Protein-Protein Interaction Network

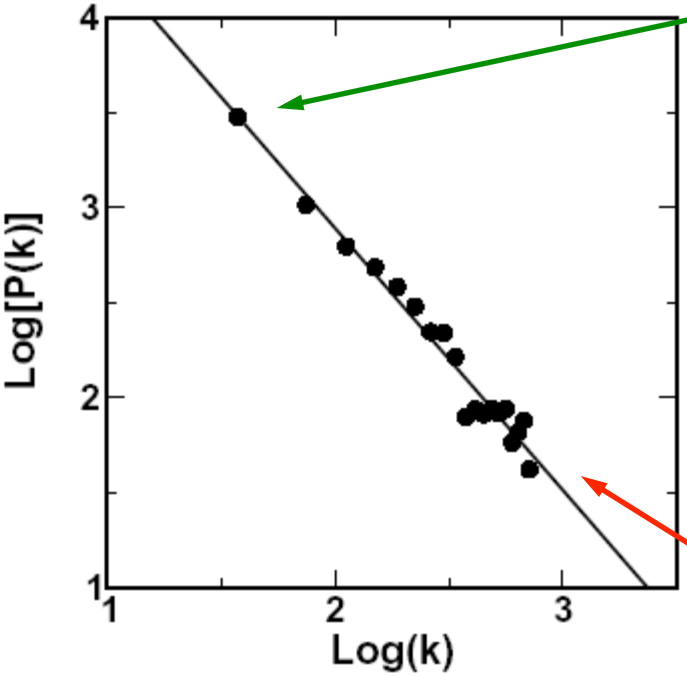
Lecture 4

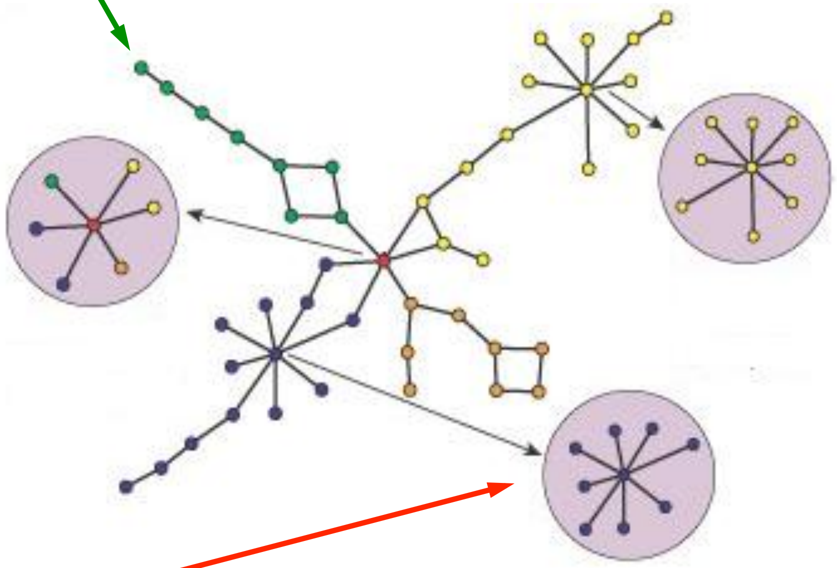
Outline

- Protein-Protein Interaction Model
- How to get PPI
  - Y2H
  - Bioinformatics
- PPI databases
- PPI network properties
- Analysis method and applications
- Integration with other omic data

Scale Free

Island





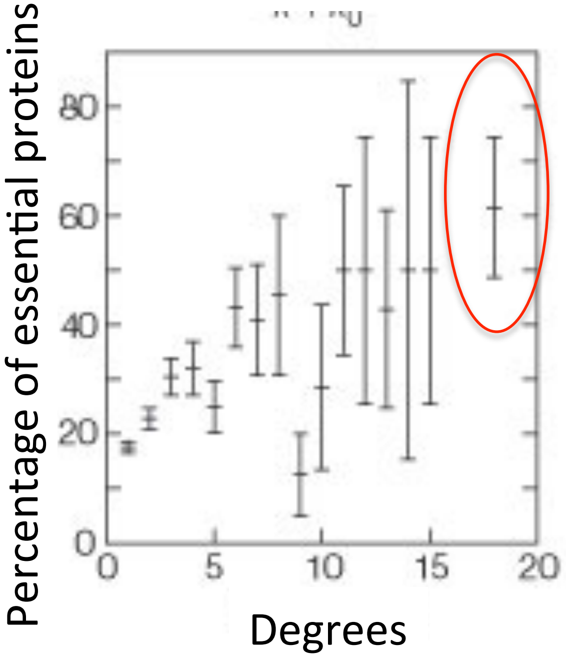
Hub

$$P(k) \sim k^{-\gamma}$$

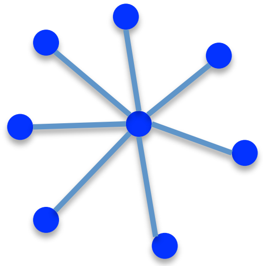
Han *et al.* Nature, 2004

Hub proteins=Essential proteins

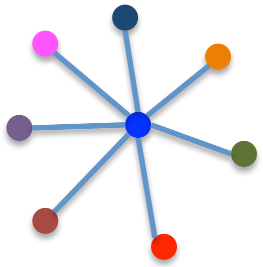
- An essential gene is one that, when knocked out, renders the cell unviable.
- Hub proteins are significantly enriched for essential proteins. (Jeong et al. 2001, Nature 411,41)



Date or Party Hubs



Party Hubs are expressed with their connection partners at same time. They will form a large protein complex. They are more essential. Most of them are house keeping genes.

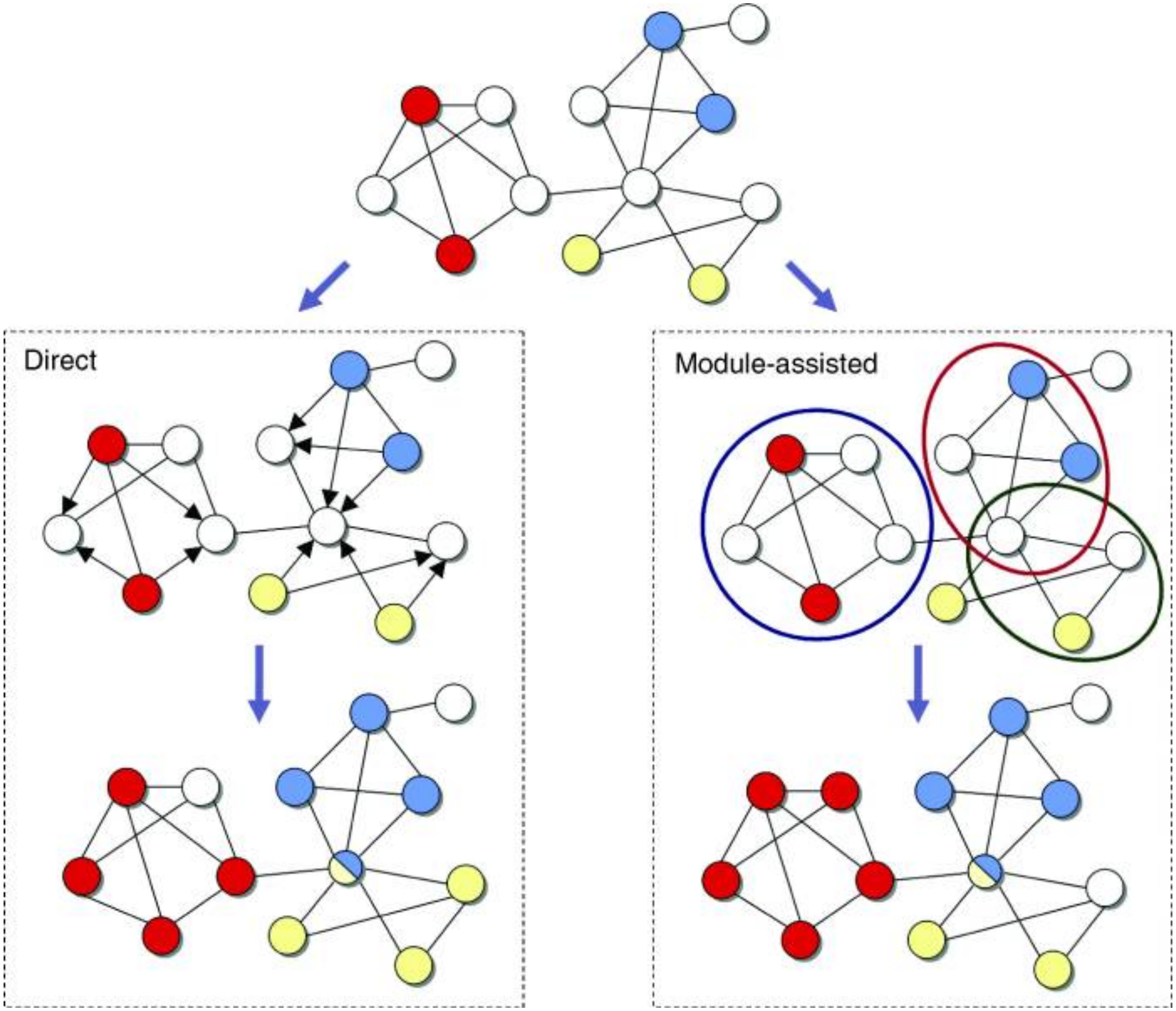


Date Hubs bind with their different connection partners at different time. They have many different binding sites. They have more disorder regions.

Outline

- Protein-Protein Interaction Model
- How to get PPI
  - Y2H
  - Bioinformatics
- PPI databases
- PPI network properties
- Analysis method and applications
- Integration with other omic data

Function prediction

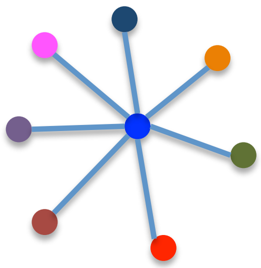


Function prediction

- **Direct Methods**
  - **Neighborhood based Methods**
  - **Graph theory methods**
  - **Probabilistic Methods**
- **Module assisted methods**
  - **General Methods**
  - **Hierarchical clustering based**
  - **Graph clustering methods**
  - **Expansion of complex seeds**

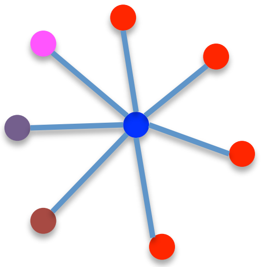
Neighborhood based methods

- Decides the function of a protein from a set of known functions of its neighbors.



Neighborhood based methods (1)

- Predicts for a given protein up to three functions most common among its neighbors.

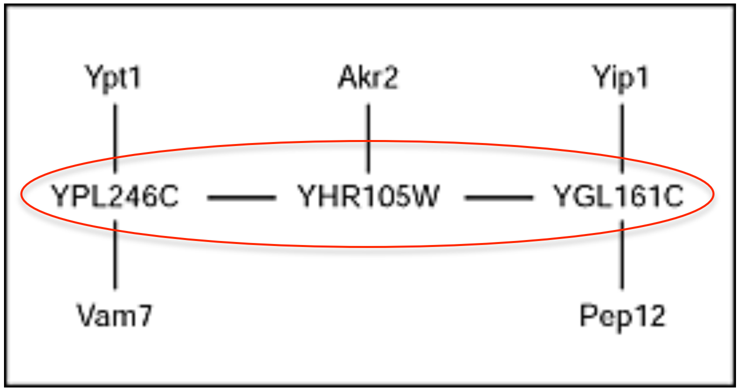


4 Red neighbors, that is larger then the threshold 3

Schwikowski et al (2000) Nature Biotech. 18, 1257.

Prediction of function by direct and indirect protein interactions

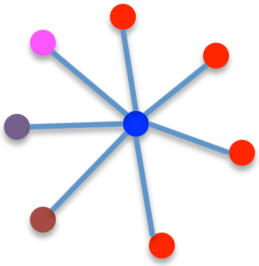
- YHR105W, YPL246C, and YGL161C are proteins of unknown function. Akr2 is a protein involved in endocytosis and therefore suggests a function for YHR105W. This potential function is supported by indirect interactions with Ypt1, Vam7, Yip1, and Pep12, which have been also implicated in vesicular transport and/or membrane fusion.



