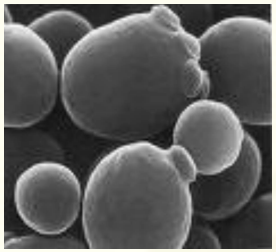
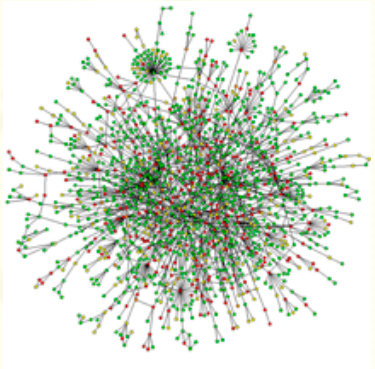
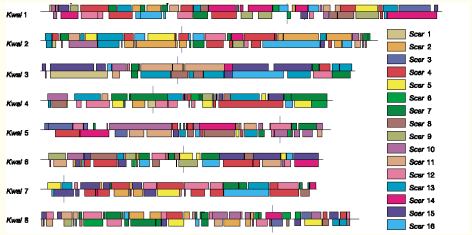
Lecture 4: Yeast as a model organism for functional and evolutionary genomics

Part II







A brief review

- **What have we discussed:**
- Yeast genome in a glance
- Gene expression can tell us about yeast functions
- Transcriptional regulation.
  - How to find transcription factor binding sites ?
  - TF sites turnover, evolution of the gene regulatory network

- **Next:**
- **Proteomics:** Protein-protein interaction and network
  - Experiments and analysis
  - Biological network analysis – a primer
- **Genetic interactions**: SGA technology
- **Gene and genome duplication** in yeast

What is Proteomics

- Proteomics (1997): large-scale study of proteins in a high-throughput manner.

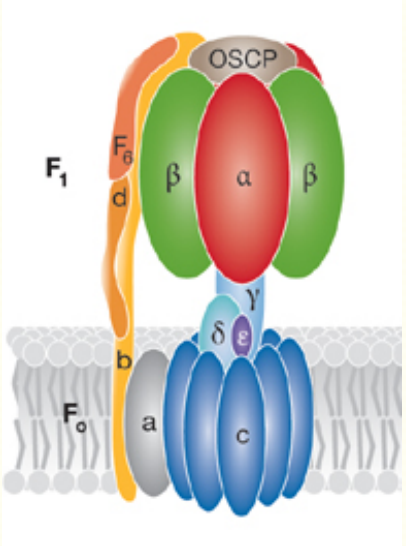
- **Protein-protein interactions, protein complexes,**
  - Yeast Two-hybrid, and mass spectrometry
- **Post-translational modifications,** e.g. phosphorylation
  - Mass spec, protein binding array
- **Protein abundance and half-life**
  - GFP (green florescence protein), microscopy or flow cytometry
- **Protein sub-cellular localization**
  - GFP and high content microscopy

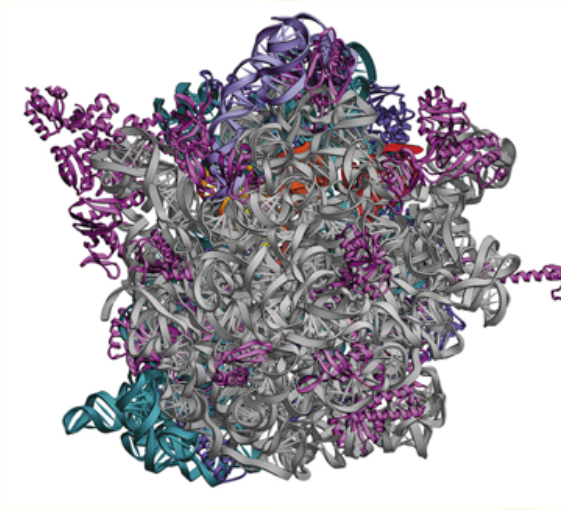
Protein-Protein interactions and Protein complexes

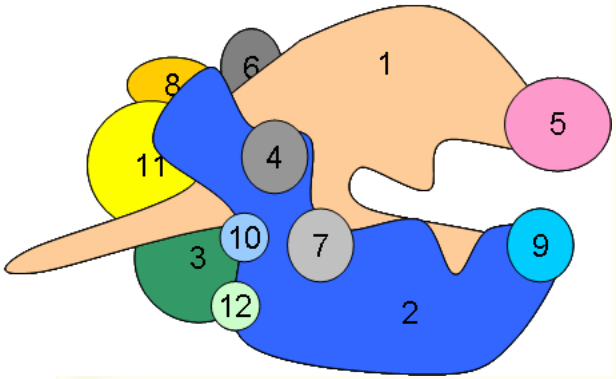
- The majority of the proteins in the cell form a complex or have stable interactions with another protein or with themselves, very few proteins work alone in the cell.

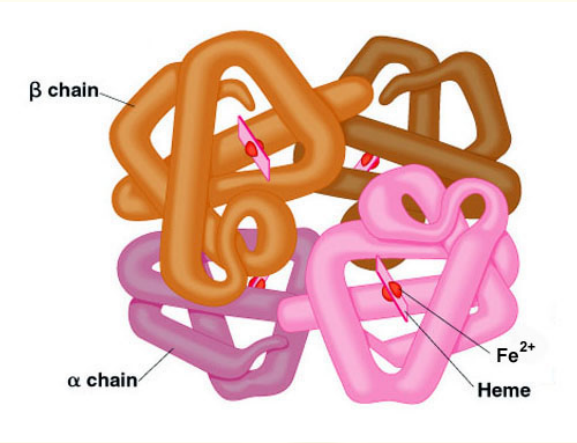
- **Permanent protein complexes:**
  - homo-dimer, hetero-dimer, trimer, tetramer, multi-subunit complex
  - Gene expression are highly correlated

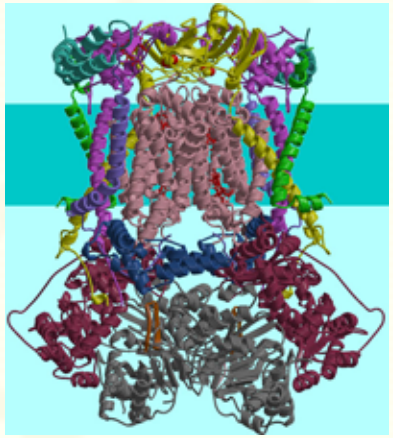
- **Transient protein-protein interactions:**
  - For example: kinase / phophatase and substrates
  - No correlation in gene expression.

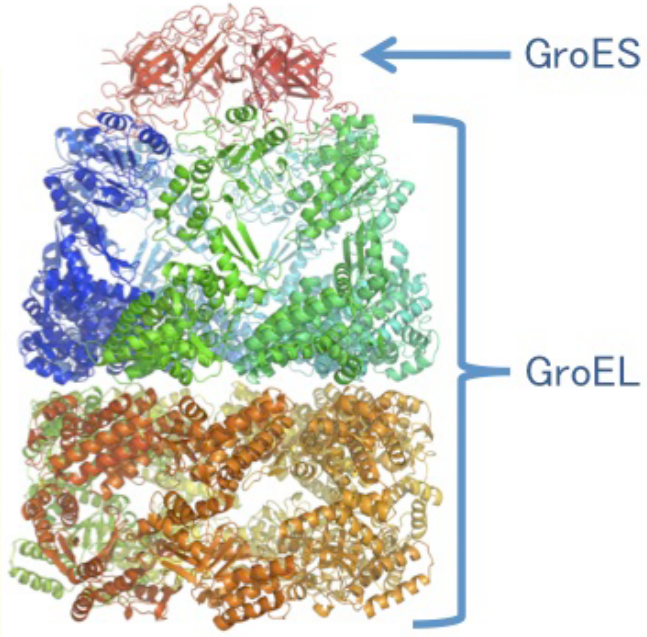








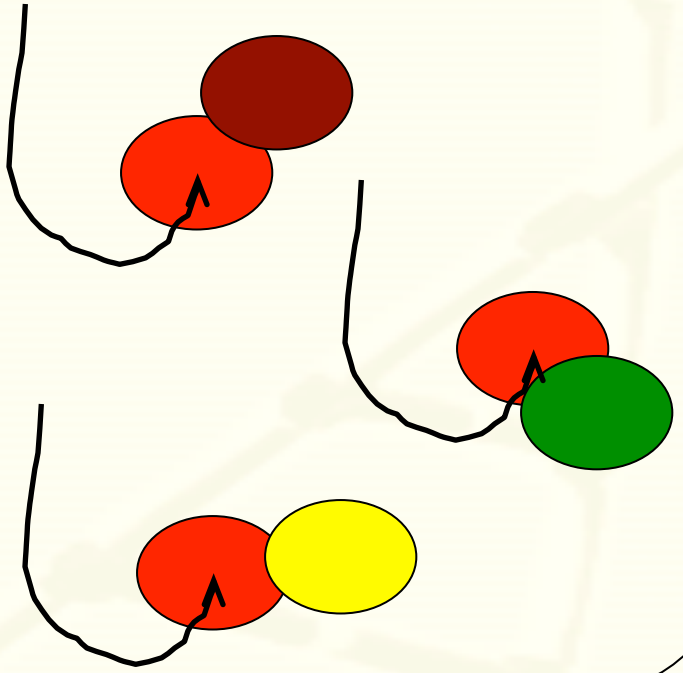




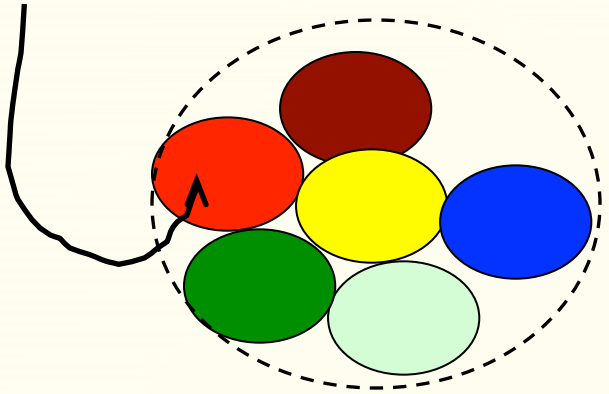
Examples of stable protein complexes

Two complementary experimental approaches

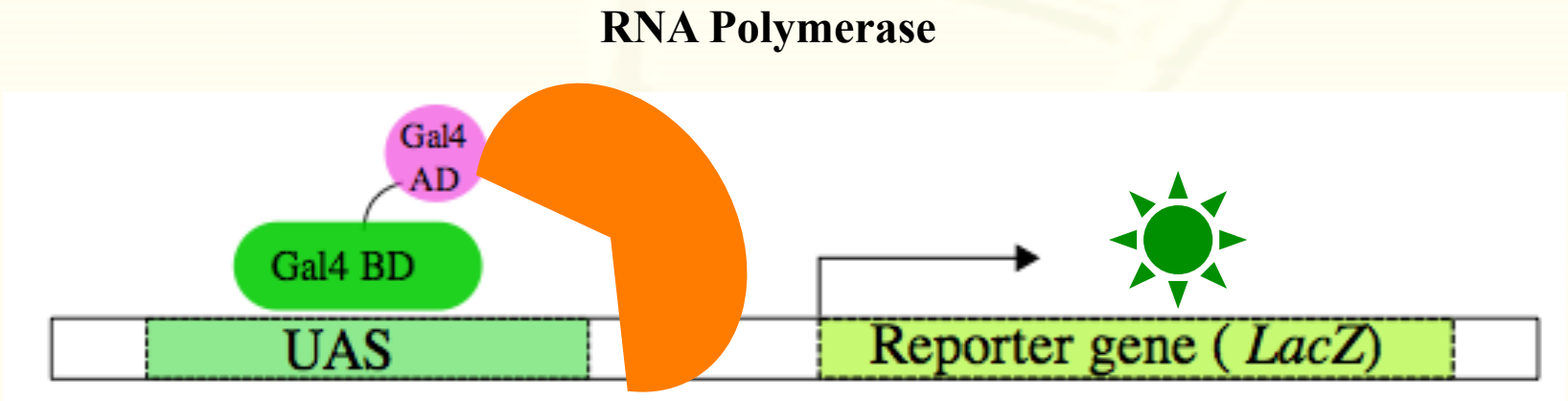
- **Yeast Two-Hybrid (Y2H)**: detect pair-wise interactions, use one protein as "bait" and detect one "prey" at a time.



- **TAP-MS** (Tandem Affinity Purification followed by Mass spectrometry): Use one protein as "bait" and identify the entire protein complex

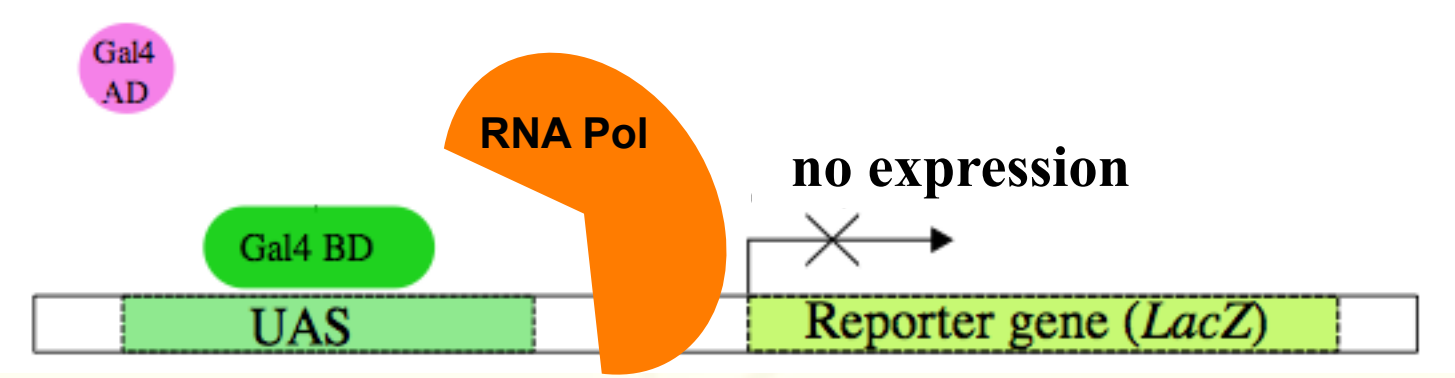


How does Yeast 2-Hybrid work ?

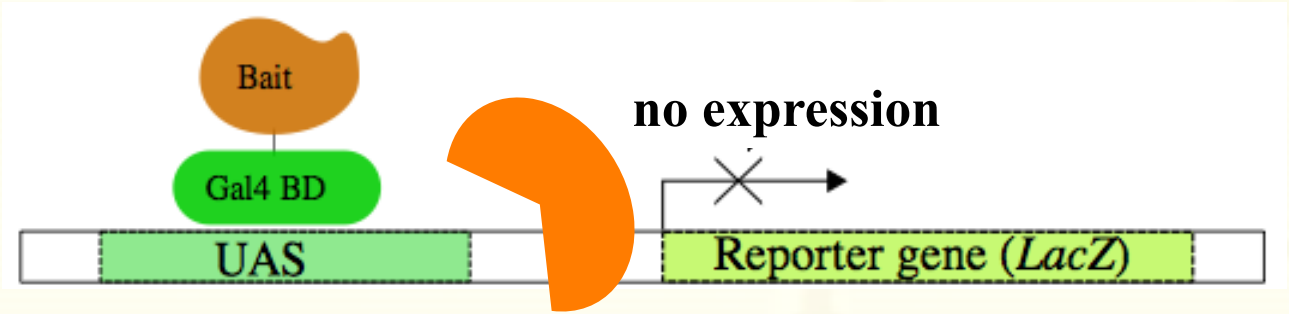


- Expression of the reporter gene (***LacZ***) depends on the binding of Gal4 transcription factor to the promoter.
- Gal4 consists of a **DNA Binding Domain (BD)** and an **Activation Domain (AD)**.
- The **BD** binds to promoter, and **AD** binds to RNA polymerase, both domains are required to trigger gene expression.

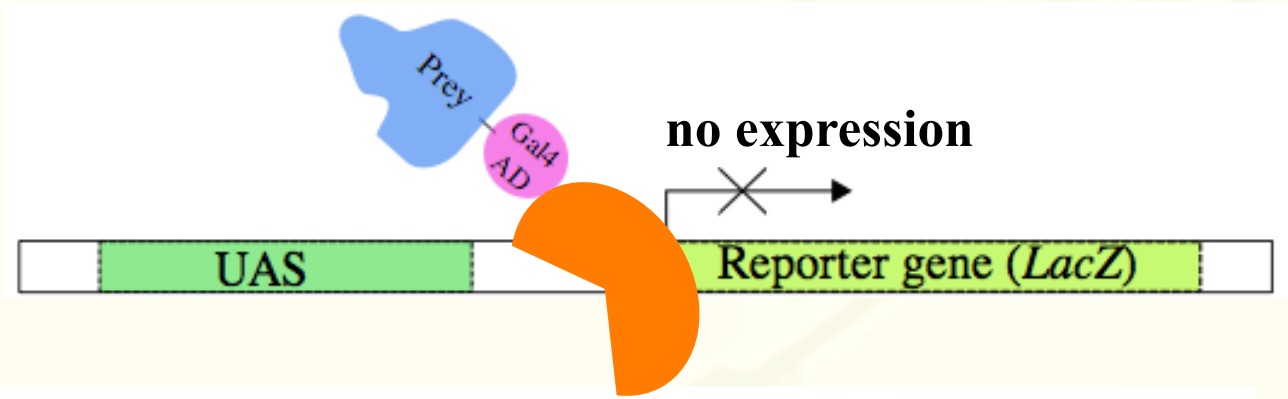
How does Yeast 2-Hybrid work ?



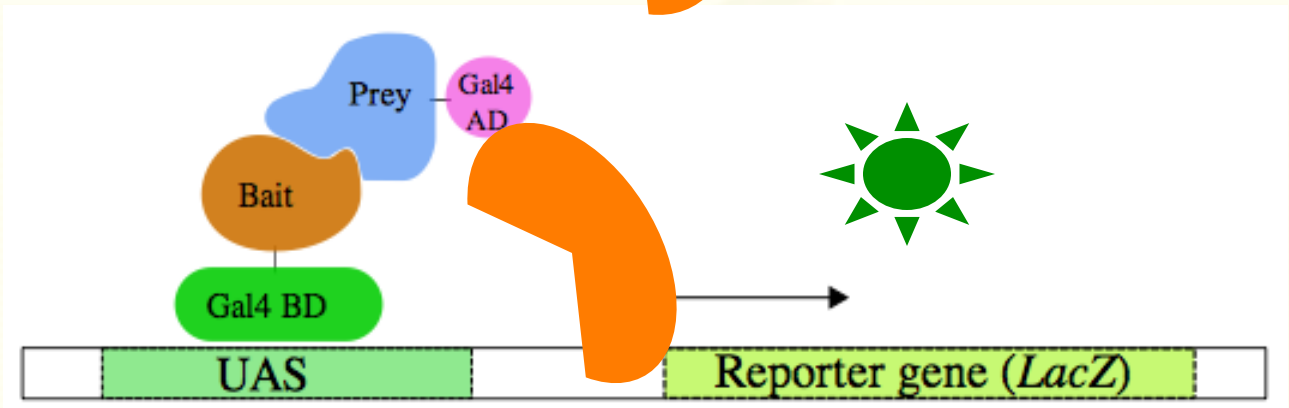
- Expression of the reporter gene (LacZ) depends on the binding of Gal4 transcription factor to the promoter.
- Gal4 consists of a **DNA Binding Domain (BD)** and an **Activation Domain (AD).**
- The **BD** binds to promoter, and **AD** binds to RNA polymerase, both domains are required to trigger gene expression.



