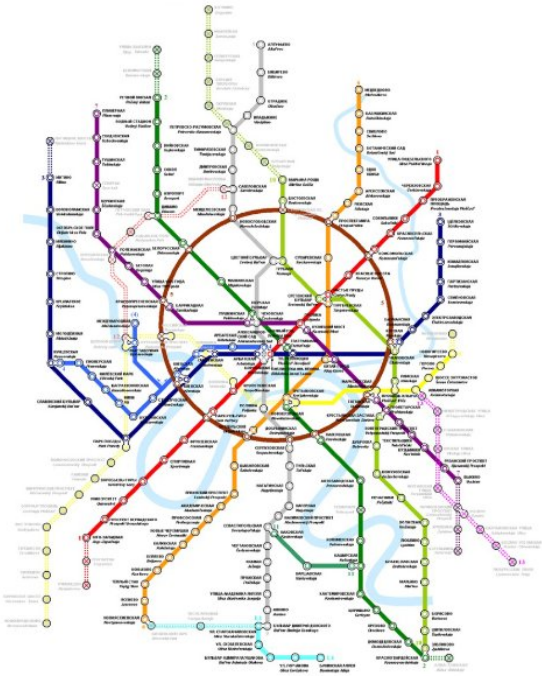


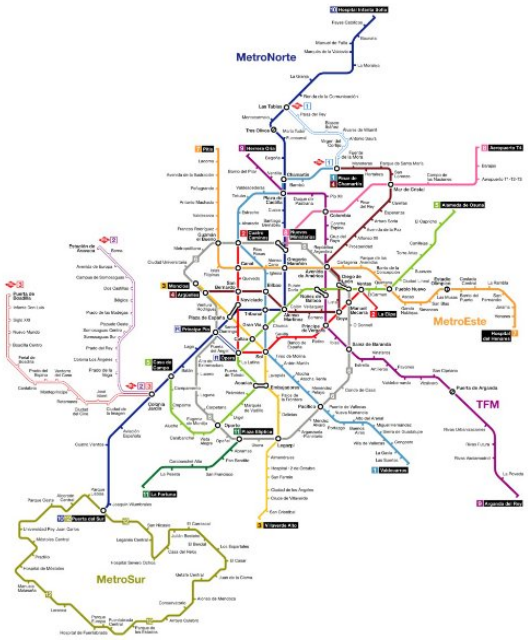
Barabasi et al., 2004

Scale-free Networks



- Biological networks are (presumably) scale-free
  - Few nodes are highly connected (hubs)
  - Most nodes have very few connections
- Also true for many other graphs: electricity networks, public transport, social networks, ...
- Evolutionary explanation
  - Growth: Networks grow by addition of new nodes
  - Preferential attachment: new nodes prefer linking to highly connec. nodes
    - Possible explanation: Gene duplication – interaction with same targets
  - Older nodes have more chances to connect to nodes
  - Hub-structure emerges naturally



