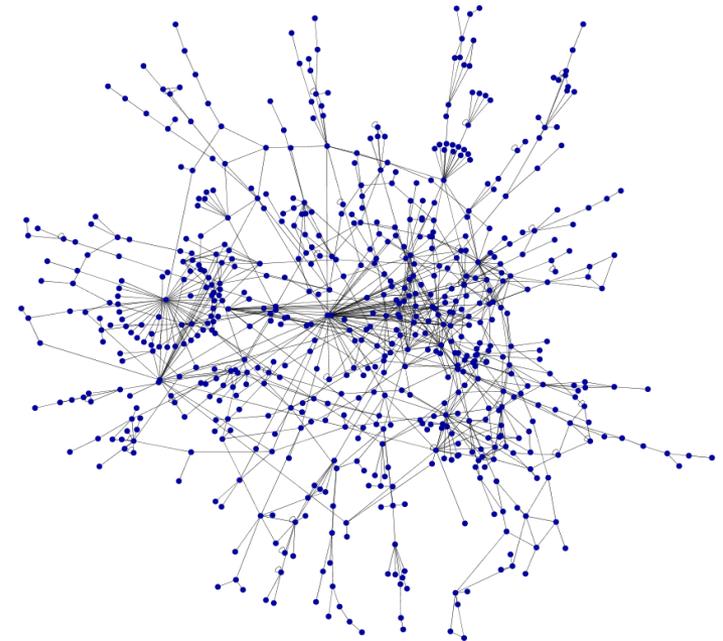
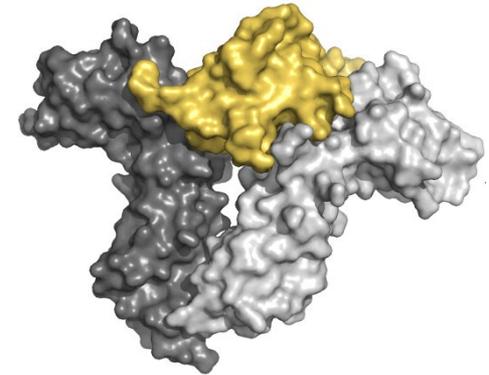


Protein-Protein-Interaction Networks

Ulf Leser

This Lecture

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



Motivation

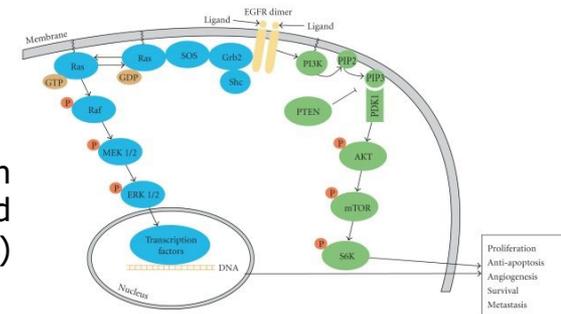
- Interaction: **Physical binding** of two or more proteins
 - E.g. signal transduction, gene regulation, protein transport, ...
 - Transient (signal) or permanent (complex formation)
 - Directed effect (regulates) or undirected (binds)
 - Specific (activates) or unspecific (binds, interacts)
- Changes 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 happening in a cell
 - Typically includes complex formation

Biological pathways

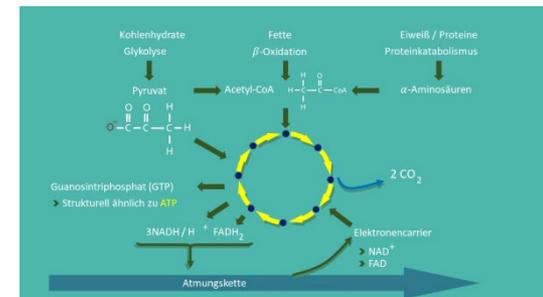
- PPI are an important step to define **pathways**: Groups of proteins performing a complex function together

- EG KRAS signaling: Controlling cell replication

Janicik et al. Clinical Relevance of KRAS in Human Cancers, Journal of Biomedicine and Biotechnology 2010(16)



- Citric acid cycle: Important for energy management and creation of inter-mediate



<https://studyflix.de/biologie/citratzyklus-2143>

Context-dependency

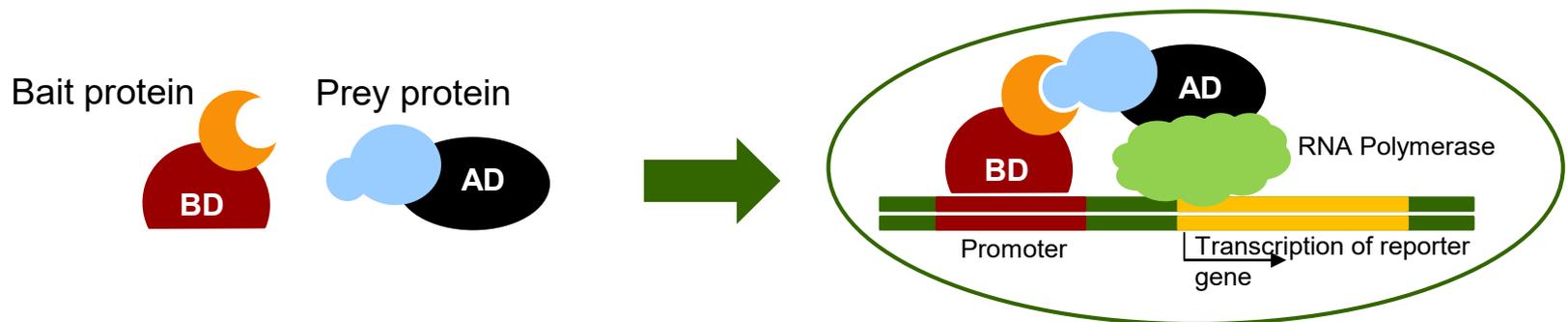
- Most PPIs are **context-dependent**
 - Cell type, cell cycle phase and state
 - Environmental conditions
 - Developmental stage
 - Protein modification
 - Presence of cofactors and other binding partners
 - Species
 - ...
- Disregarded by many PPI detection methods
- **Low quality** of typical data sets
 - Predicted / measured PPI do not happen in (most) real cells

Experimental methods

- PPIs are studied extensively using different experimental methods
- Many are **small-scale**: Two given proteins in a given condition
 - Classical biochemistry
- **High-throughput methods**
 - Yeast two-hybrid assays (Y2H)
 - Tandem affinity purification and mass spectrometry (TAP-MS)
- Many more measuring indirect “interactions”
 - Co-expressed in cell; part of same pathway; ...