Only Bait and DNA Binding Domain, no reporter expression

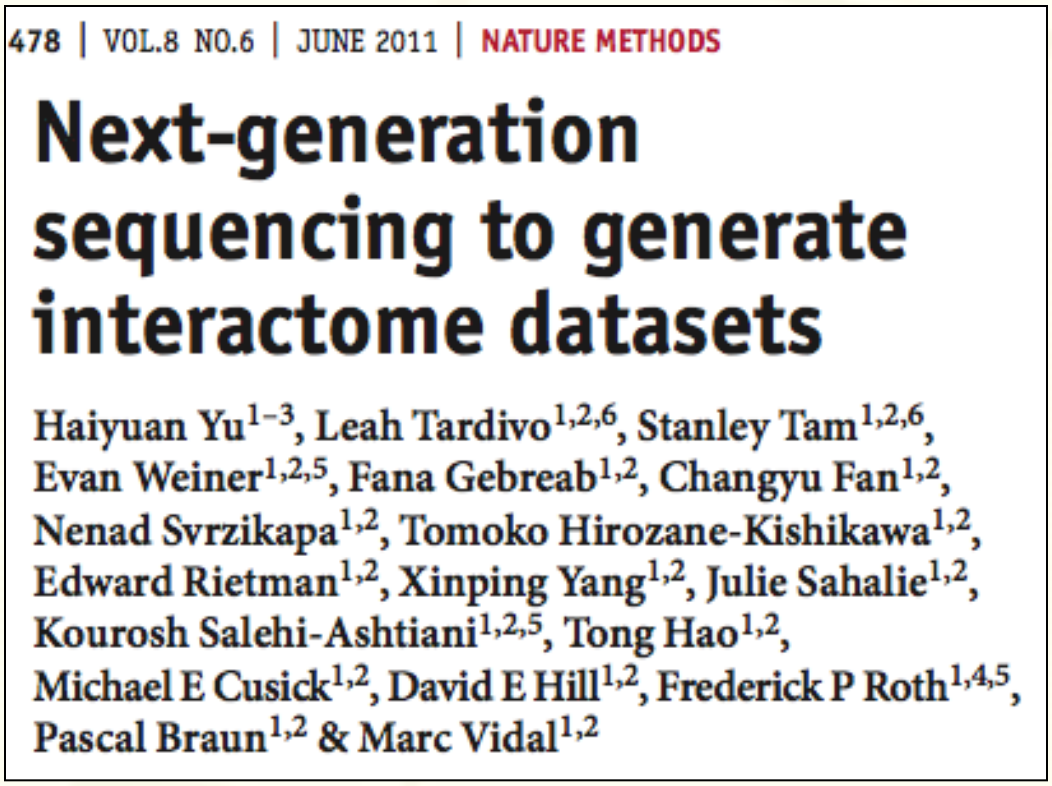


Only Prey and Activation domain, no reporter expression



Interaction between Bait and Prey triggers reporter expression

The next generation Y2H



478 | VOL.8 NO.6 | JUNE 2011 | NATURE METHODS

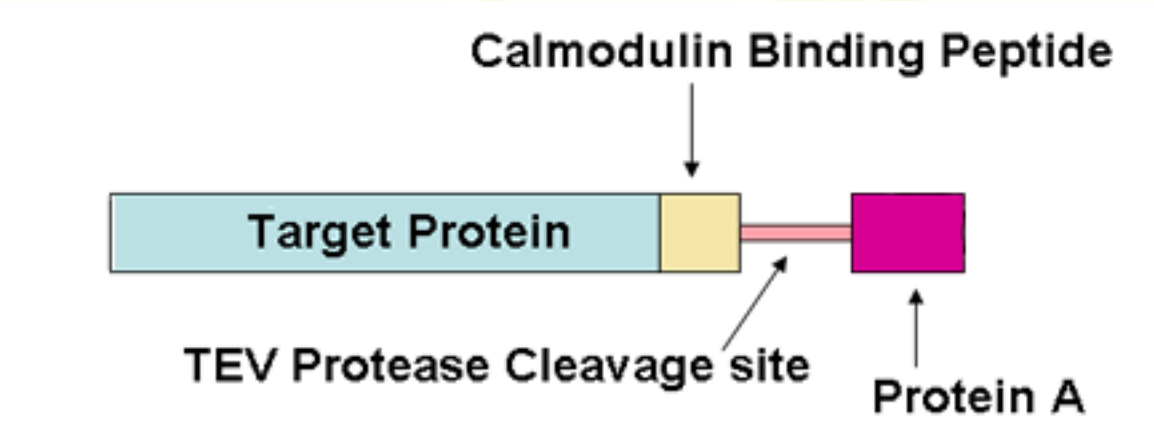
Next-generation sequencing to generate interactome datasets

Haiyuan Yu[1–3], Leah Tardivo[1,2,6], Stanley Tam[1,2,6], Evan Weiner[1,2,5], Fana Gebreab[1,2], Changyu Fan[1,2], Nenad Svrzikapa[1,2], Tomoko Hirozane-Kishikawa[1,2], Edward Rietman[1,2], Xinping Yang[1,2], Julie Sahalie[1,2], Kourosh Salehi-Ashtiani[1,2,5], Tong Hao[1,2], Michael E Cusick[1,2], David E Hill[1,2], Frederick P Roth[1,4,5], Pascal Braun[1,2] & Marc Vidal[1,2]

Use **next generation sequencing** and **barcoding** to measure gene expression level in parallel

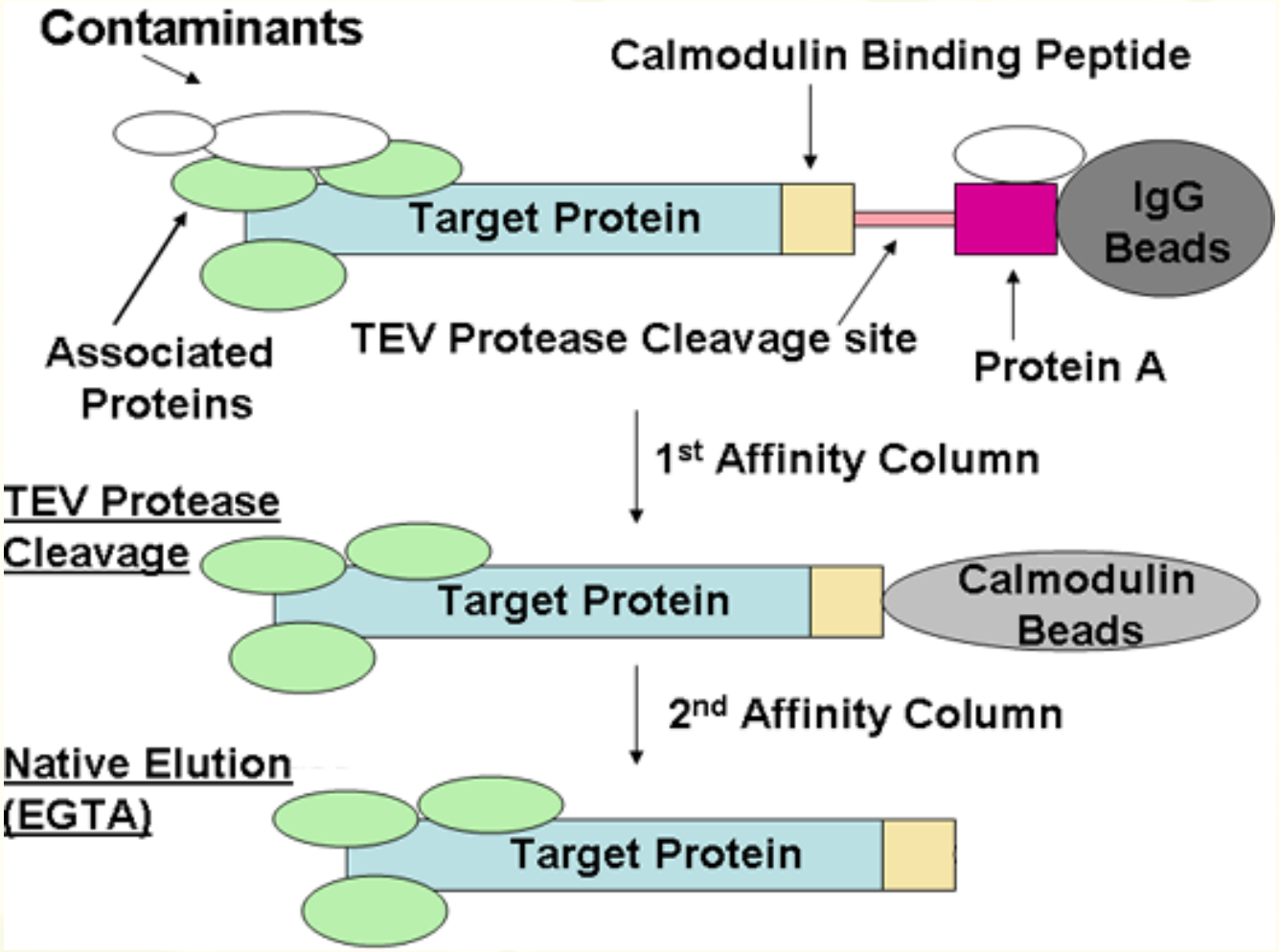
TAP-MS: Tandem Affinity Purification followed by Mass Spectrometry

- Tandem Affinity: adding two “tags” to the “bait” protein to improve purification.



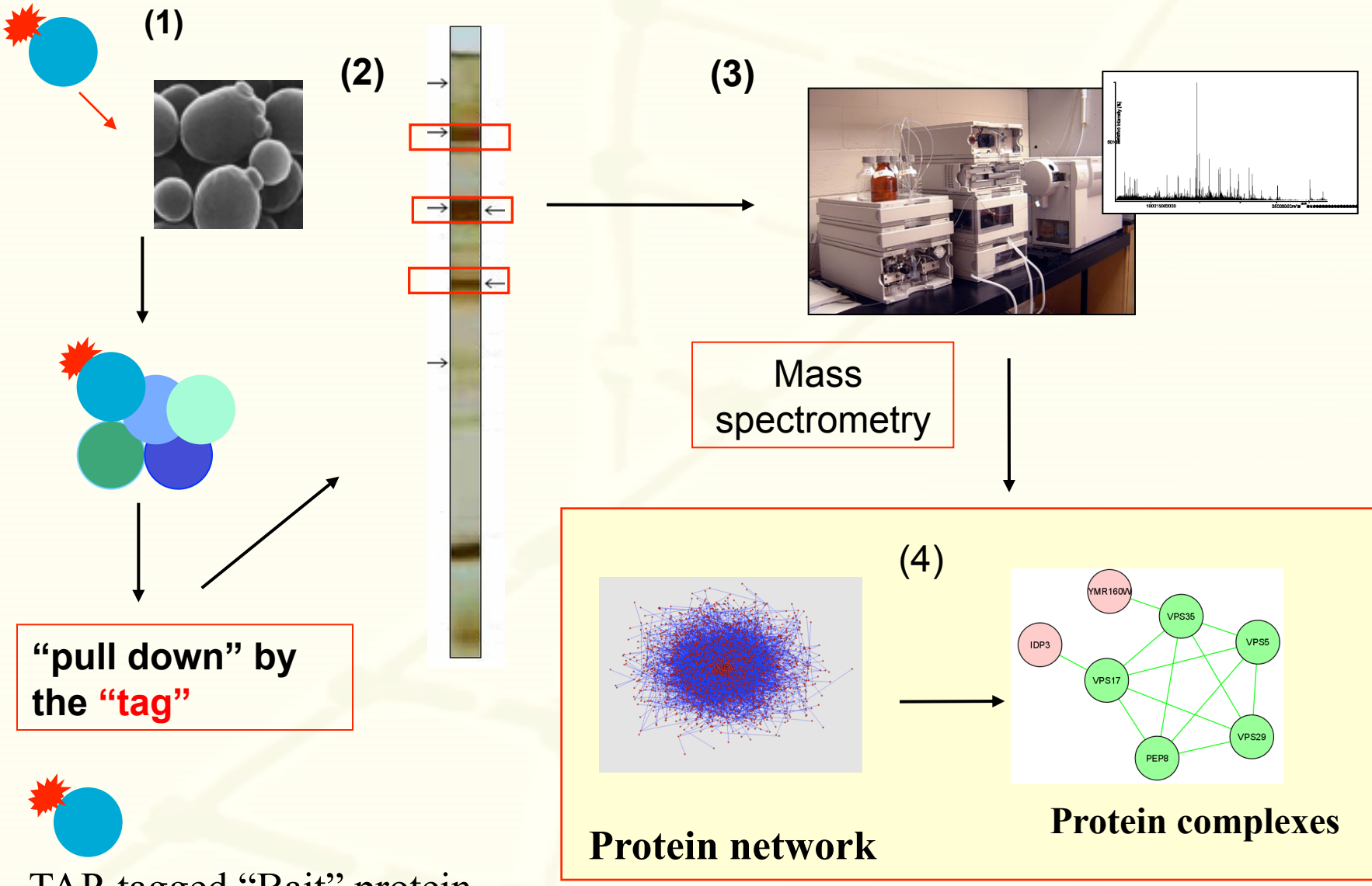
http://www.cellmigration.org

TAP-MS: Tandem Affinity Purification followed by Mass Spectrometry



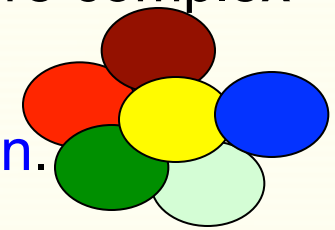
http://www.cellmigration.org

TAP-MS: Tandem Affinity Purification followed by Mass Spectrometry



Comparing Y2H and TAP-MS

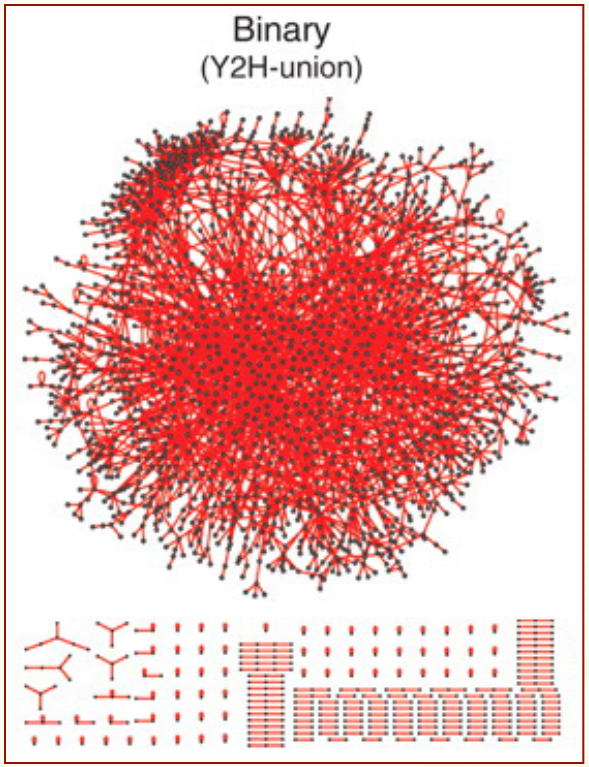
- They are two complementary and orthogonal methods
- Y2H is best for detecting pair-wise direct interactions, while TAP-MS is best for detecting the entire complex
- TAP-MS can not detect transient interaction.



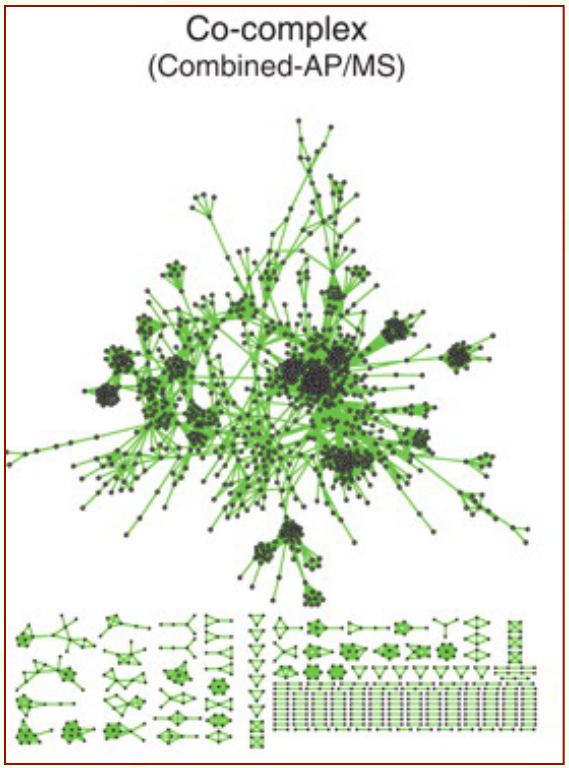
- Y2H can not detect indirect interactions
- TAP-MS can be adapted to identify RNA components of the complex.

different network topologies

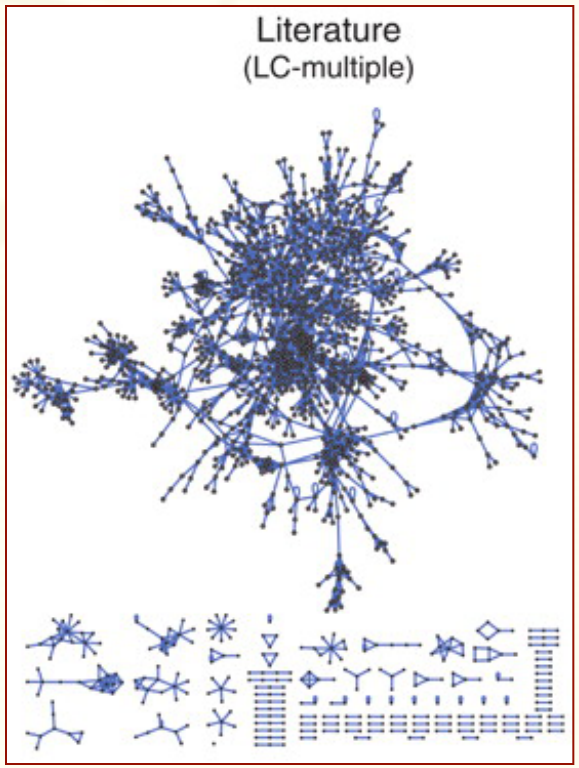
Binary Y2H



AP-MS



Literature curation



Yu, … Vidal Science d2008

Integrating Y2H and TAP-MS

- Sophisticated "machine learning" algorithms have been developed to integrate these two types of data, and other biological information (such as gene expression correlation) to reduce false positive rate, and to predict new interactions.

