Protein-Protein Interaction Network

Lecture 2

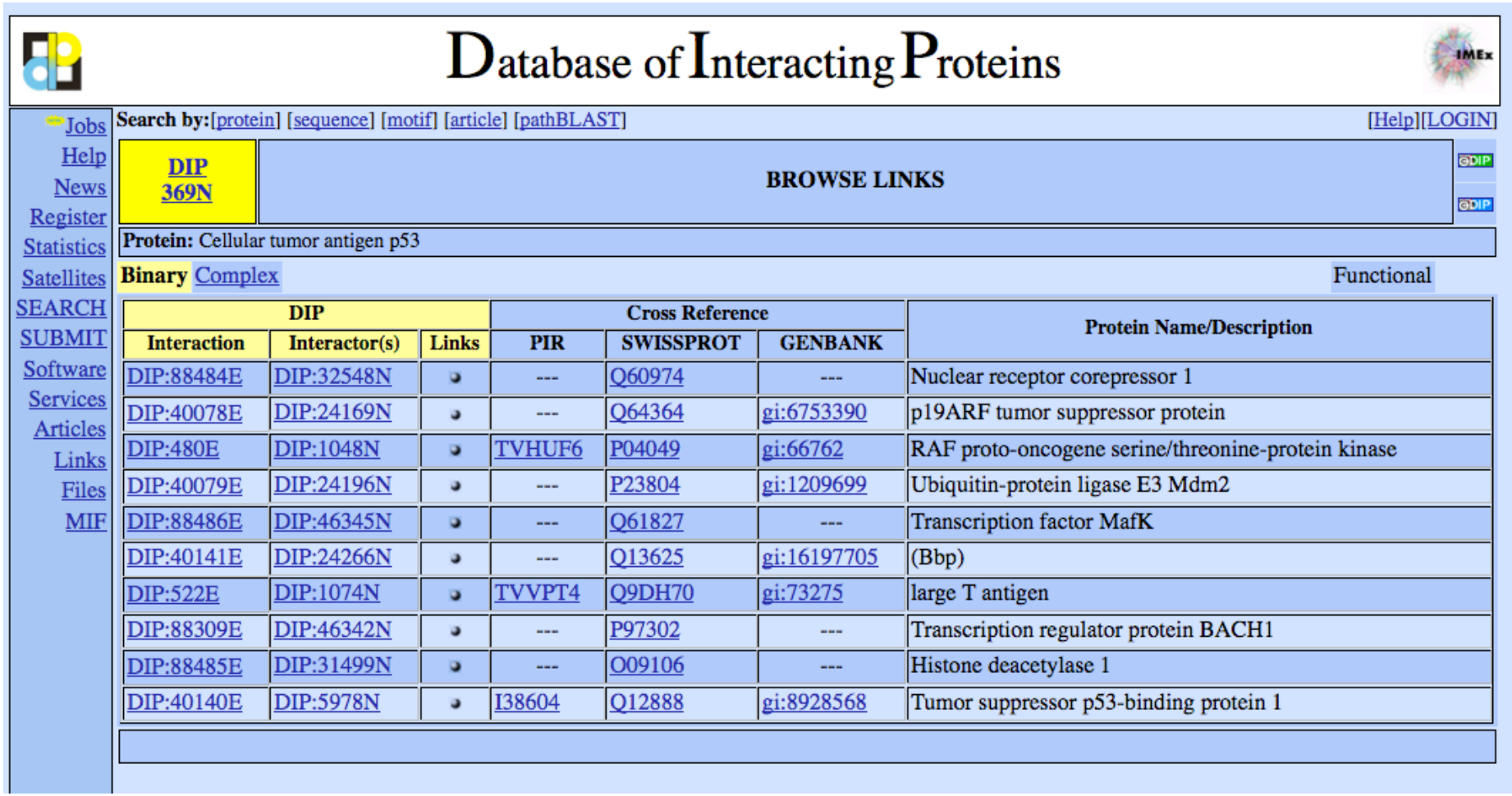
Outline

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

Databases that store interaction data

- Database of Interacting Proteins (DIP), http://dip.doe-mbi.ucla.edu/
- Biomolecular Interaction Network Database (BIND) , http://www.bind.ca/
- Molecular Interactions Database (MINT), http://160.80.34.4/mint/
- INTERACT http://www.ebi.ac.uk/intact/index.html
- PIBASE, http://alto.compbio.ucsf.edu/pibase/
- MIPS contains interaction data (both direct and clusters) for yeast
- SCOPPI, http://www.scoppi.org/
- Prolinks, http://mysql5.mbi.ucla.edu/cgi-bin/functionator/pronav

DIP



Database of Interacting Proteins

Jobs
Help
News
Register
Statistics
Satellites
SEARCH
SUBMIT
Software
Services
Articles
Links
Files
MIF

Search by: [protein] [sequence] [motif] [article] [pathBLAST] [Help][LOGIN]

DIP 369N

BROWSE LINKS

Protein: Cellular tumor antigen p53

Binary Complex Functional

DIP			Cross Reference			Protein Name/Description
Interaction	Interactor(s)	Links	PIR	SWISSPROT	GENBANK	
DIP:88484E	DIP:32548N		---	Q60974	---	Nuclear receptor corepressor 1
DIP:40078E	DIP:24169N		---	Q64364	gi:6753390	p19ARF tumor suppressor protein
DIP:480E	DIP:1048N		TVHUF6	P04049	gi:66762	RAF proto-oncogene serine/threonine-protein kinase
DIP:40079E	DIP:24196N		---	P23804	gi:1209699	Ubiquitin-protein ligase E3 Mdm2
DIP:88486E	DIP:46345N		---	Q61827	---	Transcription factor MafK
DIP:40141E	DIP:24266N		---	Q13625	gi:16197705	(Bbp)
DIP:522E	DIP:1074N		TVVPT4	Q9DH70	gi:73275	large T antigen
DIP:88309E	DIP:46342N		---	P97302	---	Transcription regulator protein BACH1
DIP:88485E	DIP:31499N		---	O09106	---	Histone deacetylase 1
DIP:40140E	DIP:5978N		I38604	Q12888	gi:8928568	Tumor suppressor p53-binding protein 1

Tumor suppressor gene P53, PID ID “<DIP:369N>”

DIP Interaction Details

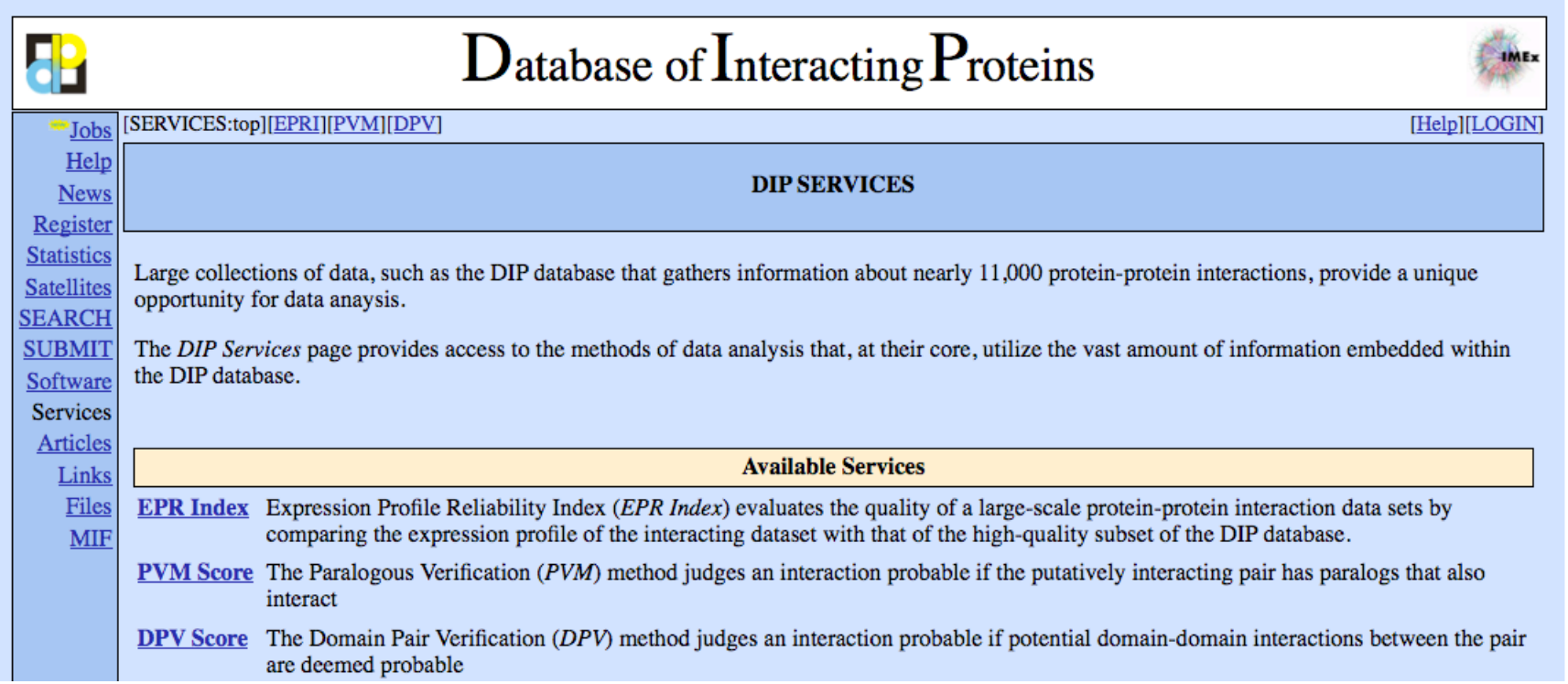
DIP LINK [-----]

DIP 88484E				
DIP 369N	**PIR** DNMS53	**SwissProt** P02340	**GenBank** gi:2144761	
	Name/Description	Cellular tumor antigen p53		
DIP 32548N	**PIR**	**SwissProt** Q60974	**GenBank**	
	Name/Description	Nuclear receptor corepressor 1		

Evidence Help

Type	Method	Details	Source	Curation	IMEx
E(d)	anti bait coimmunoprecipitation		PMID:19011633	DIP	
V	SMSC(1)	---			

DIP services



Database of Interacting Proteins

Jobs
Help
News
Register
Statistics
Satellites
SEARCH
SUBMIT
Software
Services
Articles
Links
Files
MIF

[SERVICES:top][EPRI][PVM][DPV] [Help][LOGIN]

DIP SERVICES

Large collections of data, such as the DIP database that gathers information about nearly 11,000 protein-protein interactions, provide a unique opportunity for data anaysis.

The *DIP Services* page provides access to the methods of data analysis that, at their core, utilize the vast amount of information embedded within the DIP database.