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
 - If **activated T binds to promoter** of G, G is expressed
 - Expression of G can be measured (e.g. PCR)
 - T has two domains: Binding domain (BD) and Activation Domain (AD)
 - 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

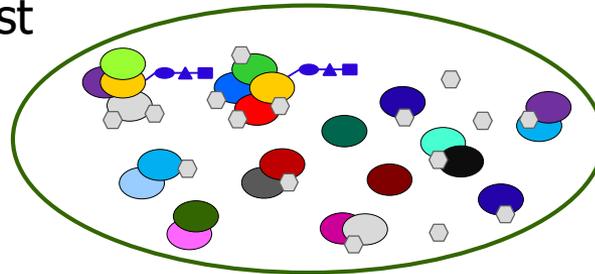
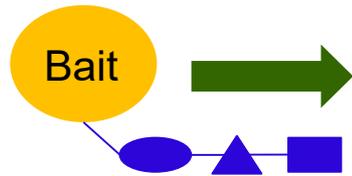
- Throughput: Many preys tested with same bait (and vice versa)
- Can be automatized in yeast cells – **high coverage** of interactome
 - Insertion of specifically designed gene complexes into yeast
- Readout **very sensitive** (PCR)

- Problems

- High **rate of false positives** (up to 50%)
 - Artificial environment and regulation: Yeast cells
 - No post-translational modifications, no spatial context
 - Unclear under which conditions the two proteins in vivo are expressed at the same time
 - ...
- Fusion influences binding behavior – **false negatives**

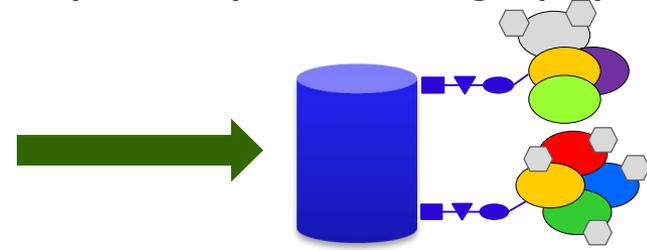
Tandem affinity purification and mass spectrometry

1. Genetically tag protein X of interest

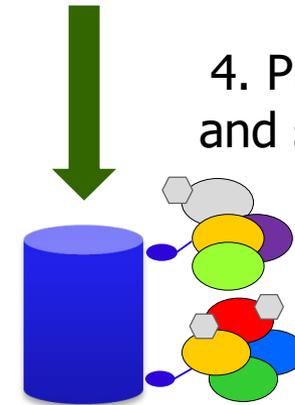


2. X binds all partners in its natural environment

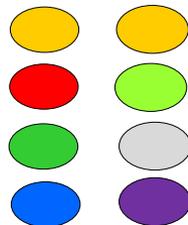
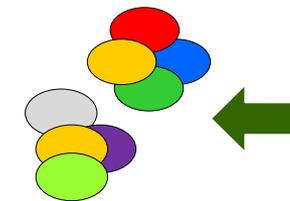
3. Complexes including X are "fished" by affinity chromatography



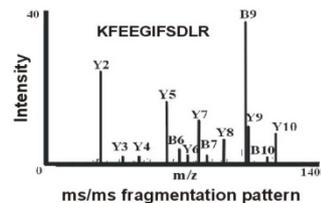
4. Purification and separation



5. Isolated purified protein complexes



6. Identification of constituents by mass spectrometry

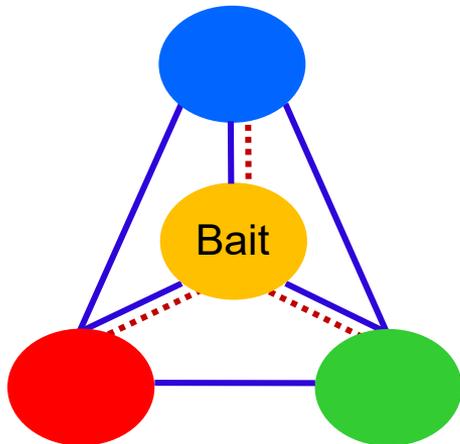


Properties

- 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**
 - MS needs to measure a mixture of different proteins (complex)
 - Internal structure of complex is not resolved
 - Who binds whom?

Matrix / Spokes Model

- General: Methods cannot discern **direct interactions** from interactions mediated by other proteins (**complexes**)
- **Matrix model**: Assume interactions between all proteins of a purified complex $\rightarrow (N*(N-1))/2$
- **Spokes model**: Assume only interactions between the bait and the co-purified proteins $\rightarrow N-1$

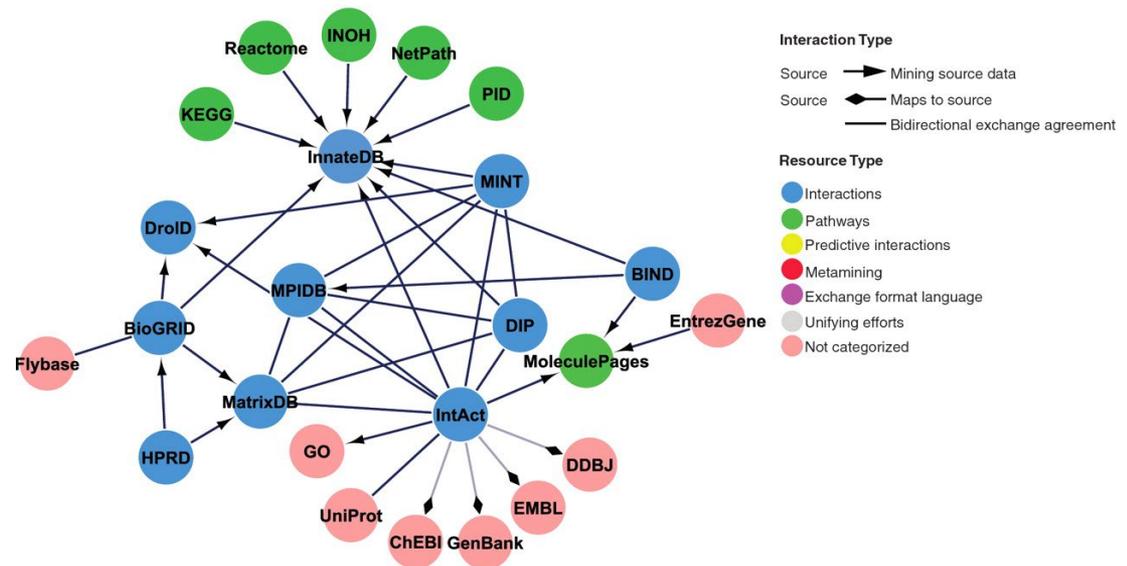


# Proteins	Matrix	Spokes
4	6	3
10	45	9
80	3540	79

PPI Databases [KP10]