Schwikowski et al (2000) Nature Biotech. 18, 1257.

Neighborhood based methods (2)

- Examine the neighborhood of a protein and compute scores for a certain function to see if this function is enriched in this neighborhood.
- For a protein, each function *f* is assigned a score $(n_f - e_f)^2/e_f$. If this score is larger than a threshold, the protein has this function.
- n_f is the number of neighbor proteins that have the function *f*
- e_f is the expectation of this number based on the frequency of *f* among the network's proteins.

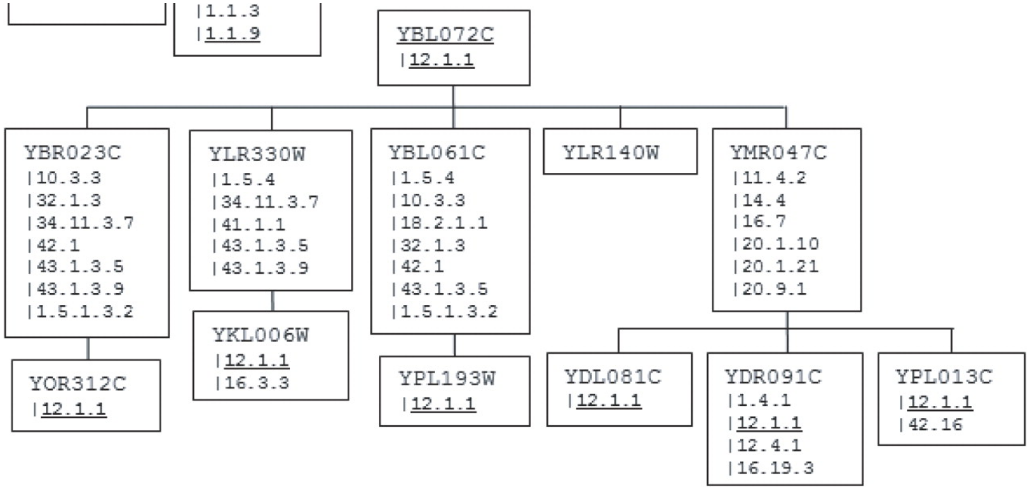


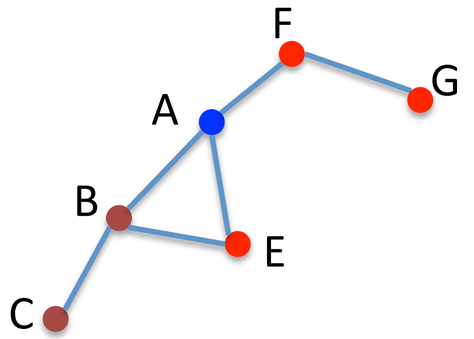
$n_f = 4$ for red function

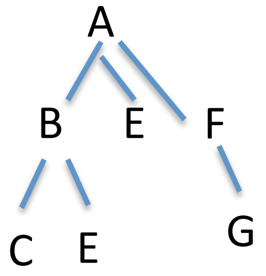
Hishigaki et al (2001) *Yeast 2001;18:523-531.*

Neighborhood based methods (3)

- Considers level 1 and level 2 neighborhood of a target protein.
- Level-1 neighbors that are also Level-2 neighbors are the highest likelihood of sharing functions







Chua et al. (2006) Bioinformatics, 22(13),1623

Function prediction

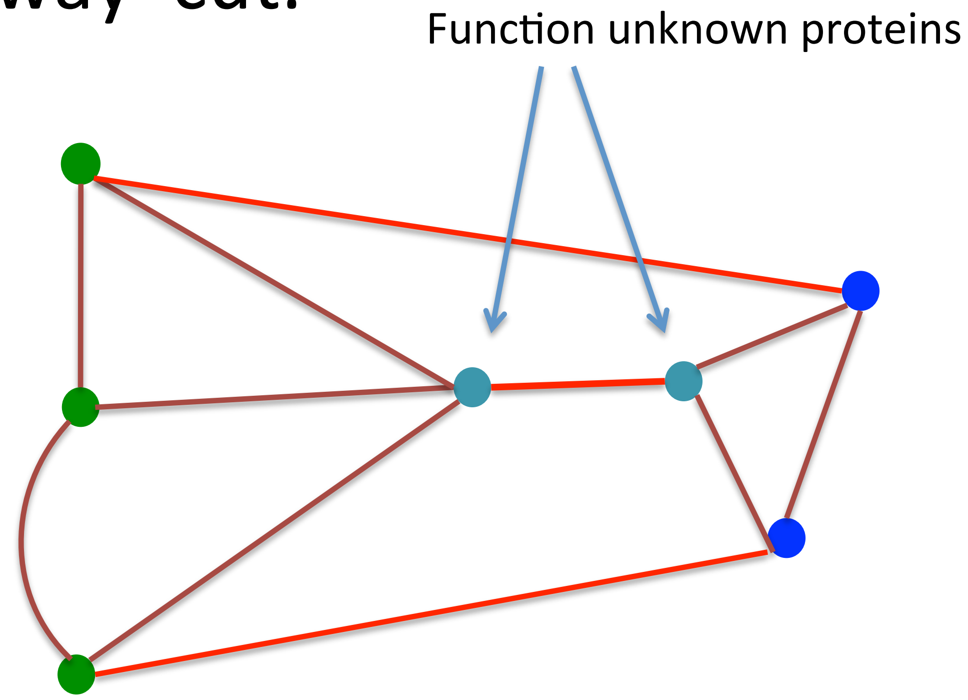
- **Direct Methods**
  - **Neighborhood based Methods**
  - **Graph theory methods**
  - **Probabilistic Methods**
- **Module assisted methods**
  - **General Methods**
  - **Hierarchical clustering based**
  - **Graph clustering methods**
  - **Expansion of complex seeds**

Graph theory Methods

- In contrast to local, neighborhood counting methods, these approaches are global, and take into account the global topology of the network.

Graph theory Methods

- Minimum multi-way cut.

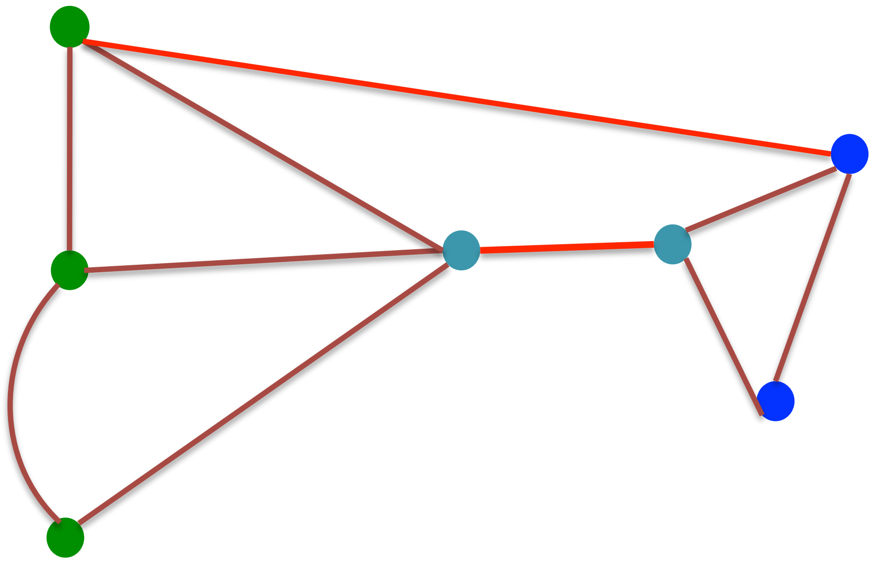


Function unknown proteins

Vazquez et al (2003) Nature Biotech, 21, 697

Graph theory Methods

- Minimum two-way cut.



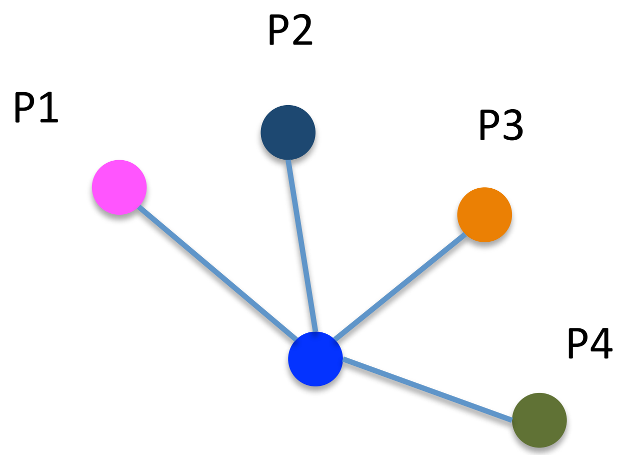
Karaoz et al (2004)

Function prediction

- **Direct Methods**
  - **Neighborhood based Methods**
  - **Graph theory methods**
  - **Probabilistic Methods**
- **Module assisted methods**
  - **General Methods**
  - **Hierarchical clustering based**
  - **Graph clustering methods**
  - **Expansion of complex seeds**

Probabilistic Methods

- Markov Random Field (MRF)
- http://en.wikipedia.org/wiki/Markov_random_field
- Similar to a Bayesian Network, but MRF is an undirected graph.
- Each node will be assigned a probabilistic score, and the probability of the unknown node will be inferred from the other nodes.