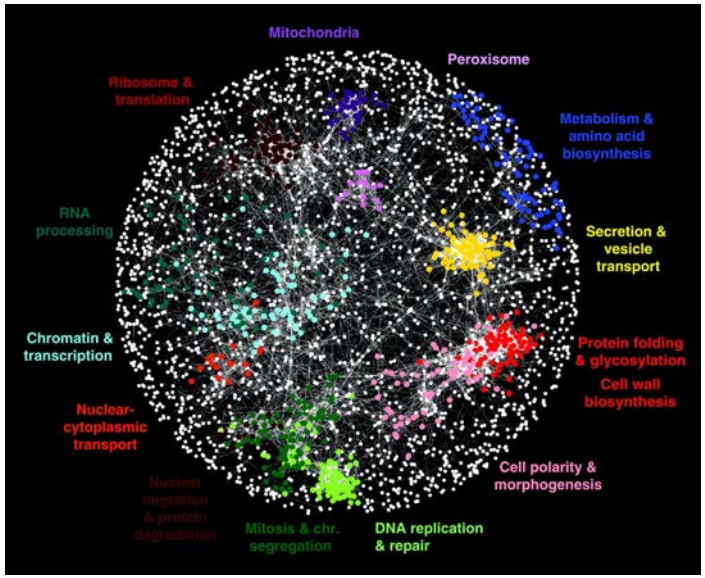

MetroNorte
MetroEste
TFM
MetroSur

Other Biological Networks

- Regulatory networks: How genes / transcription factors influence the expression of each other
  - TF regulate expression of genes and of other TFs
  - Edges semantics: activate / inhibit / regulate
  - Important, for instance, in cell differentiation
- Signal networks: Molecular reaction to external stimulus
  - Transient interactions including small molecules
  - Temporal dimension important (fast)
  - Important, for instance, in oncology
- Metabolic networks
- Protein-protein interaction networks

Modular network organization

- Cellular function is carried out by modules
  - Sets of proteins interacting to achieve a certain function
- Function is reflected in a modular network structure