- There are ~700 DBs related to PPI and pathways
 - See <http://www.pathguide.org>

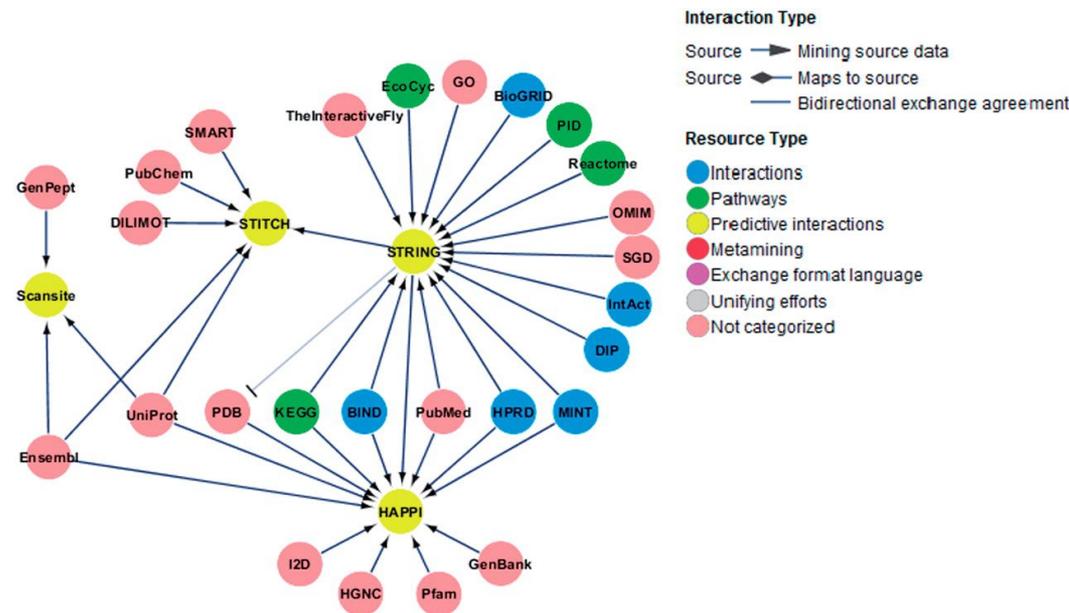
- Manually curated “source” DBs (blue)
 - Gather data from published low-throughput methods (text mining + curation)



PPI Databases

- There are ~700 DBs related to PPI and pathways
 - See <http://www.pathguide.org>

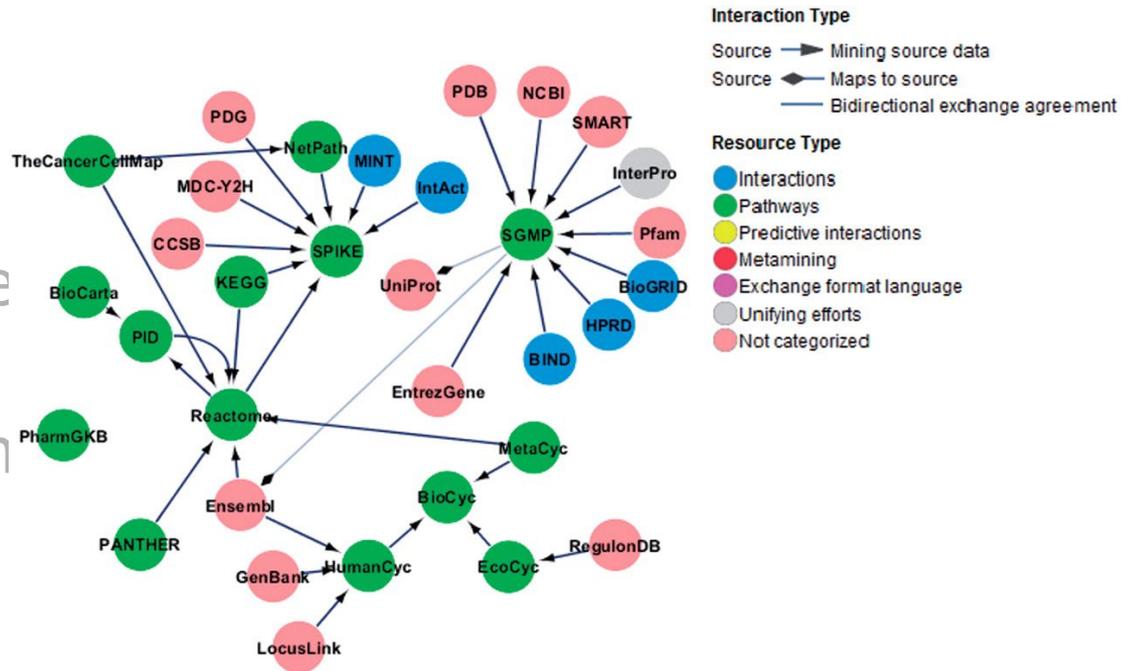
- Manually curated “source” DBs
- DBs integrating others and HT data sets
- Predicted interactions (yellow)



PPI Databases

- There are ~700 DBs related to PPI and pathways
 - See <http://www.pathguide.org>

- Manually curated “source” DBs
- DBs integrating other and HT data sets
- Predicted interaction
- Pathway DBs (green)



No standards [KP10]

- Different definitions of a PPI
 - Only binary, physical interactions?
 - Inclusion of complexes or pathways?
 - Transient interactions? **Functional associations?**
- 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 [KP10]