Available Services

EPR Index	Expression Profile Reliability Index (*EPR Index*) evaluates the quality of a large-scale protein-protein interaction data sets by comparing the expression profile of the interacting dataset with that of the high-quality subset of the DIP database.
PVM Score	The Paralogous Verification (*PVM*) method judges an interaction probable if the putatively interacting pair has paralogs that also interact
DPV Score	The Domain Pair Verification (*DPV*) method judges an interaction probable if potential domain-domain interactions between the pair are deemed probable

Expression Profile Reliability (EPR)
Homology methods -Paralogous Verification (PVM)
Domain Pair Verification (DPV)

BIND

- Designed to hold direct interaction, cluster and pathway data 81,000 interactions written in ASN.1 (Abstract Syntax Notation) for computational efficiency

Interaction 13118 Mus musculus Full BIND Record **Launch Viewer:** Select Below

Molecule	Description	Molecular Function	Cellular Component	Biological Process	Experiment(s)	Links	
P53 • Trp53;TP53	Transformation related protein 53. Tumour suppressor protein with DNA binding and transcription factor function. Role in cell cycle; mutations involve [more...]	• DNA binding • transcription factor activity • protein binding	• nucleus • cytoplasm • cytosol	• protein-nucleus import\, translocation • transcription • regulation of transcription\, DNA-dependent • apoptosis • DNA damage response\, signal transduction by p53 class mediator • negative regulation of cell cycle	• Immunoprecipitation	NCBI SeqHound	[2 Pubmed Abstracts] [Other BIND data]
MDM2 • Mdm-2	Transformed Mmouse 3T3 cell Double Minute 2; nuclear phosphoprotein; LocusID:17246	• ubiquitin-protein ligase activity • protein binding • ATP binding • ligase activity	• nucleus	• start control point of mitotic cell cycle • cell growth and/or maintenance • protein ubiquitination • protein catabolism		NCBI SeqHound	

Bader GD, Betel D, Hogue CW. (2003) Nucleic Acids Res. 31(1):248-50

Arabidopsis Databases that store interaction data

- TAIR ftp://ftp.arabidopsis.org/home/tair/Proteins/Interactome2.0/
- http://bioinformatics.psb.ugent.be/supplementary_data/stbod/athPPI/site.php
- AtPIN http://bioinfo.esalq.usp.br/atpin/atpin.pl
- AtPid http://atpid.biosino.org/

Domain-Domain interaction Database

- iPfam, http://www.sanger.ac.uk/Software/Pfam/iPfam/
- 3did (domain interactions) http://gatealoy.pcb.ub.es/3did/
- DIMA http://webclu.bio.wzw.tum.de/dima/downloads.jsp

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

Random Networks

- Uniformly random network:
  - distributes the edges uniformly among nodes.
- Probabilistic interpretation:
  - There exists a set (ensemble) of networks with given number of nodes and edges. Select a random member of this set.

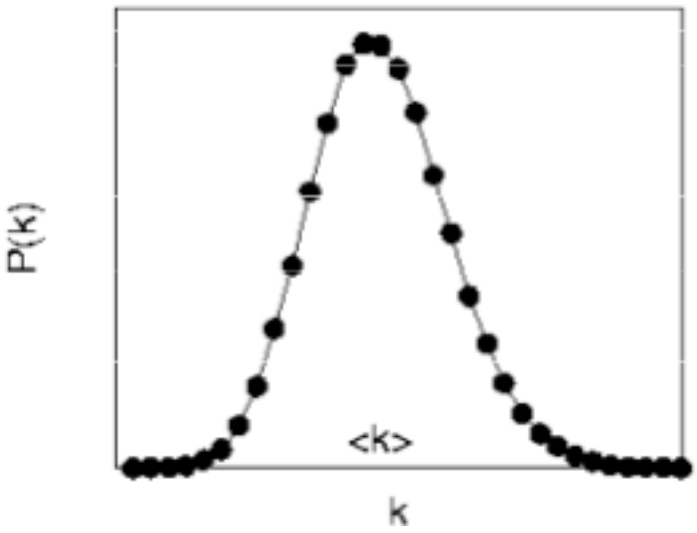
Random Networks

p = 0 p = 0.1 p = 0.15

- fixed node number *N*
- connecting pairs of nodes with probability *p*

Expected number of edges: $E = p\frac{N(N-1)}{2}$

Node degrees in random graphs



Average degree:

$$\langle k \rangle \approx p|V|$$

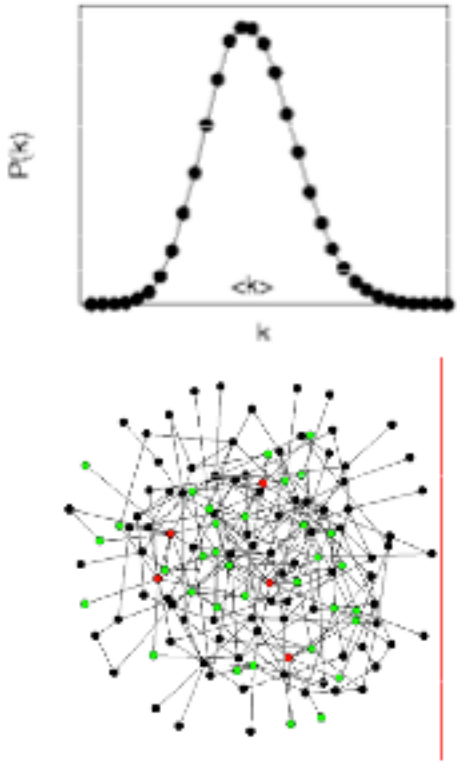
Degree distribution:

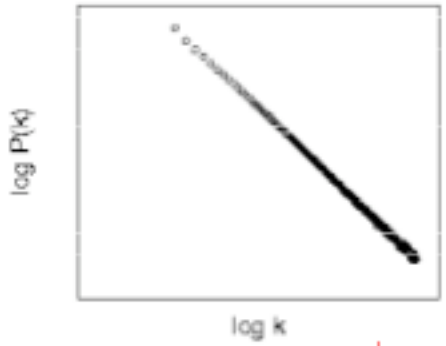
$$\mathbf{P}(k) \approx \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

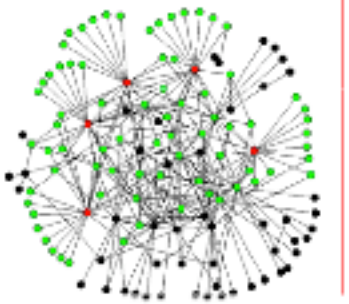
Most of the nodes have approximately the same degree. The probability of very highly connected nodes is exponentially small.

A scale free network

- Power-law degree distributions were found in diverse networks







Large variability

A scale free network

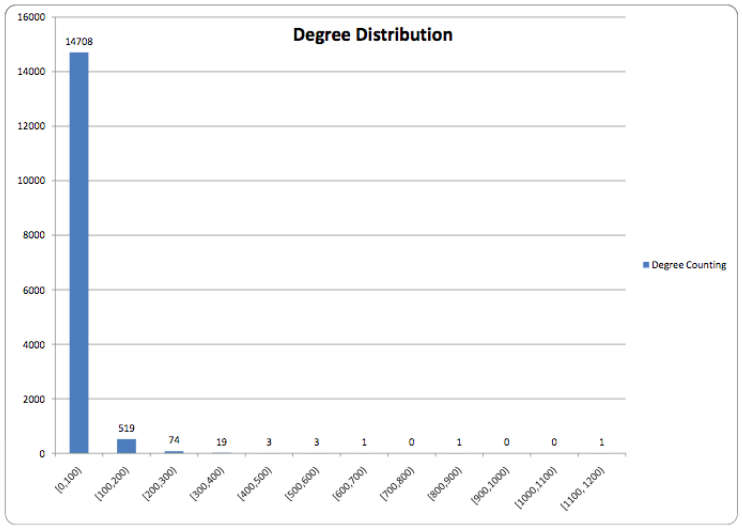
- Power-law degree distributions were found in diverse networks

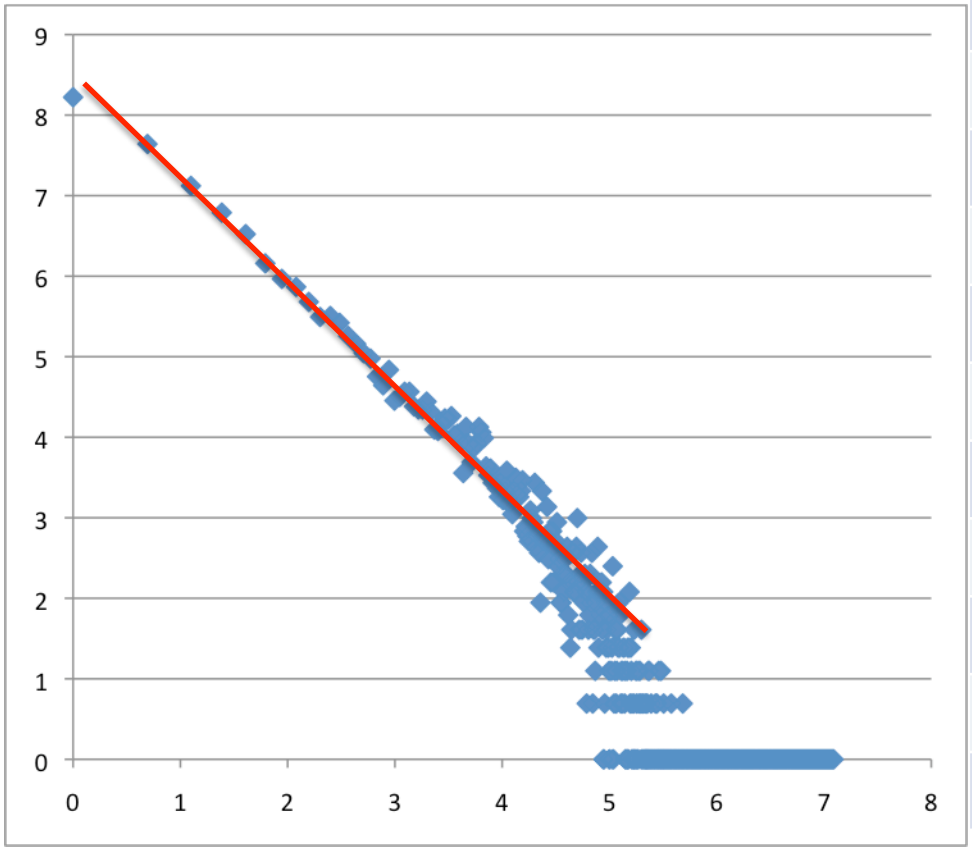
$$\log\left(\mathbf{P}(k)\right) \approx -\gamma \log(k)$$