Science 17 October 2003:
Vol. 302 no. 5644 pp. 449–453
DOI: 10.1126/science.1087361

REPORT

A Bayesian Networks Approach for Predicting Protein–Protein Interactions from Genomic Data

Ronald Jansen[1,*], Haiyuan Yu[1], Dov Greenbaum[1], Yuval Kluger[1], Nevan J. Krogan[4], Sambath Chung[1,2], Andrew Emili[4], Michael Snyder[2], Jack F. Greenblatt[4] and Mark Gerstein[1,3,†]

Two landmark MS papers

Article

Nature **440**, 637-643 (30 March 2006) | doi:10.1038/nature04670; Received 20 December 2005; ;
Accepted 23 February 2006

Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*

Nevan J. Krogan[1,2,12,11], Gerard Cagney[1,3,12], Haiyuan Yu[4], Gouqing Zhong[1],

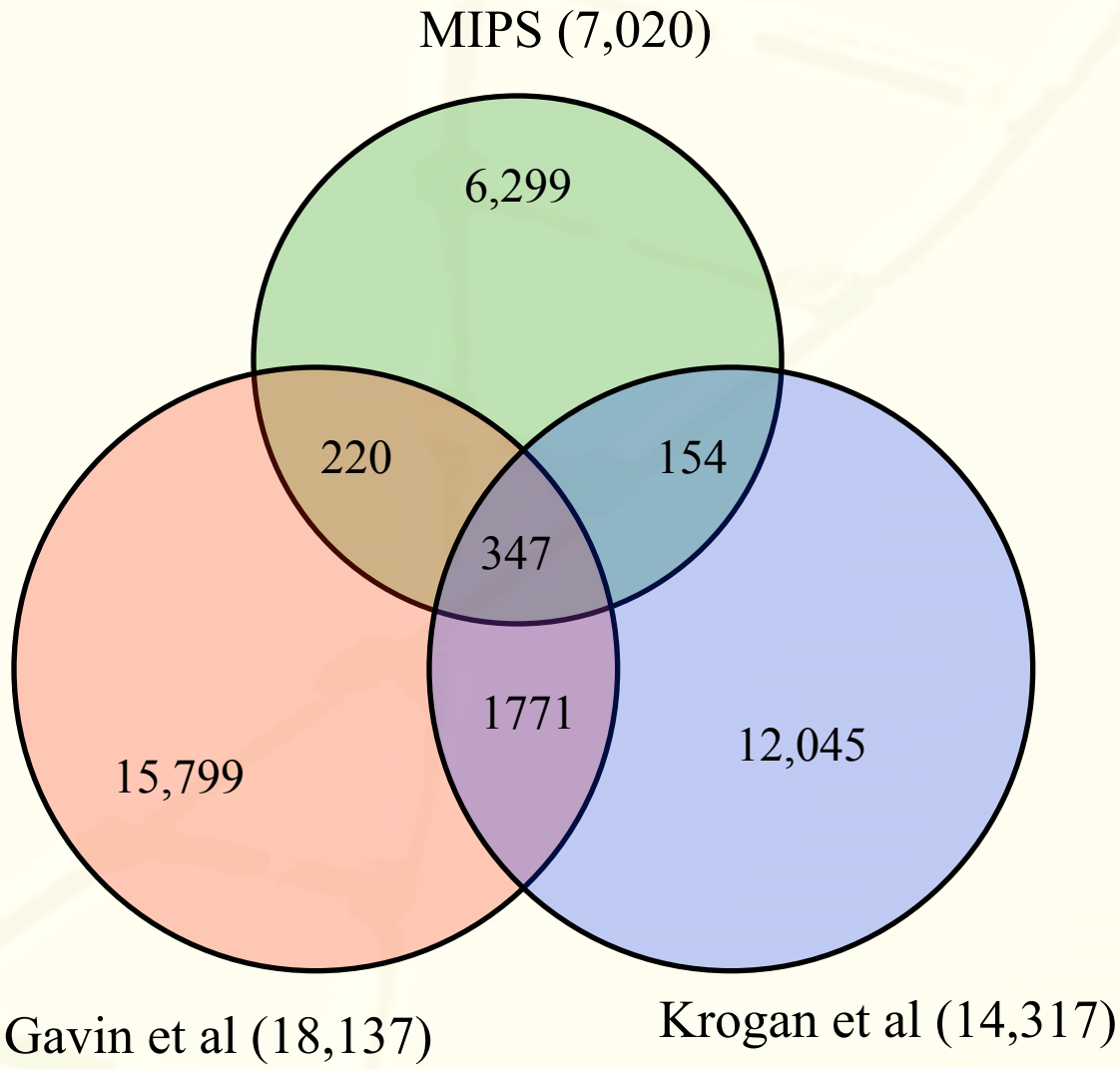
Article

Nature **440**, 631-636 (30 March 2006) | doi:10.1038/nature04532; Received 17 October 2005; ;
Accepted 15 December 2005

Proteome survey reveals modularity of the yeast cell machinery

Anne-Claude Gavin[1,6,5], Patrick Aloy[2,6], Paola Grandi[1], Roland Krause[1,3],

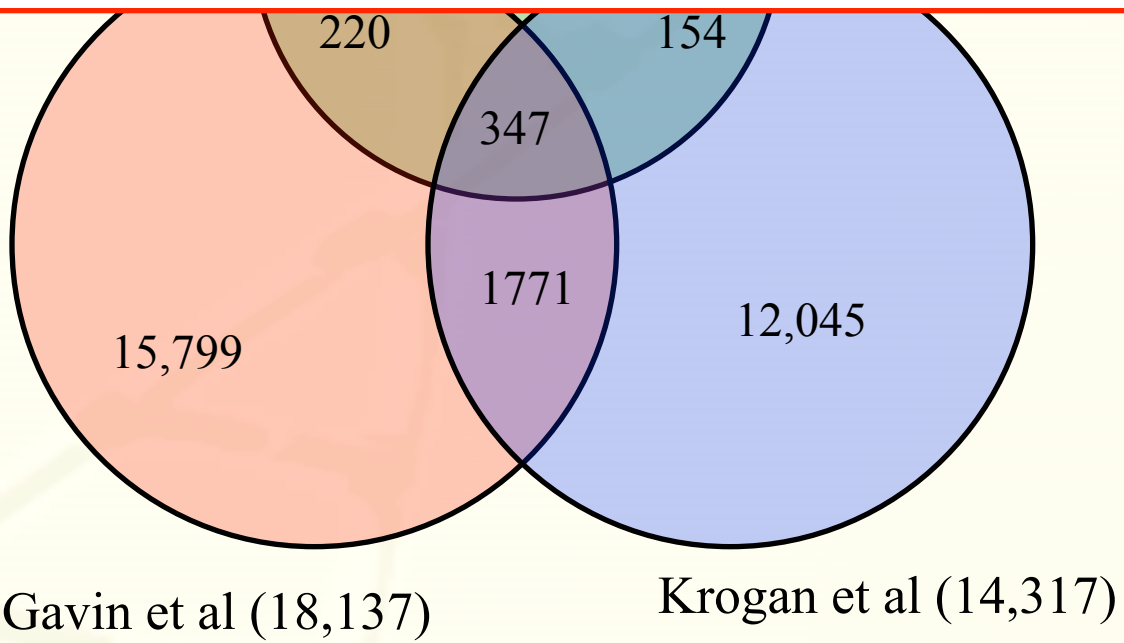
These two experiments reported very few common interactions



These two experiments reported very few common interactions

Possible reasons:

1. These two studies used different bait proteins.
2. Used different statistical threshold in "calling an interaction"
3. It is likely that these studies only surveyed a small portion of the entire "interactome"

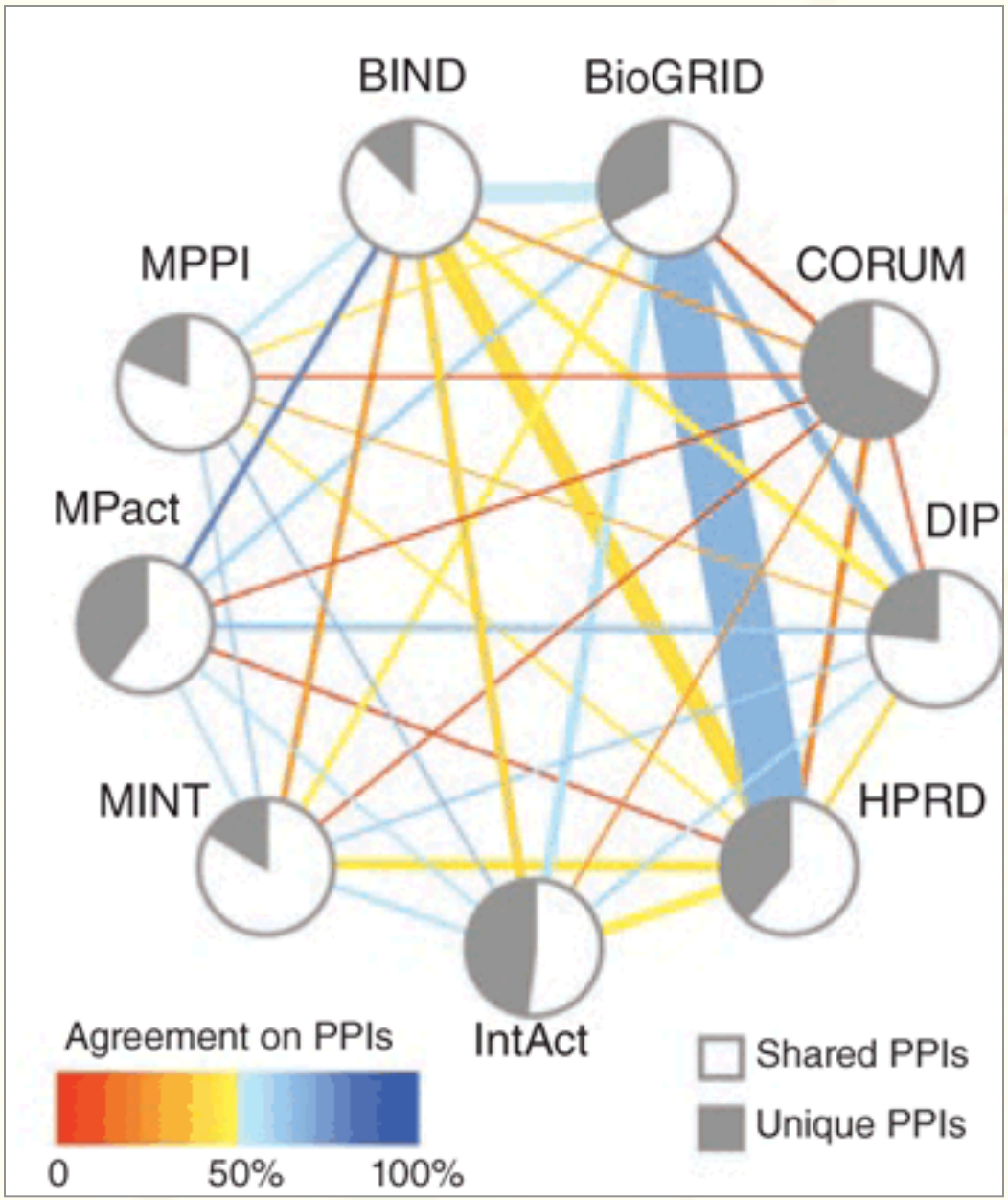


Database for protein-protein interactions

- **DIP**:http://dip.doe-mbi.ucla.edu/
- **MINT**: http://mint.bio.uniroma2.it/
- **BIND**: http://bind.ca/
- **HPRD**: http://www.hprd.org/
- **IntAct**: http://www.ebi.ac.uk/intact
- **BioGrid**: http://thebiogrid.org/
- **Ophid**:http://ophid.utoronto.ca/ophidv2.201/

- **iRefweb**: http://wodaklab.org/iRefWeb/
- **iMEx** (international Molecular Exchange Consortium)
  - http://www.imexconsortium.org/

However, these databases are not always consistent

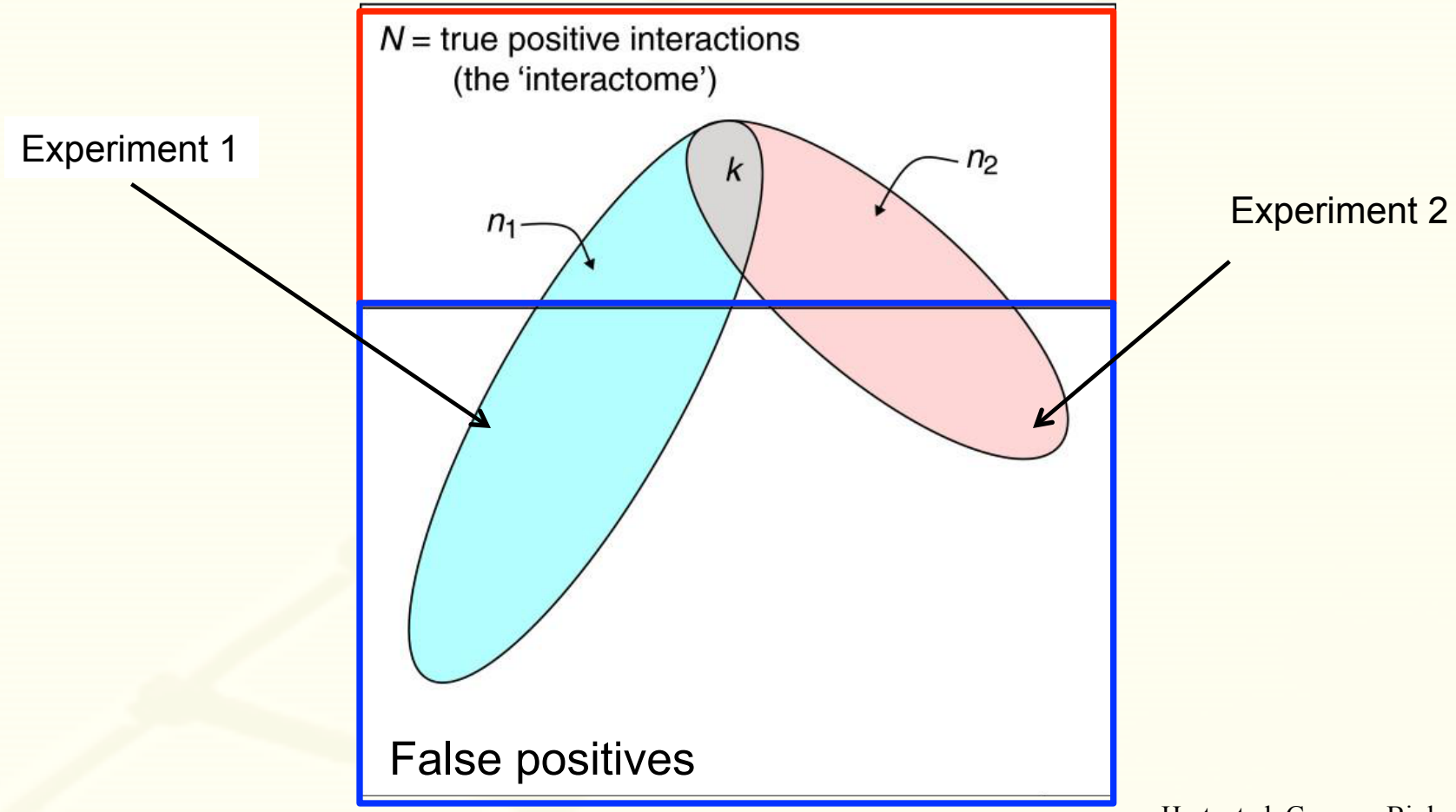


Inconsistency between interaction databases

Turinsky Nature Biotech 2011

Estimate the size of yeast interactome

- **Assumptions:** the inconsistency between different experiments is mostly because each experiment only sampled a small portion of the entire "interactome". Therefore using the sampling theory we can estimate the total number of interactions.



Hart, et al. Genome Biology 2006

Estimate the size of yeast interactome

- **Assumptions:** the inconsistency between different experiments is mostly because each experiment only sampled a small portion of the entire "interactome". Therefore using the sampling theory we can estimate the total number of interactions.

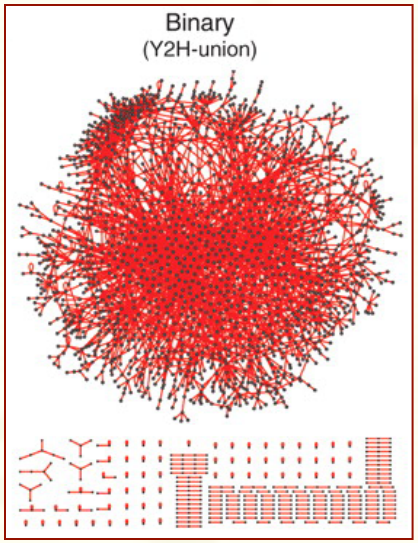
- **In Yeast:**
- Maximum possible interactions: 5800 X 5800 /2 = **16,820,000**
- Estimated interactions in budding yeast: 37,800 - **75,500**
- Current known interactions (BioGrid): **61,459**

- **In Human ***
- Maximum possible interactions: 22,000 X 22,000 /2 = 242 million
- Estimated interactions in human: **154,000-369,000**
- Current known interactions (BioGrid): **10,290**

- * ignoring alternative splicing

Hart, et al. Genome Biology 2006

A brief tutorial on biological networks



Analysis of Protein interaction network

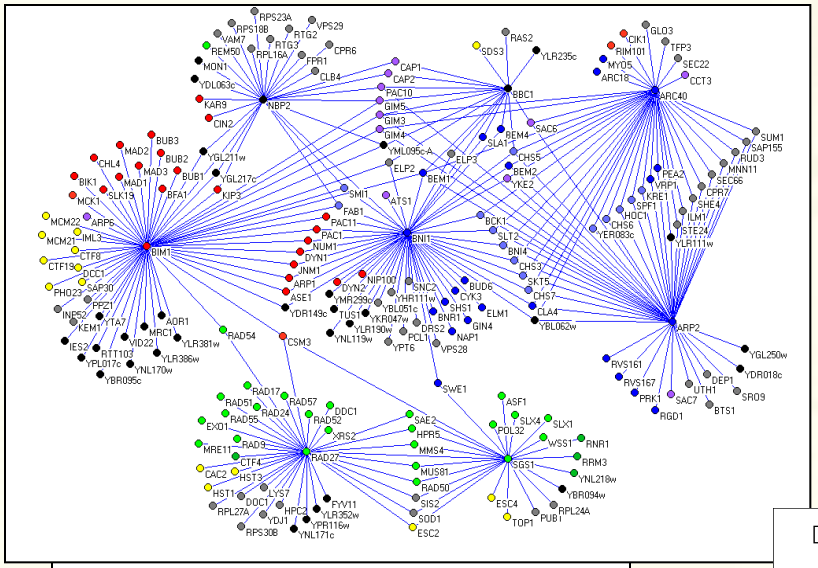
- **Topological analysis**
  - Degrees, hubs, modularity etc
- **Network dynamics**
  - Integrate network with gene expression data
- **Evolutionary analysis**
  - Conservation of protein sequence, interactions, complexes, and modules.
- **Robustness, noise buffering, evolvability, etc …**

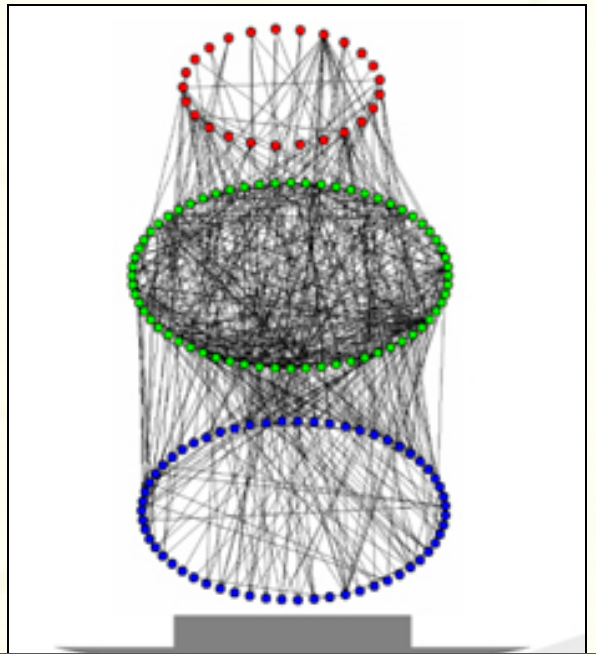
Biological networks

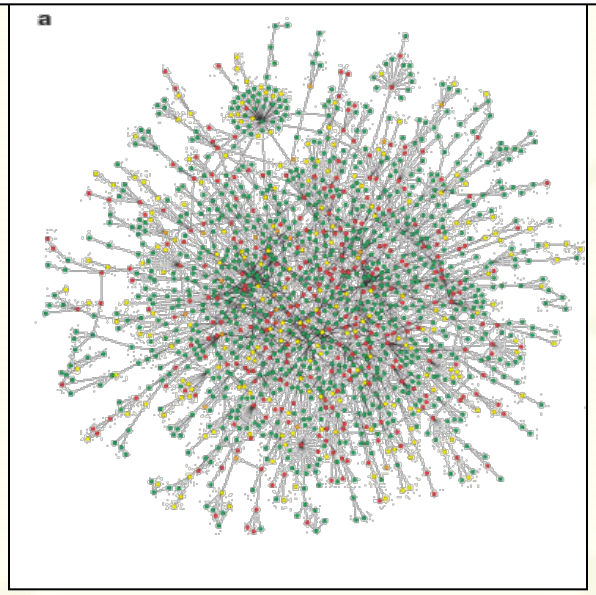
- **Why use a network approach ?**
  - Because this is how cells work !
  - Because molecules often interact with several other molecules to fulfill their functions, e.g. enzyme-metabolites, protein – proteins, protein – DNA, genetic interactions.
  - We need to network approach to understand the cellular pathways, signal transduction etc.

