Biological networks

Outline

1. Learning biological networks: experiments
2. Statistical properties of the networks
3. Understanding networks structure: motifs, modules, etc
4. From structure to function
5. Compare/align networks
6. Dynamics of networks

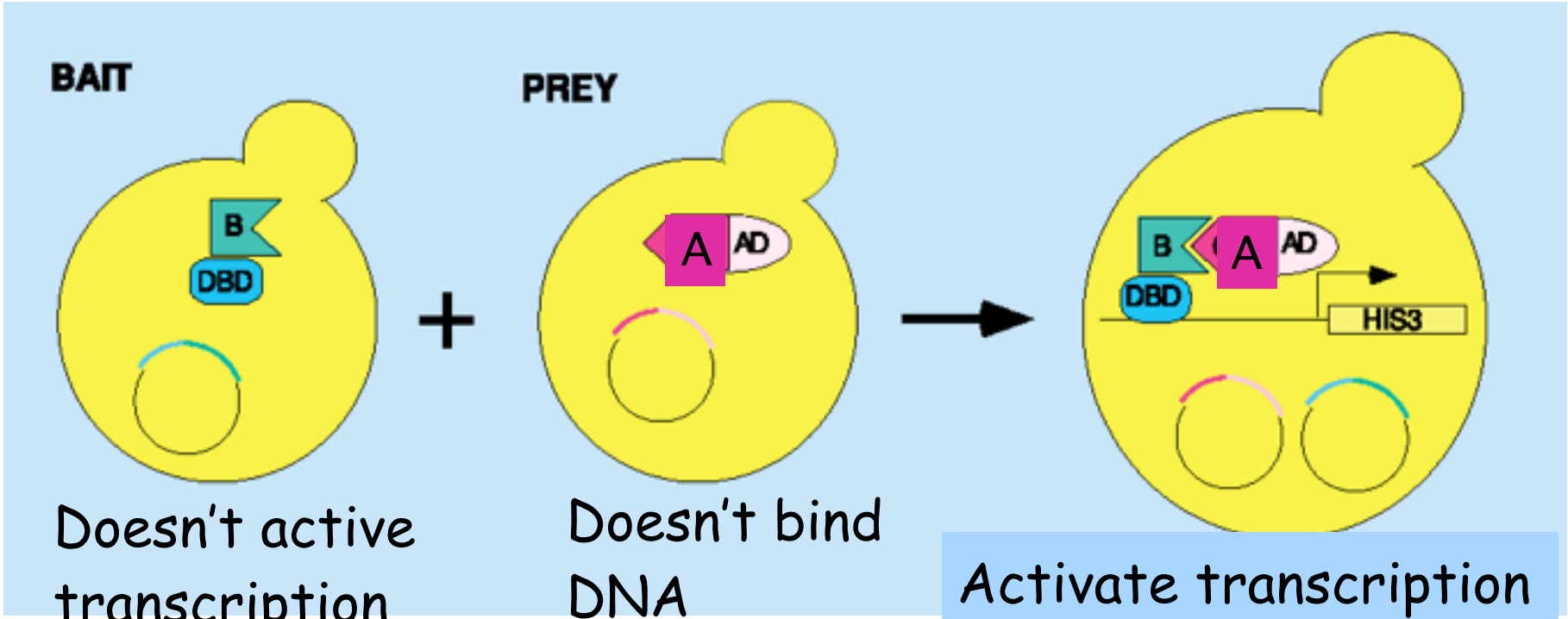
Outline

1. Learning biological networks: experiments
2. Statistical properties of the networks
3. Understanding networks structure: motifs, modules, etc
4. From structure to function
5. Compare/align networks
6. Dynamics of networks

Protein-protein Interactions

Yeast two-hybrid assay:

Does a protein A interact with B ?



Doesn't active transcription

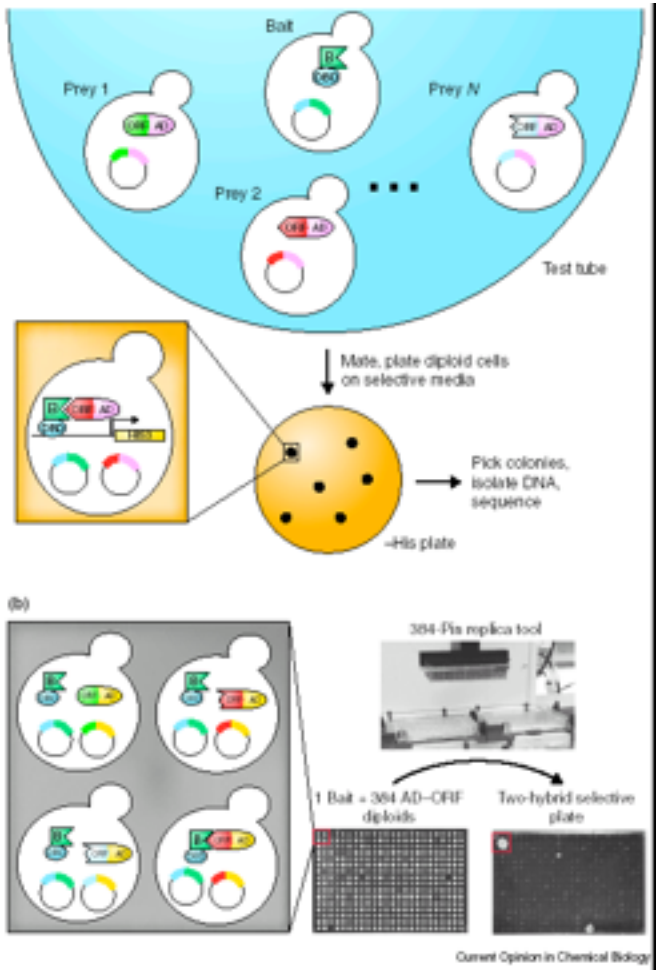
Doesn't bind DNA

Activate transcription and grow on -his media

Protein-protein Interactions

Large scale yeast two-hybrid assay:

Find pairs of interacting proteins



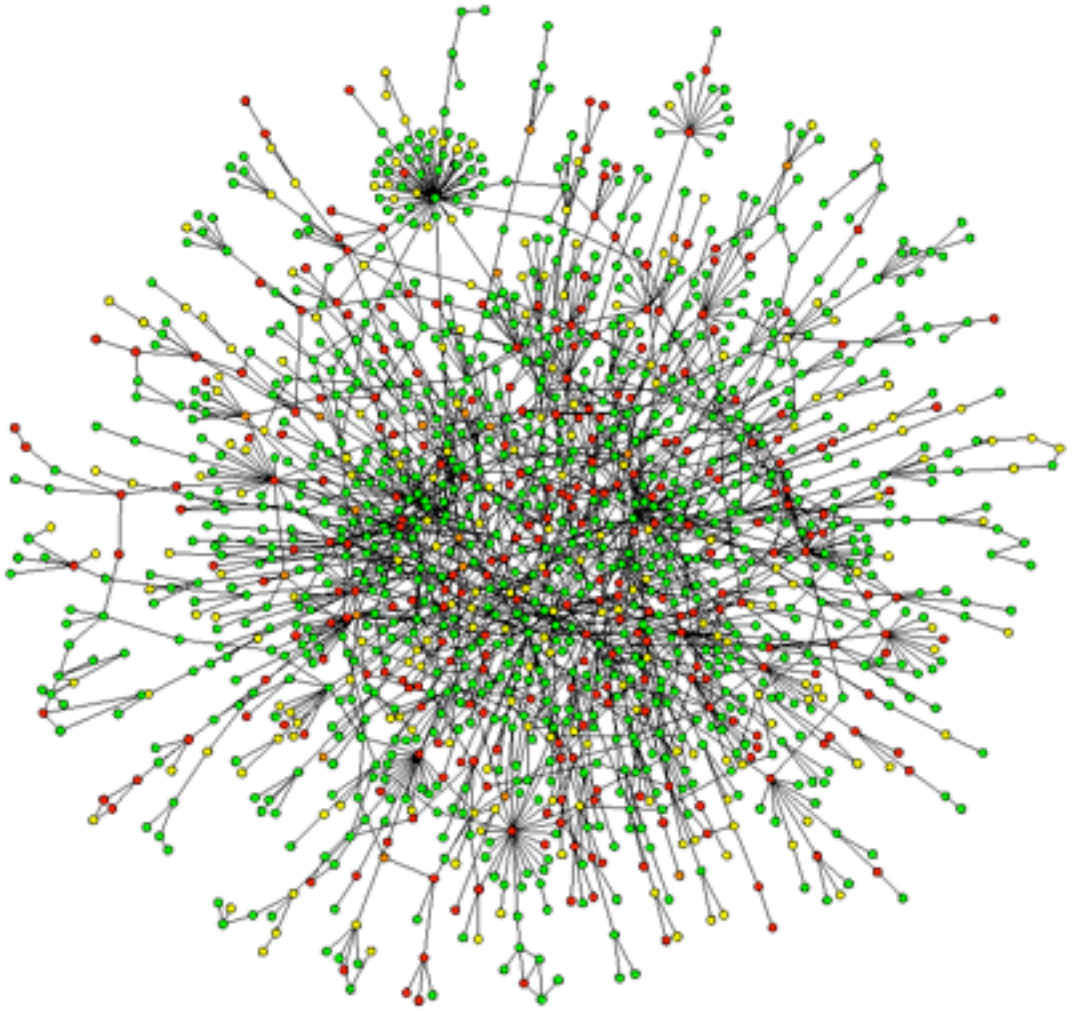
A comprehensive two-hybrid analysis to explore the yeast protein interactome.
Ito T et al, *PNAS* 2001

A comprehensive analysis of protein-protein interactions in *S. cerevisiae.*
P. Uetz et al, *Nature* 2000

Protein-protein Interactions

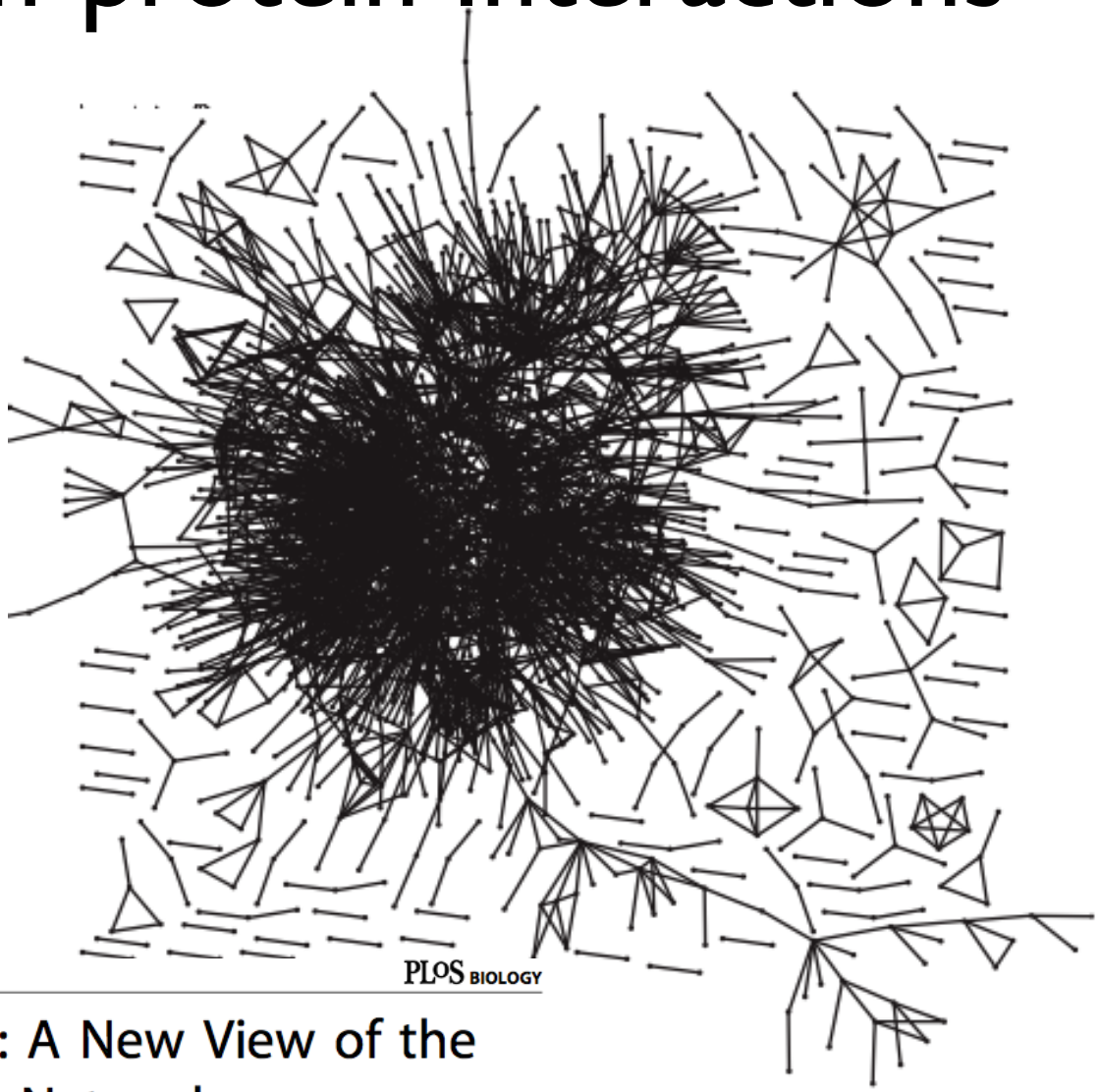
2001

- proteins m=4000
  interactions n=6500



Protein-protein Interactions

Baker's yeast:
9000 interactions
3000 proteins
(combined, 2006)



OPEN ACCESS Freely available online

PLOS BIOLOGY

Stratus Not Altocumulus: A New View of the Yeast Protein Interaction Network

Nizar N. Batada[1,2*], Teresa Reguly[1], Ashton Breitkreutz[1], Lorrie Boucher[1,3], Bobby-Joe Breitkreutz[1], Laurence D. Hurst[2*☯], Mike Tyers[1,3*☯]

High-Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping

Albertha J. M. Walhout and Marc Vidal[1]

Next-generation sequencing to generate interactome datasets

Haiyuan Yu[1–3], Leah Tardivo[1,
Evan Weiner[1,2,5], Fana Gebrea
Nenad Svrzikapa[1,2], Tomoko
Edward Rietman[1,2], Xinping
Kourosh Salehi-Ashtiani[1,2,5],
Michael E Cusick[1,2], David E Hi
Pascal Braun[1,2] & Marc Vidal[1]