$$\mathbf{P}(k) \approx c\, k^{-\gamma}$$

Power-law degree distributions

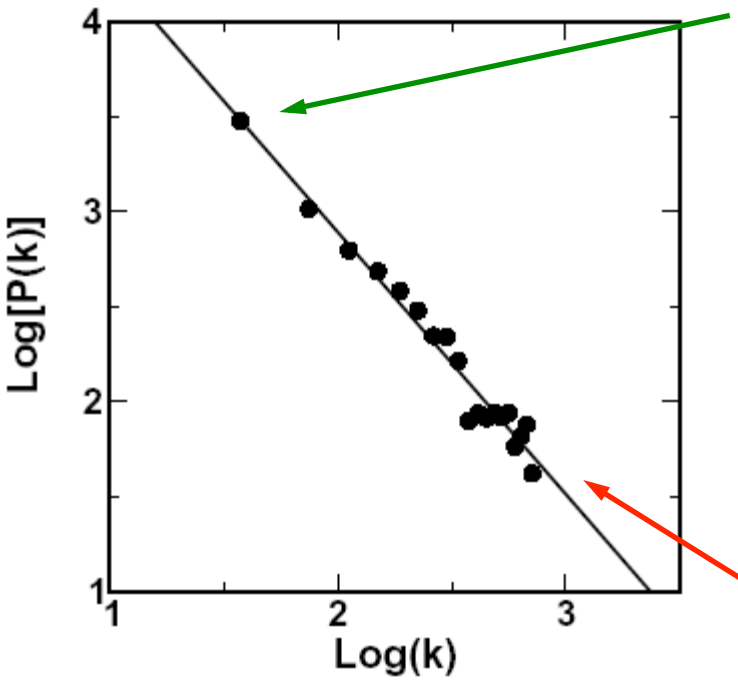
ATH PPI

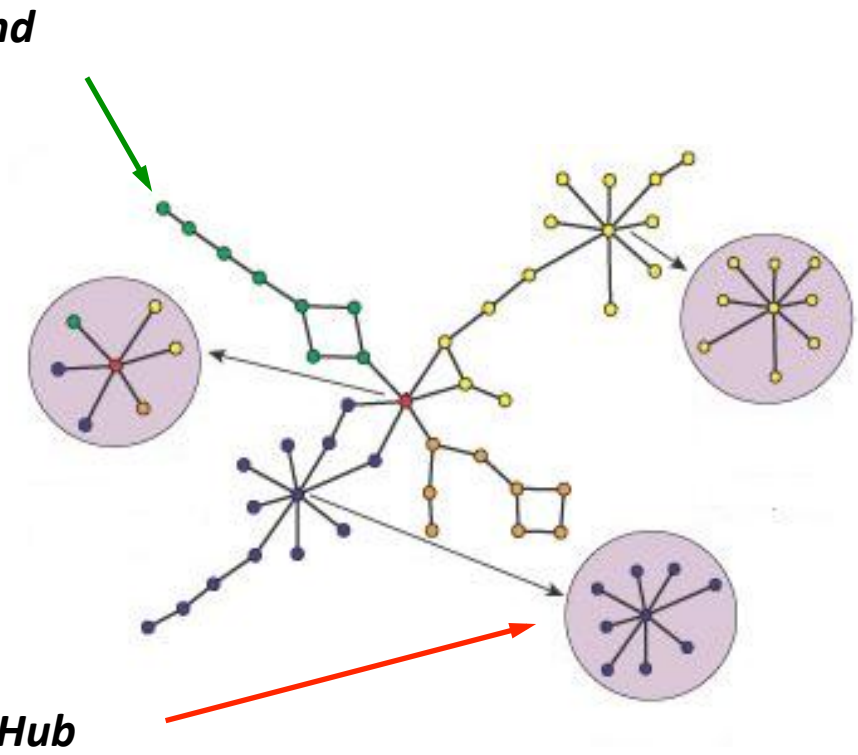




k	log(k)	P(k)	log(P(k))
1	0	3721	8.221
2	0.693	2082	7.641
3	1.098	1238	7.121
4	1.386	888	6.788
5	1.609	680	6.522
6	1.791	473	6.159
7	1.945	390	5.966
8	2.079	353	5.866
9	2.197	293	5.680
10	2.302	243	5.493
11	2.397	246	5.505
12	2.484	226	5.4205
13	2.564	192	5.257
14	2.639	174	5.159
15	2.708	155	5.043
16	2.772	145	4.9767
17	2.833	116	4.753

Scale Free



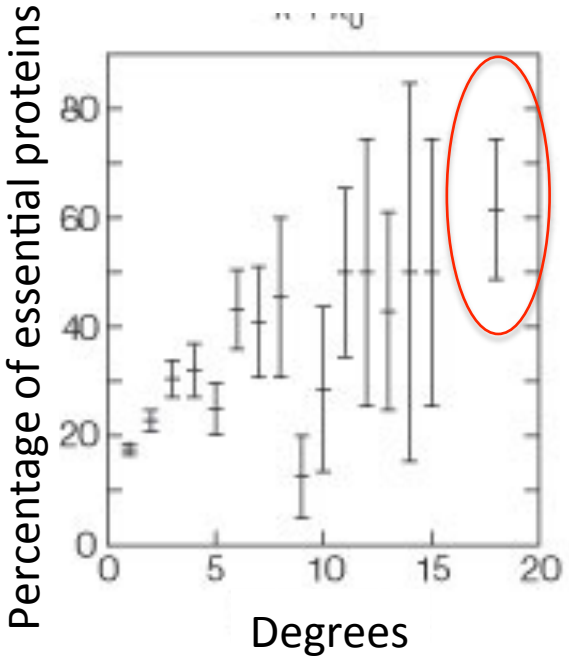


$$P(k) \sim k^{-\gamma}$$

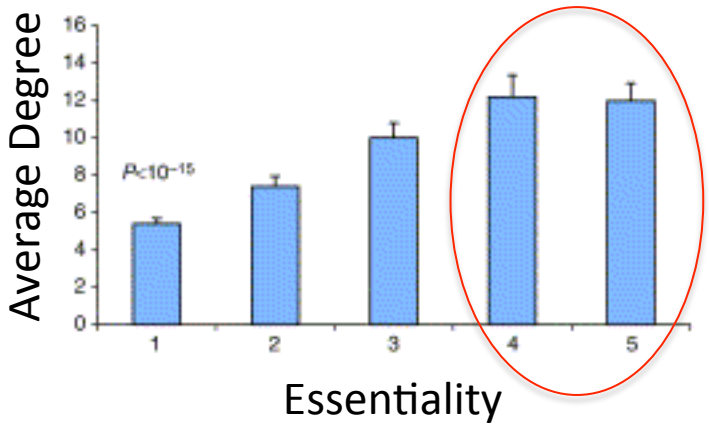
Han *et al.* Nature, 2004

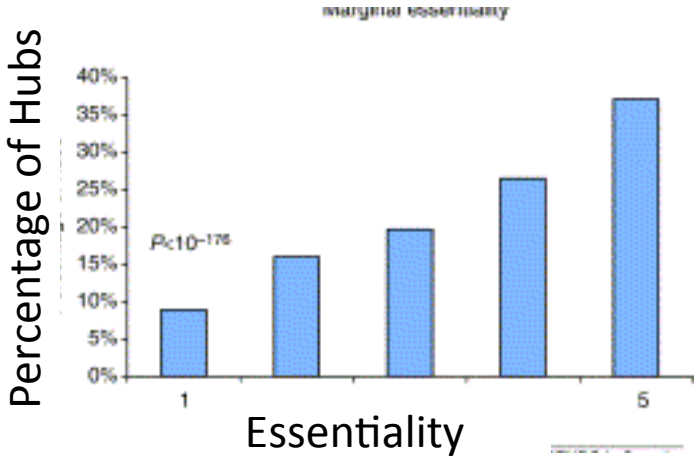
Hub proteins=Essential proteins

- 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)



Essential proteins





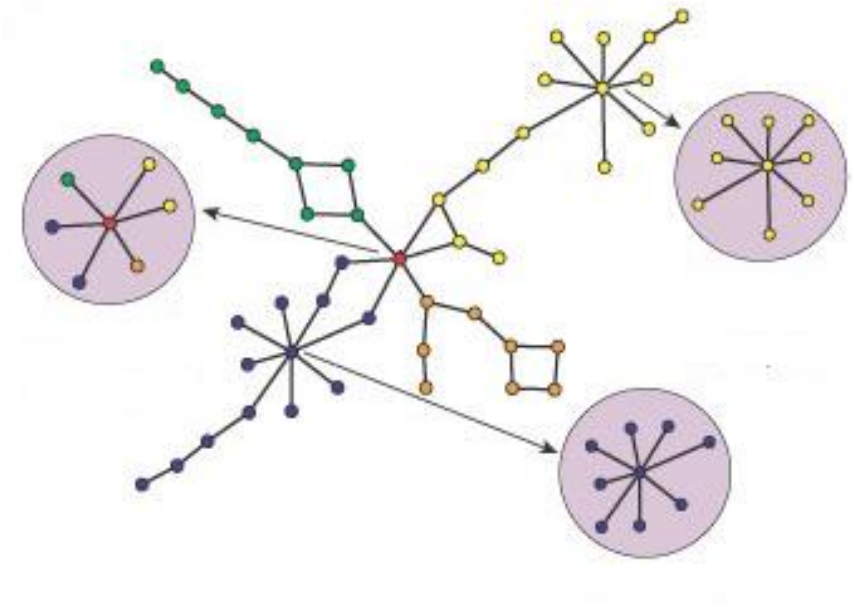
Hubs have high degrees

Essential genes have high essentiality.