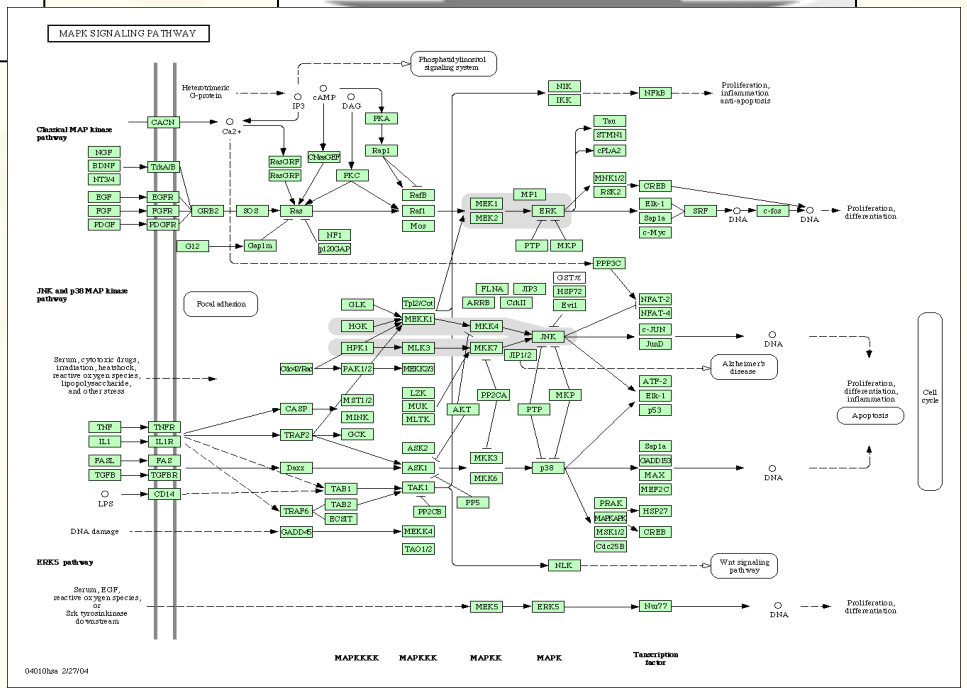
“The whole is greater than the sum of parts”.

Biological networks

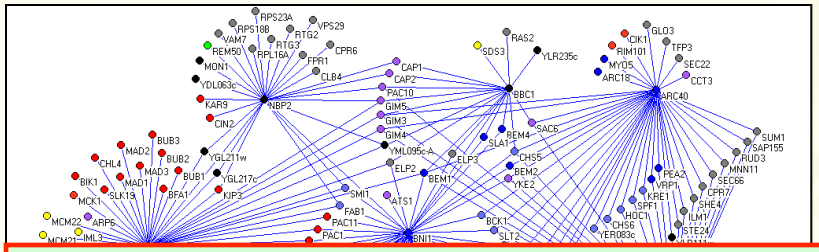


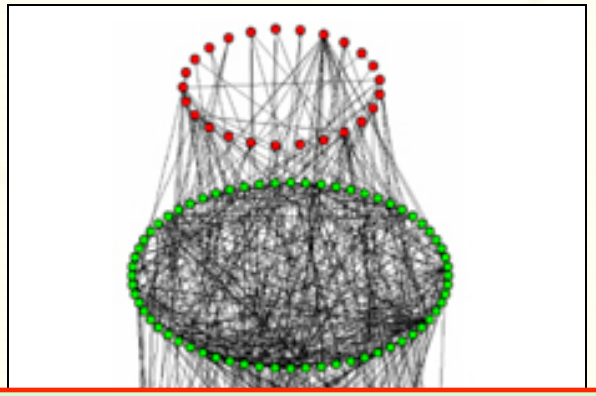




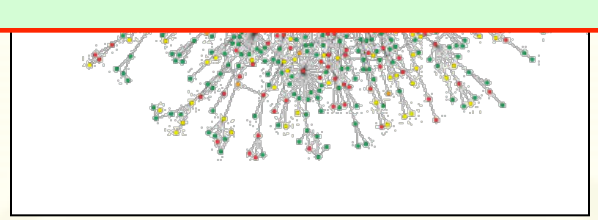


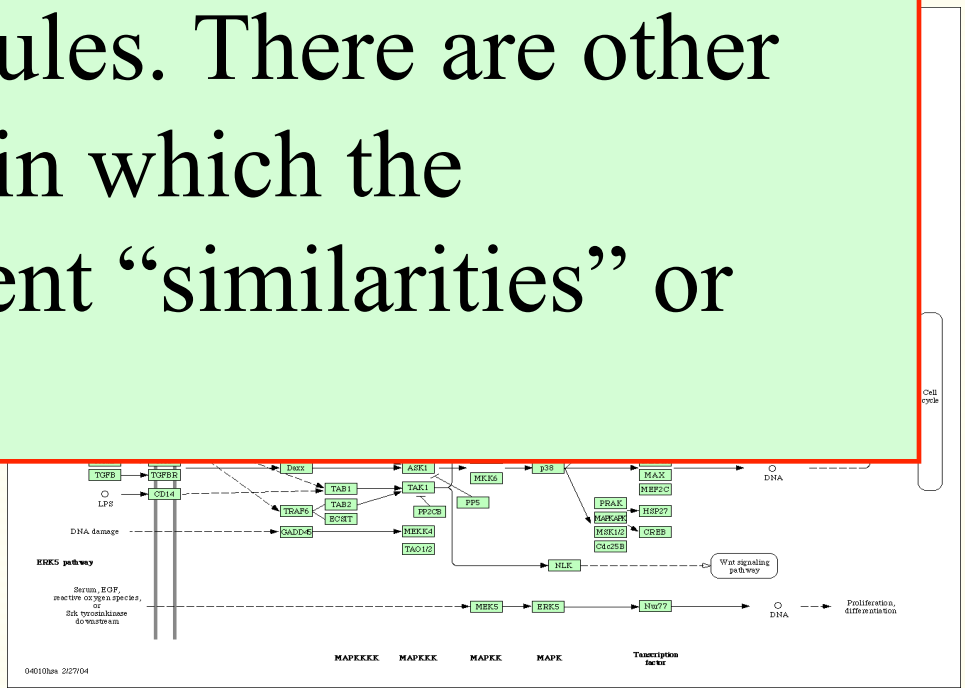
Biological networks



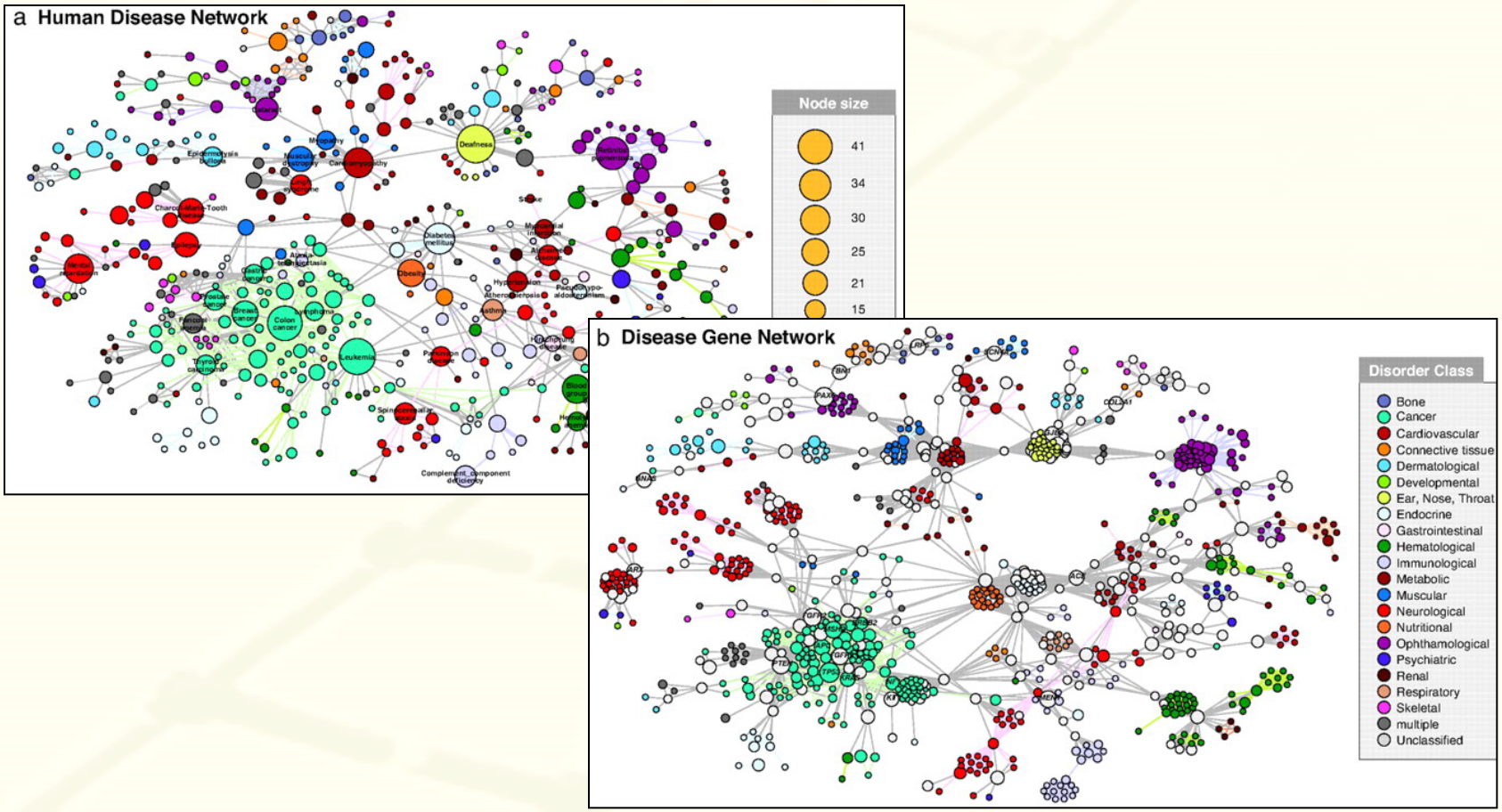


These are all real networks in which the edges represent biological interactions between bio-molecules. There are other types of networks, in which the connections represent “similarities” or “associations”.

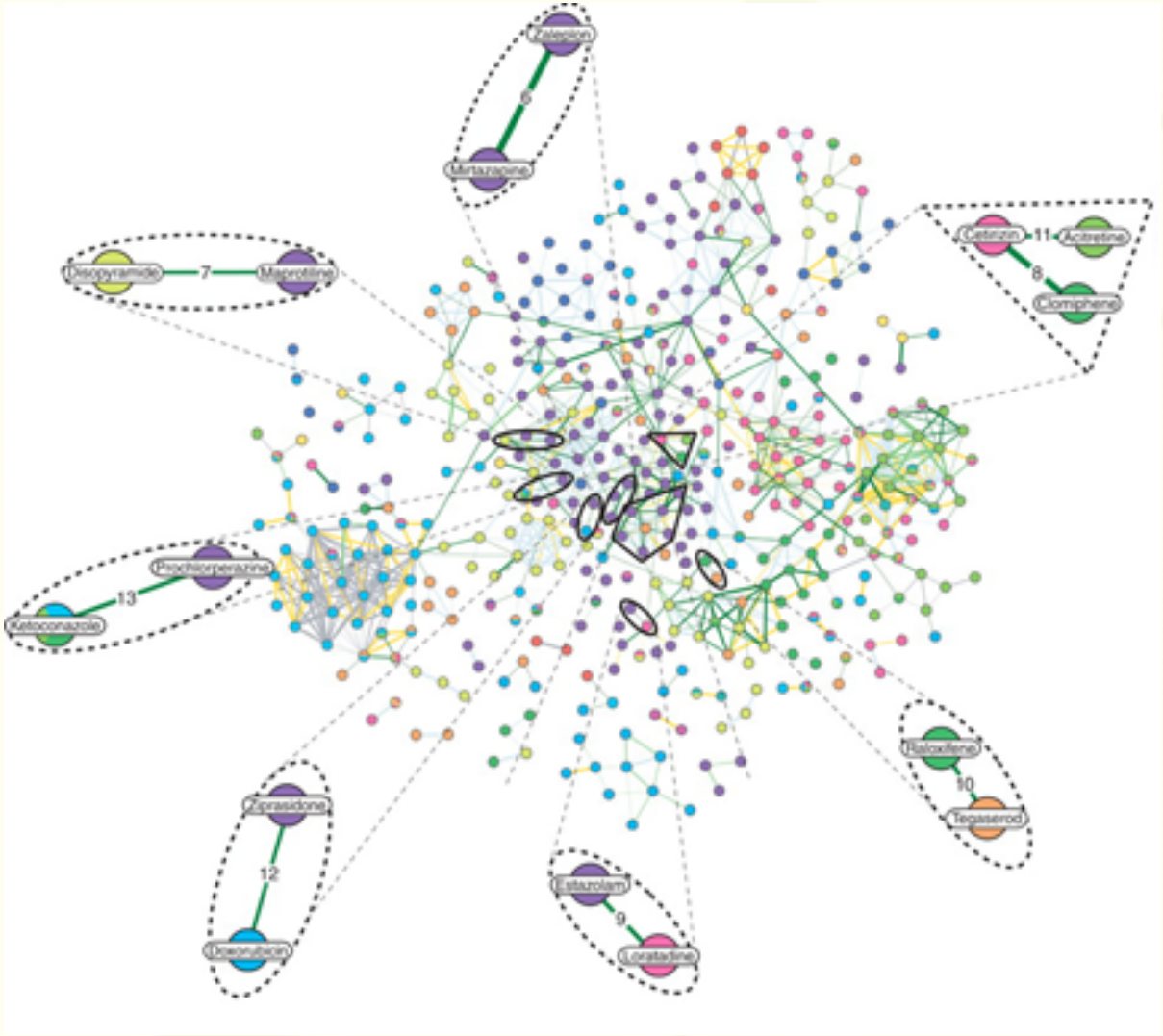




Network approach is an efficient and intuitive way to analyze and visualize relationships and similarities among a large number of subjects.

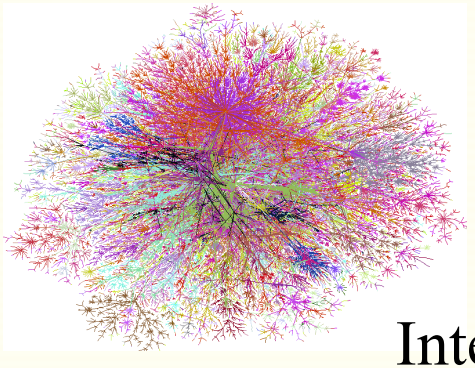


Another example: a network of drug side-effect similarities

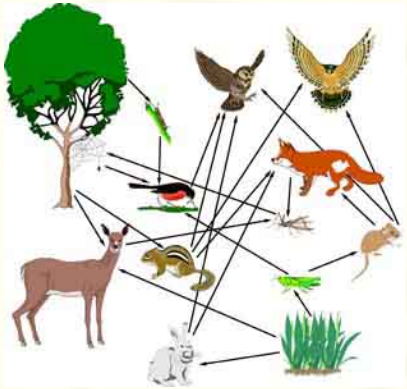


Campillos, Kuhn, Gavin, Jensen, Bork **Drug Target Identification Using Side-Effect Similarity** Science 2008

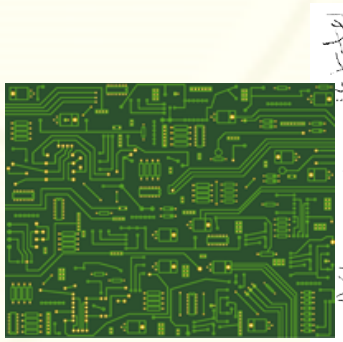
Networks are everywhere



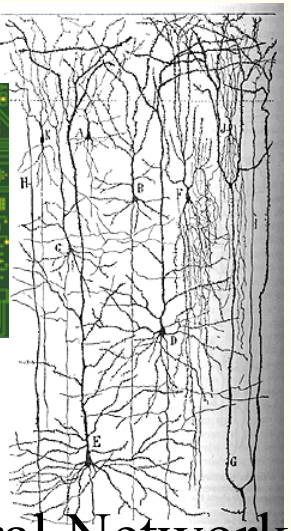
Internet
[Burch & Cheswick]



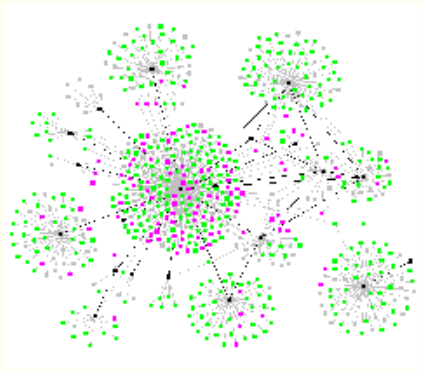
Food Web



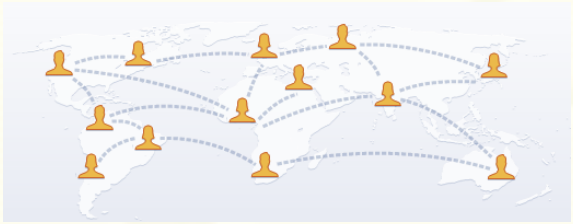
Electronic Circuit



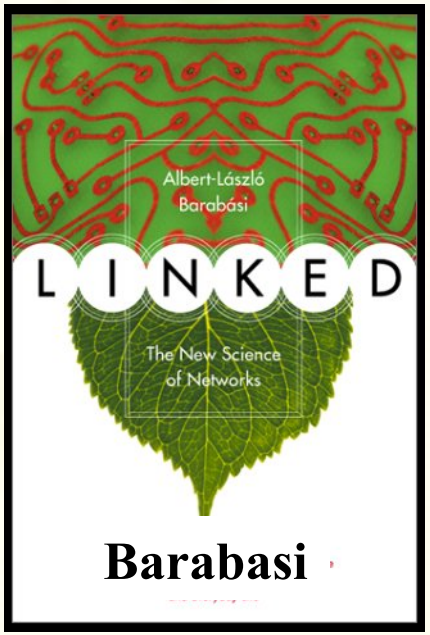
Neural Network
[Cajal]