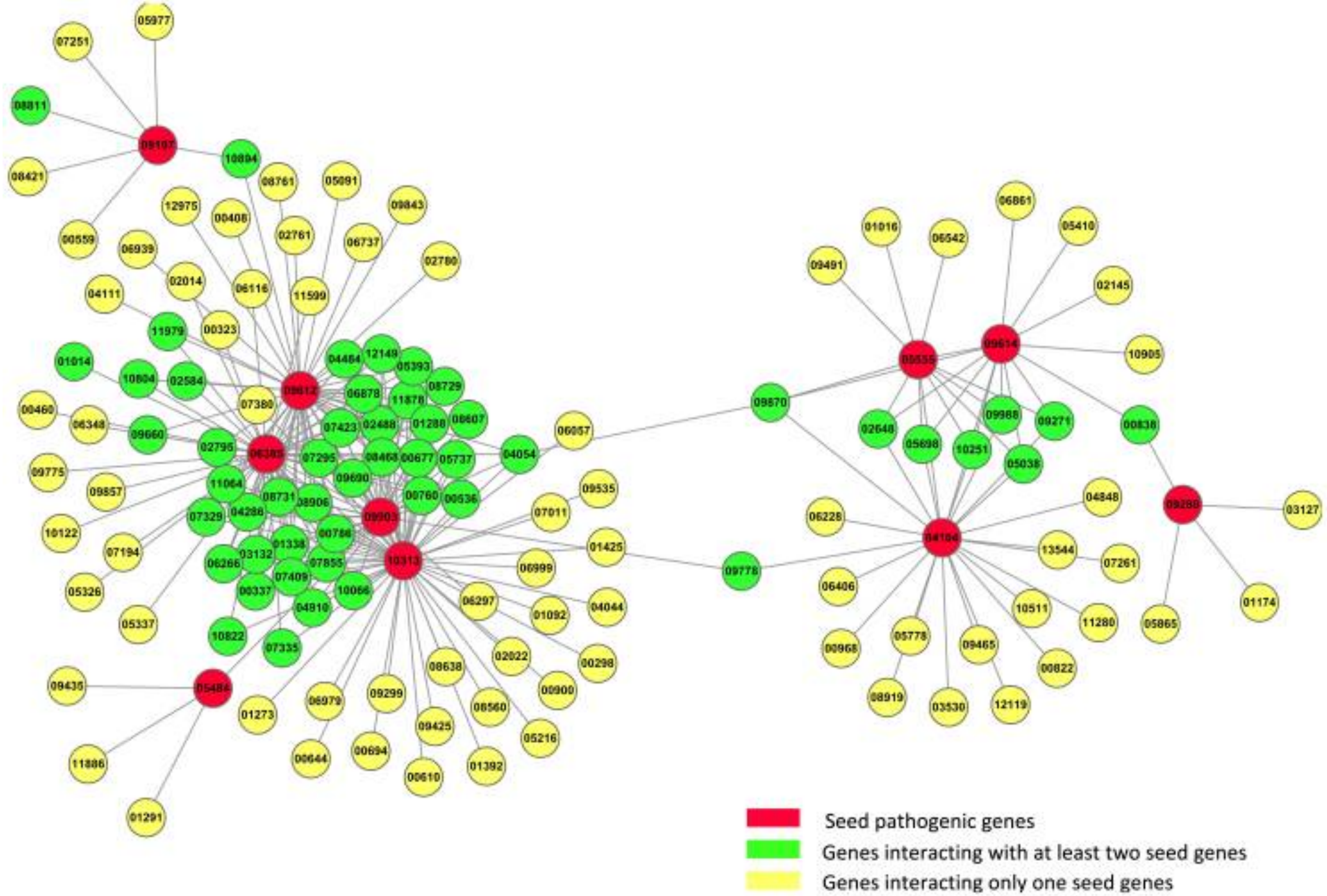


Deng et al. (2003) J. Comp. Biology 10, 6, 947

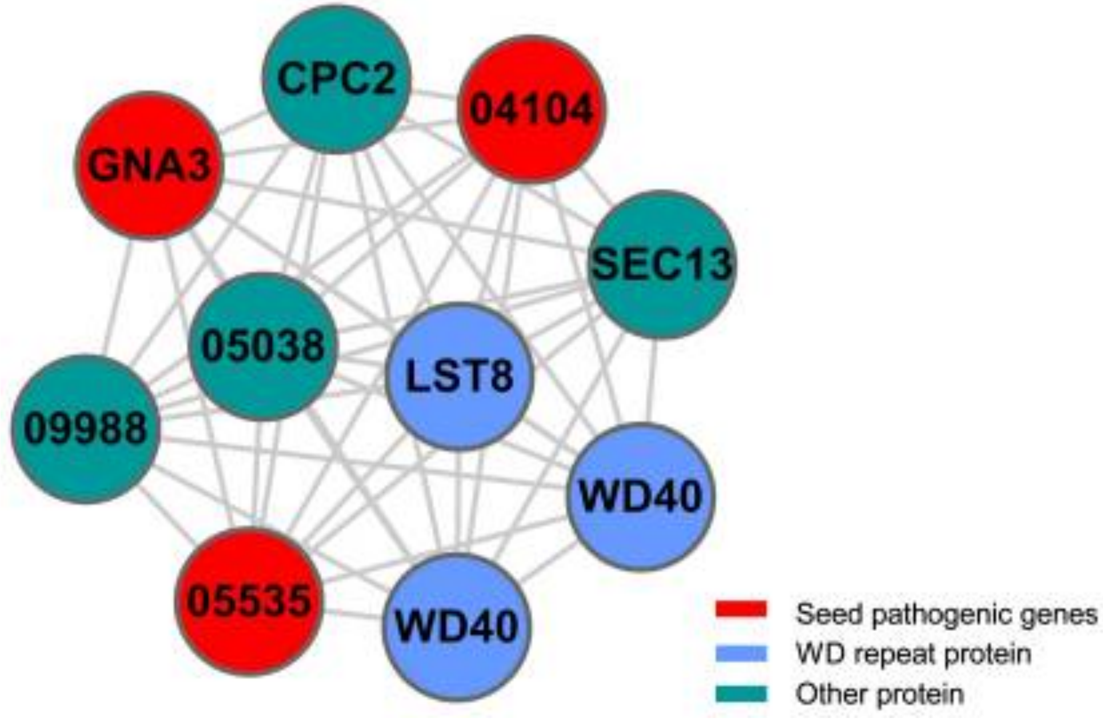
Predict pathogenic genes

- A network approach to predict pathogenic genes for *Fusarium graminearum*. (Liu et al. Plos One, 2010, 5(10))
- *Fusarium graminearum* is the pathogenic agent of Fusarium head blight (FHB), which is a destructive disease on wheat and barley
- Aim: with a network of *Fusarium* and 49 known pathogenic genes, can we predict more pathogenic genes?

Pathogenic gene interaction network



Clique in Pathogenic network

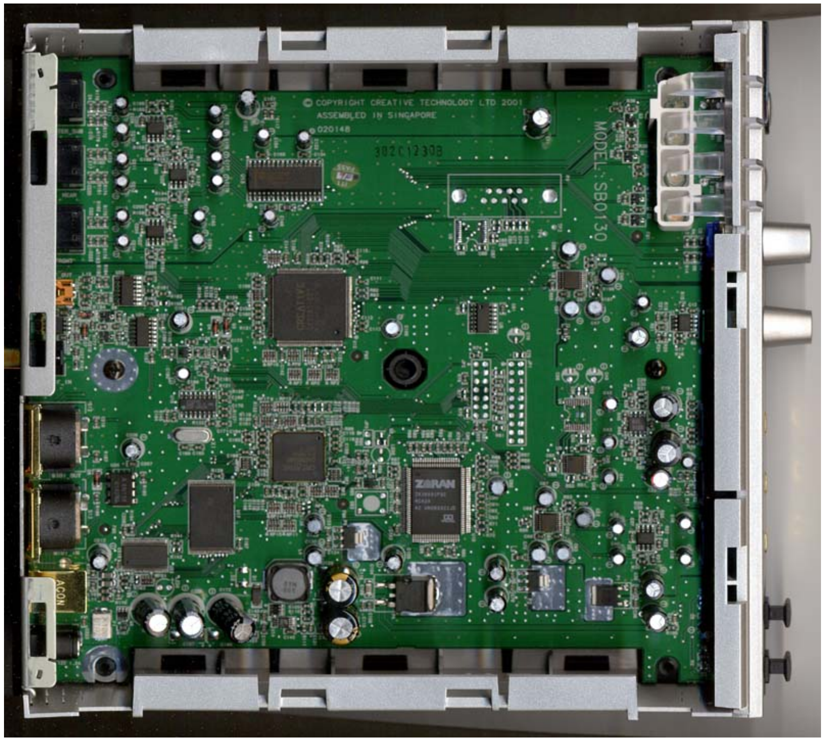


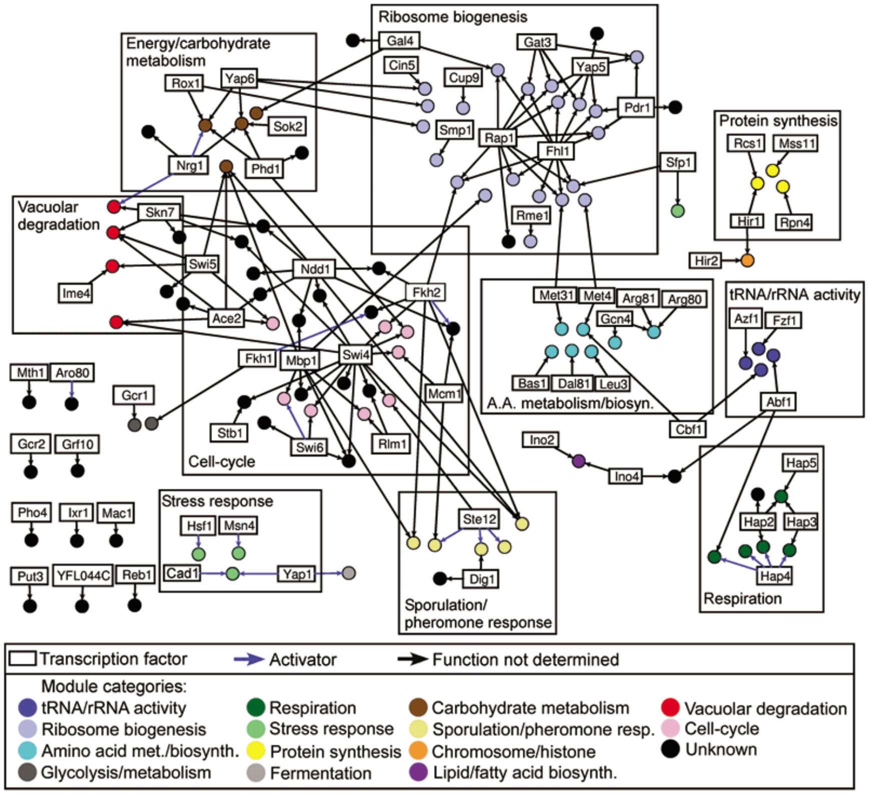
Function prediction

- **Direct Methods**
  - **Neighborhood based Methods**
  - **Graph theory methods**
  - **Probabilistic Methods**
- **Module assisted methods**
  - **General Methods**
  - **Graph clustering methods**
  - **Expansion of complex seeds**

Interaction Network Is Made of Modules

Bar-Joseph et al, *Nature Biotech*. 2003





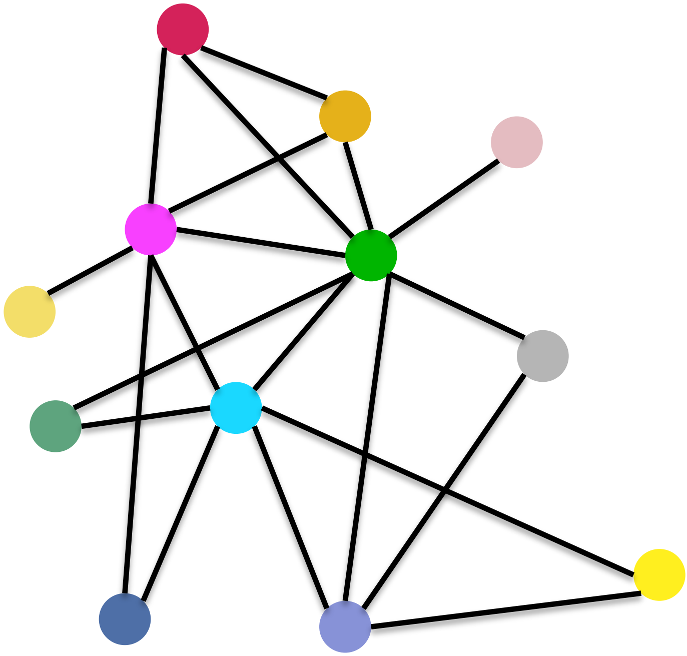
Computer Circuit Boards

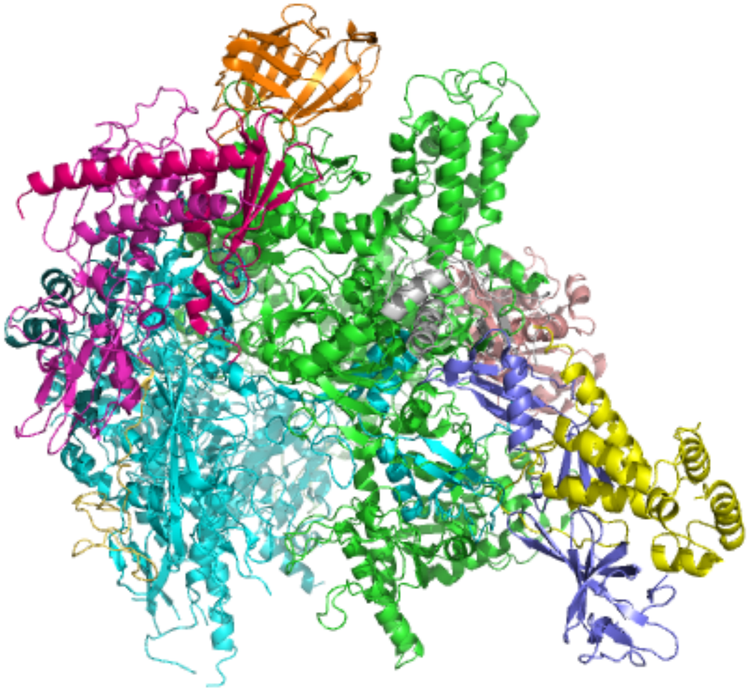
Transcriptional regulatory network

Computational prediction of modules from network

Protein Complex

- 12-subunit RNA Polymerase II





PDB: 2B8K

General Methods

- Find regions that have high clustering coefficient. **MCODE**, Bader and Hogue (2003) BMC Bioinformatics, 4:2.
- Define a Cluster property score. Starting from single nodes, clusters are gradually grown as long as the cluster property of the added nodes and the density of the cluster both exceed a certain threshold. Altaf-Ul-Amin et al (2006), BMC Bioinformatics, 7:207
- Each candidate set of proteins is a assigned a likelihood ratio score that measures its fit to a protein complex model. **NetworkBlast**, Sharan et al (2005), J. Computational Biology, 12(6), 835.