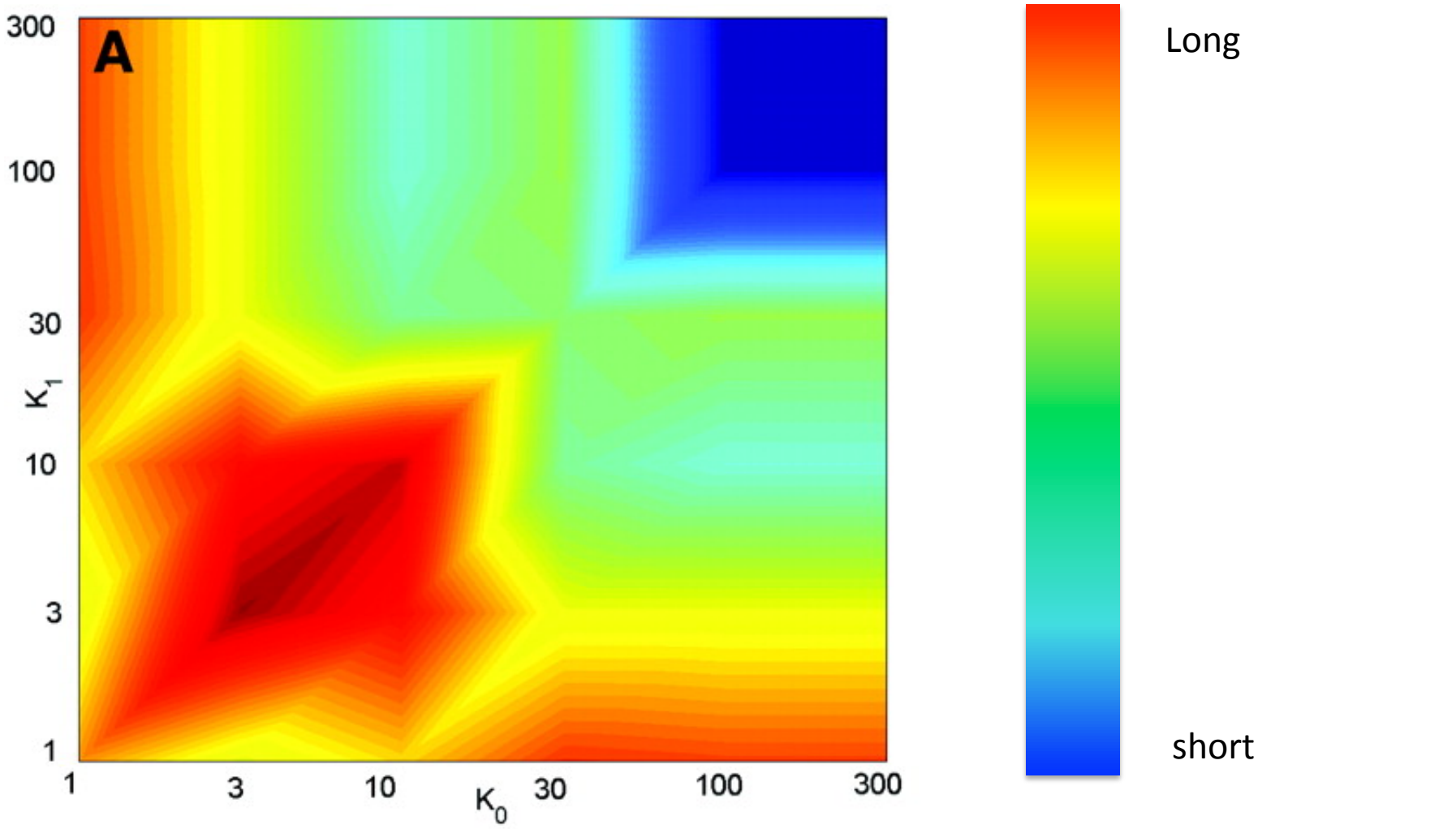
Yu (2004) Trends in Genetics, 20(6), 227

Hub proteins close to each other

- Hub proteins have lower average length of shortest path among themselves than non-hub proteins. (Moslov et al. 2002 *Science 296, 910*)

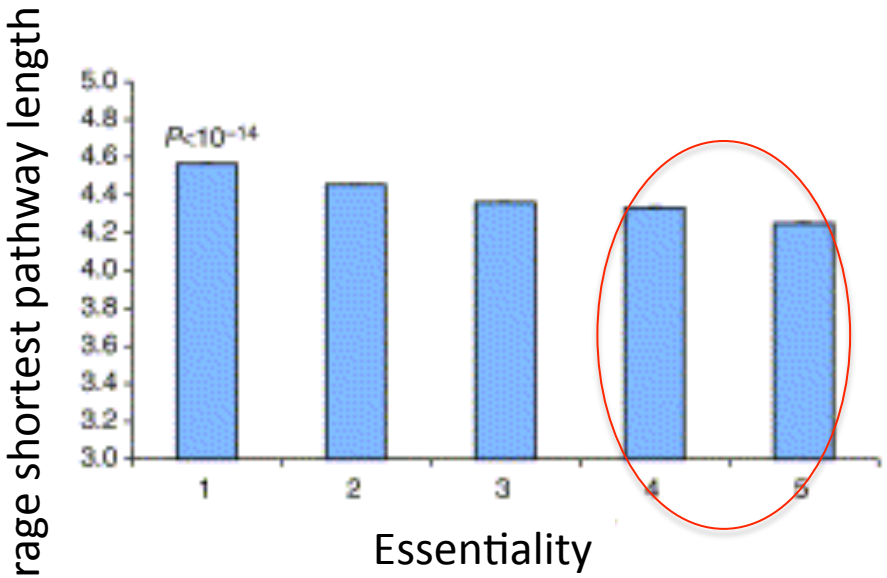


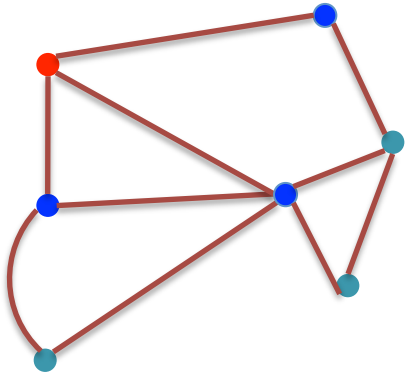
Length of shortest path

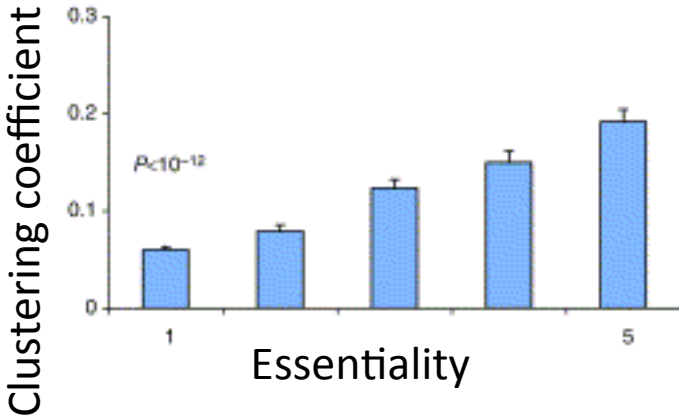


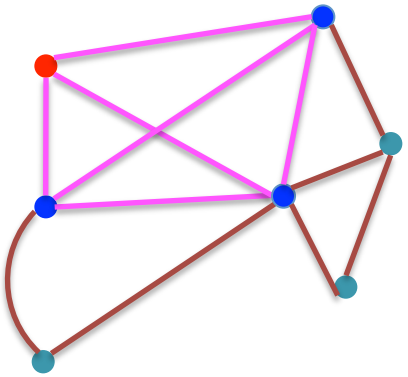
Moslov et al. 2002 *Science 296, 910*

Essential proteins





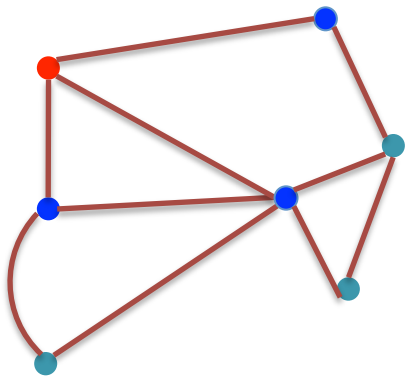




Clustering coefficient

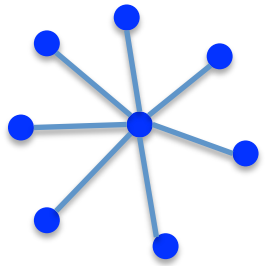
- Local clustering coefficient C_i for a vertex v_i is given by the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them.

$$C_i = \frac{\left|e_{ij}\right|}{\mathbf{V}(\mathbf{V}-1)/2}$$

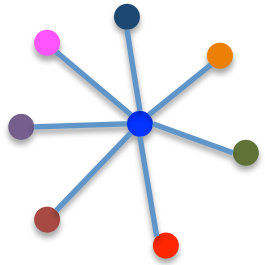


Static or Dynamic

- Combined PPI with gene expression profiles.
- Calculate co-express correlation between hubs and their neighbors.
- Two types of hubs:



Party Hub



Date Hub

Han et al. (2004) Nature 430(6995):88-93

Gene Co-expression correlation

	T1	T2	T3	T4	T5
A	2.5	2.8	3.7	4.6	1.5
B	0.2	0.8	0.3	1.5	0.6
C	1.9	1.3	0.2	0.8	1.6
D	0.8	1.4	0.7	1.6	1.7
E	1.5	1.8	0.3	0.5	1.9