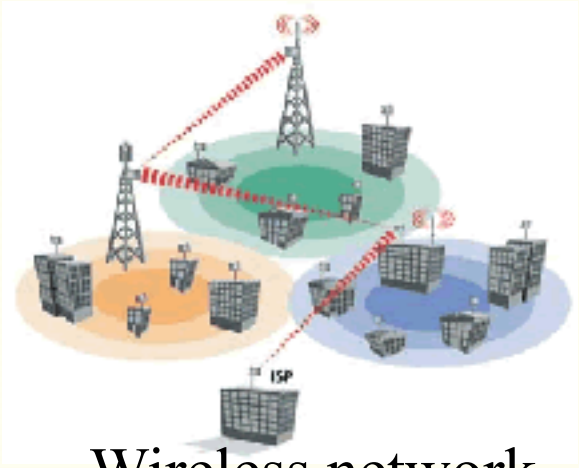


Disease Spread
[Krebs]



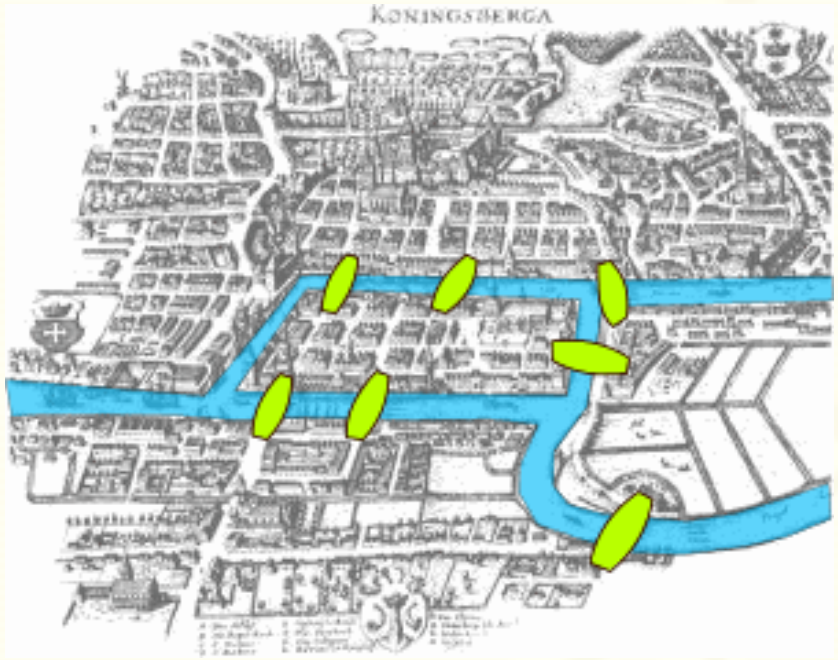
Social Network





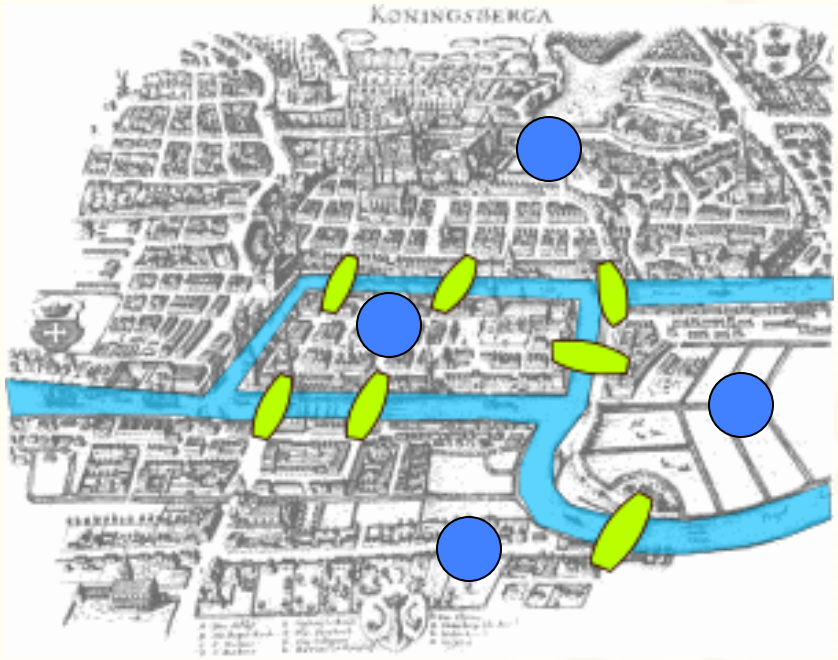
Wireless network

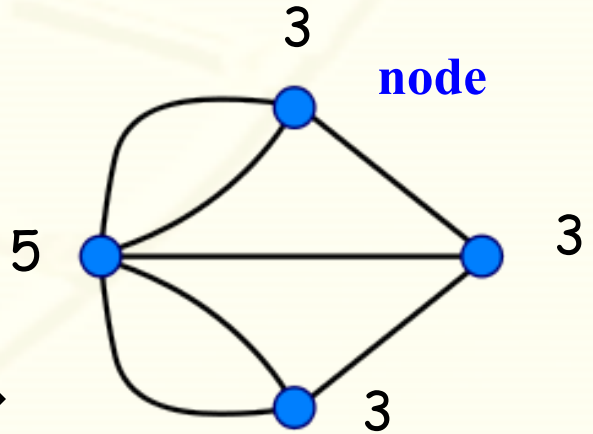
Example: The 7 bridges of Königsberg
(7桥问题）



Is it possible to walk with a route that crosses each bridge exactly once ?

Example: The 7 bridges of Königsberg
(7桥问题）

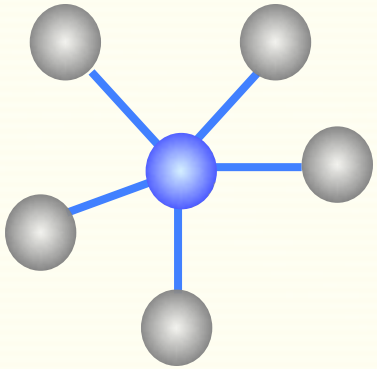




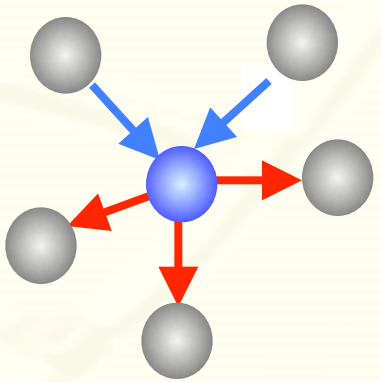
Is it possible to walk with a route that crosses each bridge exactly once ?

Solution: represent landmass as nodes, and bridges as edges

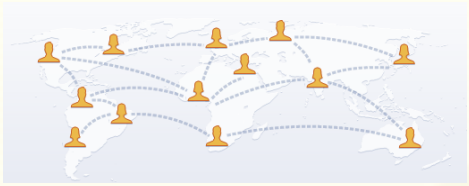
undirected vs directed graph

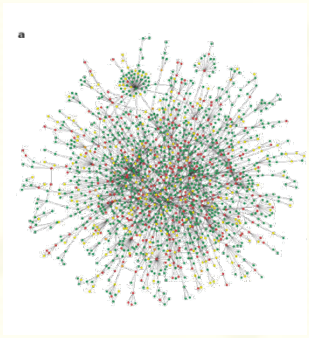


Degree = 5

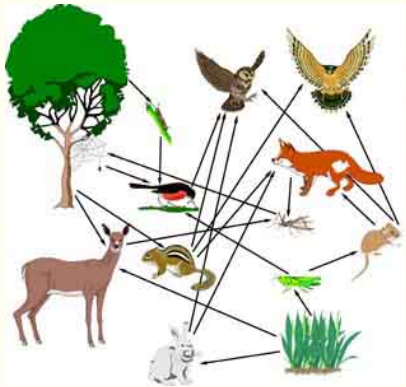


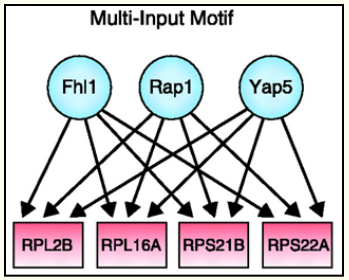
In-degree = 2
Out-degree = 3





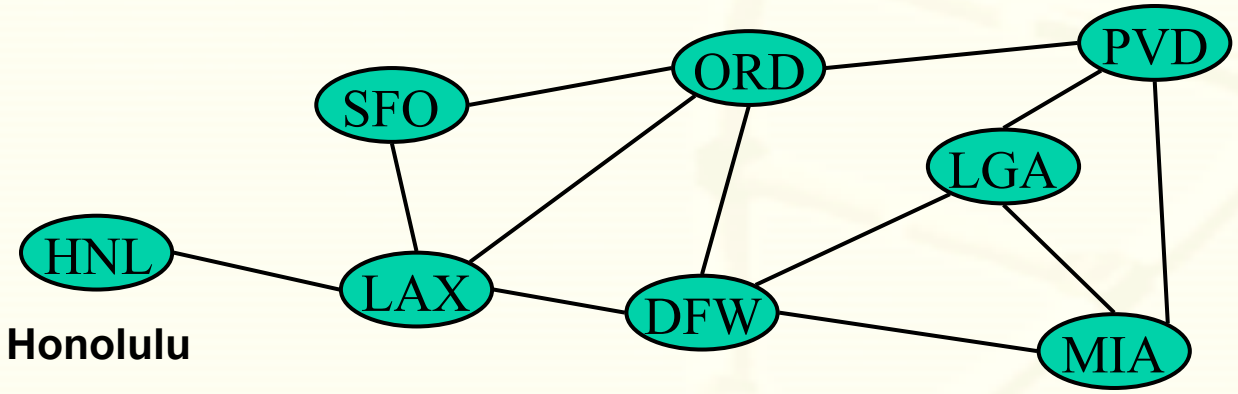
Social network,
protein interaction network