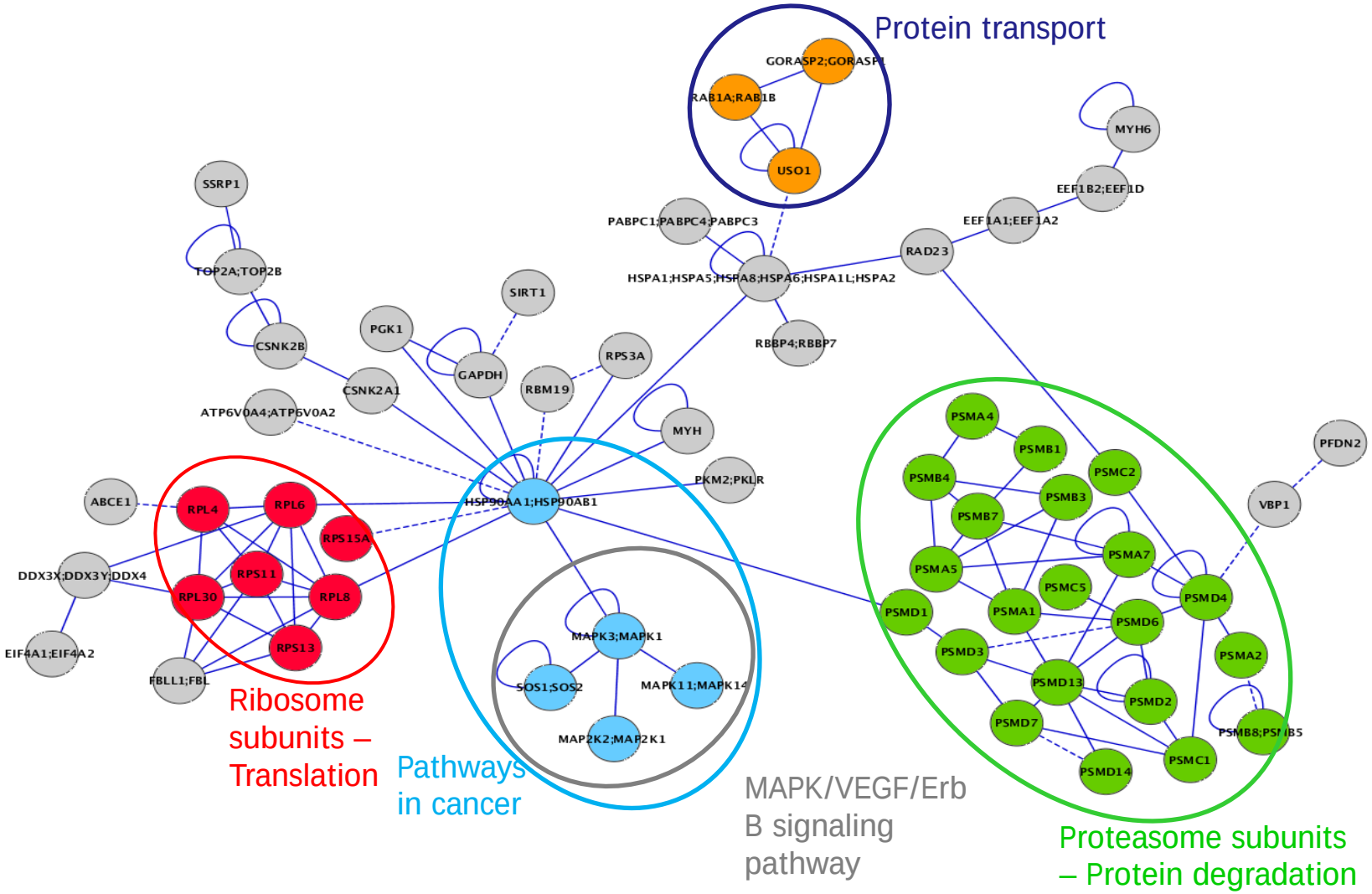


Don‘t be fooled by layout

Modules must be dense, not close

Costanzo et al., Nature, 2010

Functional Modules



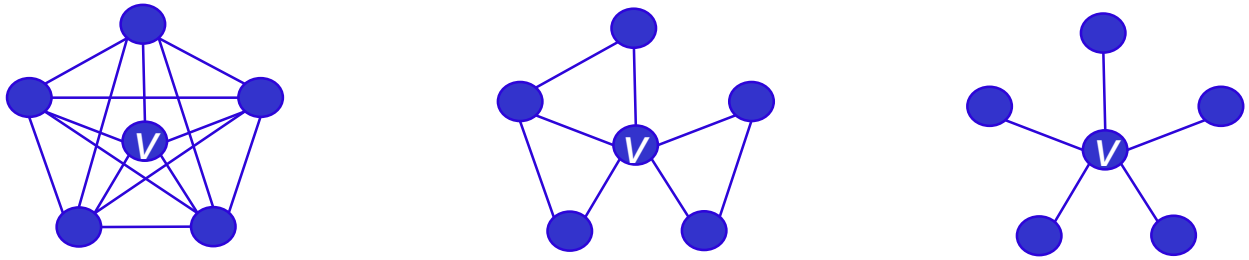
Clustering Coefficient

- Modules (clusters) are densely connected groups of nodes
- Cluster coefficient C reflects network modularity by measuring 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



$C_v = 10/10 = 1$ $C_v = 3/10 = 0.3$ $C_v = 0/10 = 0$

- Cluster coefficient C is a measure for the entire graph
- We also want to find modules, i.e., regions in the graph with high cluster coefficient
- A clique is a maximal complete subgraph, i.e., a maximal set of nodes where every pair is connected by an edge

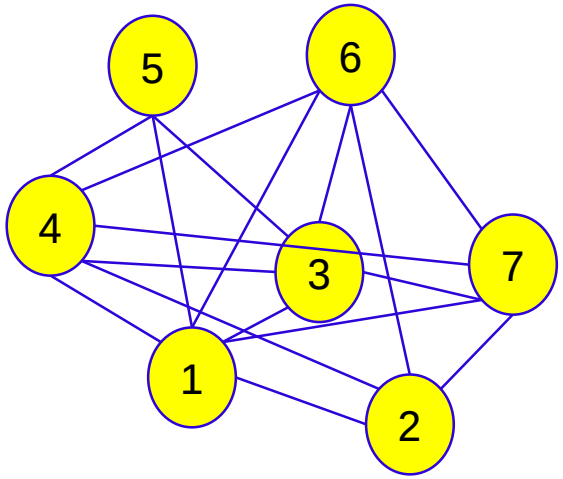
Finding Modules / Cliques

- Finding all (maximal) cliques in a graph is intractable
  - NP-complete
- Finding "quasi-cliques" is equally complex
  - Cliques with some missing edges
  - Same as subgraphs with high cluster coefficient

```
build set S_2 of all cliques of size 2
i:= 2;
repeat
  i := i+1;
  S_i := ∅;
  for j := 1 to |S_i-1|
    for k := j+1 to |S_i-1|
      T := S_i-1[j] ∩ S_i-1[k];
      if |T|=i-1 then
        N := S_i-1[j] ∪ S_i-1[k];
        if N is a clique then
          S_i := S_i ∪ N;
        end if;
      end if;
    end for;
  end for;
until |S_i| = 0:
```