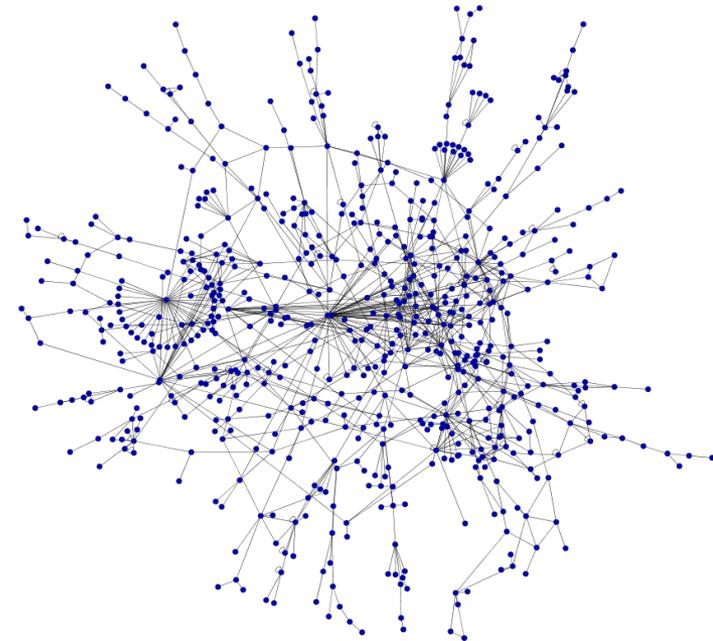
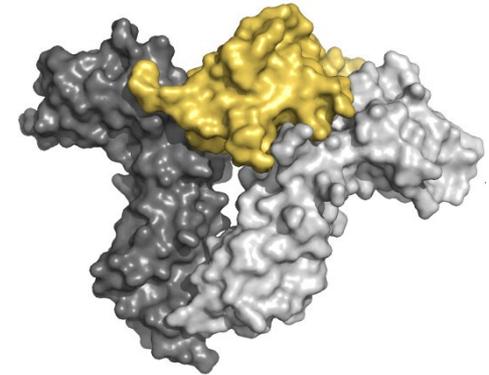
Database	Species	Proteins	Interactions
IntAct	No restriction	53.276	271.764
BioGrid	No restriction	30.712	131.638
DIP	No restriction	23.201	71.276
MINT	No restriction	31.797	90.505
HPRD	Human only	30.047	39.194
MMPPPI	Mammals		
.			
STRING	No restriction (630)	2.590.259	
UniHI	Human only		
OPID	Human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

Experimentally
verified

Experimentally
verified and / or
predicted

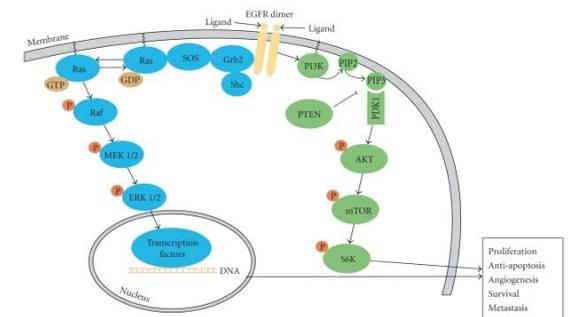
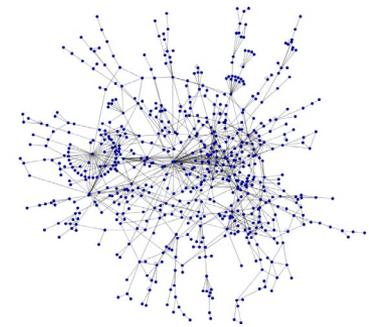
This Lecture

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



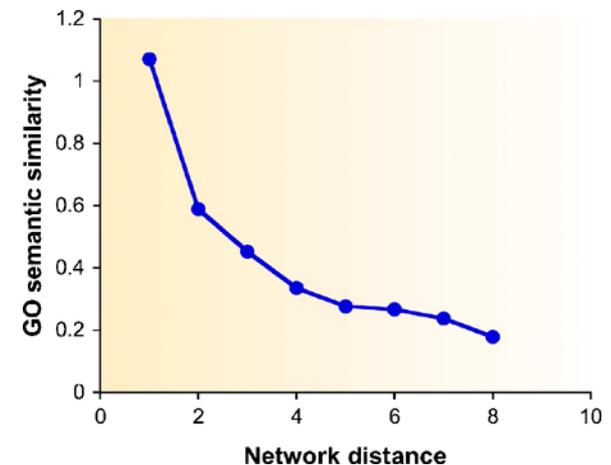
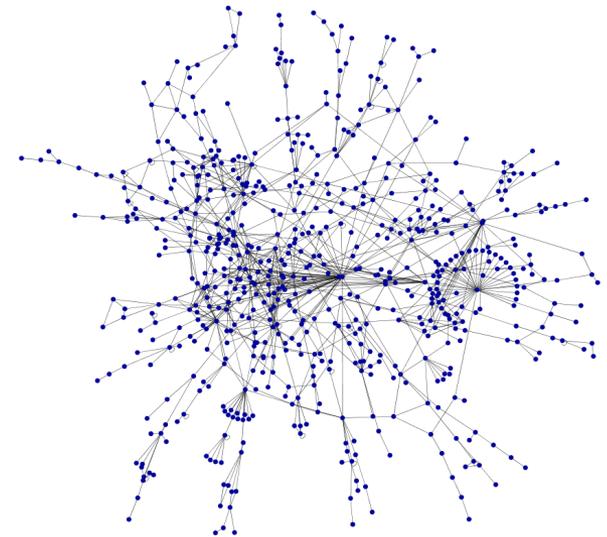
PPI Networks

- **PPI Network**: A set of PPI
- Every PPI network can be considered as a **graph**
 - Proteins: Nodes
 - PPI: Edges
- PPI networks differ vastly in size
 - Single pathway: 5-50 proteins
 - Human PPI: ~400.000 protein (isoforms)
 - Density: 2-8 interactions per protein
 - Very rough estimation
- Other **types of molecules** are ignored (cofactor, metabolites, hormones, ...)
 - Exception: Metabolic networks



Some Fundamental Observations

- Proteins that are **close in the PPI-network** of a cell **share function** more frequently than distant proteins
- **Central proteins** are vital
- Complexes form **dense subgraphs**
- **Functional modules** are subgraphs
- **Maybe**, certain small subgraphs are found significantly more often than expected by chance
 - In a random graph of density X – how often do you expect a triangle?
 - But: Are biological networks random networks?

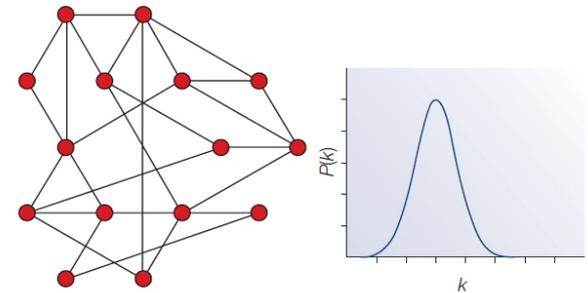


Degree distribution

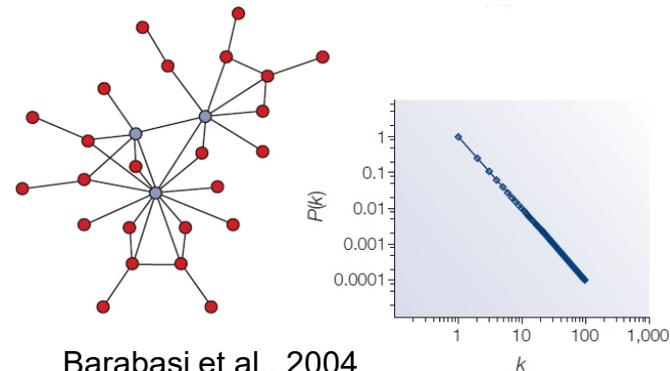
- **Degree distribution** $P(k)$: relative frequency of nodes with degree k
- Used to define different classes of networks
- Common distributions