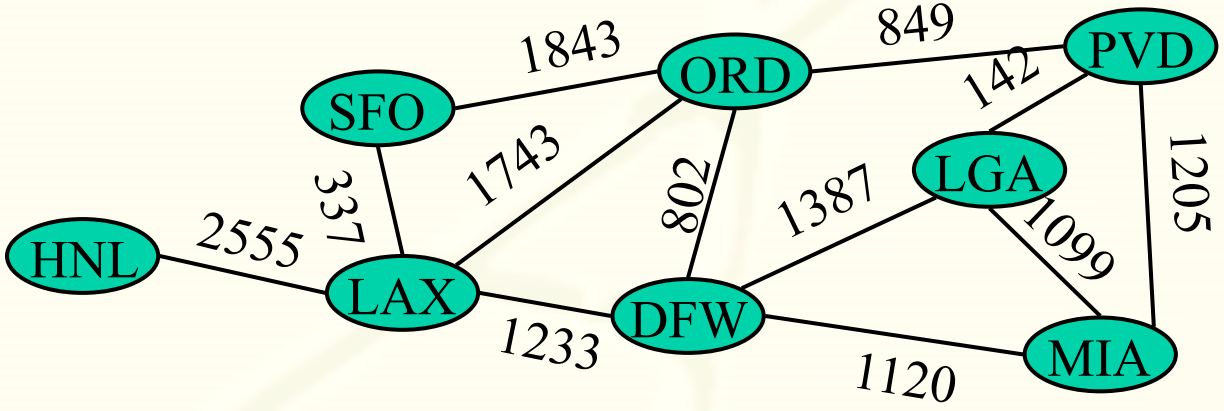


Food web,
regulatory network

Weighted vs unweighted graph

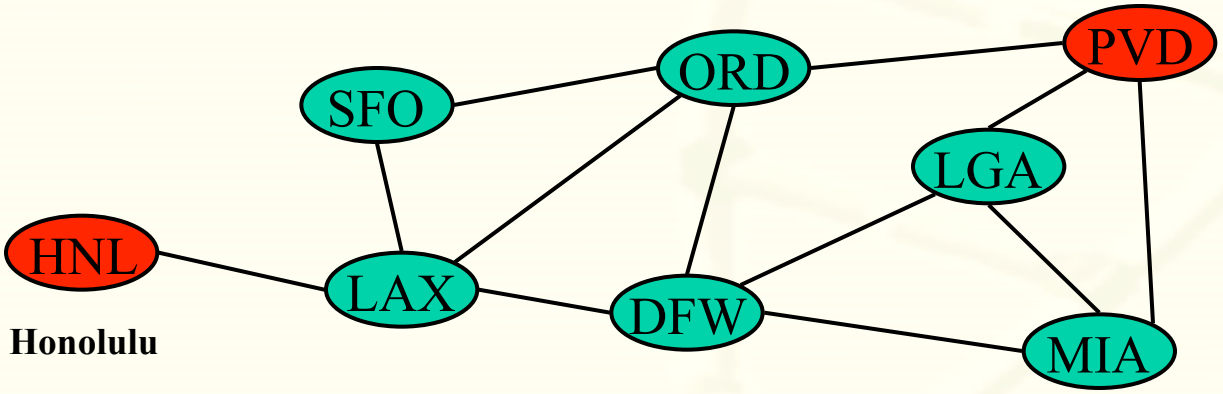


unweighted



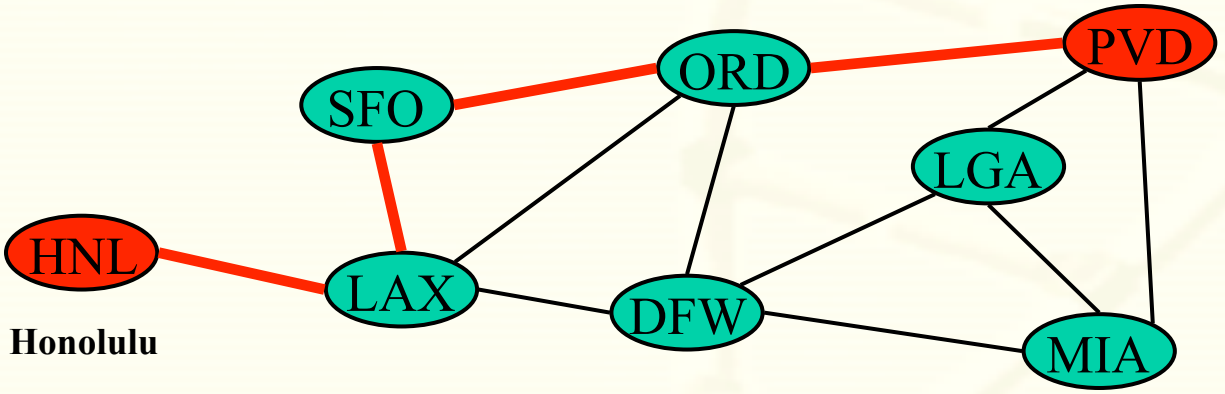
weighted

Shortest Path – unweighted graph

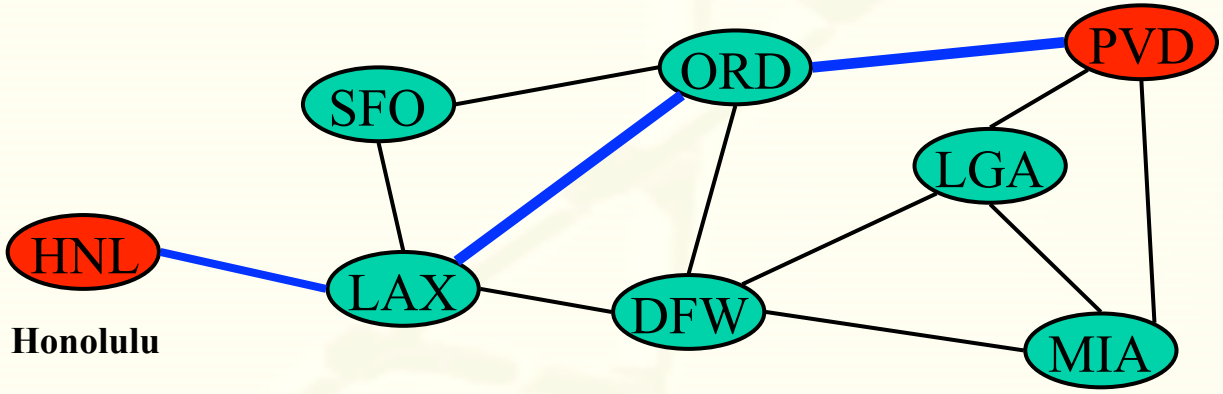


unweighted

Shortest Path – unweighted graph

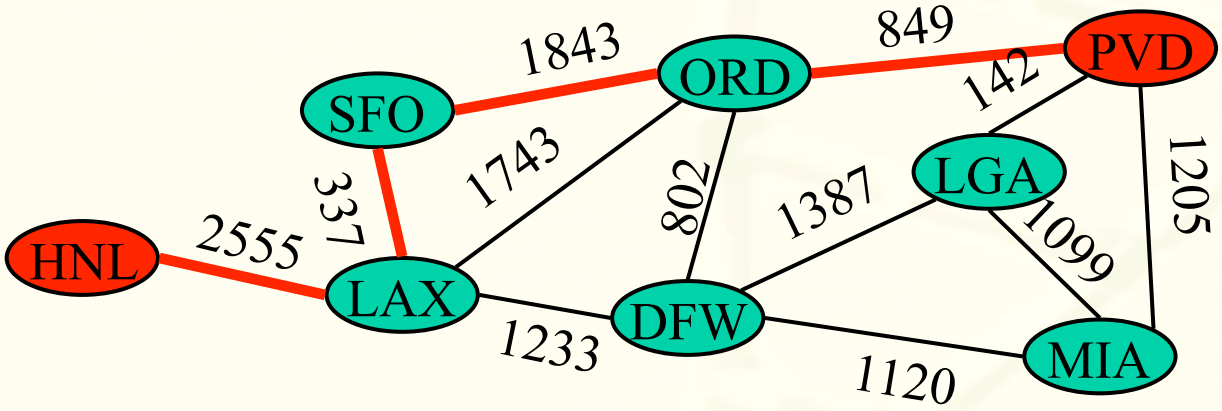


Path Length = 4

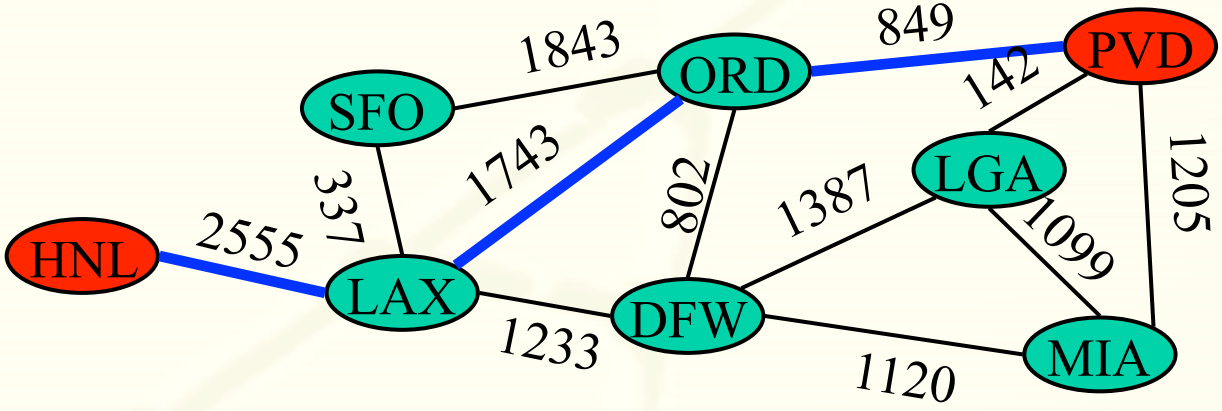


Path Length =3

Shortest Path – weighted graph



Path Length = 5584 km



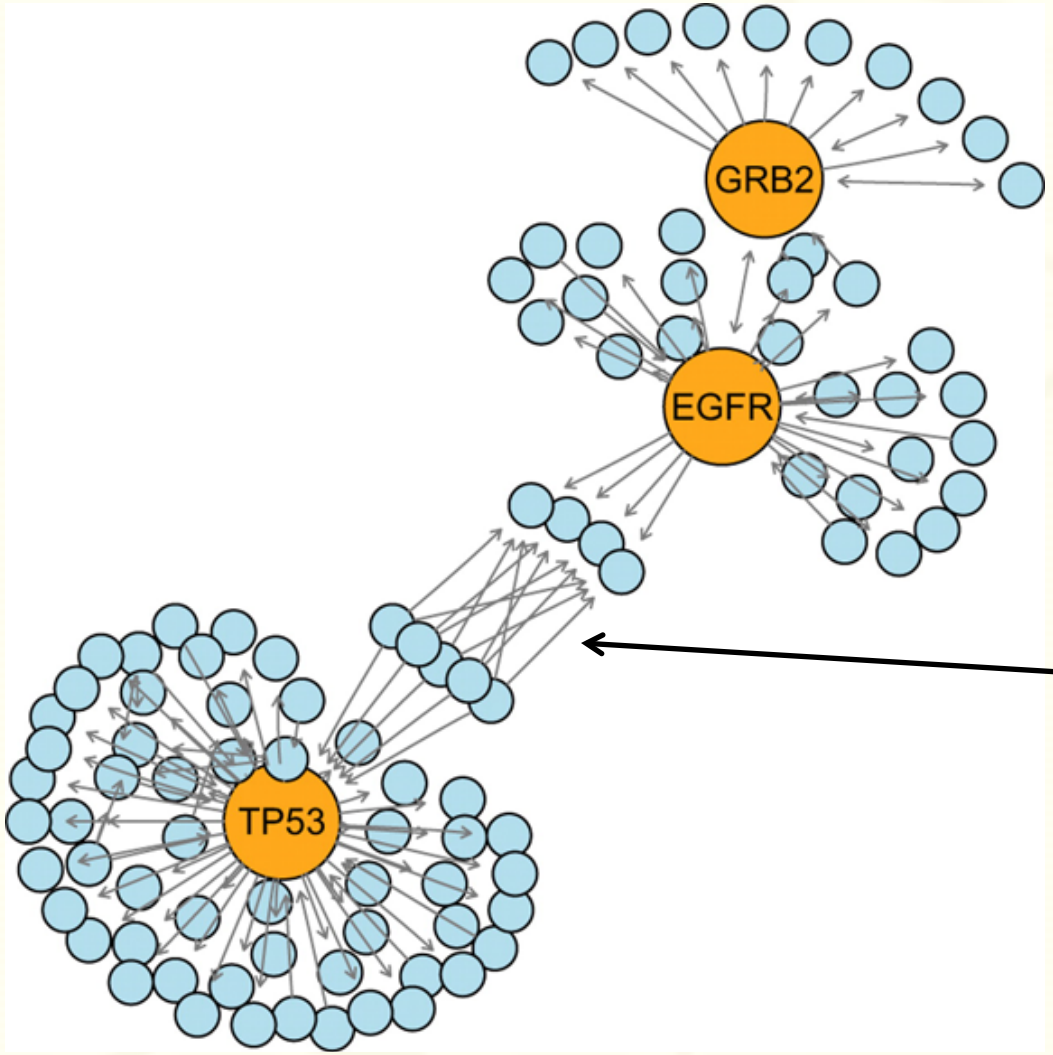
Path Length = 5147 km

Hubs and Betweenness

- **Hubs**: the nodes in the network that have the most number of connections (highest degree)

- **Betweenness:** measures the importance of a node in network communication
  - For all the possible node pairs, we determine the shortest path between them.
  - Then for each node, we ask what fraction of these shortest paths pass through this node.

Hubs and Betweenness: who are more important ?



hubs

Nodes of high betweenness

Zhang et al Bioinformatics 2009

hubs in PPI network are important

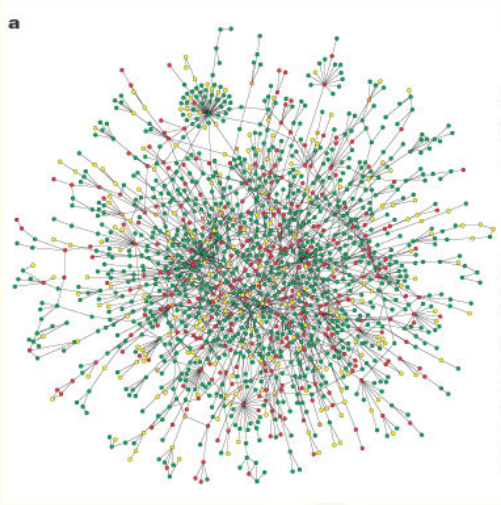
- **Quiz:** how to test whether a yeast gene is important ?
- **Hint:** we discussed this in the beginning of lecture 3.

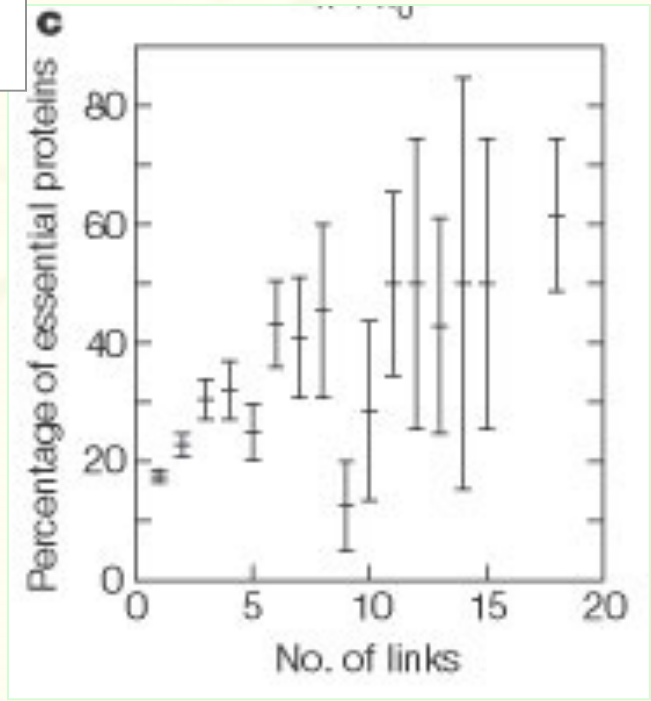
Hub proteins are more likely to be essential

Nature **411**, 41-42 (3 May 2001) | doi:10.1038/35075138

Lethality and centrality in protein networks

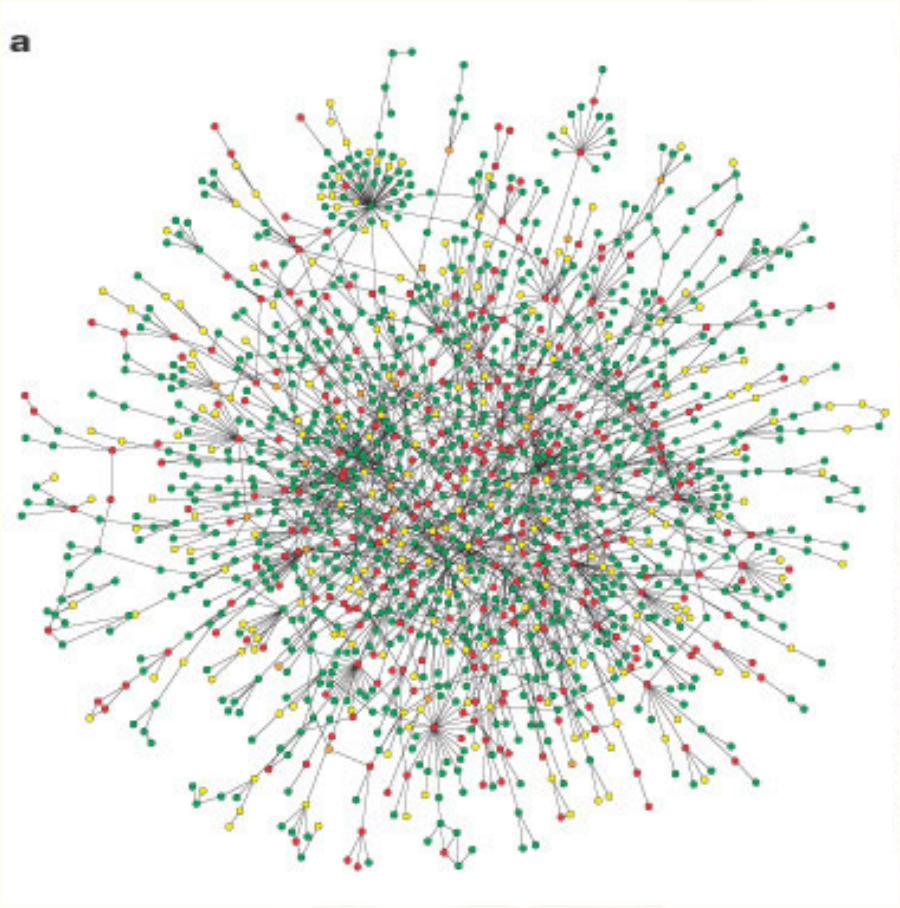
H. Jeong[1], S. P. Mason[2], A.-L. Barabási[1] & Z. N. Oltvai[2]

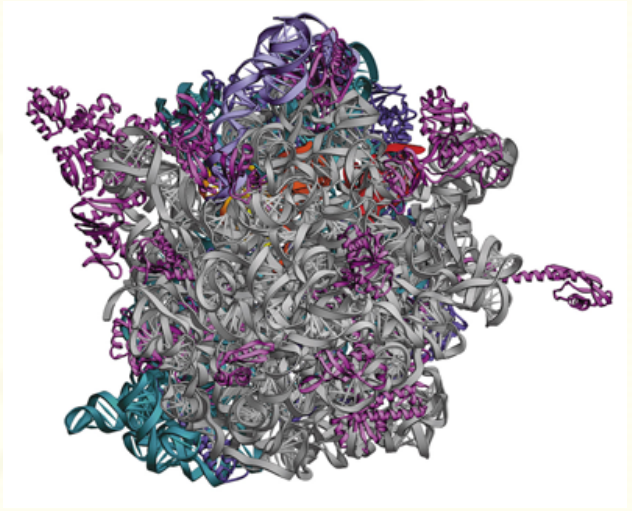




0.7% of the yeast proteins have > 15 links, 62% of them are essential.

But be careful of ribosome !





Such global topological analysis can be dominated by **heavy hitters** such as ribosome or polymerase.

“Party hub” and “Date hub”

- **Party Hubs:** which interact with most of their partners at the same time.

- **Date Hubs**: which bind their different partners **at different times or different locations**

“Party hub” and “Date hub”

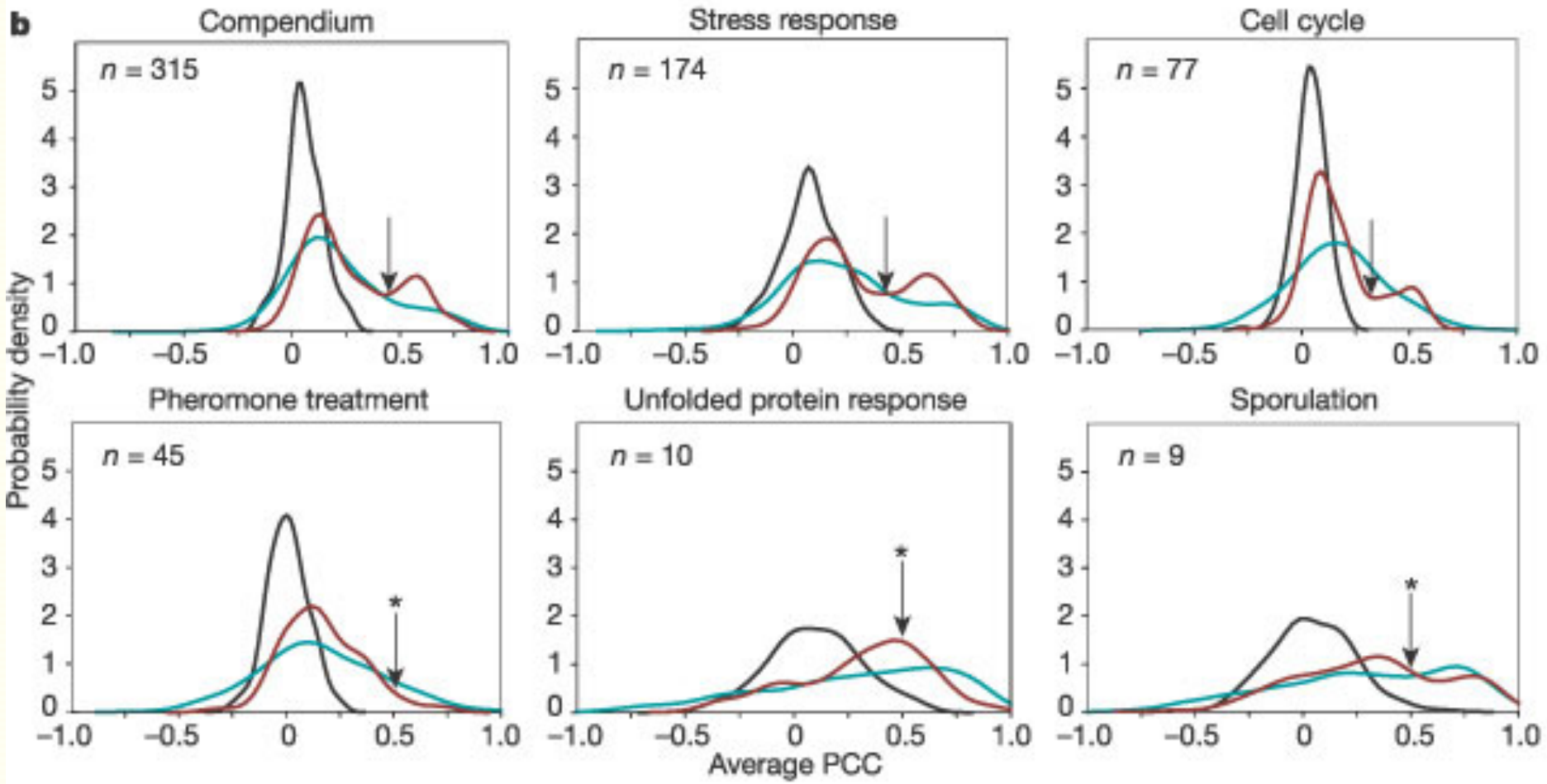
Nature **430**, 88-93 (1 July 2004) | doi:10.1038/nature02555; Received 16 December 2003; Accepted 6 April 2004; Published online 9 June 2004

Evidence for dynamically organized modularity in the yeast protein–protein interaction network

Jing-Dong J. Han[1], Nicolas Bertin[1], Tong Hao[1], Debra S. Goldberg[2], Gabriel F. Berriz[2], Lan V. Zhang[2], Denis Dupuy[1], Albertha J. M. Walhout[1,3], Michael E. Cusick[1], Frederick P. Roth[2] & Marc Vidal[1]

- These authors mapped the yeast gene expression data sets (cell cycle, environmental perturbation) to the interaction network, and observed two peaks.

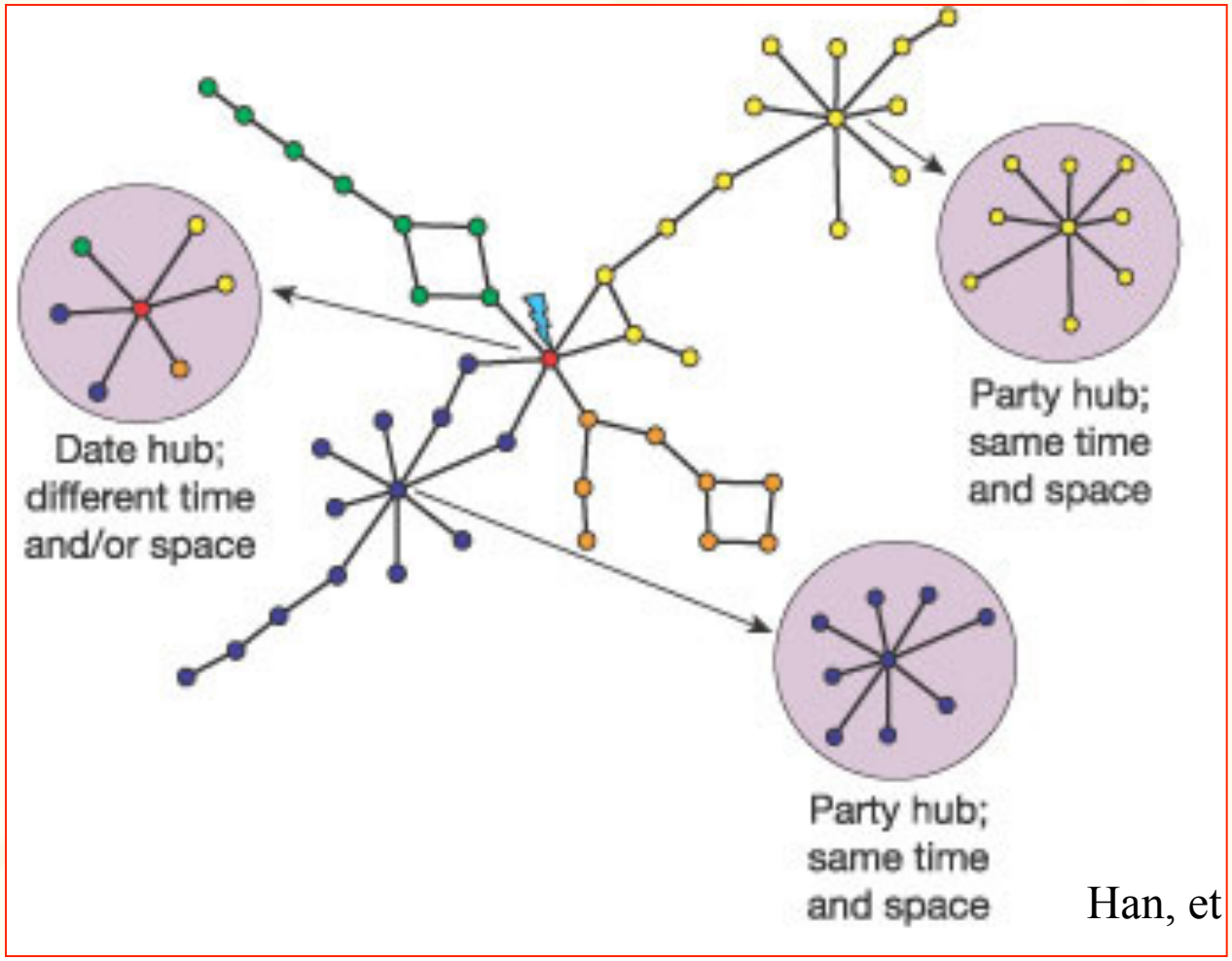
“Party hub” and “Date hub”



Han, et al Vidal Nature 2004

“Party hub” and “Date hub”

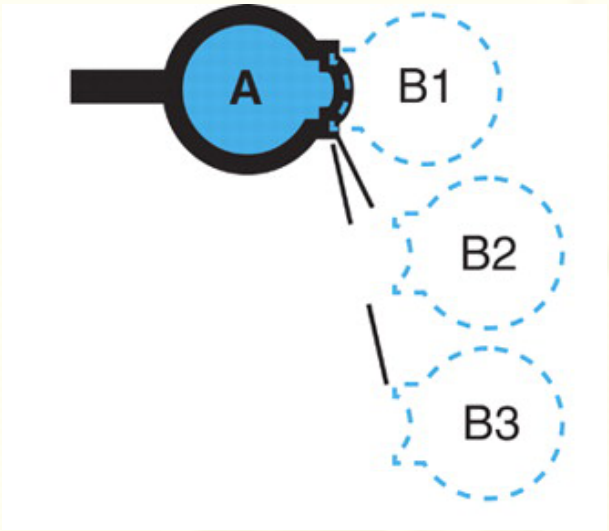
- “ …support a model of **organized modularity** in which date hubs organize the proteome, connecting biological processes—or modules —to each other, whereas party hubs function inside modules.”