- Various heuristics
  - E.g. a good quasi-clique probably contains a (smaller) clique

Example



- 4-cliques: (1,3,4,5) – (1,3,4,6) – (1,3,4,7) - ...
- Merge-Phase

$|(1,3,4,6) \cap (1,3,4,7)| = 3$

$(1,3,4,6) \cup (1,3,4,7) = (1,3,4,6,7)$

Edge (6,7) exists

5-clique

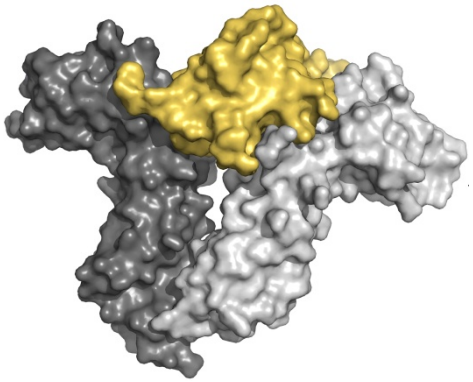
$|(1,3,4,5) \cap (1,3,4,6)| = 3$

$(1,3,4,5) \cup (1,3,4,6) = (1,3,4,5,6)$

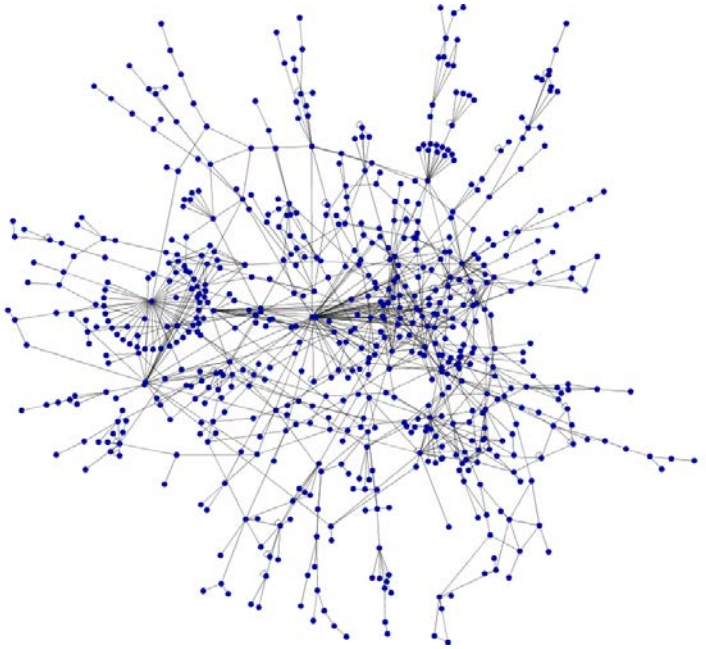
Edge (5,6) does not exists

No clique

This Lecture

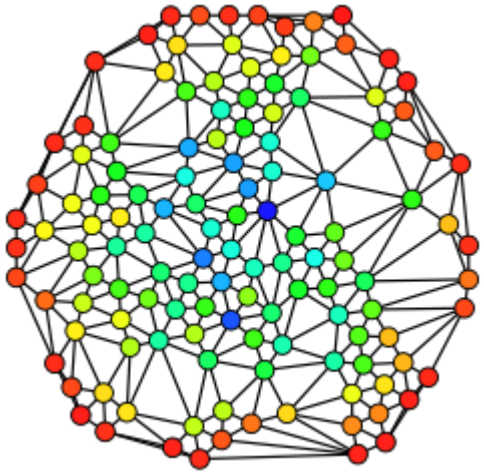


- Protein-protein interactions
- Biological networks
  - Scale-free graphs
  - Cliques and dense subgraphs
  - Centrality and diseases