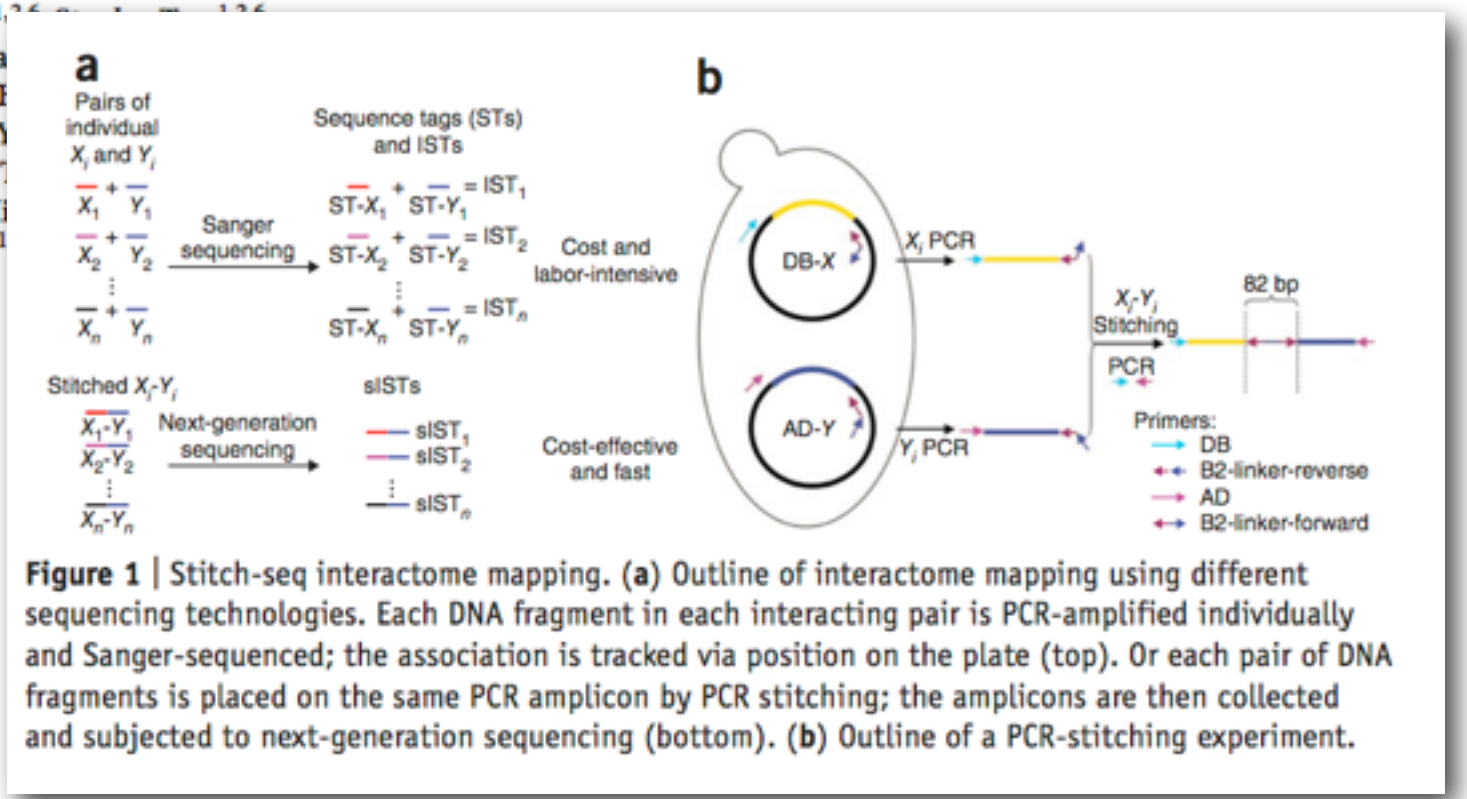


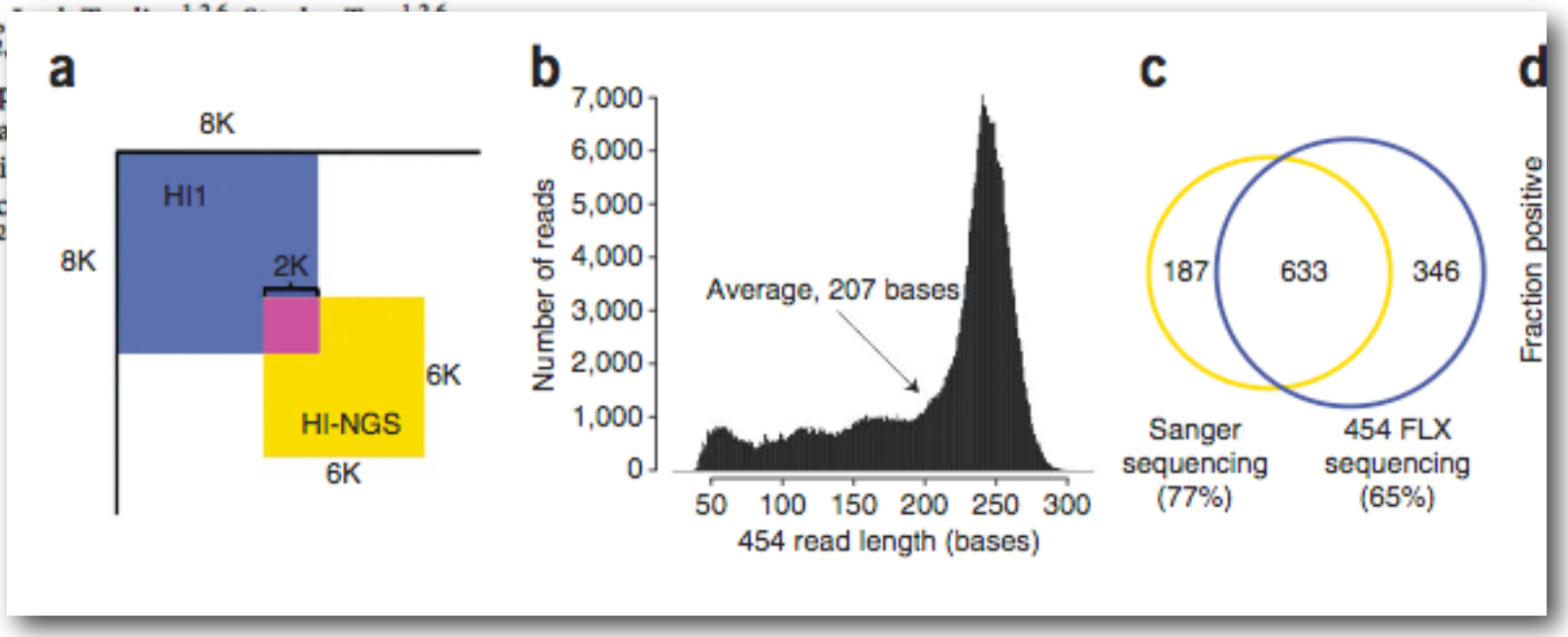
Figure 1 | Stitch-seq interactome mapping. (**a**) Outline of interactome mapping using different sequencing technologies. Each DNA fragment in each interacting pair is PCR-amplified individually and Sanger-sequenced; the association is tracked via position on the plate (top). Or each pair of DNA fragments is placed on the same PCR amplicon by PCR stitching; the amplicons are then collected and subjected to next-generation sequencing (bottom). (**b**) Outline of a PCR-stitching experiment.