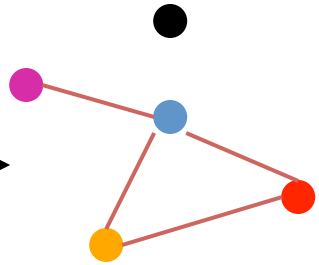
pair-wise

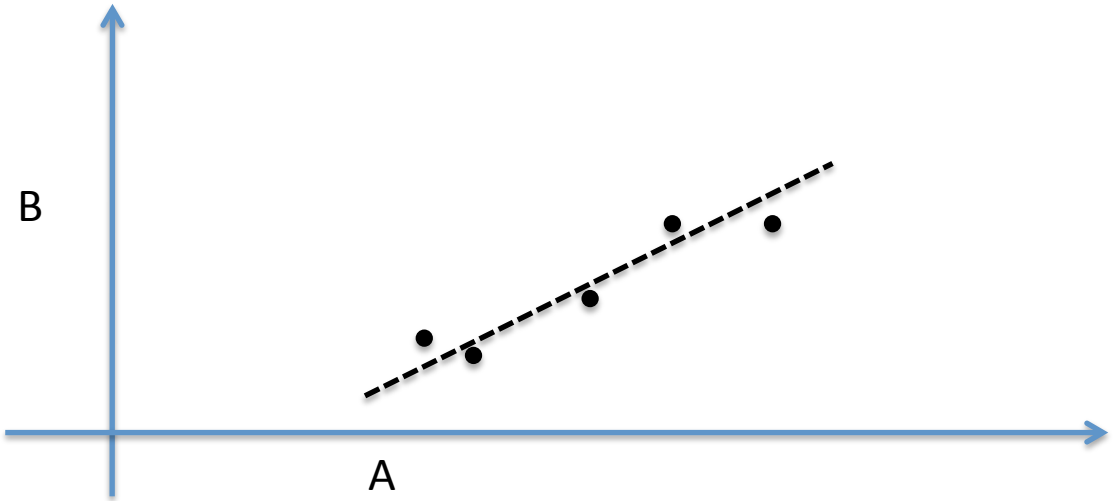
correlation

	C.C
A-B	0.76
A-C	0.90
A-D	0.50
…	0.83
D-E	0.42

cutoff

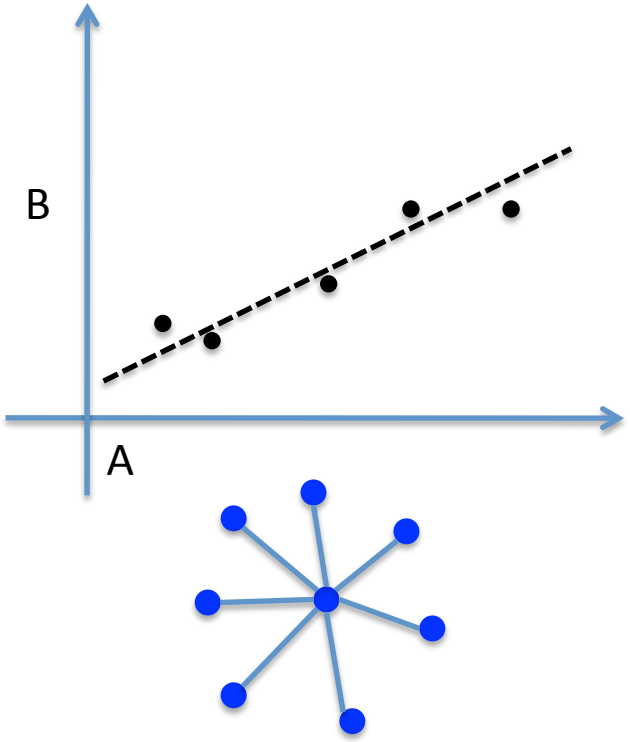
>= 0.6



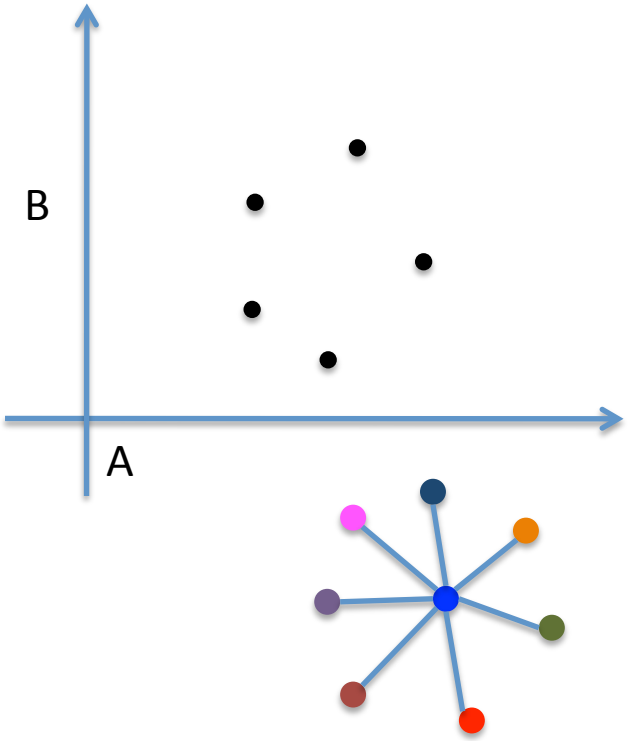


Hub Co-expression correlation

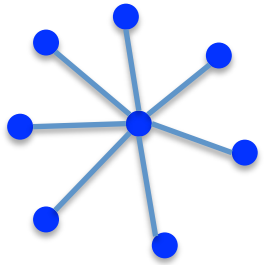
	T1	T2	T3	T4	T5
A	2.5	2.8	3.7	4.6	1.5
B	2.4	2.8	3.6	4.7	1.6
C	1.9	2.0	3.2	4.2	1.3
D	2.8	3.0	4.1	5.0	2.5
E	1.5	1.8	3.0	4.0	1.2



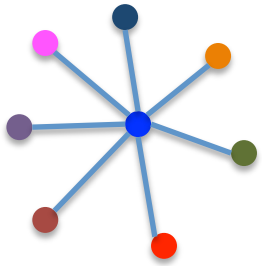
	T1	T2	T3	T4	T5
A	2.5	2.8	3.7	4.6	1.5
B	5.4	0.8	1.6	4.7	3.6
C	1.0	5.0	1.2	2.2	3.3
D	4.8	0.3	0.1	6.0	1.5
E	1.0	2.8	3.4	0.0	1.2



Date or Party Hubs

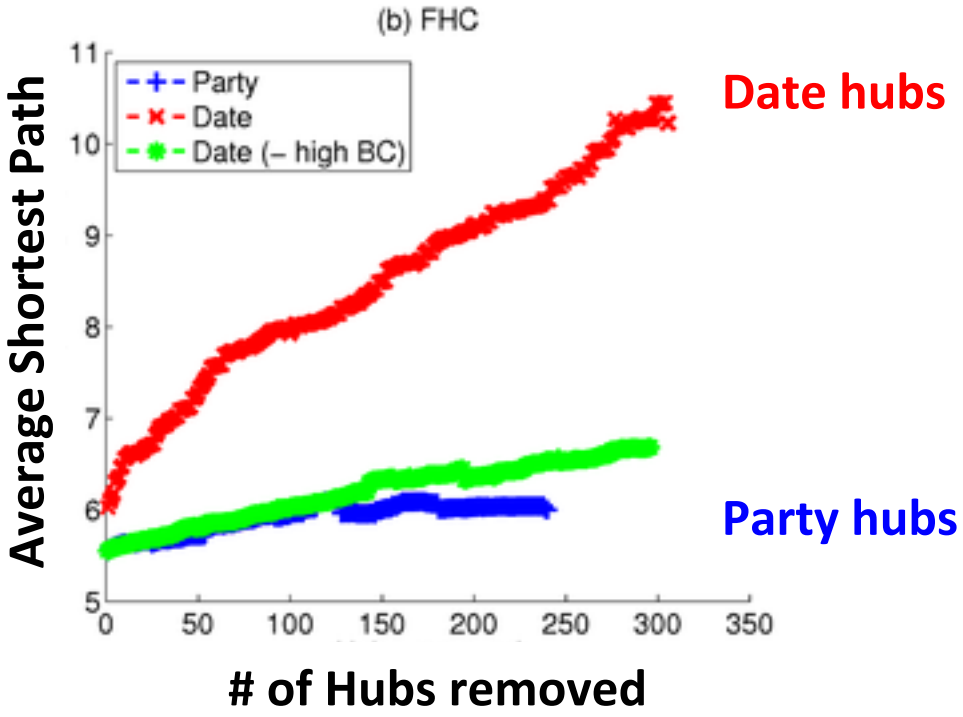


Party Hubs are expressed with their connection partners at same time. They will form a large protein complex. They are more essential. Most of them are house keeping genes.



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

Network topology of hubs



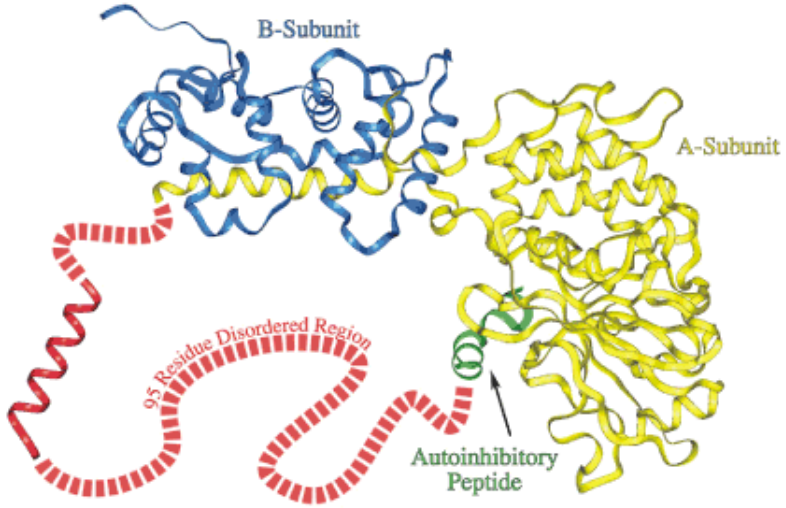
Date hubs

Party hubs

Hub proteins

- Multiple and repeated domains are enriched in hub proteins

- Long disordered regions are common in hubs.

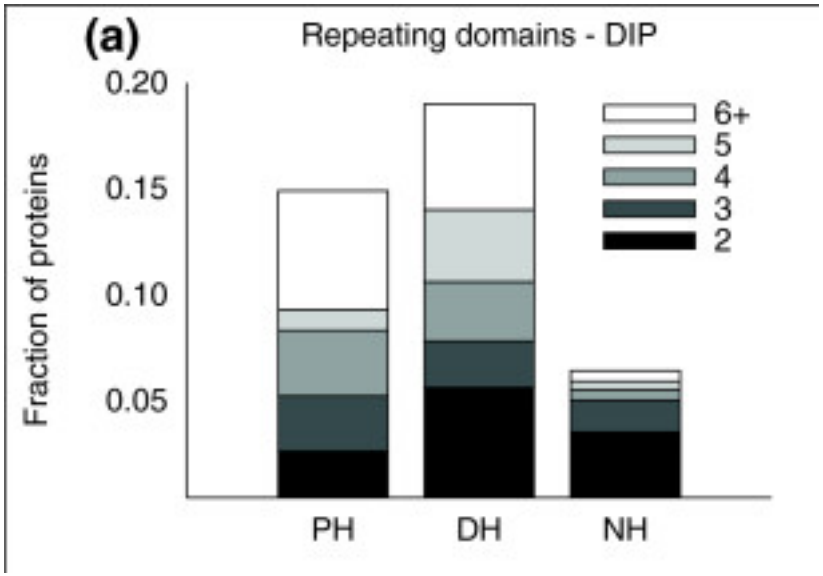


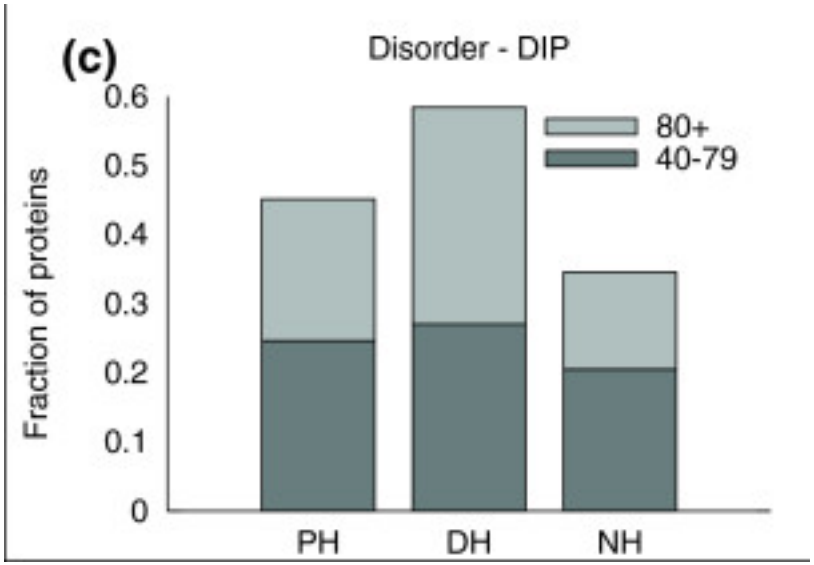
(Image adapted from: Kissinger CR, et al. 1995. "Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex." Nature 378:641-4.)

disordered regions are typically involved in regulation, signaling and control pathways in which interactions with multiple partners and high-specificity/low-affinity interactions are often requisite.