Graph clustering methods

- Use shortest path length between proteins as a distance, and conduct the clustering procedure. Arnau et al (2005) Bioinformatics, 21, 364.
- Superparamagnetic clustering (SPC). Spirin and Mirny (2003) PNAS, 100, 12123.
- highly connected subgraphs (HCS) algorithm. Przulj et al (2004), Bioinformatics, 20, 340
- The restricted neighborhood search clustering (RNSC) algorithm. King et al. (2004), 20, 3013
- The Markov clustering (MCL) algorithm. Enright et al. (2002), Nucleic Acid Research, 30, 1575

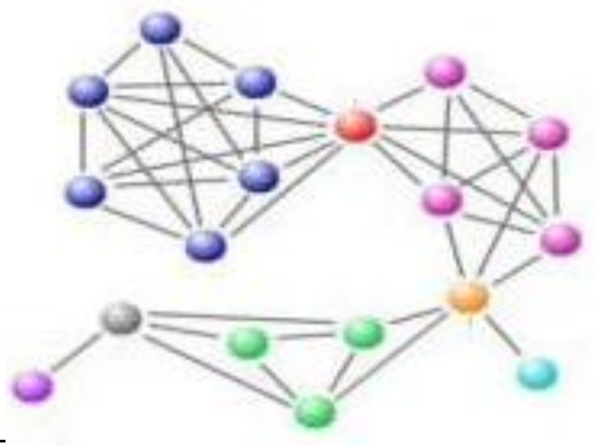
Expansion of complex seeds

- In contrast to finding complexes *de novo* in the protein interaction network, several works attempted prediction of new members for partially known protein complexes.
- SEEDY: constructs complexes by adding proteins to a given seed, as long as the reliability of the most reliable path from a candidate to the seed does not fall below a given threshold. Bader (2003) Bioinformatics, 19, 1869
- Complexpander: start from a particular ‘core' set of proteins and produces a list of candidate proteins, ranked by the probability of membership in the complex. Asthana et al (2004) Genome Research 14, 1170
- For a given “seed”, the algorithm expands it through a breadth-first-search graph traversal. Wu and Hu (2005) IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology 135.

Three examples of PPI applications

- Clique merging to identify functional modules
- Plant signature domain graph
- To predict domain functions with domain sharing network

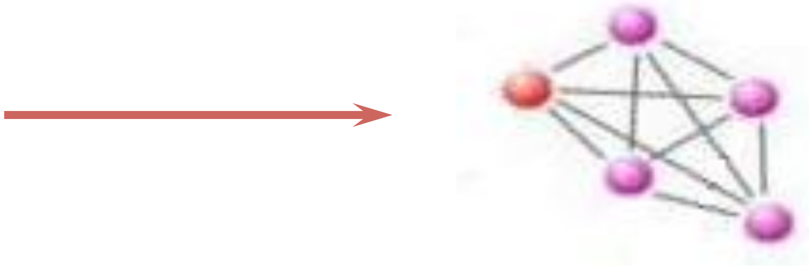
Clique merging method



Connection Density *Q*:

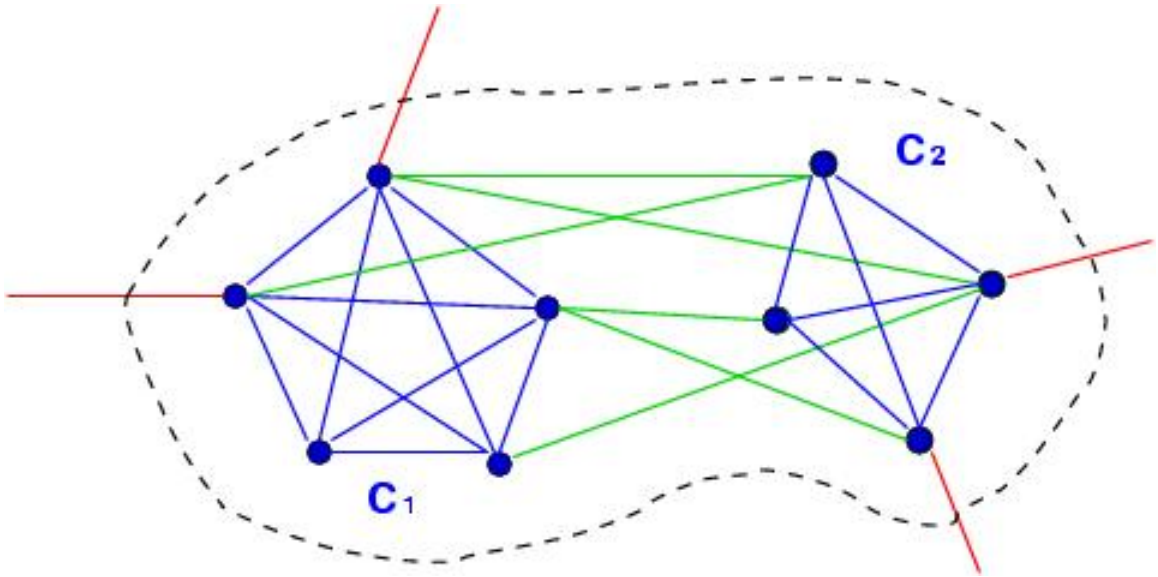
$$Q = \frac{\#\text{of Connections}}{\text{Max.}\#\text{of possible Connections}} = \frac{E}{V(V-1)/2}$$

Clique: Fully connected sub-graph (*Q*=1)



A Module in a Network

- high connection density
- more inner connections than outer connections



- A clique has the highest connection density
- Can we merge cliques to get a module?

Outputs

84 modules with sizes from 4 to 69 proteins.

Are they biological modules?

Proteins within a module

Co-functioning? (shared a common bio-process term annotated in Gene Ontology database)

Co-localized? (shared a common location term annotated in Gene Ontology database)

Co-expressed? (Co-function <-> co-expressed)

Co-functioning? YES

80/84 (95%) modules are enriched with a specific GO process term, significantly (P-value <10^{-3}).

The average homogeneity in function is 82%.

		#		GO		
#	**Size**	**Merges**	homo%	Level	Description	P-value
1	**69**	**287**	38	7	Microtubule-based process	$6x10^{-29}$
2	**34**	**68**	91	5	Cytoplasm organization/biogenesis	$2x10^{-39}$
3	**31**	**50**	77	7	proteolysis	$8x10^{-32}$
4	**30**	**24**	83	6	mRNA metabolism	$6x10^{-33}$
5	**22**	**21**	91	8	Transcription from Pol II promoter	$3x10^{-24}$
6	**21**	**17**	71	7	Protein-nucleus import	$3x10^{-27}$
7	**21**	**18**	71	4	Cell organization/biogenesis	$6x10^{-5}$
8	**20**	**34**	85	8	Nuclear mRNA splicing,via spliceosome	$2x10^{-27}$
9	**20**	**43**	95	5	DNA metabolism	$6x10^{-19}$

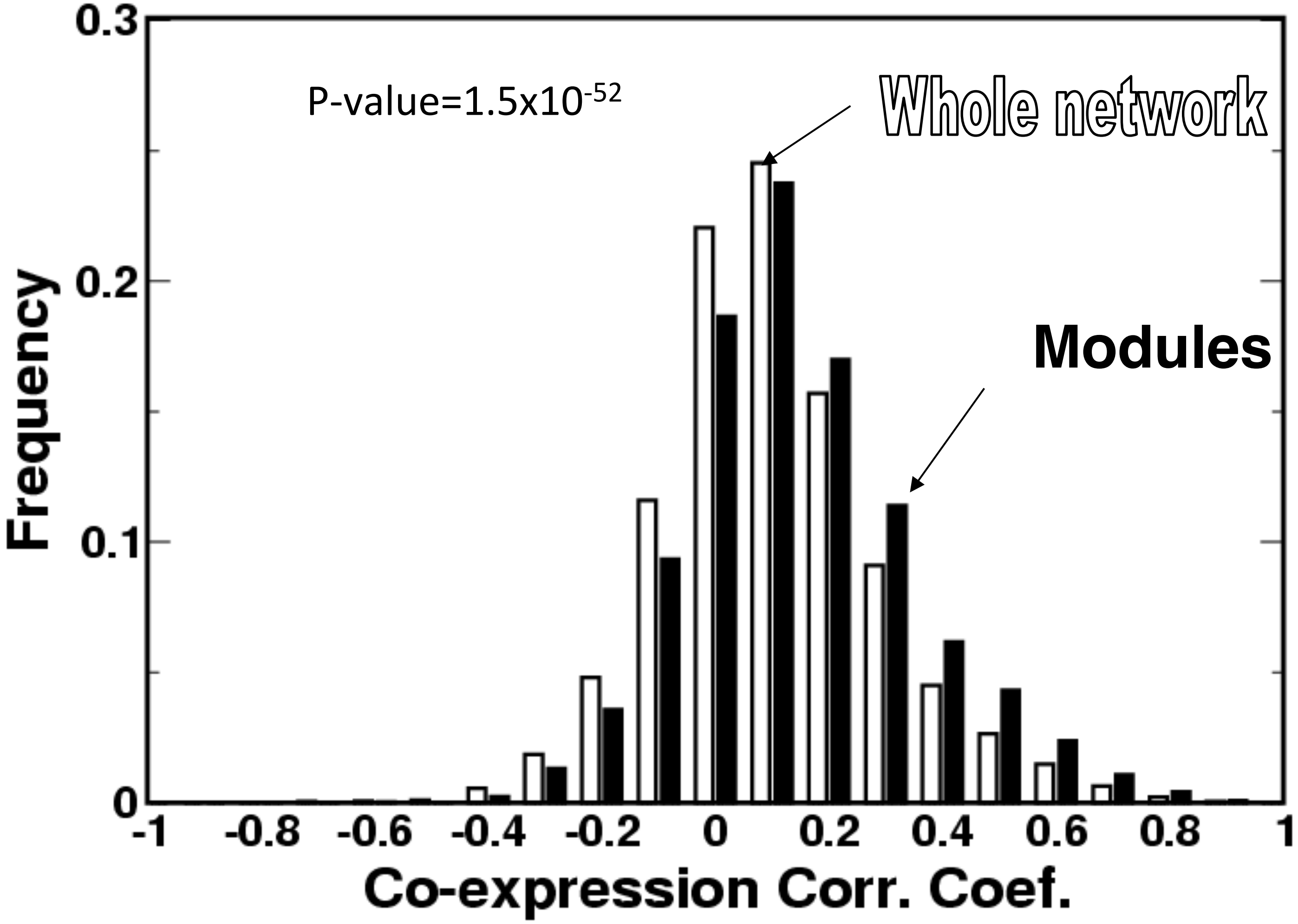
Co-localized? YES

71/84 (85%) modules are enriched with a specific cellular component term, significantly (P-value <10^{-3}).

The average homogeneity in location is 78%.