High-Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping

Albertha J. M. Walhout and Marc Vidal[1]

Next-generation sequencing to generate interactome datasets

Haiyuan Yu[1–3],
Evan Weiner[1,2]
Nenad Svrzikap
Edward Rietma
Kourosh Salehi
Michael E Cusic
Pascal Braun[1,2]

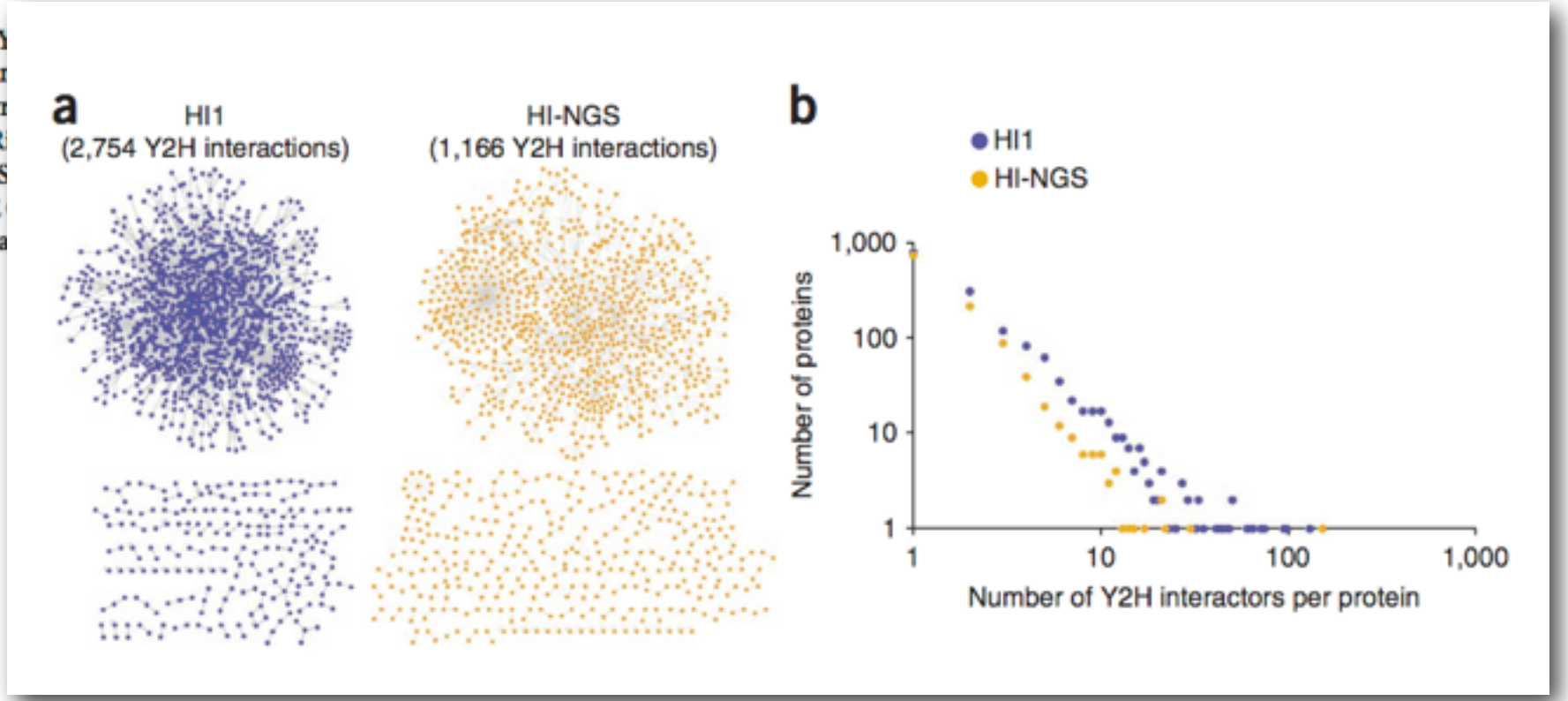


High-Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping

Albertha J. M. Walhout and Marc Vidal[1]

Next-generation sequencing to generate interactome datasets

Haiyuan Y
Evan Weir
Nenad Svr
Edward R
Kourosh S
Michael E
Pascal Bra



Evidence for Network Evolution in an *Arabidopsis* Interactome Map

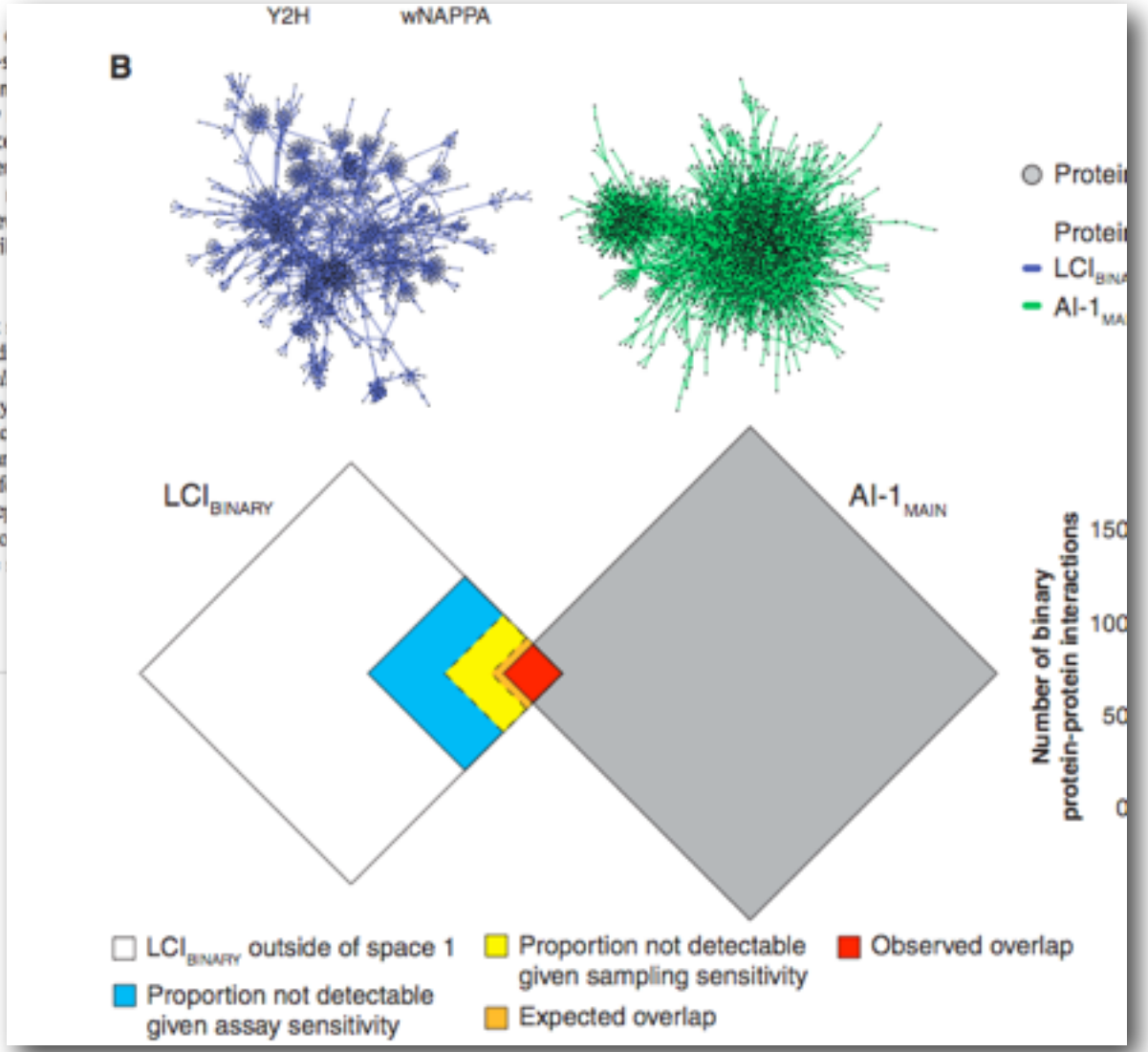
Arabidopsis Interactome Mapping Consortium*†

Plants have unique features that evolved in response to their
account of the complex cellular networks that underlie plant-
describe a proteome-wide binary protein-protein interaction
the plant *Arabidopsis thaliana* containing about 6200 highly
2700 proteins. A global organization of plant biological proc
analyses of the resulting network, together with large numbe
links between proteins and pathways. We observe a dynamic
gene duplication events, providing evidence for a model of e
networks. This and future plant interactome maps should faci
understand plant biology and improve crops.