- Binomial
$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

- **Random network** with n nodes and edge probability p
- For large n : Poisson

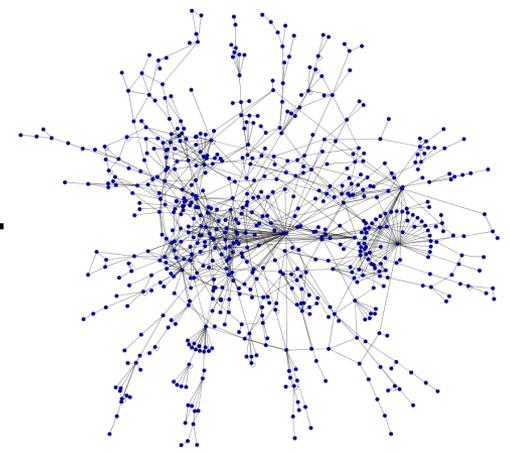


- Power-law
$$P(k) \sim k^{-\gamma}$$
- **Scale-free network** with free parameter γ



Barabasi et al., 2004

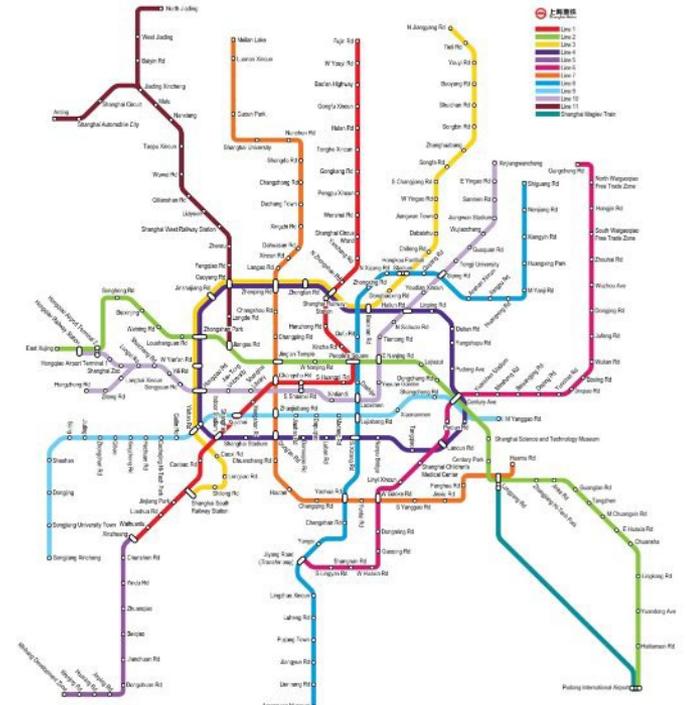
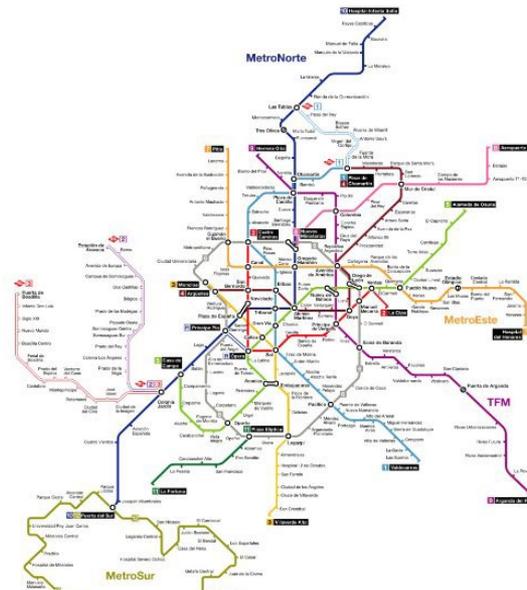
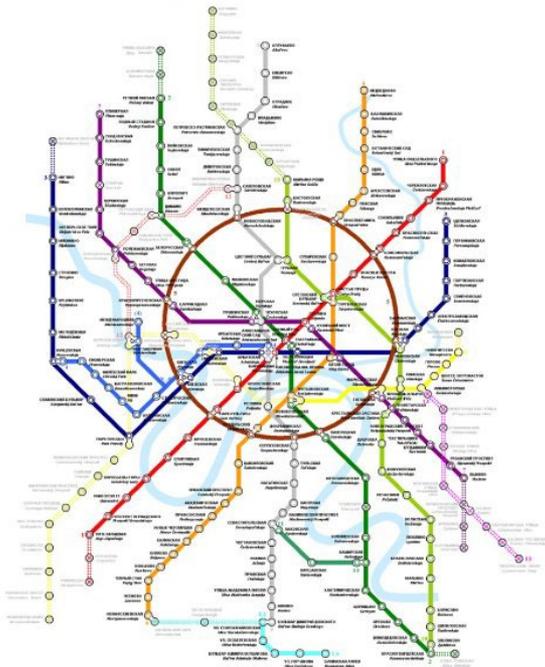
Scale-free Networks



- Unbiased biological networks are (presumably) **scale-free**
 - Unbiased: No human selection of nodes/edges
 - Few nodes are highly connected (**hubs**)
 - Most nodes have very few connections
- **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

More Scale-Free Networks

Also true for many other graphs: electricity networks, **public transport**, social networks, ...