#	Size	# Merges	homo%	GO Level	Description	P-value
1	**69**	**287**	30	4	Microtubule cytoskeleton	$5x10^{-21}$
2	**34**	**68**	85	4	Nucleolus	$3x10^{-33}$
3	**31**	**50**	81	3	Proteasome complex (sensu Eukaryota)	$3x10^{-50}$
4	**30**	**24**	60	5	U4/U6xU5 tri-snRNP complex	$4x10^{-35}$
5	**22**	**21**	82	6	Mediator complex	$4x10^{-46}$
6	**21**	**17**	67	5	Nuclear core	$2x10^{-24}$
7	**21**	**18**	19	4	Golgi apparatus	$3x10^{-2}$
8	**20**	**34**	70	5	snRNP U1	$1x10^{-33}$
9	**20**	**43**	95	3	Nucleus	$5x10^{-9}$

Co-expressed? YES

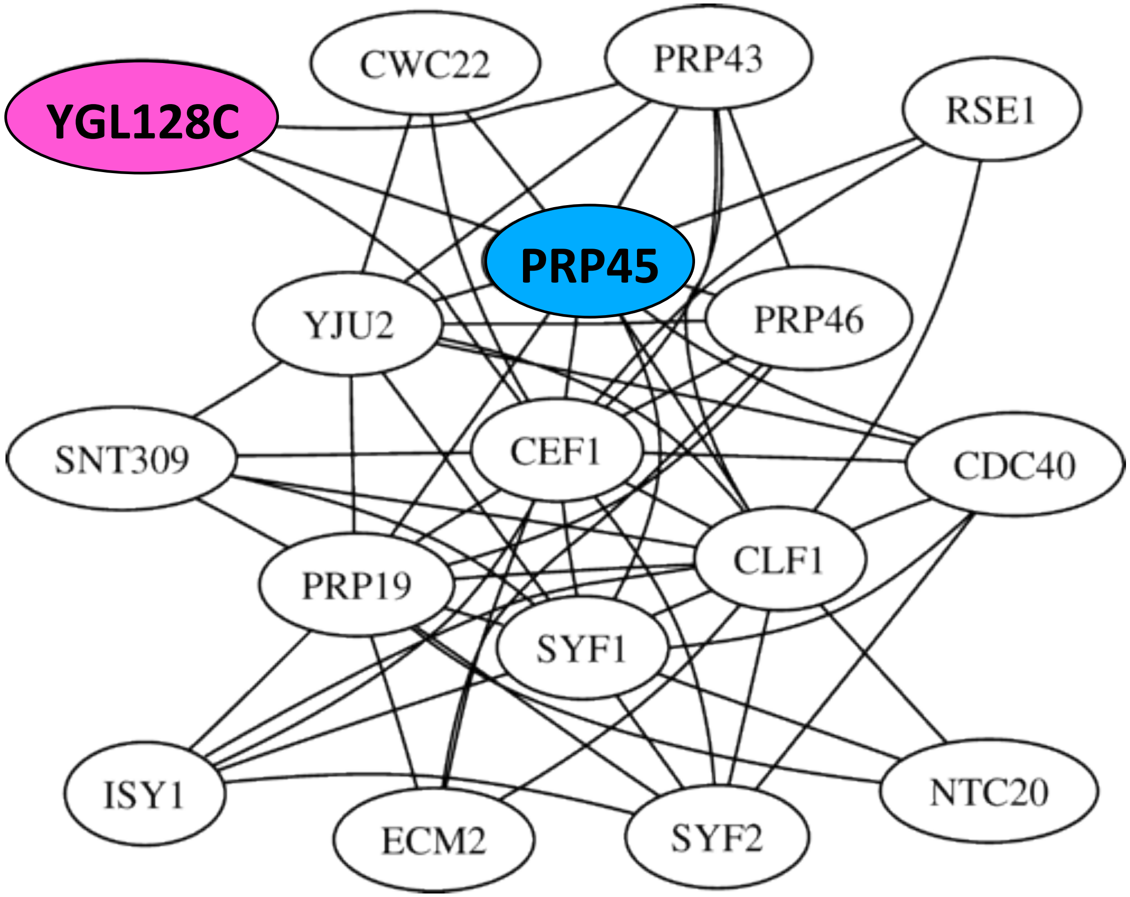


Rosetta Compendium Micro-Array Data. Hughes *et al. Cell* 2000

Module #15: Nuclear mRNA Splicing

- *PRP45* is also linked to this bio-process. (Albers *et al.* RNA 2003)

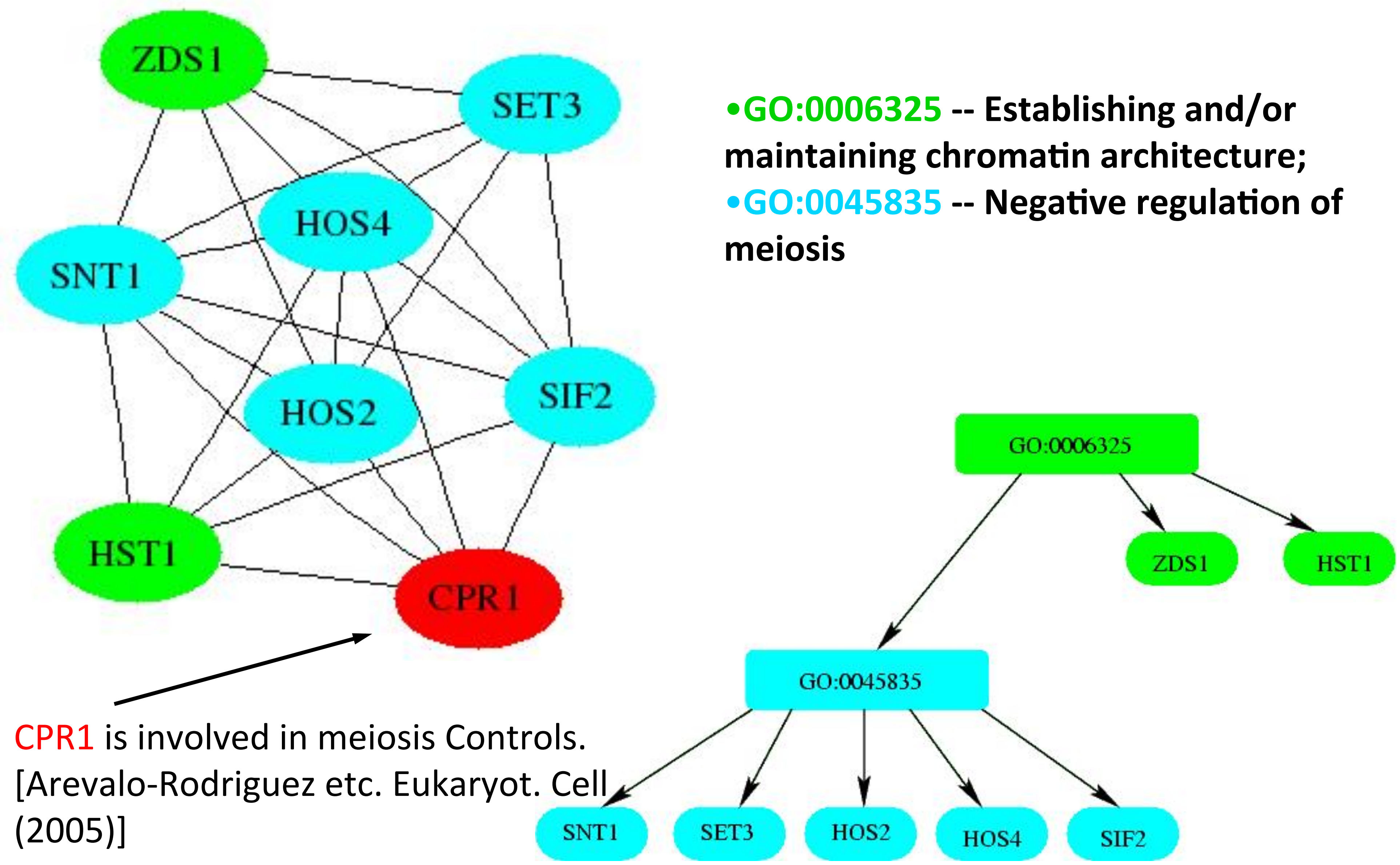
- YGL128C is an unknown gene, and hence it may be annotated.



Module #37:

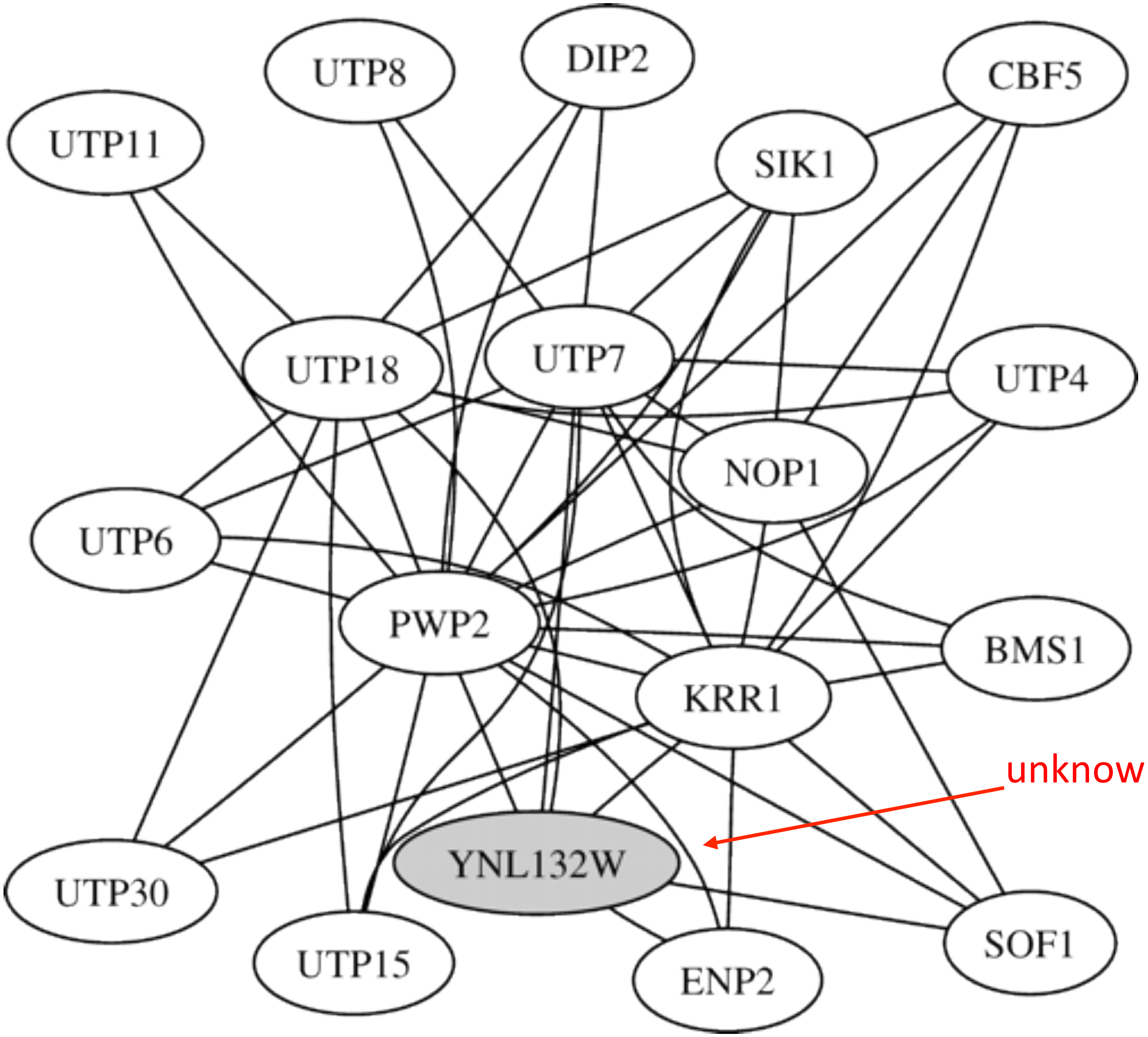
•GO:0006325 -- **Establishing and/or maintaining chromatin architecture;**

•GO:0045835 -- **Negative regulation of meiosis**



CPR1 is involved in meiosis Controls. [Arevalo-Rodriguez etc. Eukaryot. Cell (2005)]