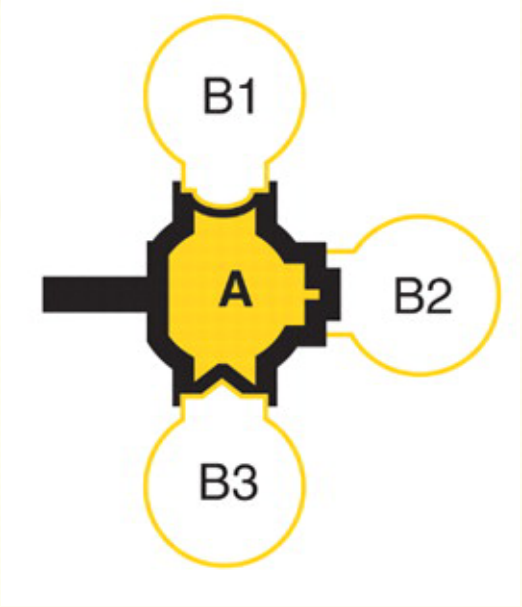


Han, et al Vidal Nature 2004

Distinct structure properties of Party Hub and Date Hub

- Hypothesis: if the Party Hubs interact many partners simultaneously, and Date Hubs interact many partners at different time, then the Party Hubs should have more structure interfaces





Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights

Philip M. Kim[1], Long J. Lu[1], Yu Xia[4,5] and Mark B. Gerstein[1,2,3,*]

Global properties of protein-protein interaction network

Many large networks such as protein interaction network, internet have the following properties:

- **Scale-free network, i.e. Power-law degree distribution**:
  - Small number of nodes have may connections while majority of nodes have few connections

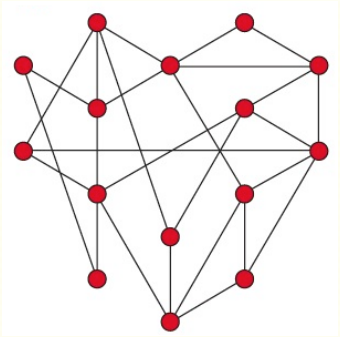
- **Small World property:**
  - A small average node-to-node path, i.e. most nodes can be reached from every other node by a small number of steps

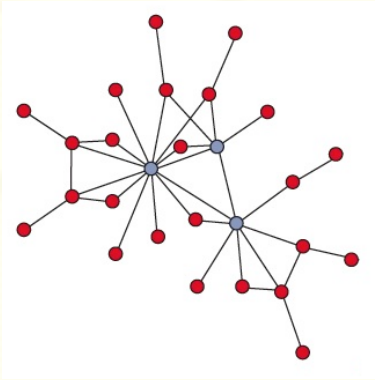
- **Robustness:**
  - Resilient and have strong resistance to failure on random attacks but vulnerable to targeted attacks

Scale-free network: power-law degree distribution

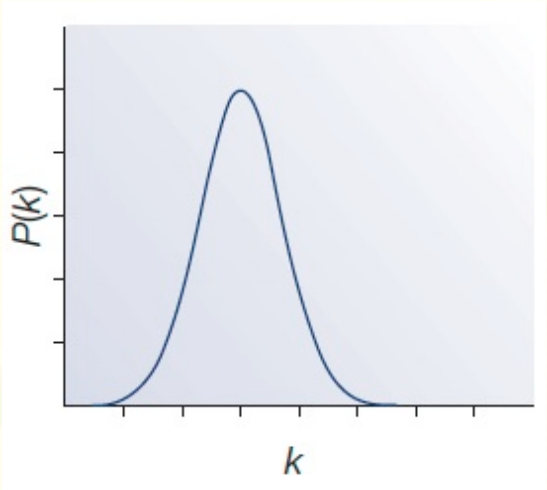
- Small number of nodes are highly connected (**hubs**), while majority of nodes have few connections.
- For example, ribosome, chaperones, Google, Yahoo

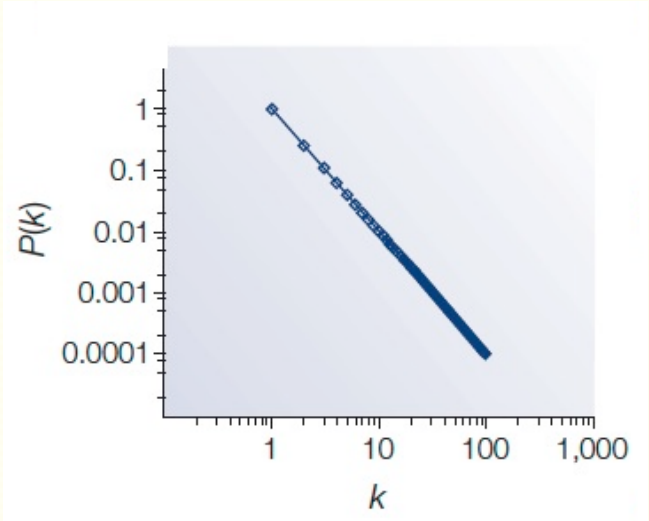
Random network





Scale-free

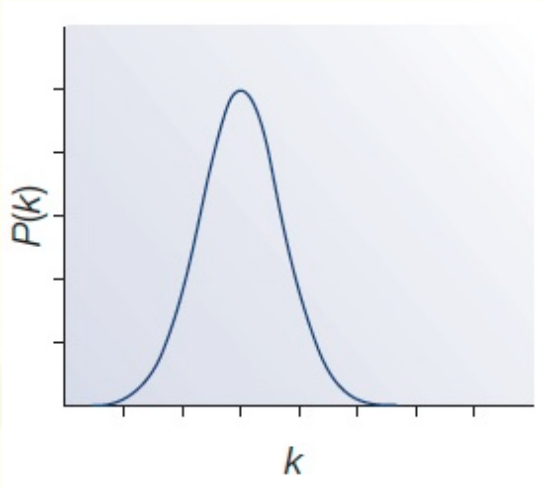


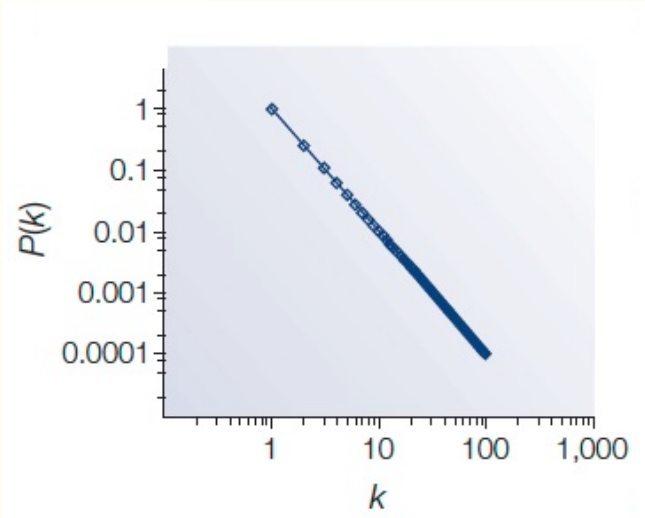


Scale-free network: power-law degree distribution

- Small number of nodes are highly connected (**hubs**), while majority of nodes have few connections.
- For example, ribosome, chaperones, Google, Yahoo

- Other examples of power-law distribution: personal wealth distribution in a society, size of companies in a free market





Scale-free network: power-law degree distribution

- Small number of nodes are highly connected (**hubs**), while majority of nodes have few connections.
- For example, ribosome, chaperones, Google, Baidu.

- Other examples of power-law distribution: personal wealth distribution in a society, size of companies in a free market

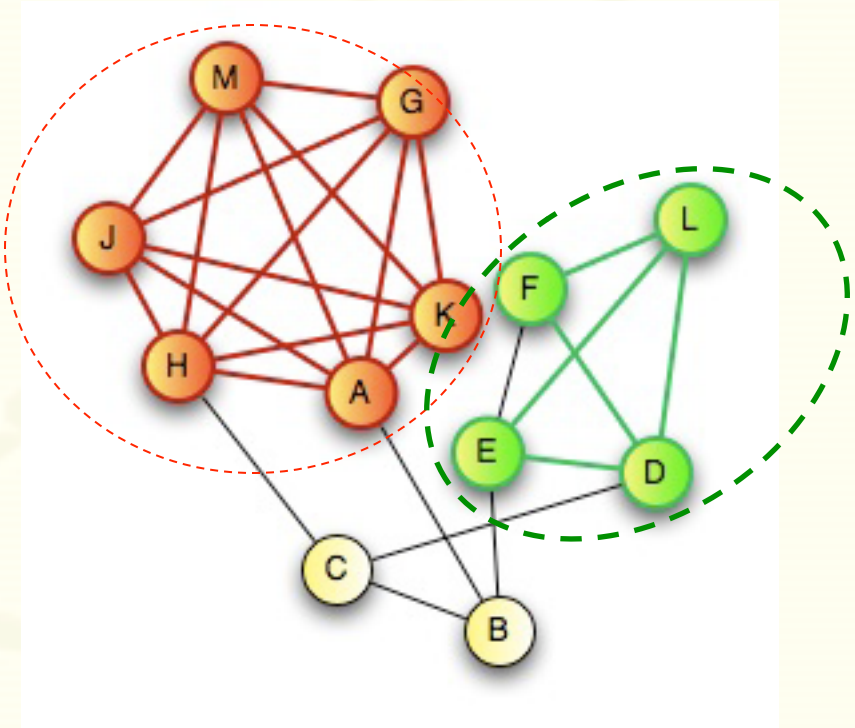
Mechanism:

- In a society: “rich getting richer”
- In protein interaction network: “preferential attachment”, i.e. hub proteins are likely to gain more interacting partners (by duplication of hub or nodes.)

Small World Network

- **Small world network**: most nodes are not neighbors of one another, but most nodes can be reached from every other **by a small number of steps** (edges).

- Small world network tend to contain cliques, i.e. a protein complex, a group of densely connected nodes.



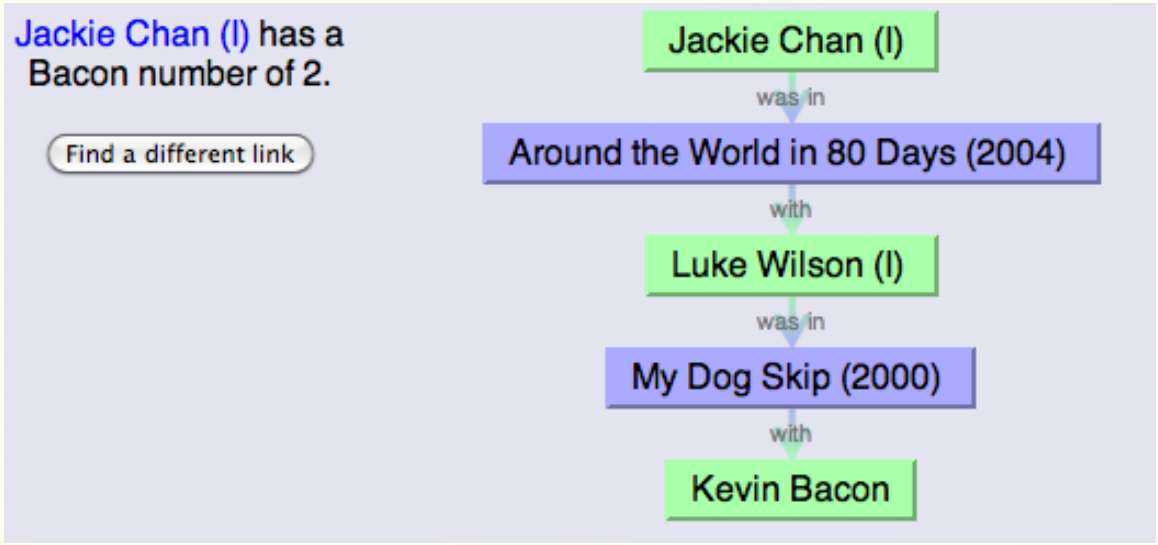
Example of Small World Network

- **Six degrees of separation:** everyone is on average approximately six steps away from any other person on Earth
  - Result of the evolution of human society and communications technology. This is probably not true 2000 years ago.

Example of Small World Network

- **Six degrees of separation:** everyone is on average approximately six steps away from any other person on Earth
  - Result of the evolution of human society and communications technology. This is probably not true 2000 years ago.

- **Six Degrees of Kevin Bacon:** any actor can be linked to Kevin Bacon in 6 steps (movies).



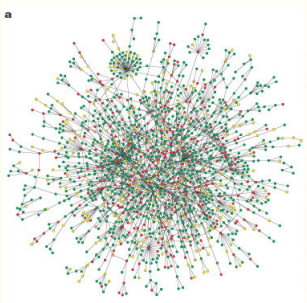
Protein interaction network is robust

Robust: immune to gene mutations and deletions.

- **Scale-free network:** the network has a few hubs and many sparsely connected nodes
- **Small World:** most of the nodes communicate to each other through the hubs.

- Therefore, a random mutation (attack) will most likely hit on a non-hub protein, and will not interfere with the communications between mosr of the nodes on the network.

- However, a clever invader such as a virus can initiate targeted attack on important nodes such as hubs, and disable the host network.



But a word of caution

- Biological data is much more complex than other type of networks
  - We only surveyed a very small % of yeast and human protein-protein interaction network, and 60% of the yeast genetic network
  - The quality of the data is improving but still noisy
  - The interactions could be biased by the experimental methodology used.

- Lack of dynamics or temporal data in biological network
  - Yeast PPi and SGA are all done in lab rich media condition
  - Almost all the human PPi are done in HeLa or HEK293 cells

- A lot of the earlier analysis papers were published when only less than 10% of the network is know.

The Temporal Dynamics of Protein Complex

- **Question:** How do protein interactions or protein complex memberships change when the cells are under different environmental conditions, or when they undergo cell cycle ?

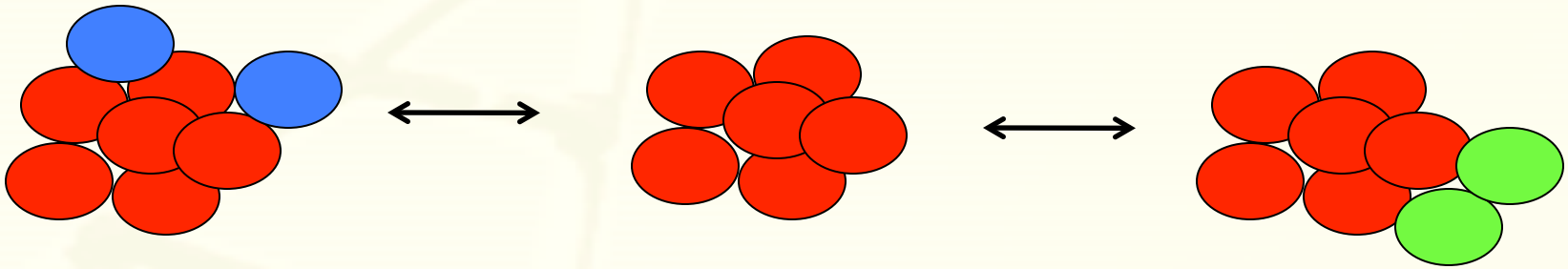
- **Rational:** The Yeast 2-Hybrid and Mass Spec experiments were all conducted in a single “non-physiological” condition in the lab.

- **Approach:** mapping the gene expression profiles onto these observed interactions.

Dynamic Complex Formation During the Yeast Cell Cycle

- **De Lichtenberg et al Science 2005**
- "… we integrated data on protein interactions and gene expression... We discovered that most complexes consist of **both periodically and constitutively expressed subunits**, which suggests that the former control complex activity by a mechanism of just-in-time assembly."

- Translation: most of the protein complexes have a "core sub complex" that never changes, and additional subunits are added to the complex at different time point.



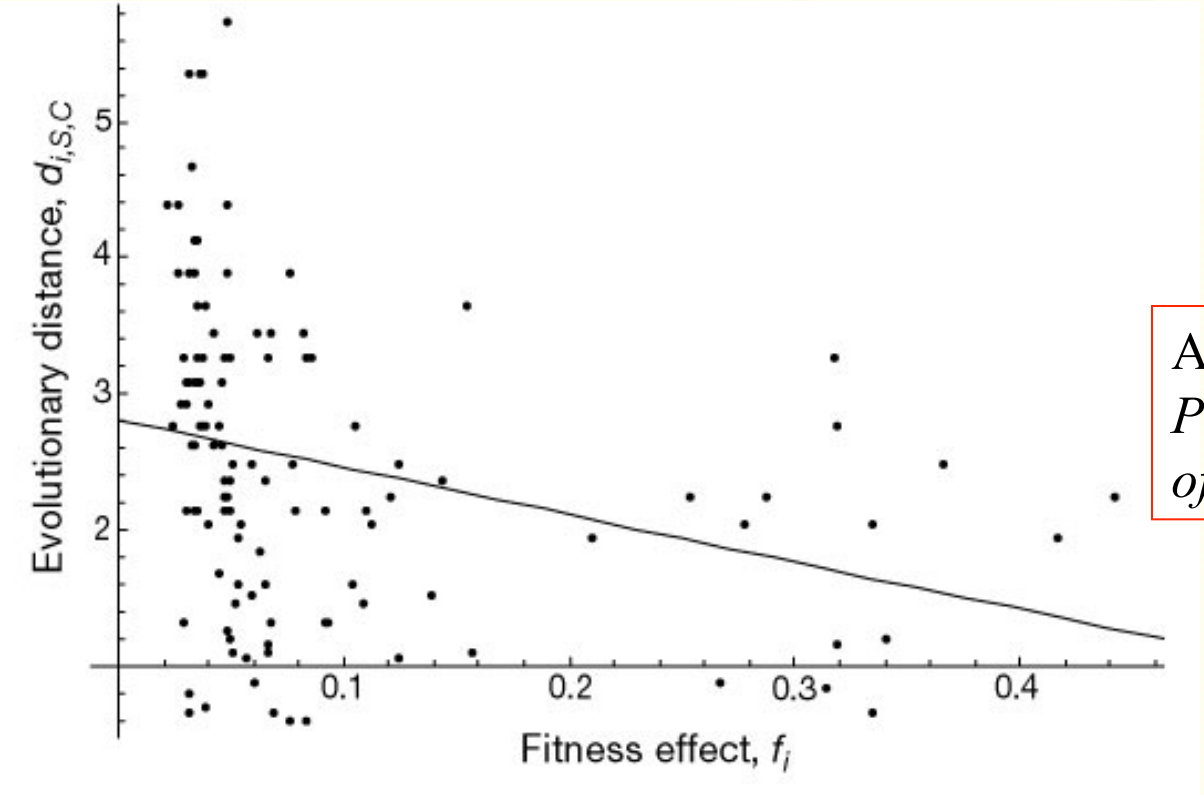
Evolutionary analysis of protein-protein interaction network

- Do protein-protein interactions have any evolutionary constraints on protein sequence evolution ? In other words, do hubs evolve at the same rate as non-hubs ?