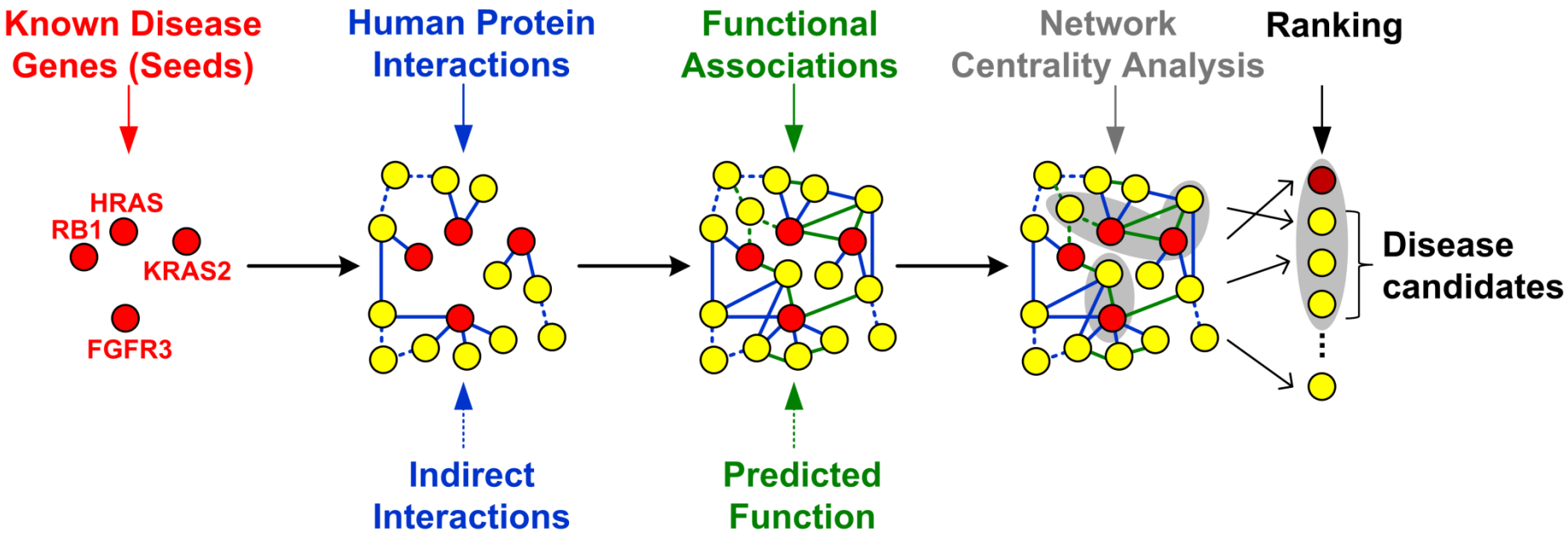


Network centrality

- Central proteins exhibit interesting properties
  - Essentiality – knock-out is lethal
  - Much higher evolutionary conservation
  - Often associated to (certain types of) human diseases
- Various measures exist
  - Degree centrality: Rank nodes by degree
  - Betweenness-centrality: Rank nodes by number of shortest paths between any pair of nodes on which it lies
  - Closeness-centrality: Rank nodes by their average distance to all other nodes
  - PageRank
  - ...