Module #13: rRNA processing

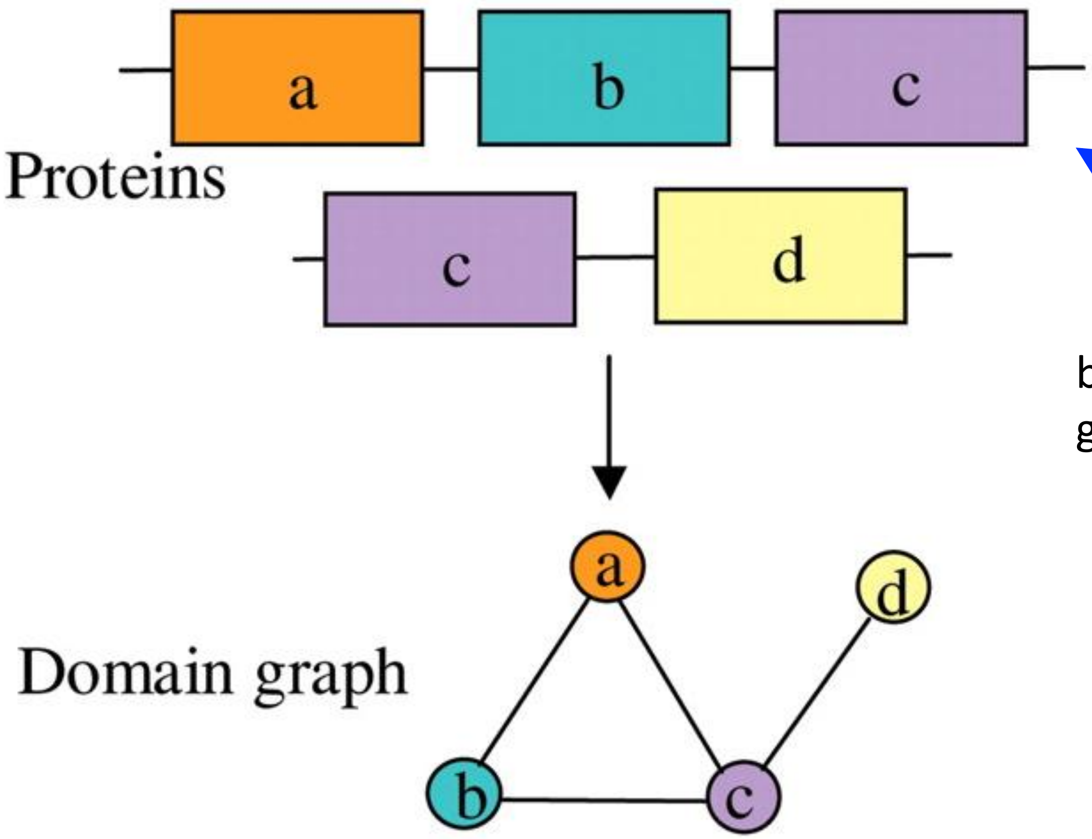


unknown gene

Three examples

- Clique merging
- Plant signature domain graph
- To predict domain functions with domain sharing network

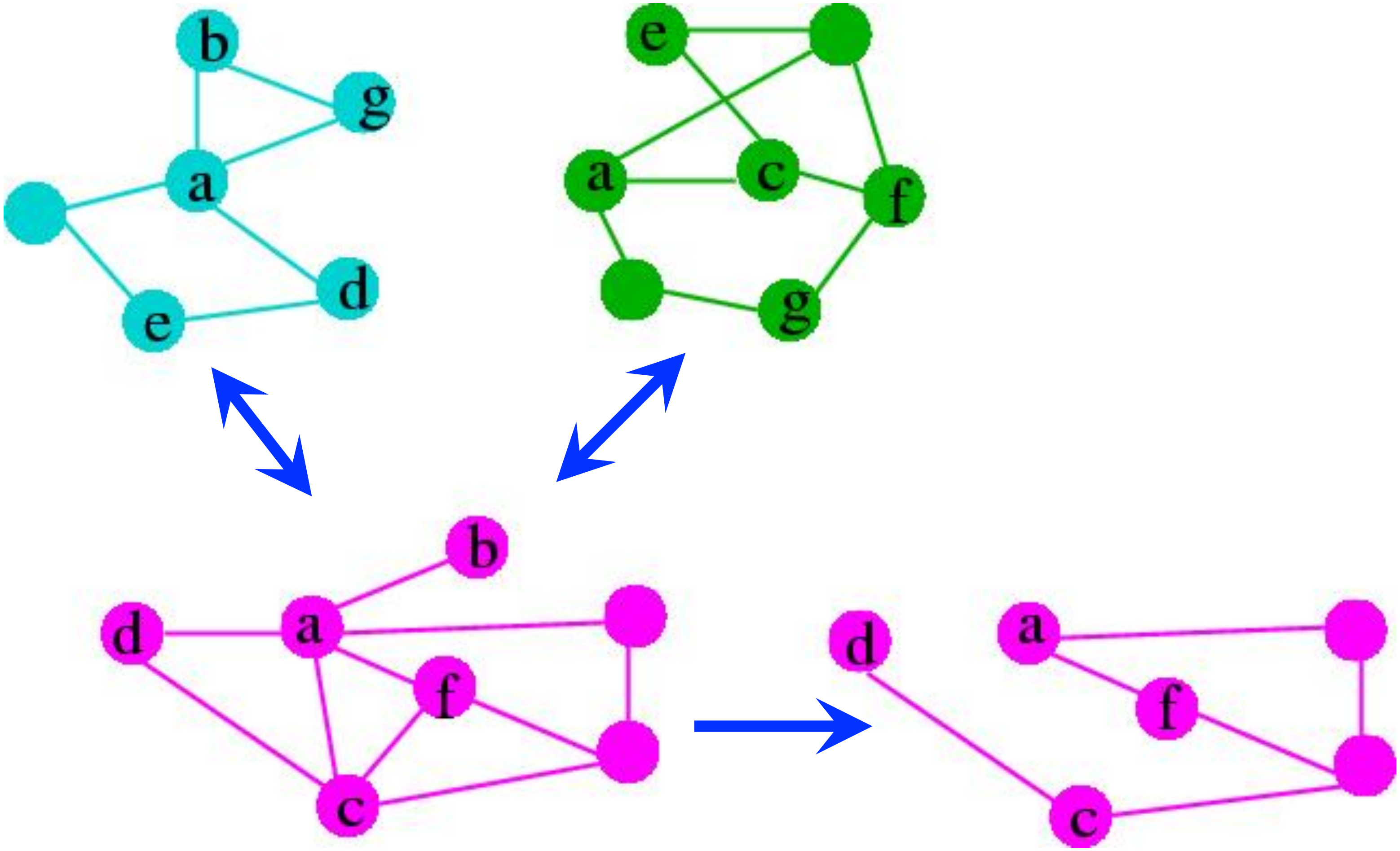
Domain Graph



by pfam domain index with whole genome sequence analysis

Ye & Godzik, *Genome Res.* 2004

Signature Domain Graph



signature domain graph

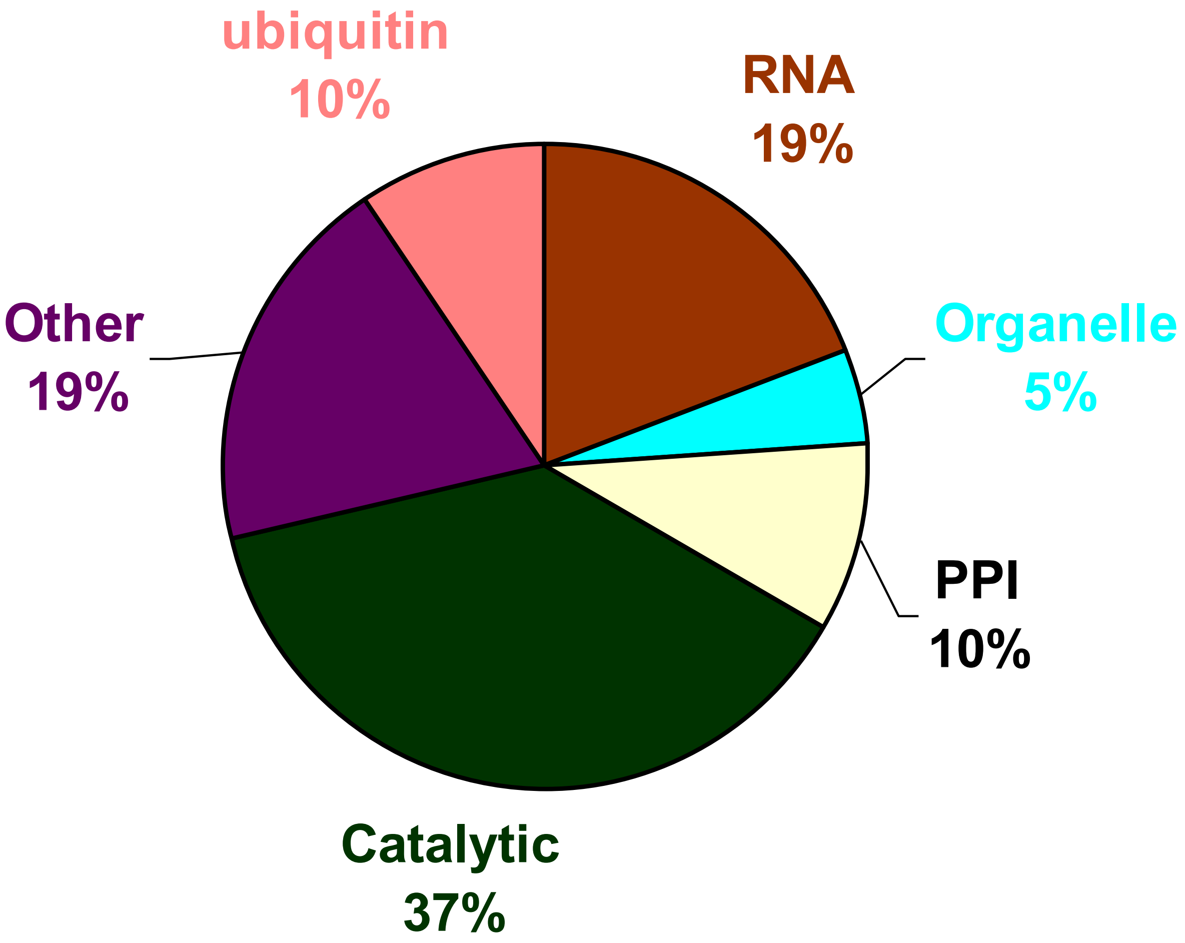
Application: *Arabidopsis thaliana*

- *Arabidopsis thaliana* (a model plant): 2454 domain (vertices), 1277 domain combination (edges)

- Evolution analysis: domain graph comparison against 10 eukaryuotic, 30 bactorial, 16 archaeal proteomes.

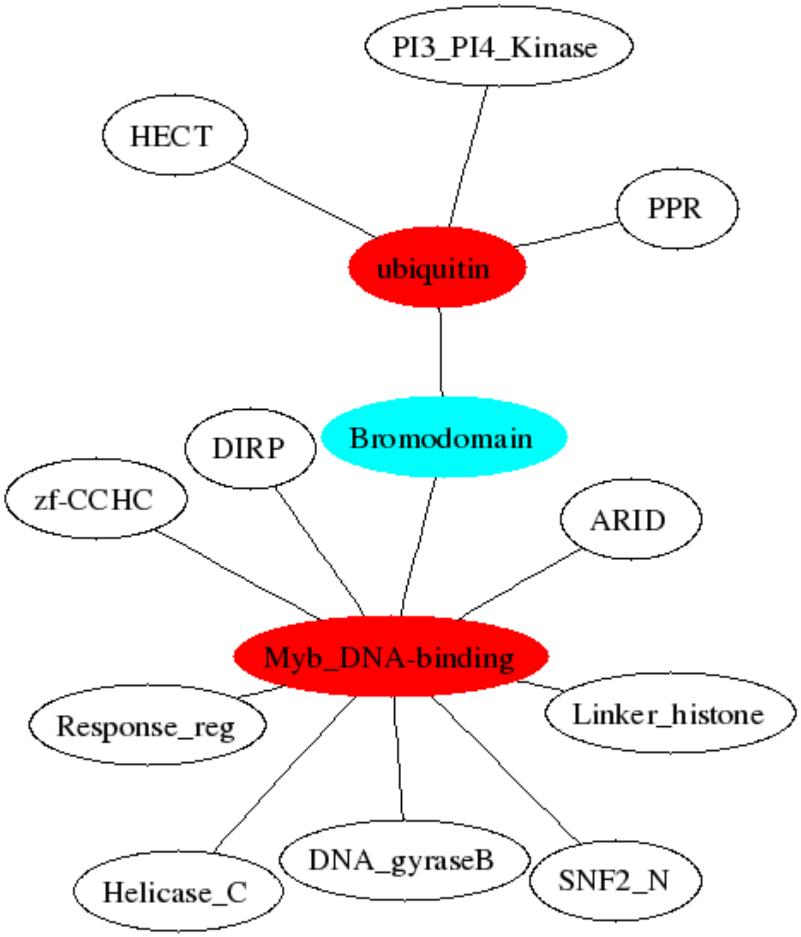
- Signature domain graph analysis methods:
  1. Link neighbor analysis
  2. Shortest path analysis