- How well are protein-protein interactions conserved in related organisms ?

Protein essentiality and evolution rate

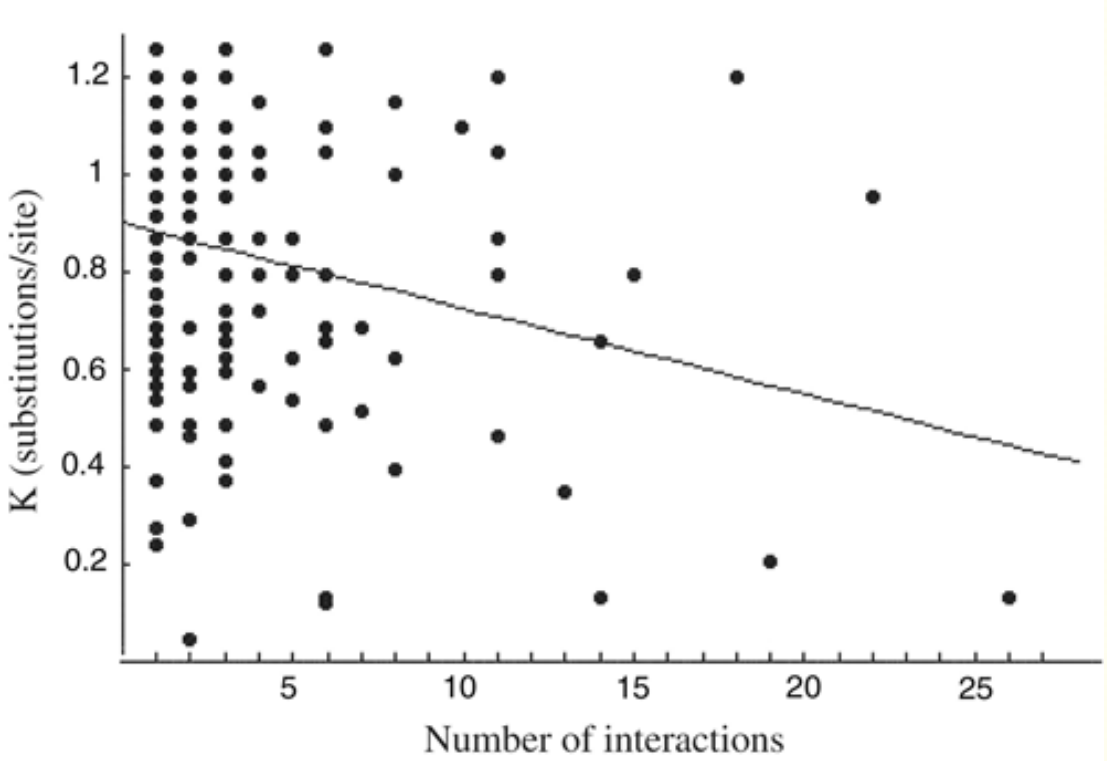
- "Our analysis reveals a highly significant relationship between protein dispensability and evolutionary rate" "The relationship is highly conserved, so that protein dispensability in yeast is also predictive of evolutionary rate in a nematode worm.



Aaron Hirsh, Hunter Fraser, *Protein dispensability and rate of evolution* **Nature** 2001

Evolutionary Rate in the Protein Interaction Network

- "connectivity of well-conserved proteins in the network is negatively correlated with their rate of evolution."



Fraser, Hirsh, et al Science 2002

Evolutionary Rate in the Protein Interaction Network

- “Proteins with more interactors evolve more slowly not because they are more important to the organism, but because a greater proportion of the protein is directly involved in its function.”
-

Fraser, Hirsh, et al Science 2002

but this is not the end of story …

BMC Evol Biol. 2003 Jan 6;3:1. Epub 2003 Jan 6.

No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly.

Jordan IK, Wolf YI, Koonin EV.

BMC Evol Biol. 2003 Oct 2;3:21.

Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets.

Bloom JD, Adami C.

BMC Evol Biol. 2004 May 27;4:13.

Evolutionary rate depends on number of protein-protein interactions independently of gene expression level.

Fraser HB, Hirsh AE.

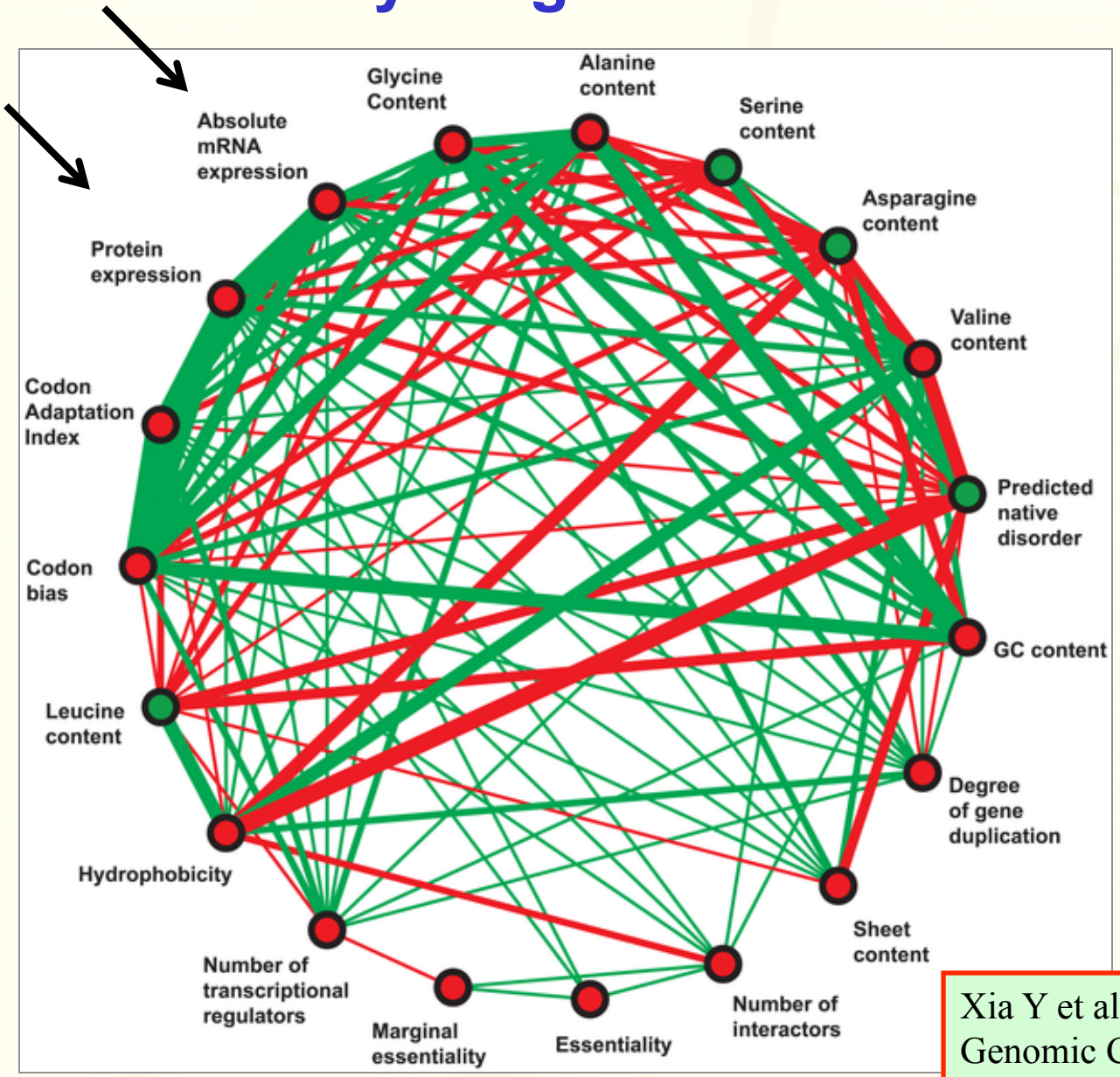
BMC Bioinformatics. 2006 Mar 13;7:128.

Protein protein interactions, evolutionary rate, abundance and age.

Saeed R, Deane CM.

Department of Statistics, Oxford OX1 3TG, UK. saeed@stats.ox.ac.uk

Everything is correlated to each other



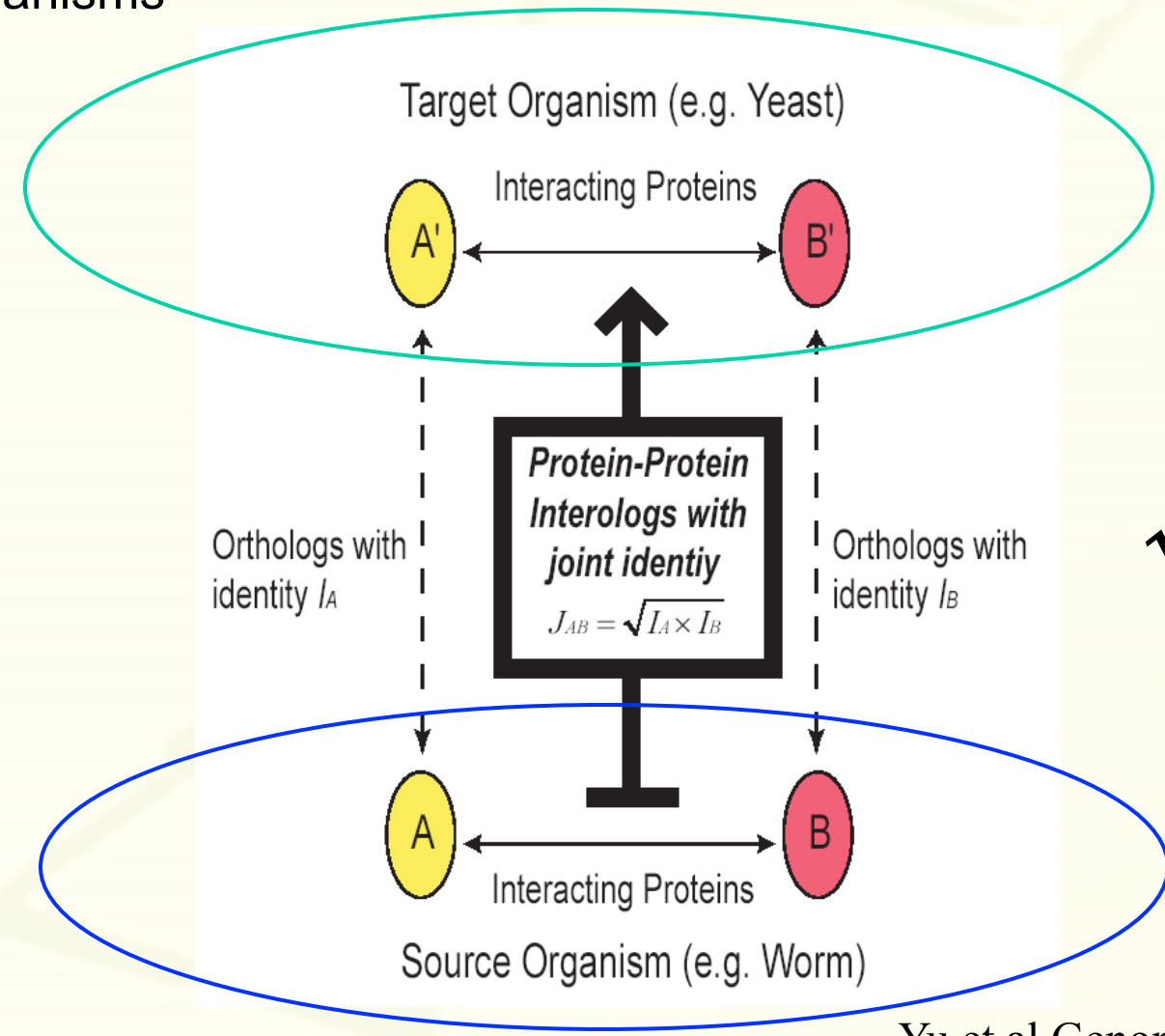
Positive correlation

Negative correlation

Xia Y et al, Integrated Assessment of Genomic Correlates of Protein Evolution, PLOS Comp Biol 2009

Conservation of protein interactions

- Interologs: orthologous pairs of interacting proteins in different organisms



$$\sqrt{I_A \times I_B}$$

Yu et al Genome Research 2004

Conservation of protein interactions

- Interologs: orthologous pairs of interacting proteins in different organisms,

Annotation Transfer Between Genomes: Protein–Protein Interologs and Protein–DNA Regulogs

Haiyuan Yu,[1] Nicholas M. Luscombe,[1] Hao Xin Lu,[1] Xiaowei Zhu,[1] Yu Xia,[1] Jing-Dong J. Han,[2] Nicolas Bertin,[2] Sambath Chung,[1] Marc Vidal,[2] and Mark Gerstein[1,3]

- "We find that protein–protein interactions can be transferred between organisms when a pair of proteins has a joint sequence identity >80%"

Enough prediction, we need some real data …

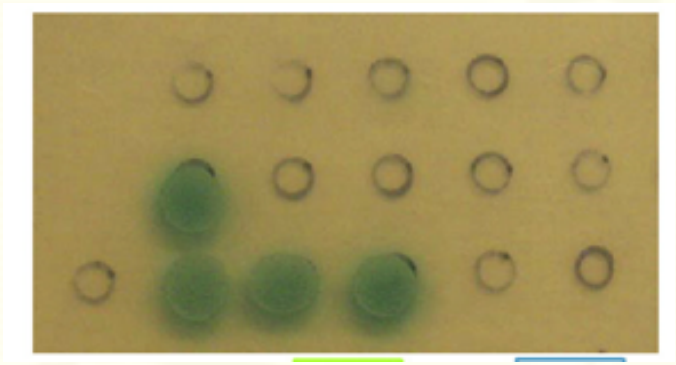
- " we experimentally examine 87 potential interactions between Kluyveromyces waltii proteins, whose one to one orthologs in the related budding yeast Saccharomyces cerevisiae have been reported to interact. "

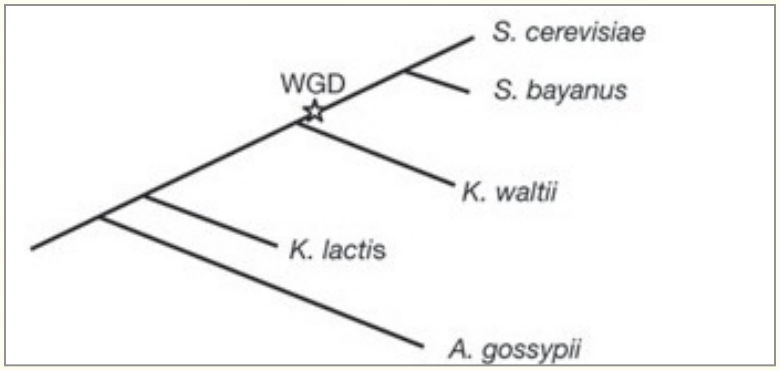
Proc Natl Acad Sci U S A. 2011 May 24;108(21):8725-30. Epub 2011 May 9.

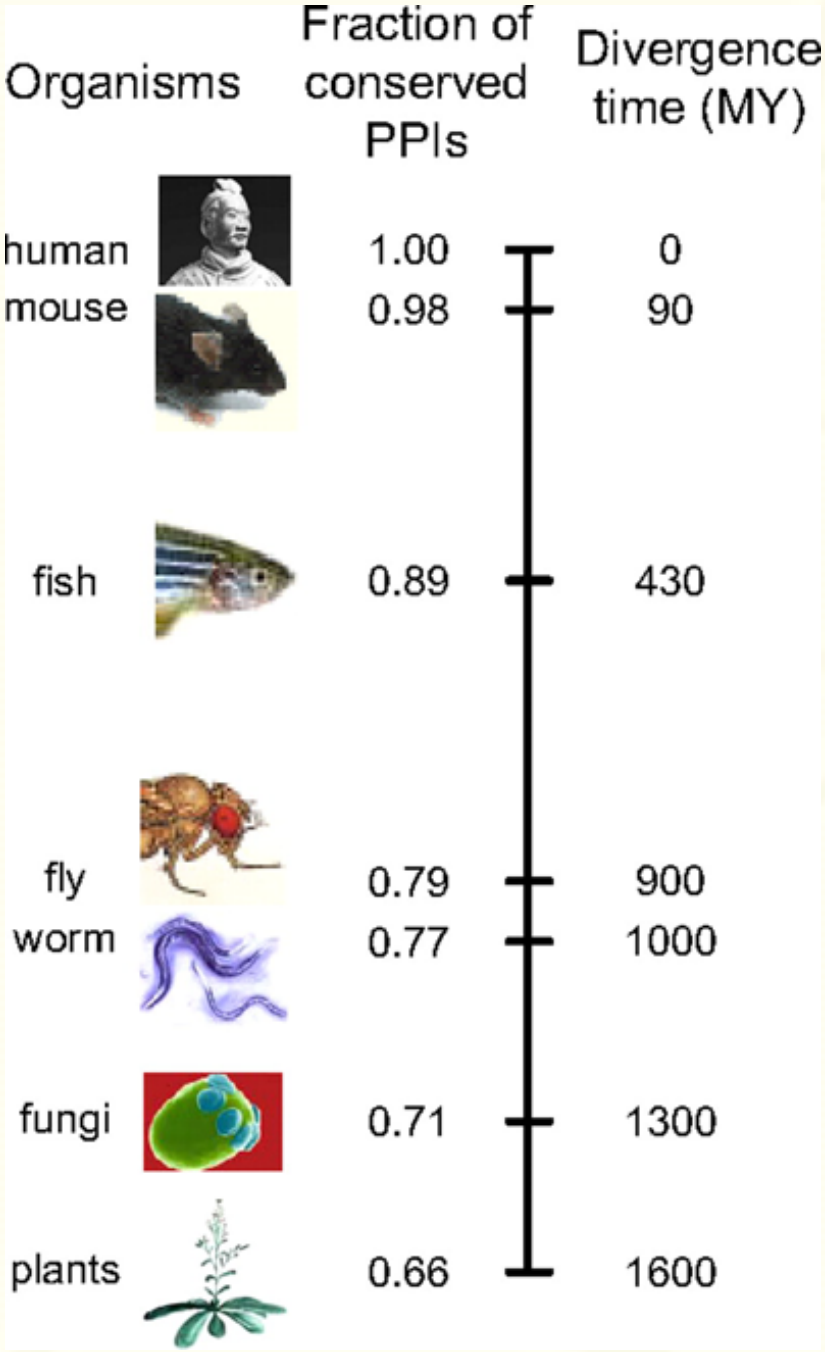
Measuring the evolutionary rate of protein-protein interaction.

Qian W, He X, Chan E, Xu H, Zhang J.

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109.







- "… we estimate that the evolutionary rate of protein interaction is (2.6 ± 1.6) × 10^{-10} per PPI per year, which is three orders of magnitude lower than the rate of protein sequence evolution "
- "The extremely slow evolution of protein molecular function may account for the remarkable conservation of life at molecular and cellular levels and allow for studying the mechanistic basis of human disease in much simpler organisms."

End of Protein-Protein interactions

- coffee break