(Ekman et al. 2006 Genome Biol. 7(6): R45)

Hub proteins





PH: Party Hubs
DH: Date Hubs
NH: Non-hubs

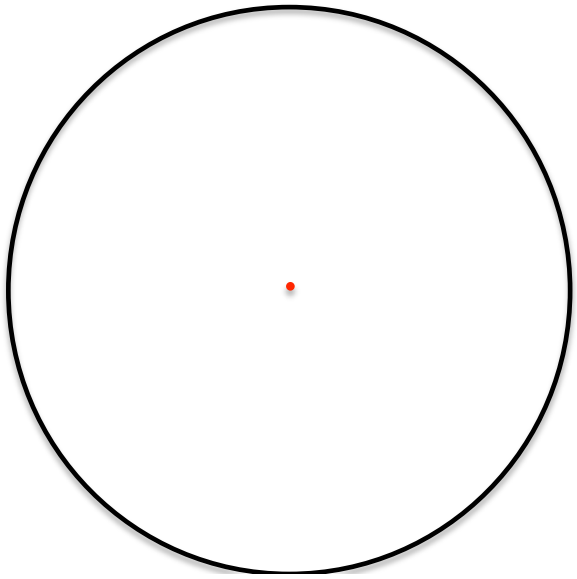
Centrality of PPI

- In yeast, worm, and fly PPI networks, the number of degrees and the centrality of proteins in the networks have similar distributions.
- Essential proteins have significant centrality.
- Proteins that have a more central position in all three networks, regardless of the number of direct interactors, evolve more slowly and are more likely to be essential for survival.

Hahn et al. (2004) Molecular Biology and Evolution, 22(4) 803.

Centrality

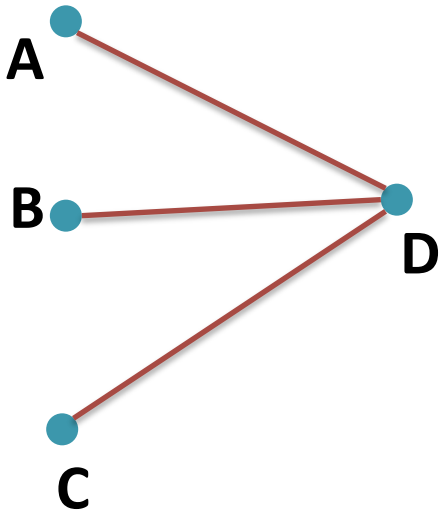
- Measure of the **centrality** of a vertex within a graph that determine the relative importance of a vertex within the graph.
  - Closeness centrality
  - Betweenness centrality



Closeness centrality

- It is defined as the average distance between a vertex *v* and all other vertices reachable from it.
- For a graph *G*: = (*V*,*E*) with *n* vertices, the degree centrality $C_c(v)$ for vertex *v* is

$$C_c = \frac{\sum_i \mathbf{dis}(vi)}{n-1}$$



A node is important if it has a small closeness centrality, because it is close to any other node.

Betweenness centrality

- Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not.
- For all node pairs (*i, j),* find the number of shortest paths between them, σ(i,j), and determine how many of these pass through node *k* - $\sigma_k(i,j)$

$$C_k = \sum_{i,j} \frac{\sigma_k(i,j)}{\sigma(i,j)}$$

A node is important if it has a large Betweenness centrality, because many shortest paths pass it.

Essentiality and Centrality

		Yeast	Worm	Fly
Betweeness Centrality	Essential	0.0009	0.0017	0.0007
	Non-Essential	0.0007	0.0009	0.0004
1／Closeness Centrality	Essential	0.244	0.183	0.238
	Non-Essential	0.239	0.175	0.221
Degrees	Essential	19.3	8.2	9.8
	Non-Essential	15.8	5.6	5.7

Hahn et al. (2004) Molecular Biology and Evolution, 22(4) 803.

Essentiality, Centrality, slow evolution rate

correlation	Yeast	Worm	Fly
D_n - Betweeness	-0.174	-0.118	-0.071
D_n - Closeness	-0.085	-0.114	-0.064
D_n - Degrees	-0.161	-0.027	-0.053

- Identified orthologs of the proteins in the yeast, worm, and fly networks in the related species *S. paradoxus*, *C. briggsae*, and *D. pseudoobscura*, respectively.
- D_n = the number of nonsynonymous differences per nonsynonymous site. (that changes amino acid). This is proportional to the evolution rate.
- Essential genes are house-keeping genes, have slow evolution rate.

Hahn et al. (2004) Molecular Biology and Evolution, 22(4) 803.

Evolution Rates of party or date hubs

	Date Hubs	Party Hubs
Dn	0.7597	0.5652
Ds	2.3133	2.4254
Dn/Ds	0.3631	0.2627

- The lowering of evolutionary rate of the party hub proteins than the date hub proteins.
- Party hubs form a big protein complex; they are more essential.

Dn: non-synonymous distance (changes amino acid)
Ds: Pairwise synonymous (do not change amino acid)

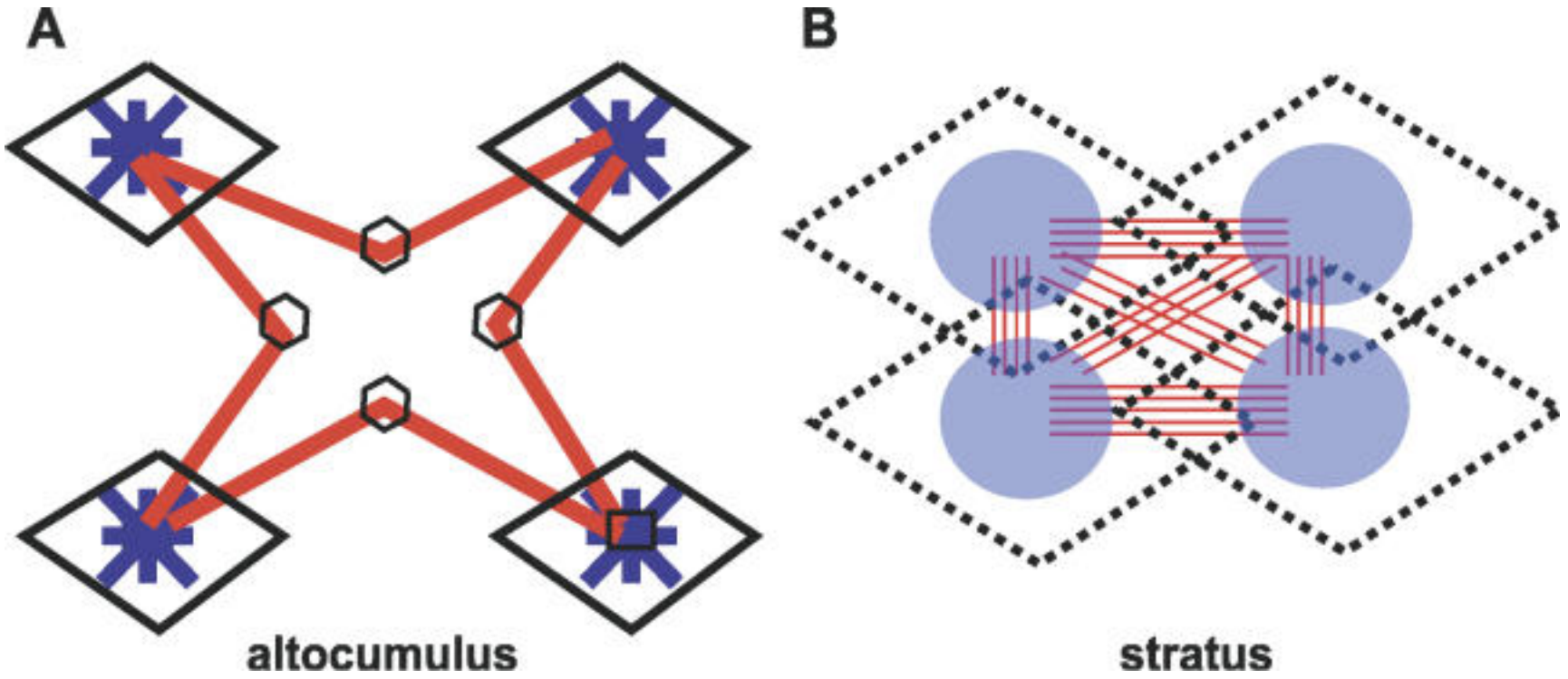
Kahali Et al (2009) Gene, 429, 18

PPI Network topology

- Global protein interaction network is highly interconnected and hence interdependent, more like the continuous dense aggregations of stratus clouds than the segregated configuration of altocumulus clouds.

Batada et al. (2006) PloS Biology, 4(10), e137

Altocumulus or Stratus



highly interconnected and
hence interdependent

Fault tolerance of PPI Networks

- Whether there exist alternative pathways that can perform some required function if a gene essential to the main mechanism is defective, absent or suppressed.

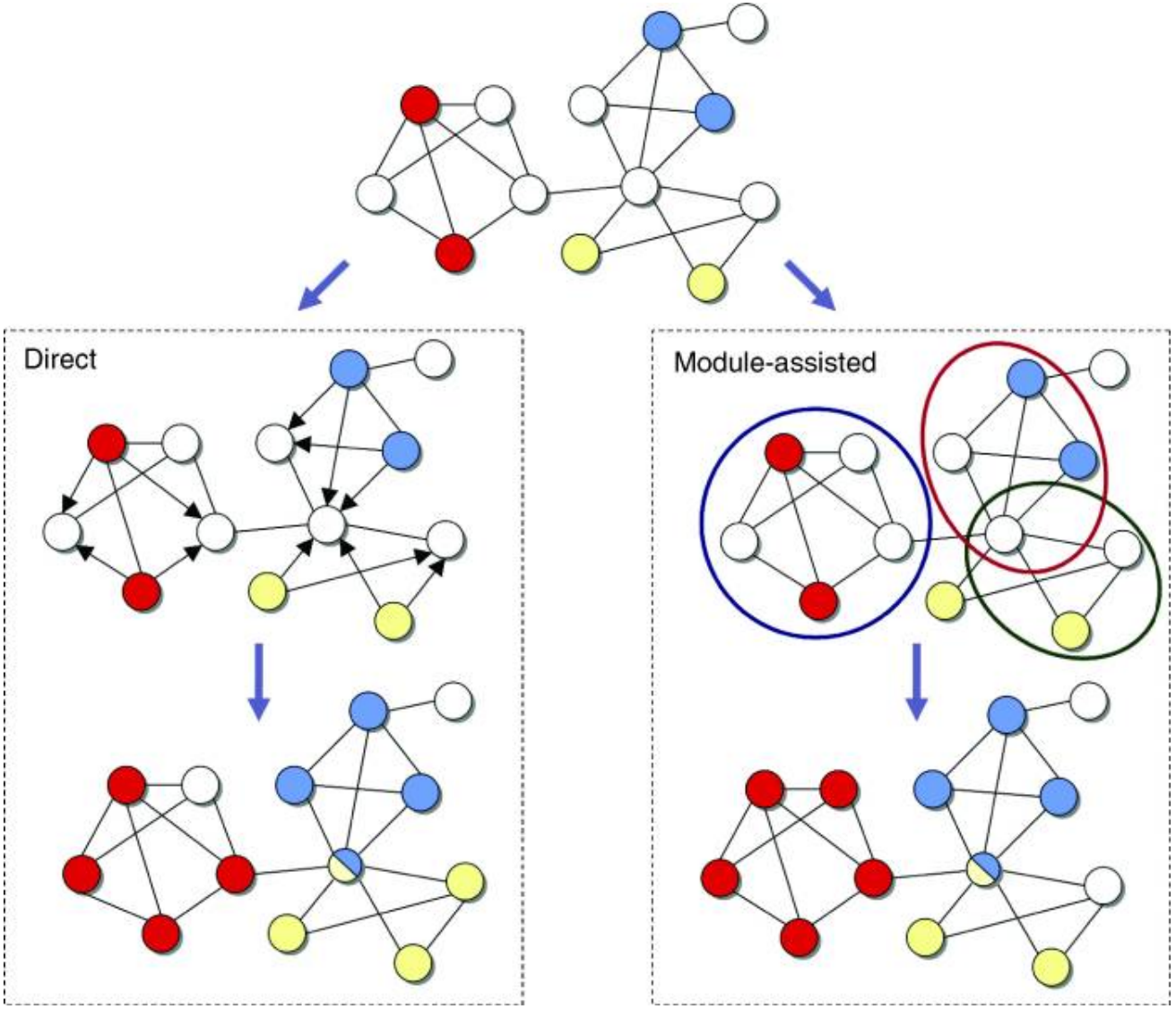
http://www.ncbi.nlm.nih.gov/pubmed/19399174