Types of Biological Networks

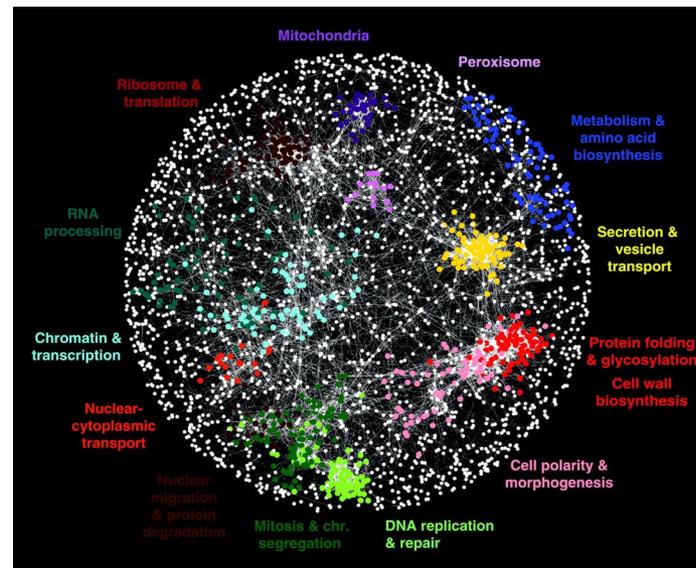
- **Regulatory networks:** How genes / transcription factors influence the **expression of each other**
 - TF regulate expression of genes and of other TFs
 - TFBS, binding regions, enhancers, miRNA, chromatin structure ...
 - Edges semantics: activate / inhibit / regulate
- **Signal networks:** Molecular reaction to external stimulus
 - Transient interactions including small molecules
 - Temporal dimension important (fast)
- **Metabolic networks**
 - Stoecheometric networks: Quantitative flow of atoms in reactions
- **Kinetic networks:** Include energy consumption
- **Very fuzzy:** Protein-protein interaction networks

Systems Biology / Systems Medicine

- **Systems Biology**: Research in the (qualitative, quantitative) nature of biological networks
 - Most research in e.coli, yeast etc.: Easy to manipulate genetically; no ethical issues; producing in masses; fast reproduction cycle; large-scale quantitative measurements ...
- **Systems Medicine**: Applications of Systems Biology in Medicine
 - Much more difficult to study, but higher reward for society
- Vision: **Simulation** instead of experimentation

Modular network organization

- Function is carried out by **modules / pathways / complexes**
 - Sets of proteins interacting in a cell to achieve a certain function
- Then, function should be reflected in a **modular network structure**



Don't be
fooled by
layout

**Modules
must be
dense, not
close**

Costanzo et al., Nature, 2010

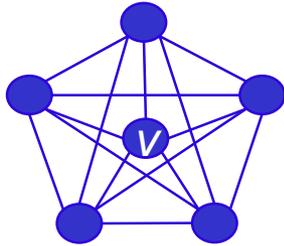
Clustering Coefficient

- Modules are densely connected groups of nodes (clusters)
 - Densely: More dense than expected by chance / than the others
- **Cluster coefficient** C reflects **network modularity** by measuring tendency of nodes to cluster ('triangle density')

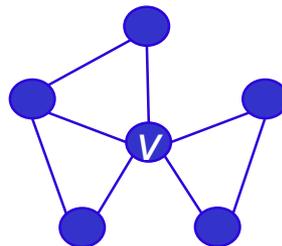
$$C_v = \frac{2E_v}{d_v(d_v - 1)} \longrightarrow C = \frac{1}{|V|} \sum_{v \in V} C_v$$

- 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

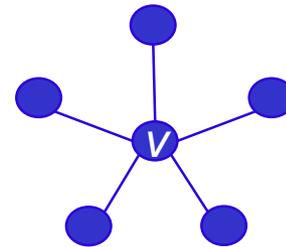
Example



$$C_v = 10/10 = 1$$



$$C_v = 3/10 = 0.3$$



$$C_v = 0/10 = 0$$

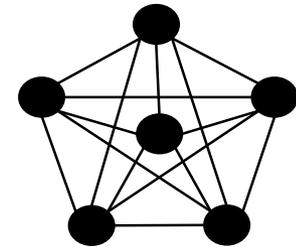
- Cluster coefficient C is a measure for the entire graph
- We also want to find modules, i.e., **regions in the graph** with high cluster coefficient
- A **clique** is a **maximal complete subgraph**, i.e., a maximal set of nodes where every pair is connected by an edge

Finding Modules / Cliques

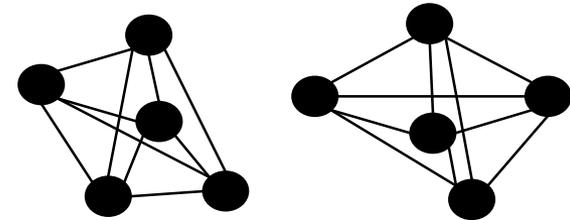
- Finding all **maximal cliques** in a graph is intractable
 - NP-complete

```
build set  $S_2$  of all cliques of size 2
 $i := 2$ ;
repeat
   $i := i+1$ ;
   $S_i := \emptyset$ ;
  for  $j := 1$  to  $|S_{i-1}|$ 
    for  $k := j+1$  to  $|S_{i-1}|$ 
       $T := S_{i-1}[j] \cap S_{i-1}[k]$ ;
      if  $|T|=i-2$  then
         $N := S_{i-1}[j] \cup S_{i-1}[k]$ ;
        if  $N$  is a clique then
           $S_i := S_i \cup N$ ;
        end if;
      end if;
    end for;
  end for;
end for;
until  $|S_i| = 0$ :
```

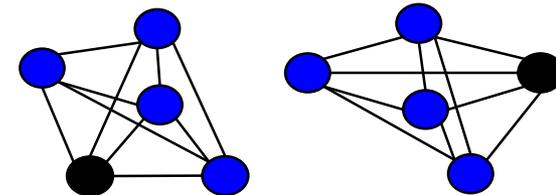
A clique of size 6



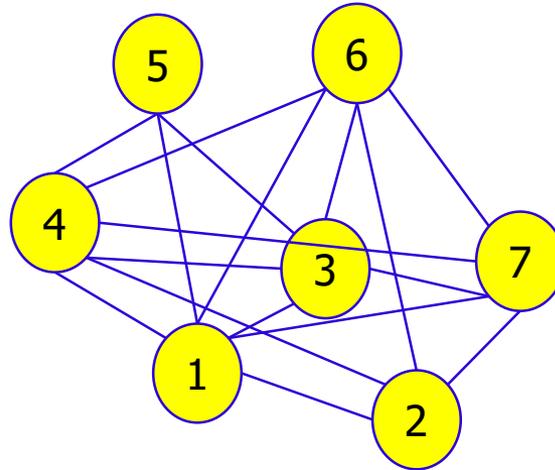
... must contain 6
cliques of size 5



... which must
overlap in $6-2=4$
nodes



Example



- 4-cliques: $(1,3,4,5) - (1,3,4,6) - (1,3,4,7) - \dots$
- 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 exist