Link Neighbor Analysis: PPR

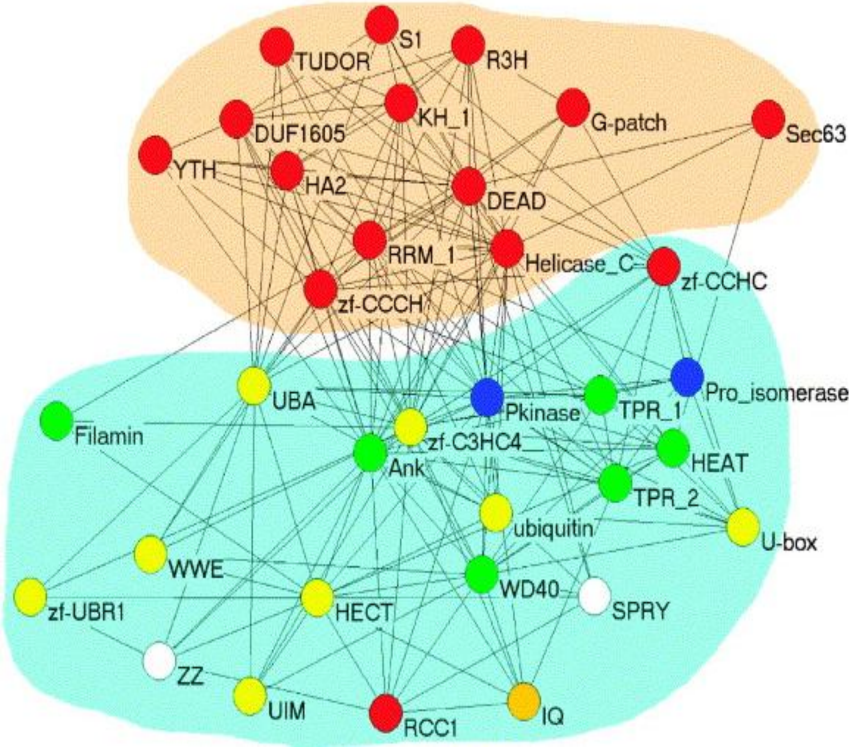


PPR might be involved in the **catalytic process** and **RNA metabolism.**

Shortest Path Analysis



Lucas *et al.* J. Mol. Biol, 2006

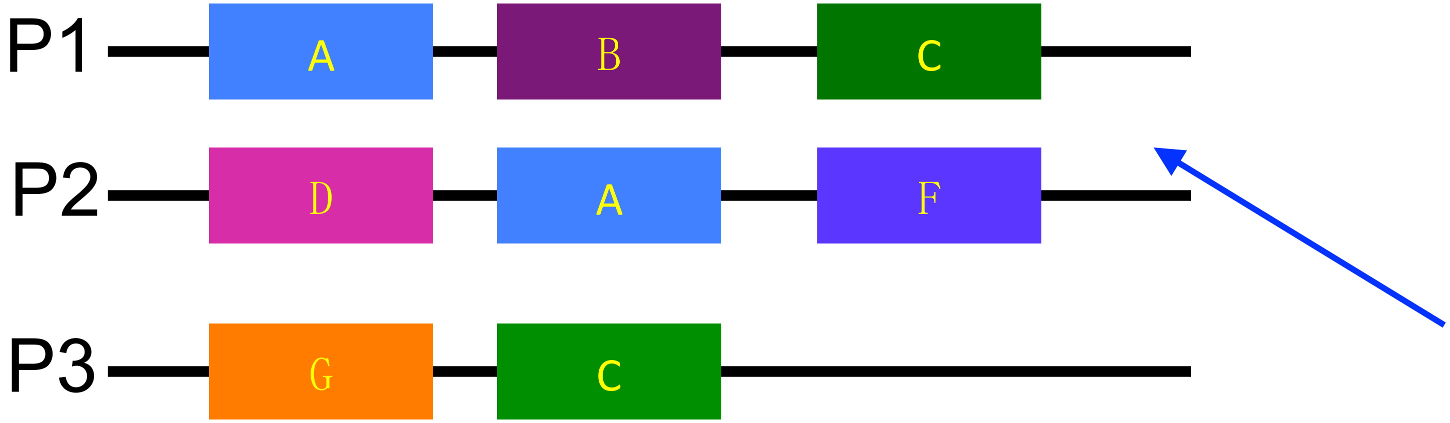


Association between transcription and ubiquitination

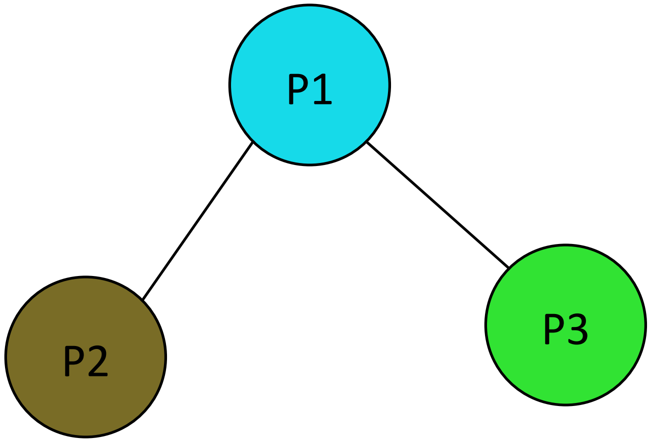
Three examples

- Clique merging
- Plant signature domain graph
- To predict domain functions with domain sharing network

Domain sharing network



by pfam domain index with whole genome sequence analysis

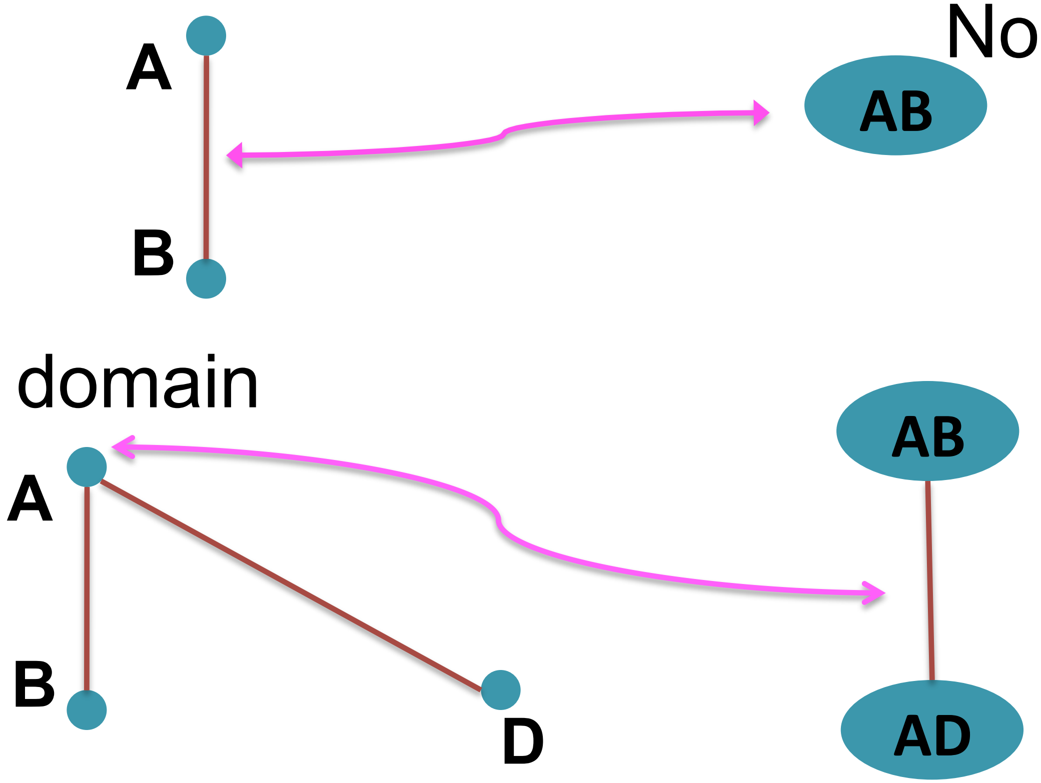


Line graph

Domain graph

Domain sharing graph

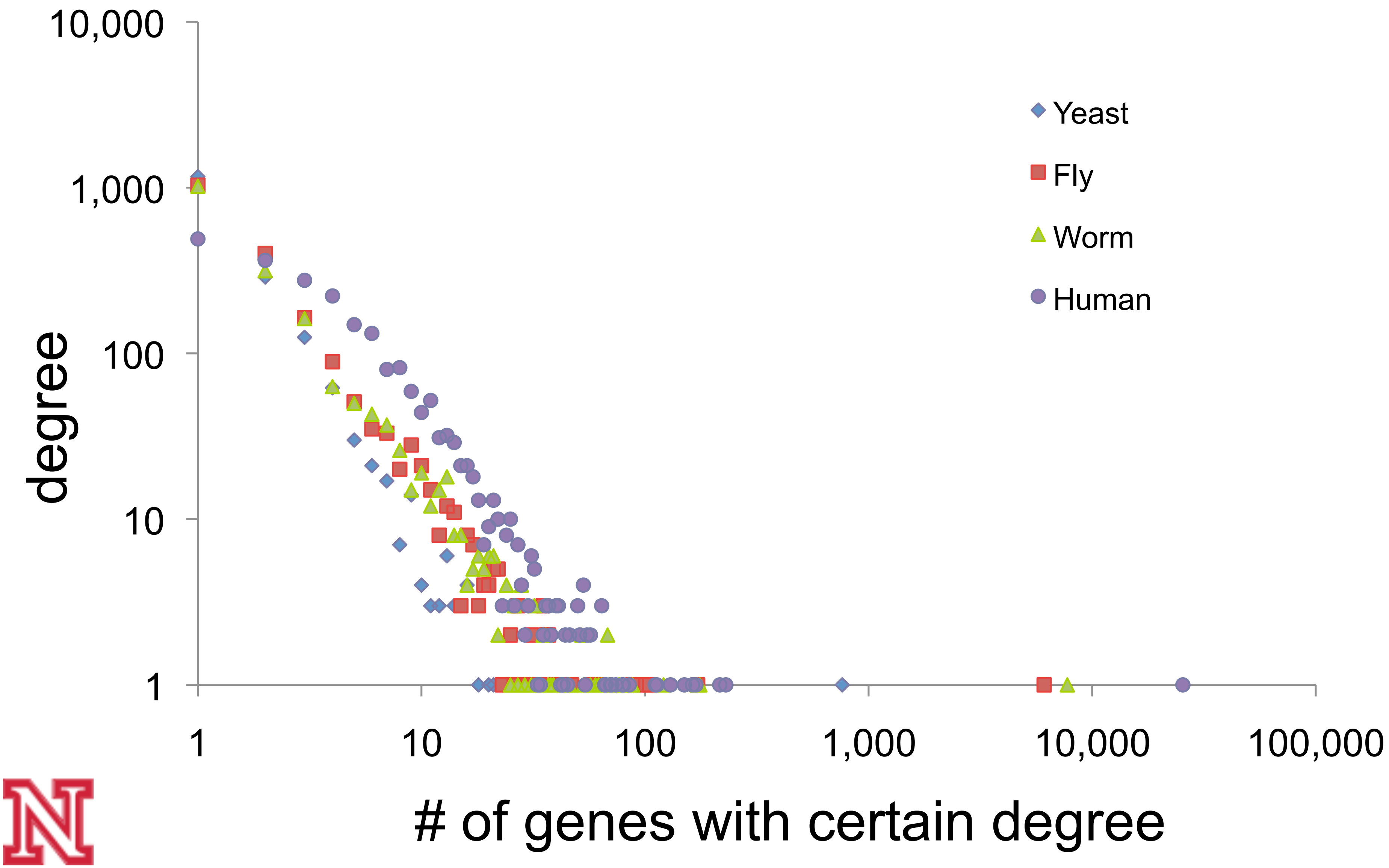
G(V,E) → G'(E,V)