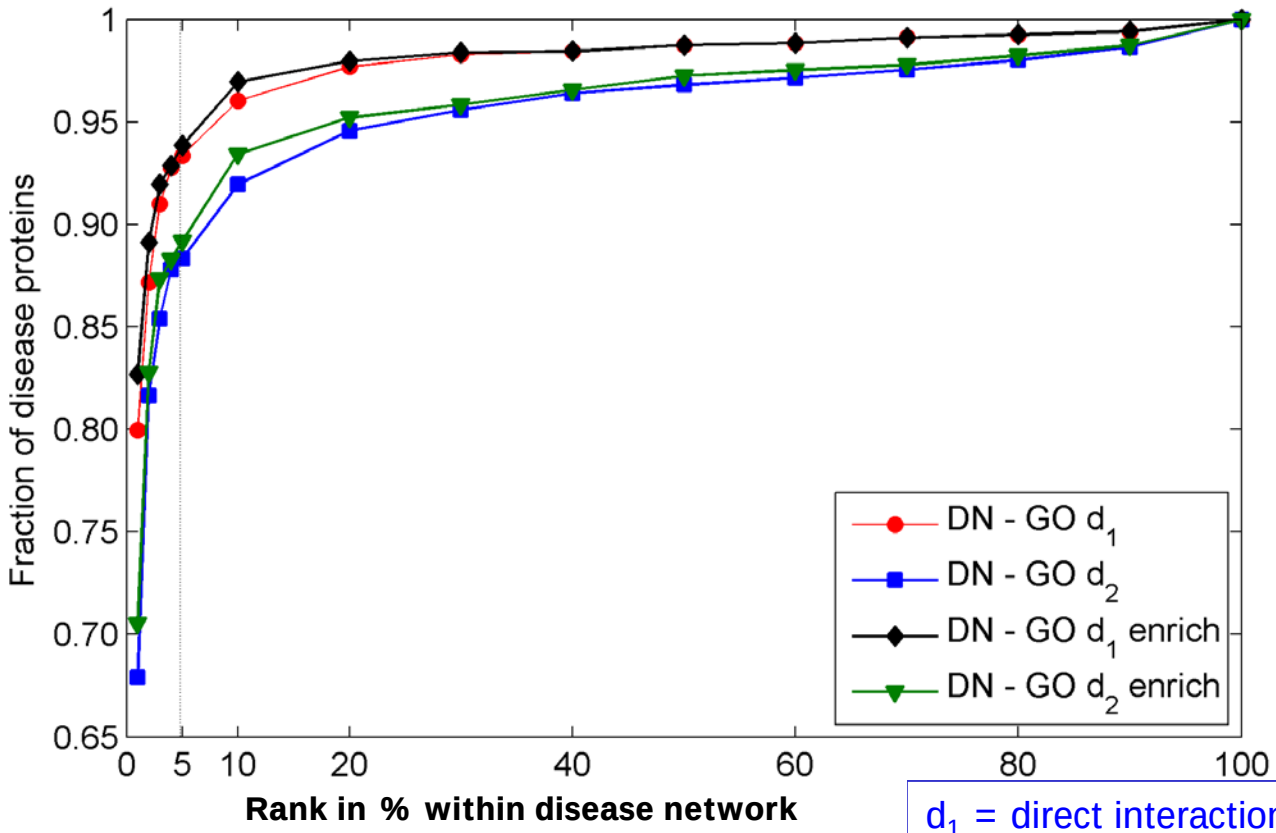


Network-based Disease Gene Ranking



Centrality of Seeds in (OMIM) Disease Networks

Fraction of seeds among top k% proteins; ~600 diseases from OMIM

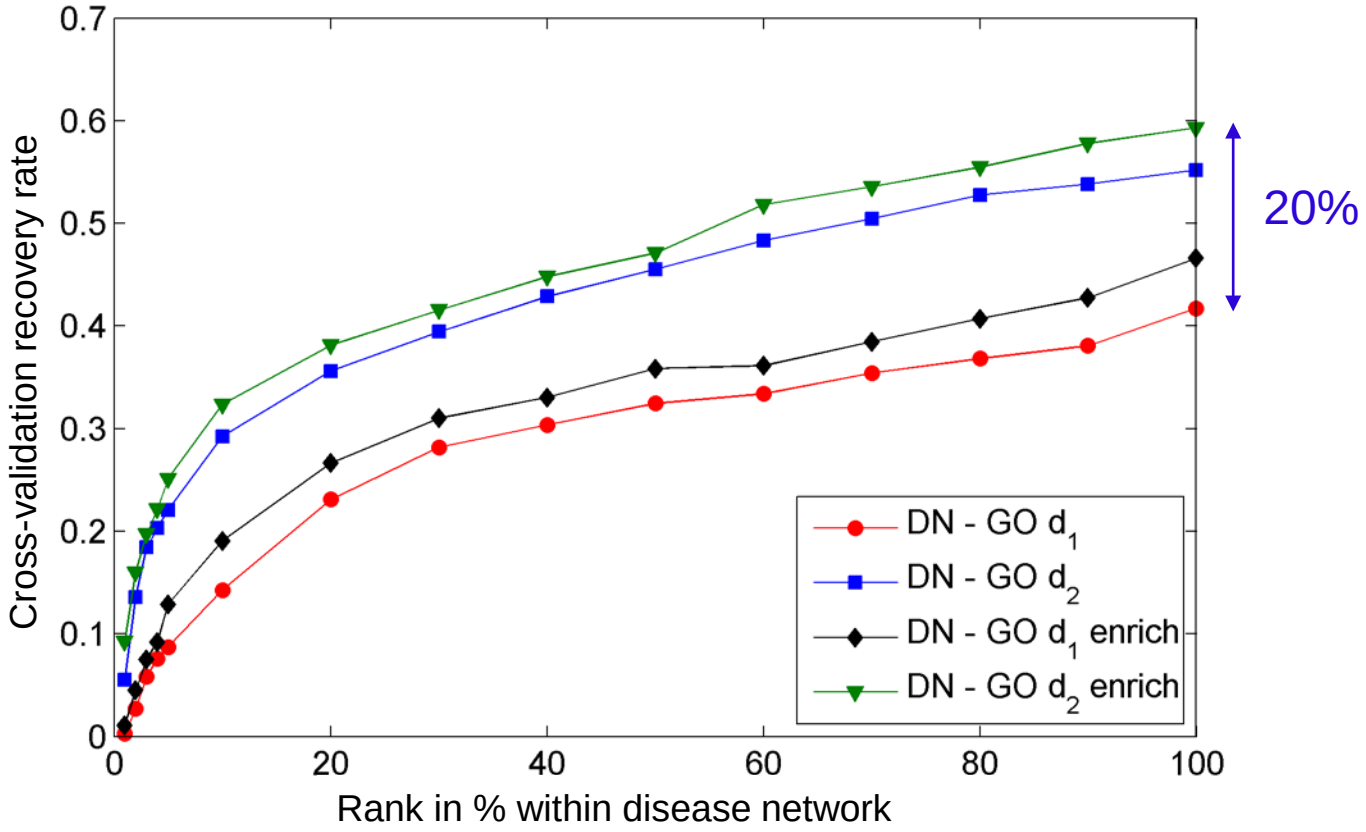


d_1 = direct interactions

d_2 = direct and indirect interactions

Cross-Validation

- If a disease gene is not yet known – can we find it?



Further Reading

- Jaeger, S. (2012). "Network-based Inference of Protein Function and Disease-Gene Associations". Dissertation, Humboldt-Universität zu Berlin.
- Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. and Barabasi, A. L. (2007). "The human disease network." Proc Natl Acad Sci U S A 104(21): 8685-90.
- Ideker, T. and Sharan, R. (2008). "Protein networks in disease." Genome Res 18(4): 644-52.
- Barabasi, A. L. and Oltvai, Z. N. (2004). "Network biology: understanding the cell's functional organization." Nat Rev Genet 5(2): 101-13.