Brady et al. (2009) Plos One, 4(4) e5364

Outline

- Protein-Protein Interaction Model
- How to get PPI
  - Experimental methods
  - Bioinformatic methods
- PPI databases
- Network properties
- Analysis method and applications

Function prediction

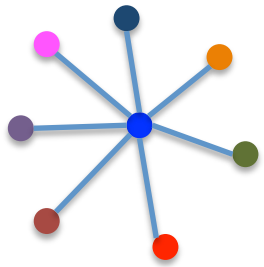


Function prediction

- **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**

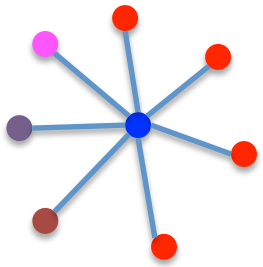
Neighborhood based methods

- 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.

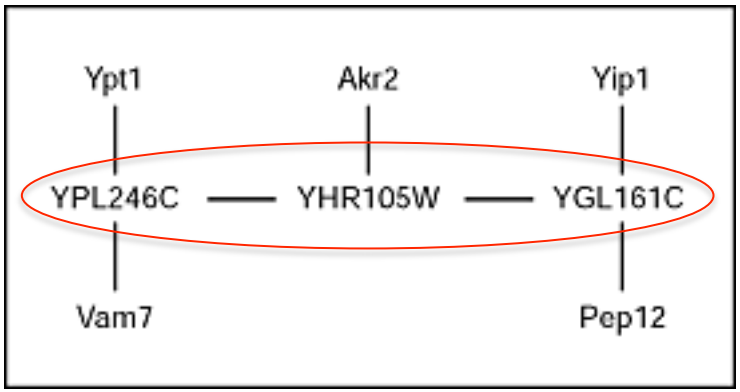


4 Red neighbors, that is larger then the threshold 3

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

Prediction of function by direct and indirect protein interactions

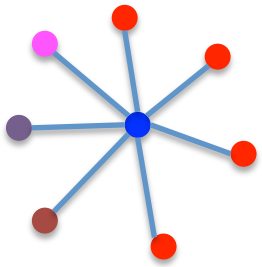
- YHR105W, YPL246C, and YGL161C are proteins of unknown function. Akr2 is a protein involved in endocytosis and therefore suggests a function for YHR105W. This potential function is supported by indirect interactions with Ypt1, Vam7, Yip1, and Pep12, which have been also implicated in vesicular transport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

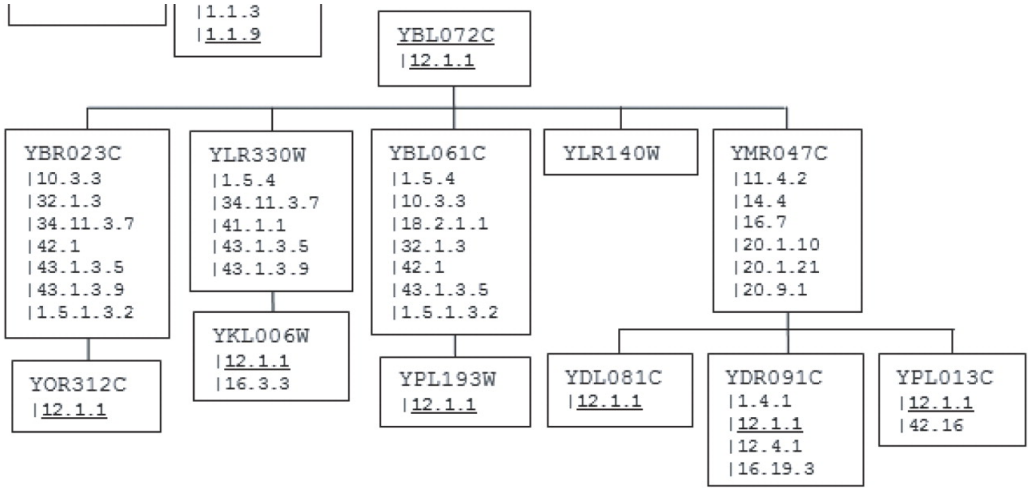


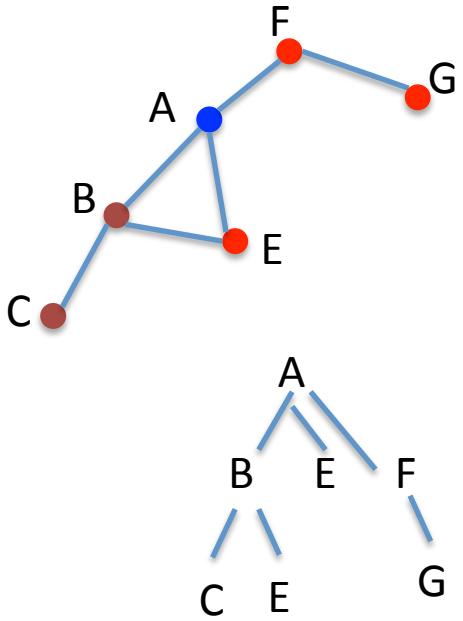
$n_f = 4$ for red function

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

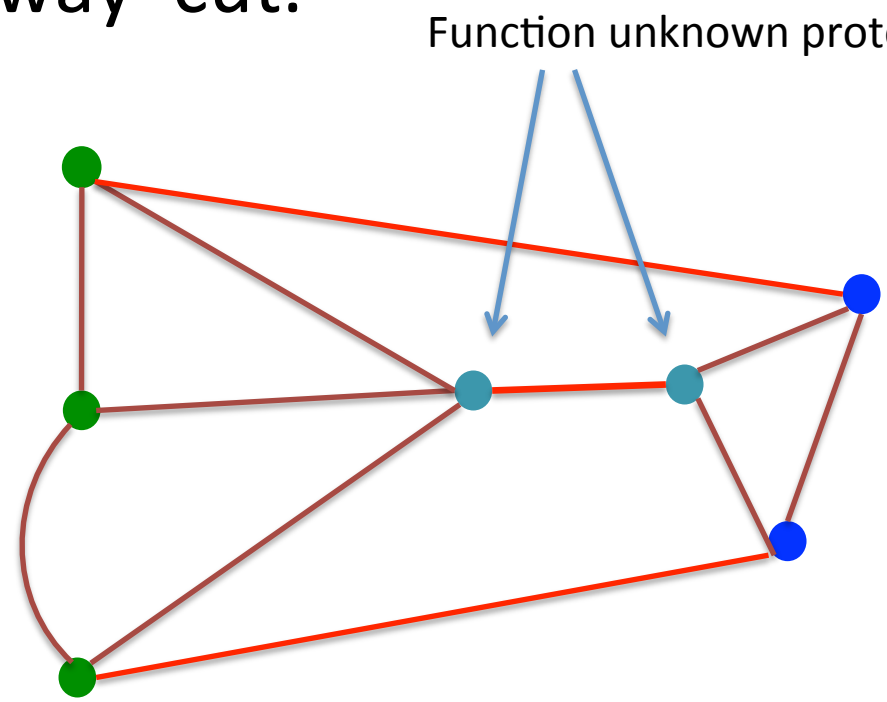
- **Direct Methods**
  - **Neighborhood based Methods**
  - **Graph theory methods**
  - **Probabilistic Methods**
- **Module assisted methods**
  - **General Methods**
  - **Hierarchical clustering based**
  - **Graph clustering methods**
  - **Expansion of complex seeds**

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

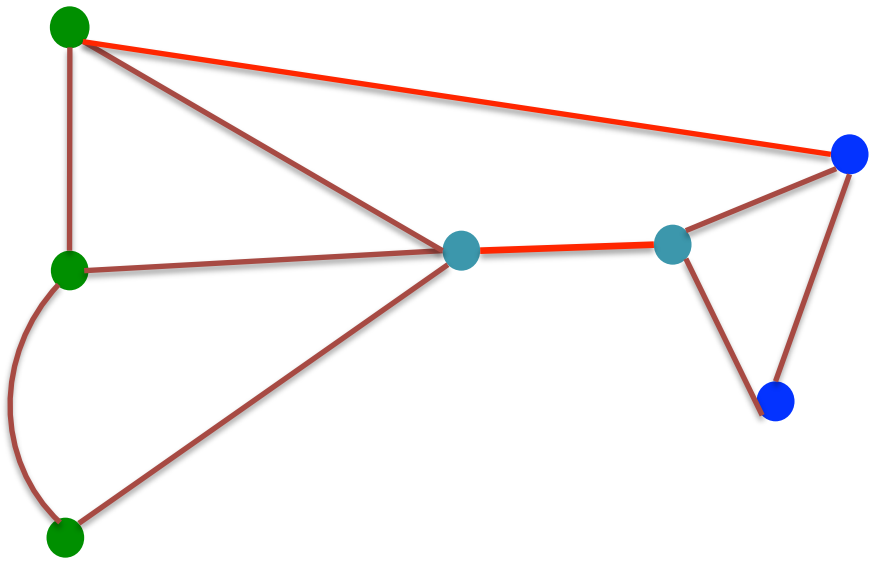
- Minimum multi-way cut.



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

Graph theory Methods

- Minimum two-way cut.