Classical genetic and molecular approaches have provided fundamental understanding of processes such as growth control or development and molecular descriptions of genotype-to-phenotype relationships for a varie-

ty of plant
protein-cod
dopsis thal
functionally
the biologic
complex a
is lacking f
and S2), de
genotype-to
ated at the

*All authors with their affiliations and contributions are listed at the end of the paper.
†To whom correspondence should be addressed. E-mail: marc_vidal@dfci.harvard.edu; ecker@salk.edu; pascal_braun@dfci.harvard.edu; david_hill@dfci.harvard.edu



5664 interactions
between 2661 proteins

Evidence for Network Evolution in an *Arabidopsis* Interactome Map

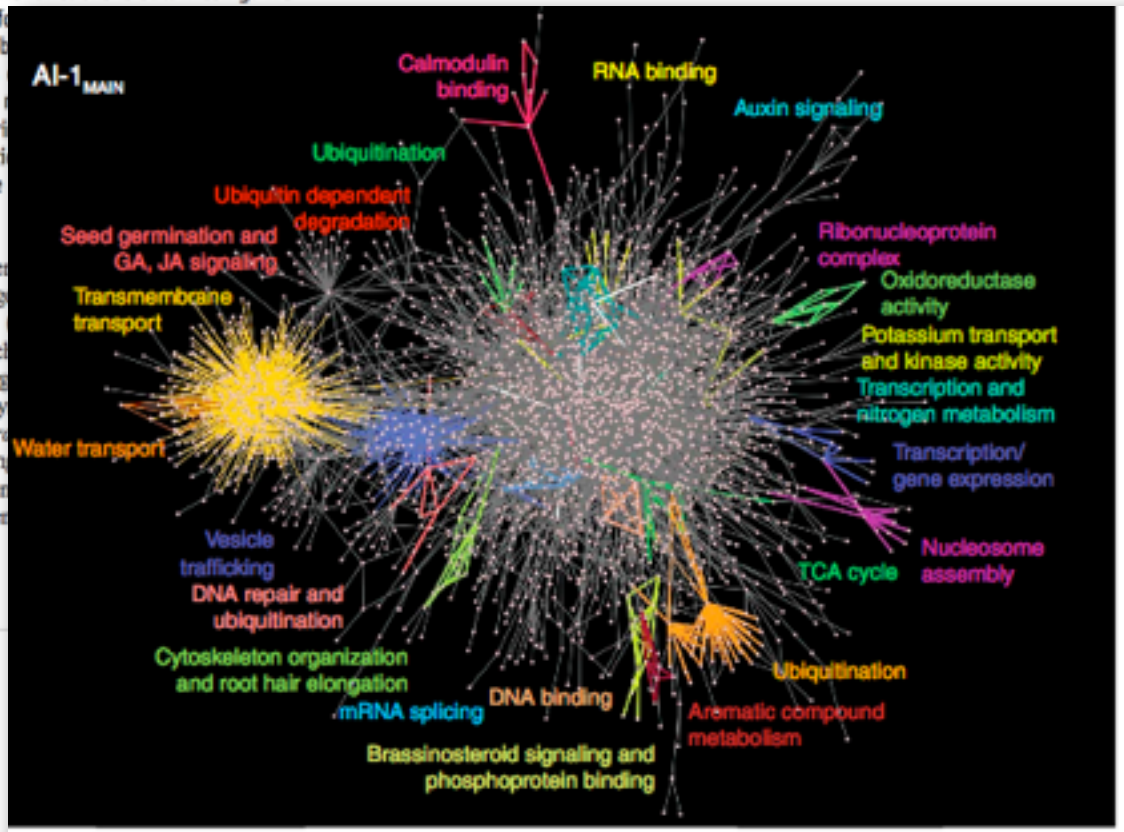
Arabidopsis Interactome Mapping Consortium*†

Plants have unique features that evolved in response to their environments and ecosystems. A full account of the complex cellular networks that underlie plant-specific functions is still missing. We describe a proteome-wide binary protein-protein interaction map f the plant *Arabidopsis thaliana* containing about 6200 highly rellab 2700 proteins. A global organization of plant biological processes analyses of the resulting network, together with large numbers of links between proteins and pathways. We observe a dynamic rewir gene duplication events, providing evidence for a model of evoluti networks. This and future plant interactome maps should facilitate understand plant biology and improve crops.

Classical genetic and molecular approaches have provided fundamental understanding of processes such as growth control or development and molecular descriptions of genotype-to-phenotype relationships for a varie-

ty of plant syste protein-coding g *dopsis thaliana* functionally unc the biological or complex and dy is lacking for *Ar* and S2), deprivin genotype-to-phen ated at the syste

*All authors with their affiliations and contributions are listed at the end of the paper.
†To whom correspondence should be addressed. E-mail: marc_vidal@dfci.harvard.edu; ecker@salk.edu; pascal_braun@dfci.harvard.edu; david_hill@dfci.harvard.edu



ig. 3. Communities in AI-1_{MAIN} (bottom) and in a typical randomized network (top left) (fig. S9). Only he largest connected component of each network is shown. Colored regions indicate communities nriched in GO annotations summarized by the indicated terms (table S10). (Upper right) Distribution f randomized networks as a function of the total number and number of GO annotation enriched communities they contain. White arrow, position of the shown randomized network; red dot and rrow, position of AI-1_{MAIN}. GA, gibberellic acid; JA, jasmonic acid; TCA, tricarboxylic acid.

LETTER

doi:10.1038/nature11354

Interaction landscape of membrane-protein complexes in *Saccharomyces cerevisiae*

Mohan Babu[1,2]*, James Vlasblom[3,4]*, Shuye Pu[3], Xinghua Guo[1], Chris Graham[1], Björn D. M. Bean[5], Helen E. Burston[5], Franco J. Vizeacoumar[1], Jamie Snider[1], Sadhna Phanse[1], Vincent Fong[1], Yuen Yi C. Tam[5], Michael Davey[5], Olha Hnatshak[1], Navgeet Bajaj[1], Shamanta Chandran[1], Thanuja Punna[1], Constantine Christopolous[1], Victoria Wong[1], Analyn Yu[1], Gouqing Zhong[1], Joyce Li[1], Igor Stagljar[3,4,6], Elizabeth Conibear[5], Shoshana J. Wodak[3,4,6], Andrew Emili[1,6] & Jack F. Greenblatt[1,6]