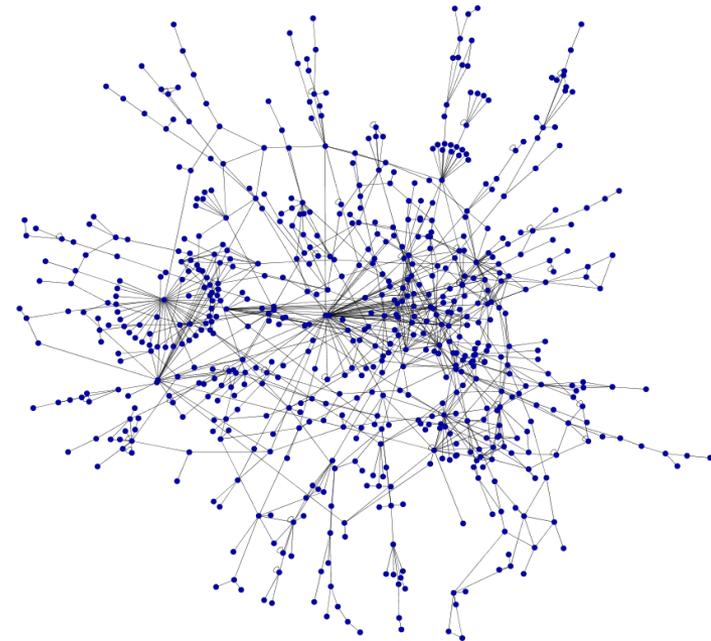
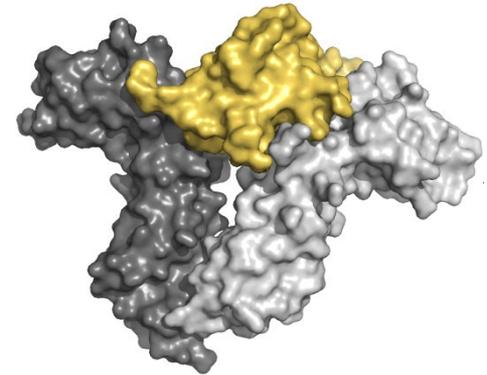
No 5-clique

Quasi-Cliques

- A **quasi-clique** is a “clique” with some missing edges
 - “Some” – parameter of methods (e.g. 20%)
 - Same as subgraphs with high cluster coefficient
 - More **realistic problem** given the high level of noise in real networks
- Finding quasi-cliques is even more complex
- Various heuristics
 - E.g. a good quasi-clique probably contains a smaller clique

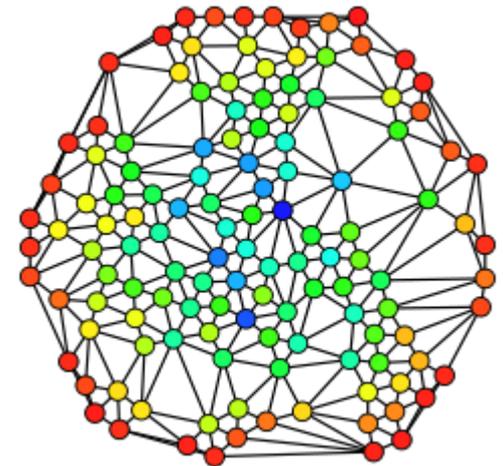
This Lecture

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

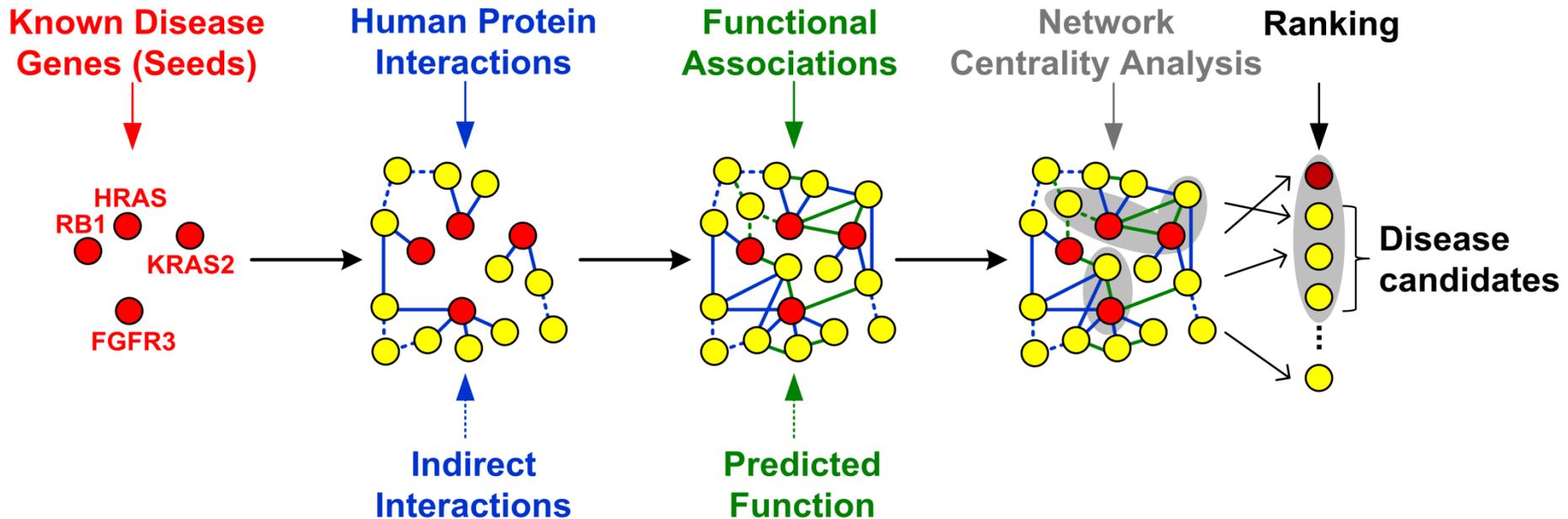


Network centrality

- **Central proteins** exhibit interesting properties
 - Essentiality – knock-out is lethal
 - Much higher **evolutionary conservation**
 - Often associated to human diseases
- Various measures for “centrality” 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
 - ...