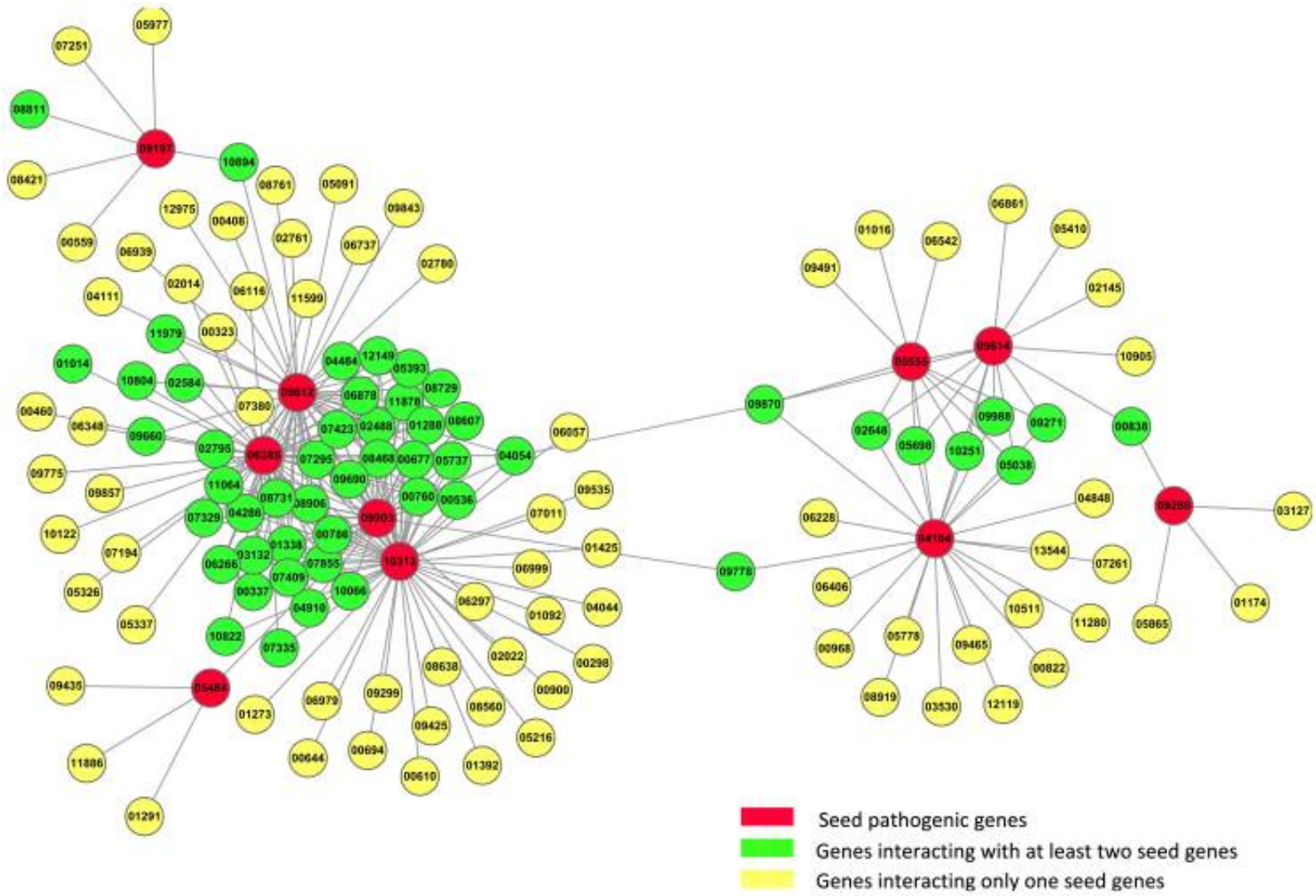


Karaoz et al (2004)

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

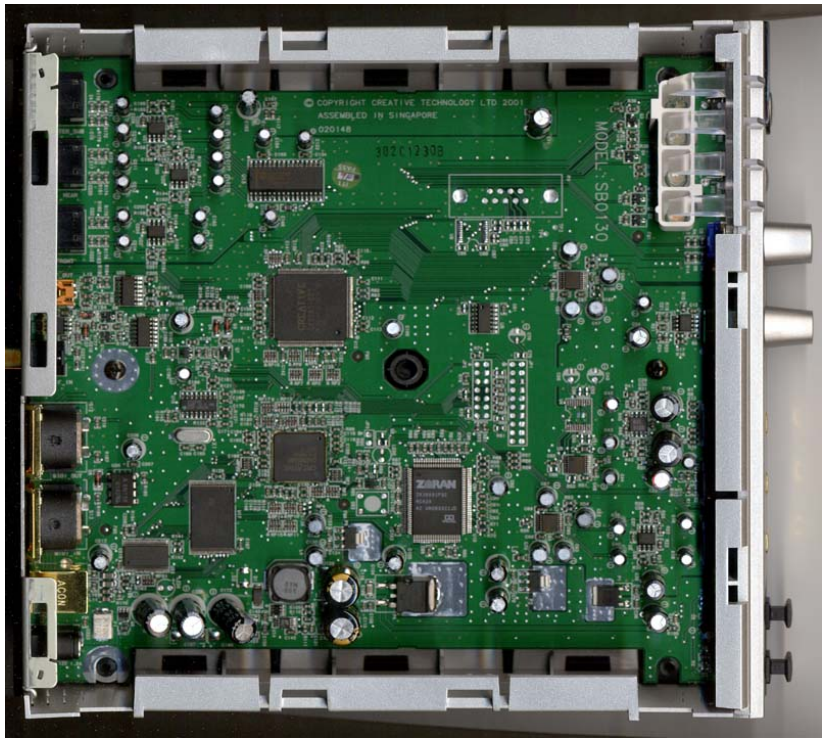


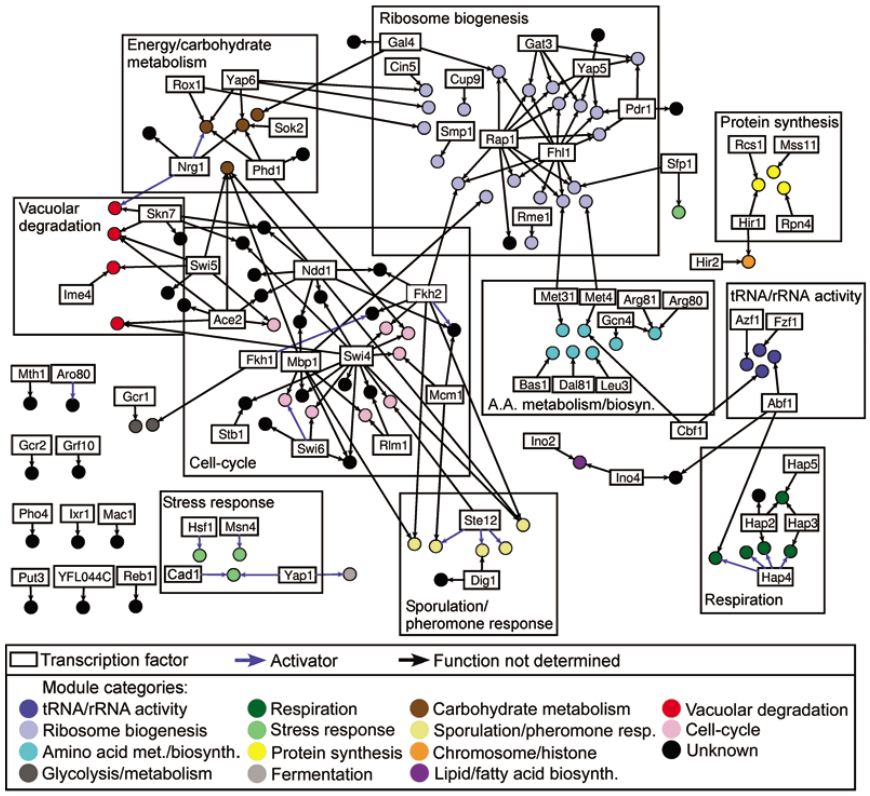
Function prediction

- **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





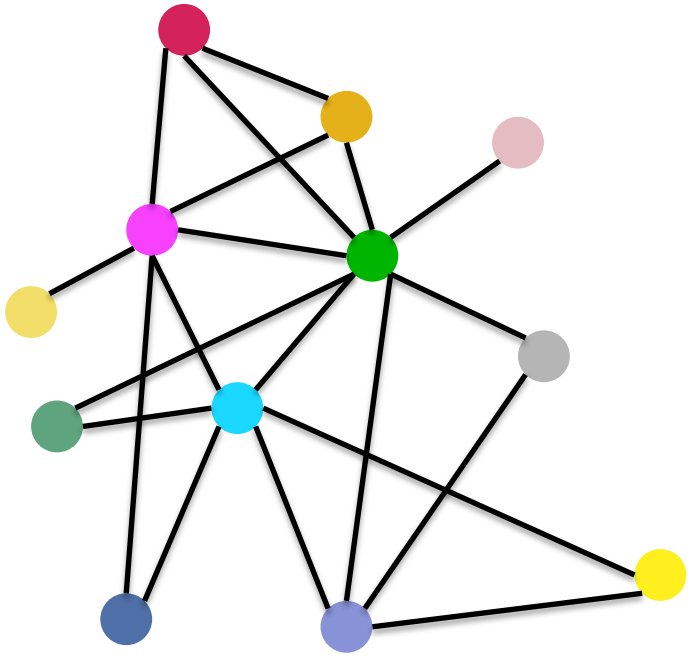
Computer Circuit Boards

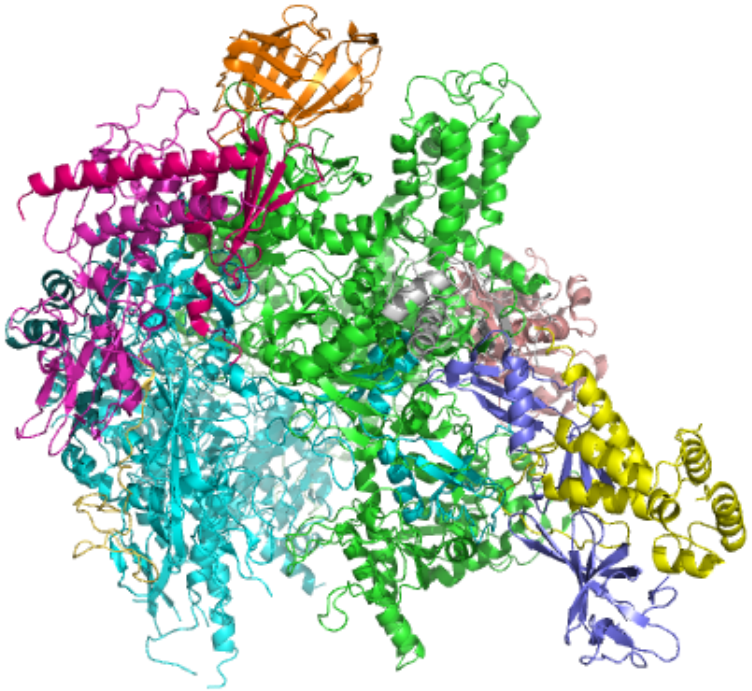
Transcriptional regulatory network

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.