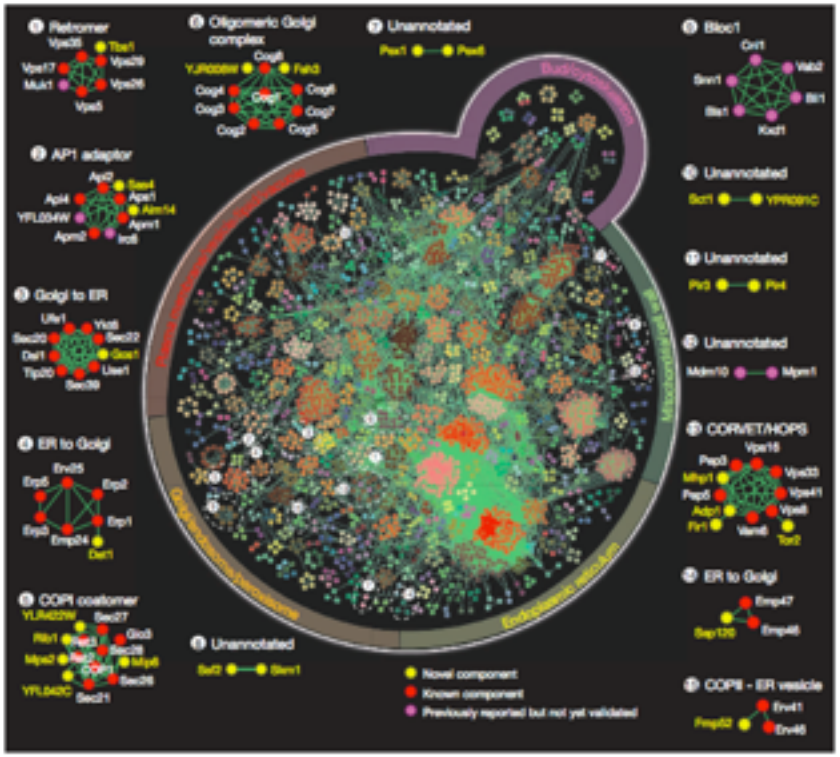


Figure 2 | Global organization of yeast MP complexes. Predicted MP clusters (subunits shown as similarly coloured nodes) inferred from the integrated network of high-confidence PPI (edges), demarcated according to primary compartment annotations. Representative complexes at the periphery highlight some of the findings of our study, including novel complexes and known complexes with new components. Our purifications were most successful for MPs localized to the Golgi and endoplasmic reticulum, a bias reflected in the highlighted examples. For each complex, previously reported components (red nodes), novel subunits (yellow nodes) and previously reported but not yet validated interactors (pink nodes) are displayed.

Agreement between databases

Literature curation of protein interactions: measuring agreement across major public databases

Andrei L. Turinsky[1], Sabry Razick[2,3], Brian Turner[1], Ian M. Donaldson[2,4] and Shoshana J. Wodak[1,5,6,*]

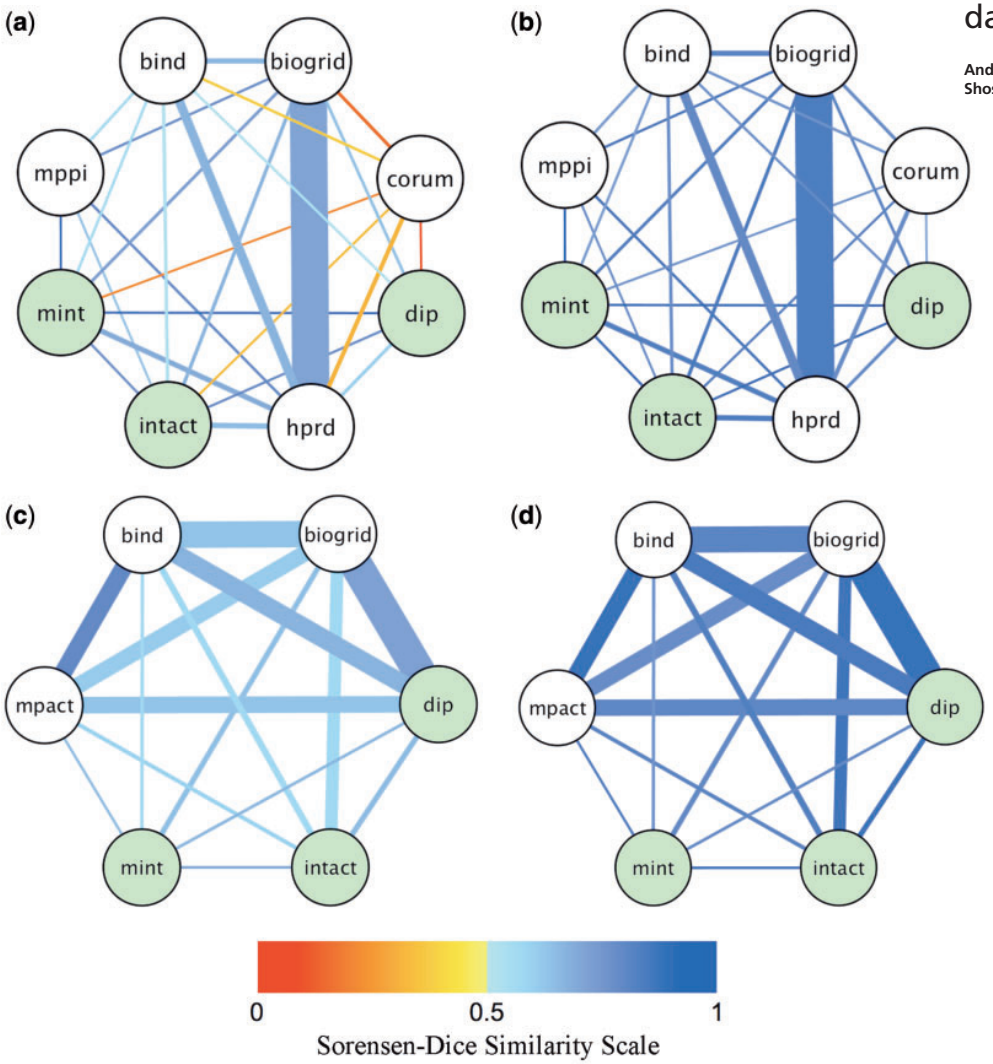
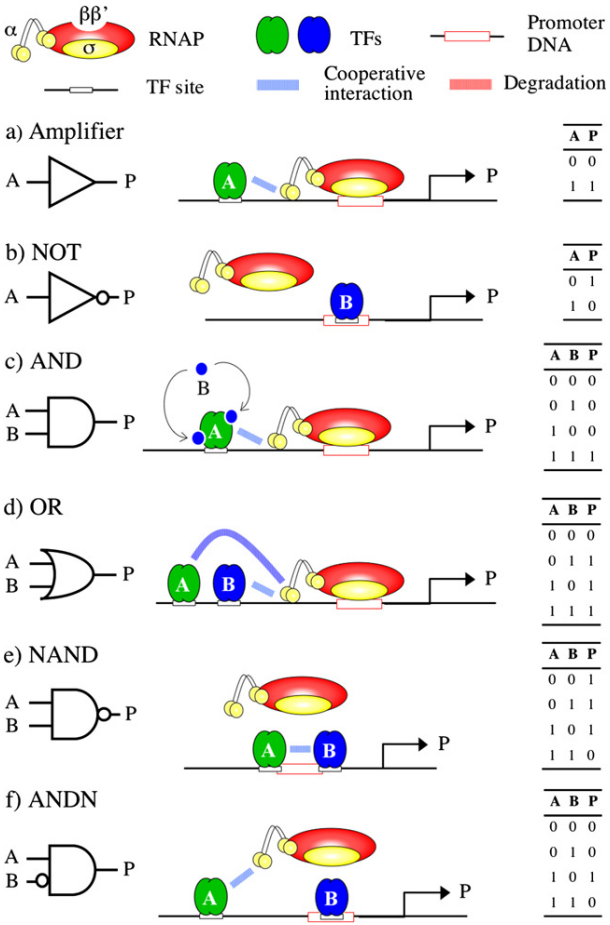
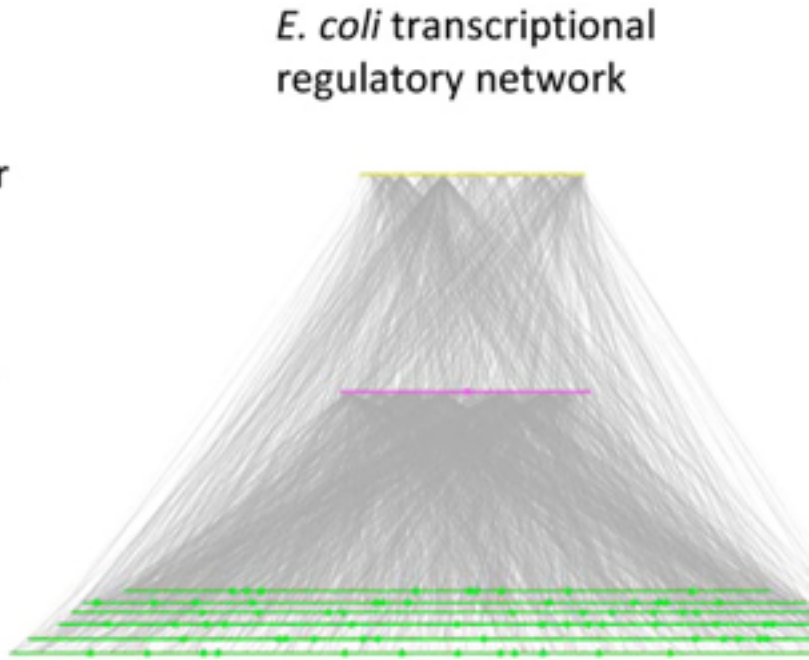