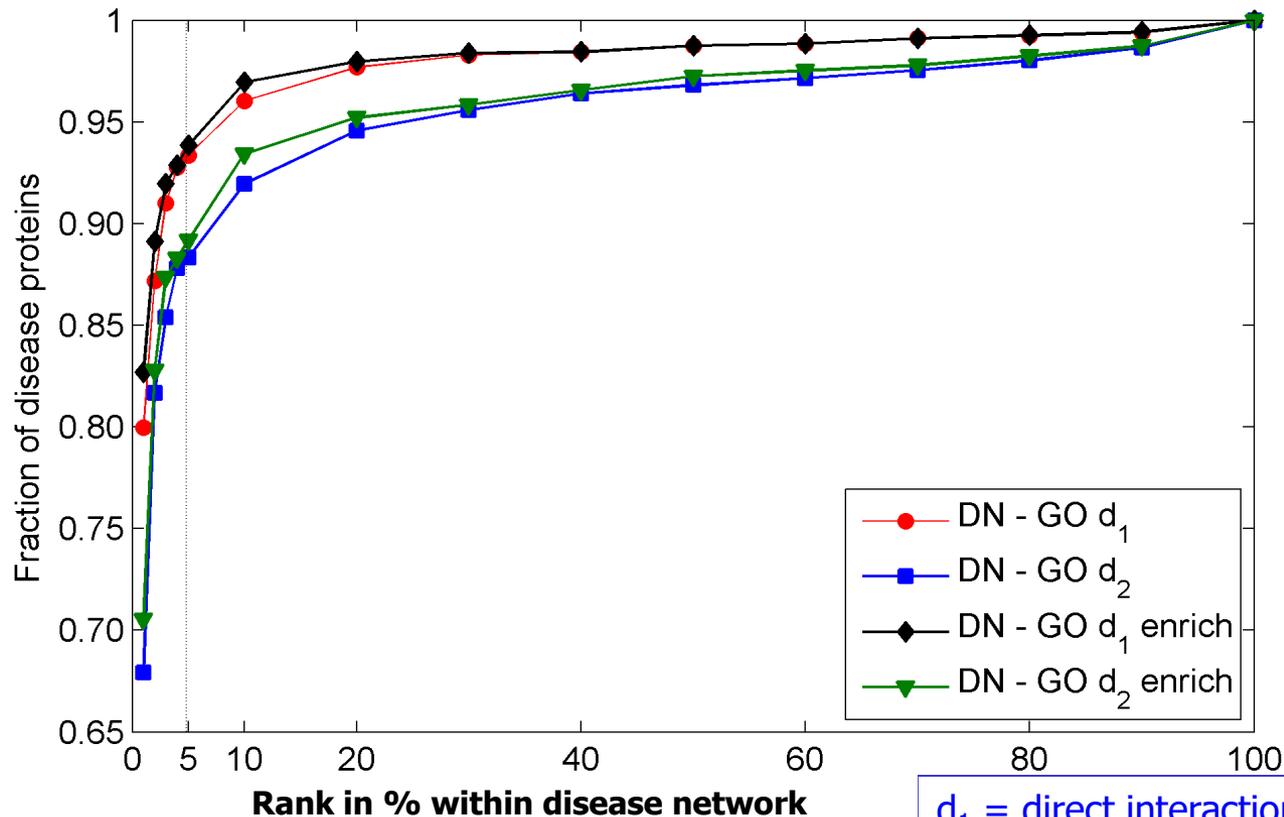


Network-based Disease Gene Ranking



Centrality of Seeds in (OMIM) Disease Networks

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

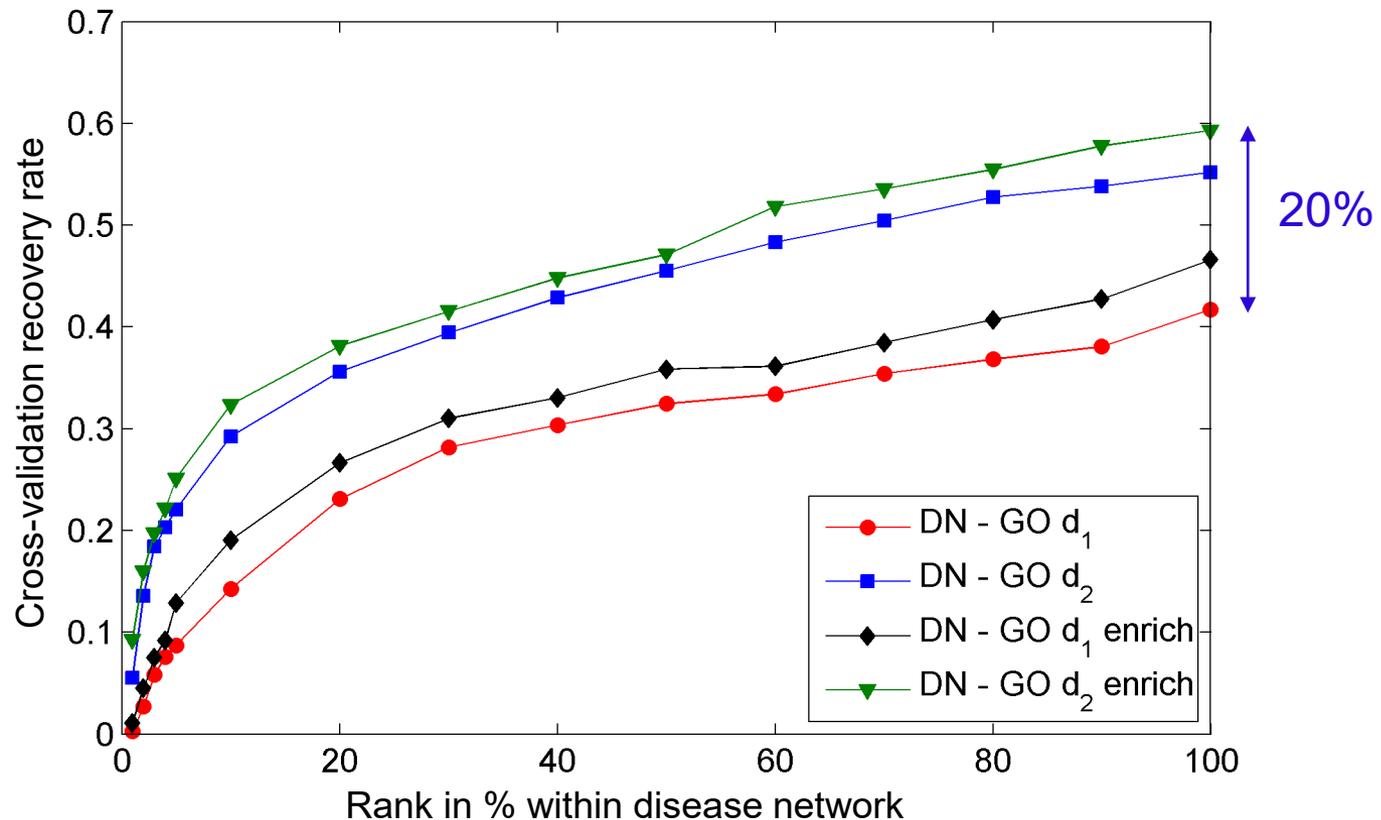


d₁ = direct interactions

d₂ = 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.