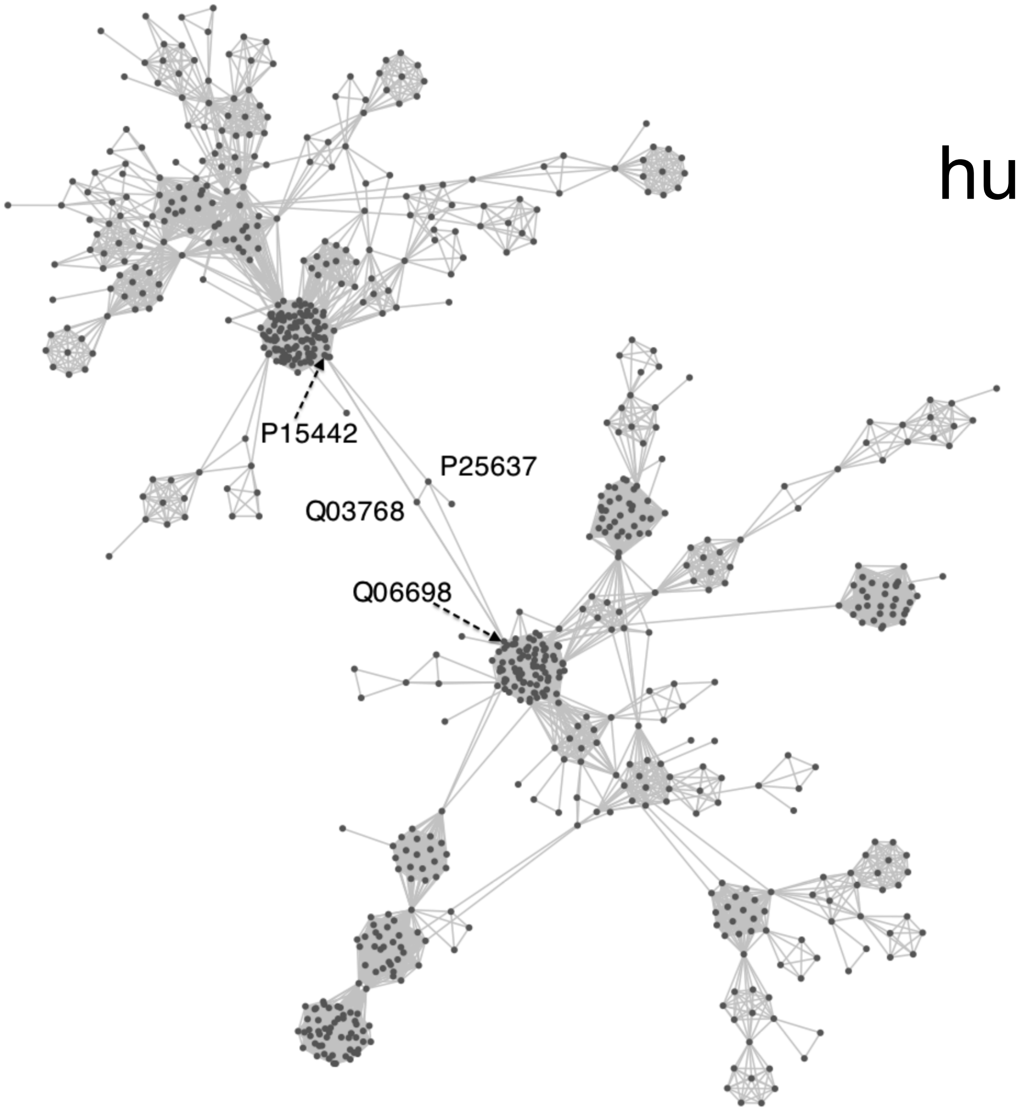
Node is a protein

Node is a domain

Scale free networks



Network example



human

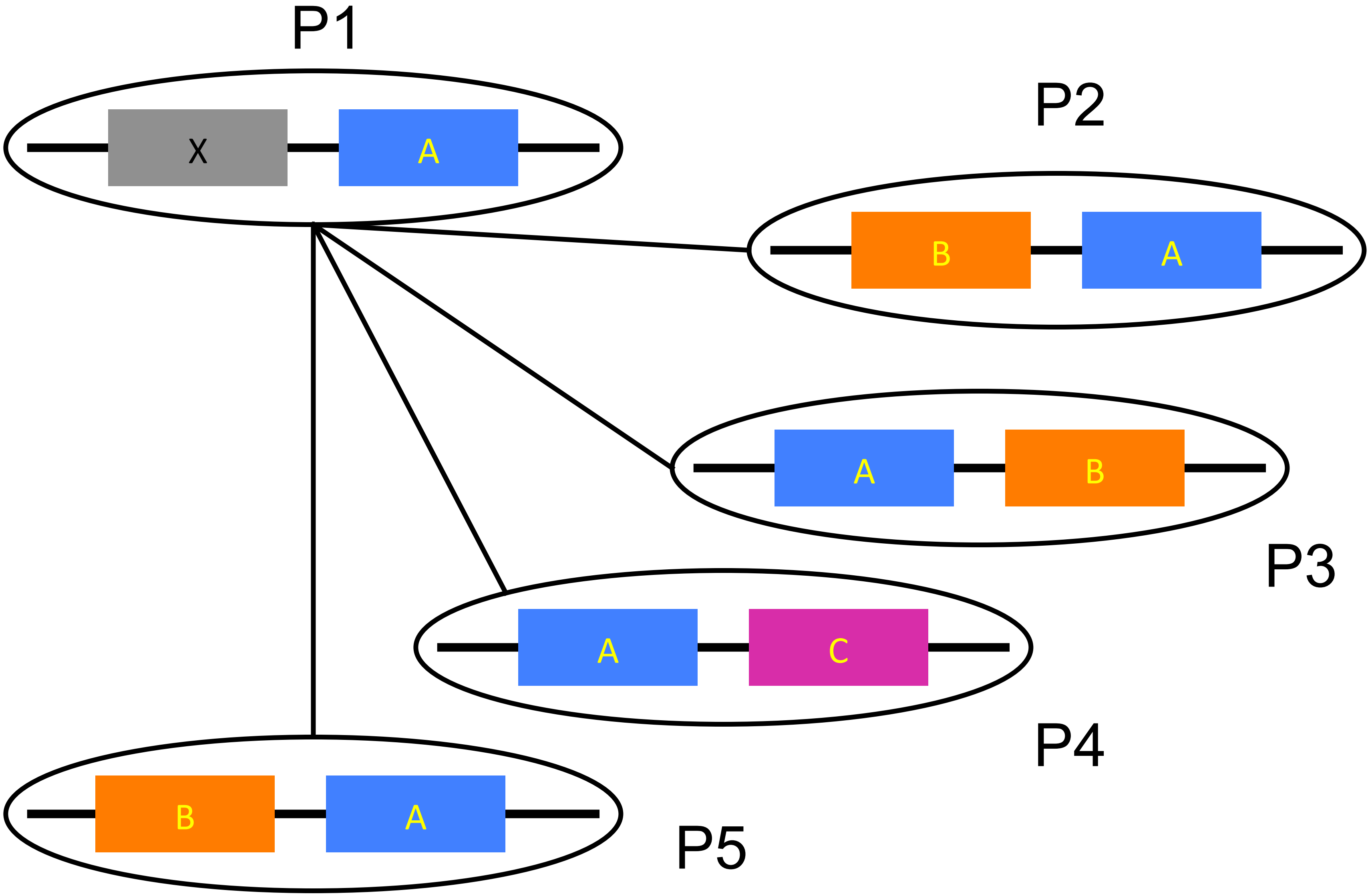
Components in human network

Module	# genes	# domains	Most popular function	%
1	25,455	1,478	*Protein binding function*	30.4
2	230	2	*Integral to membrane*	100
3	216	16	*Mental iron binding*	47.2
4	169	10	*Nucleus*	100

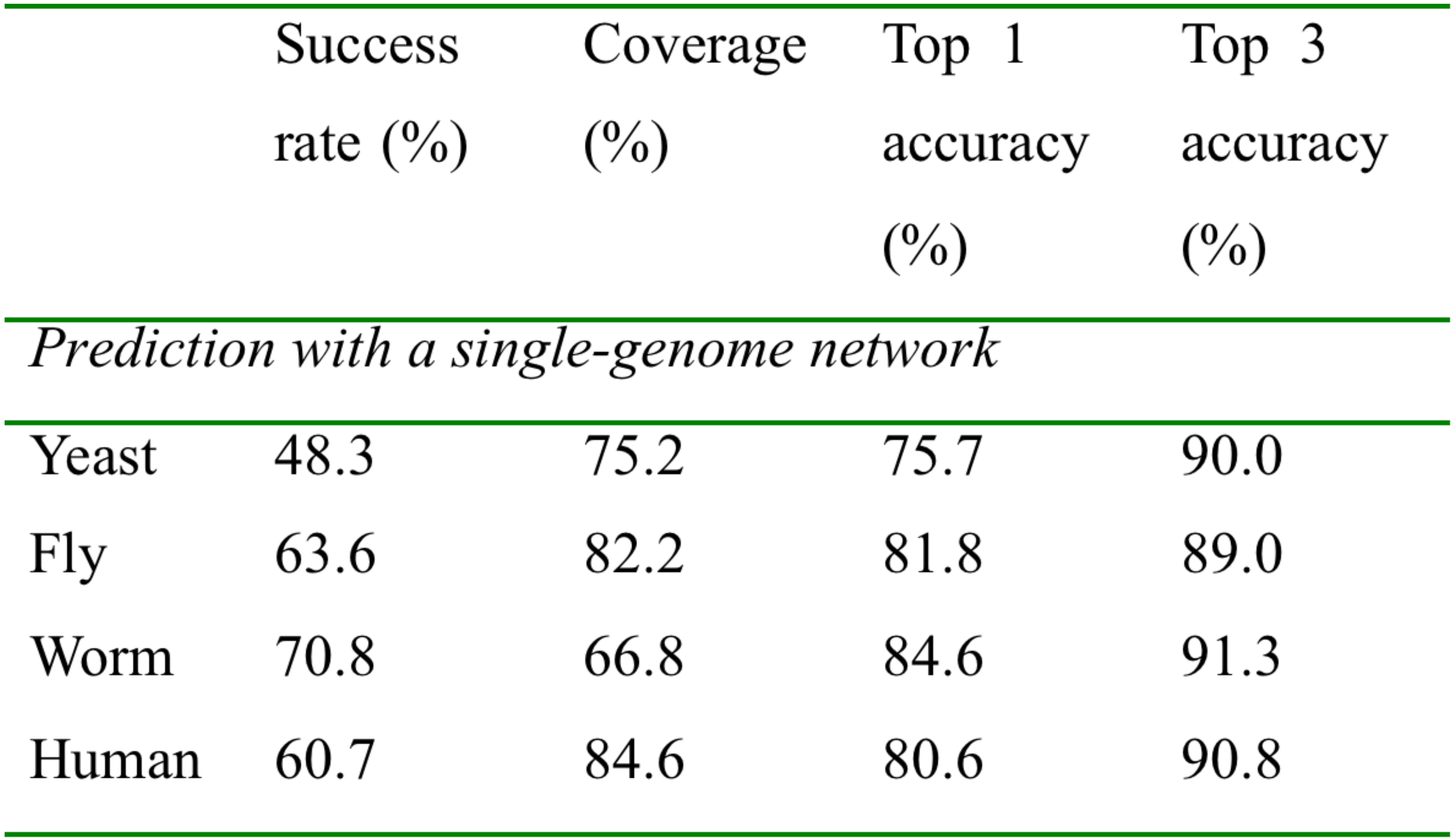
Network properties

	No. of genes	No. of domains	Mean value		
			Degree	Path distance	Clustering coefficient
Yeast	762	237	38.4	5.65	0.94
Fly	6,099	915	101.9	4.52	0.93
Worm	7,742	745	121.3	5.02	0.95
Human	25,455	1,478	312.5	4.14	0.92

Function prediction



Prediction



	Success rate (%)	Coverage (%)	Top 1 accuracy (%)	Top 3 accuracy (%)
Prediction with a single-genome network				
Yeast	48.3	75.2	75.7	90.0
Fly	63.6	82.2	81.8	89.0
Worm	70.8	66.8	84.6	91.3
Human	60.7	84.6	80.6	90.8

Random Success rates = (4.9%, 6.7%, 6.5%, and 9.9%)

Summary

1. For protein interaction network, we developed a module discovery algorithm, and the identified modules can be used for gene annotation
2. We also extracted the signature domain graph for Arabidopsis by comparing with other species, and the domain function can be inferred by network analysis.
3. Domain sharing networks were constructed, and domain combination information was used to the network to predict domain functions.