Figure 5. Pairwise agreement between databases for yeast-only and human-only co-citations. Shown is a pictorial summary of

Regulatory network





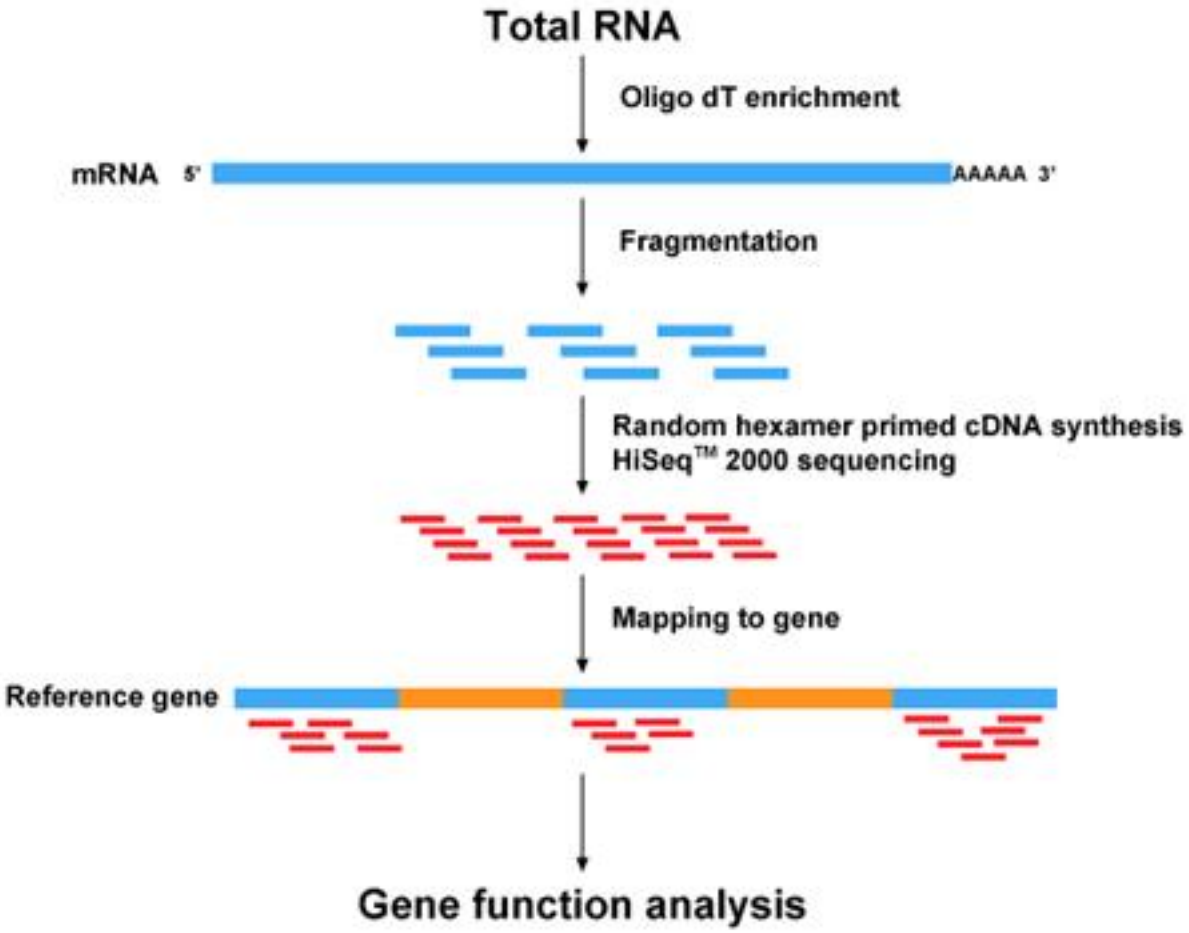
Mining logic gates in prokaryotic transcriptional regulation networks

Rafael Silva-Rocha, Víctor de Lorenzo*

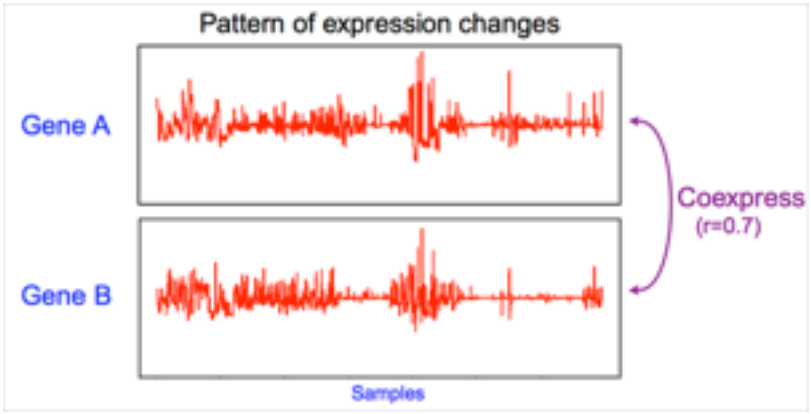
Centro Nacional de Biotecnología, CSIC, Campus de Cantoblanco, Madrid 28049, Spain

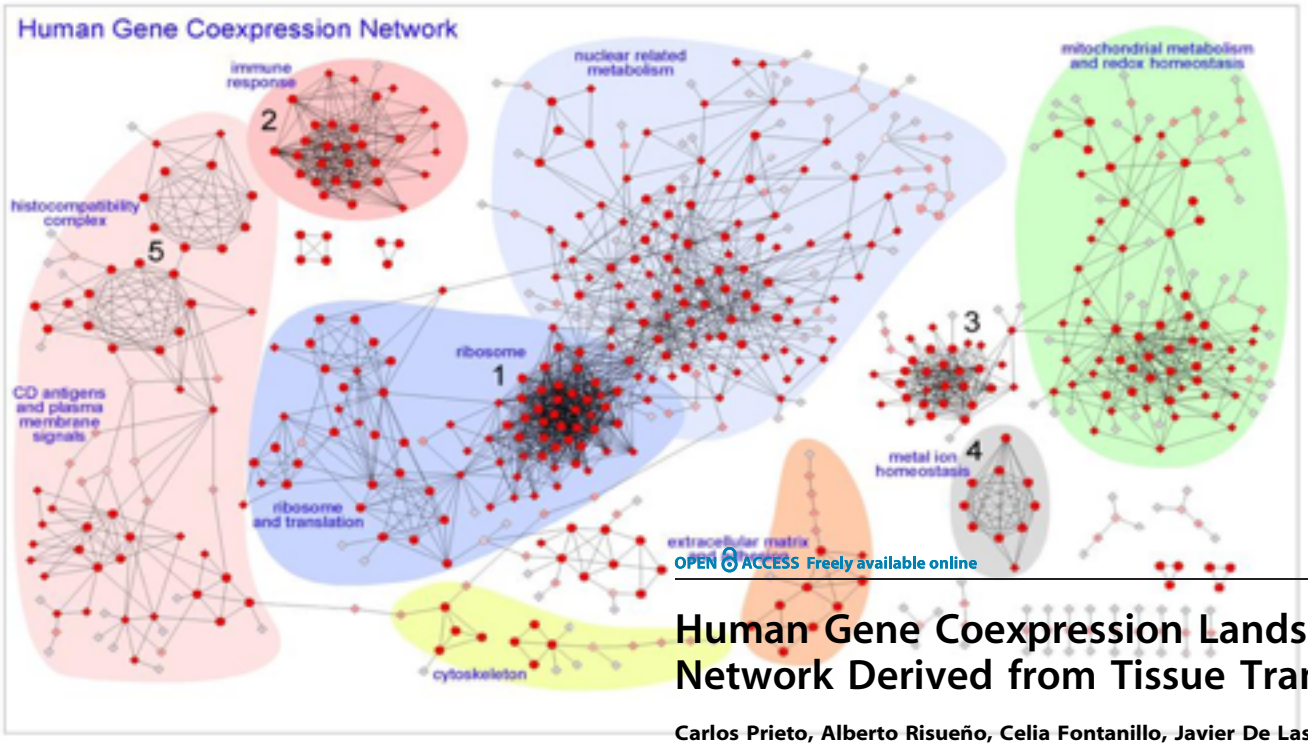
Received 10 January 2008; accepted 28 January 2008

Measuring gene expression



Network of co-expression





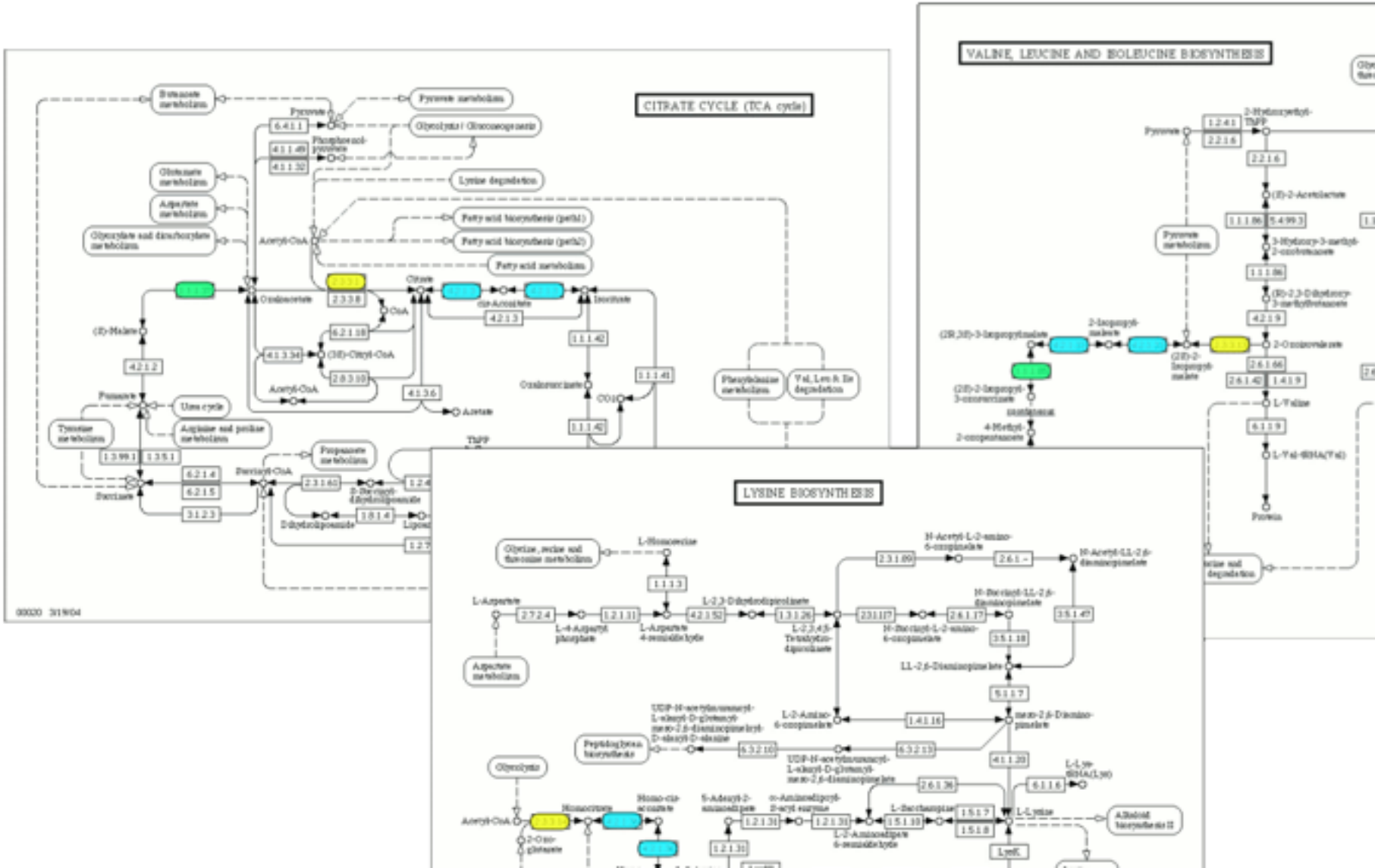
OPEN ACCESS Freely available online

PLoS one

Human Gene Coexpression Landscape: Confident Network Derived from Tissue Transcriptomic Profiles

Carlos Prieto, Alberto Risueño, Celia Fontanillo, Javier De Las Rivas*

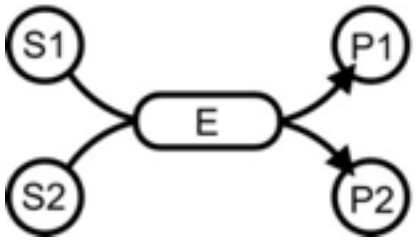
Metabolic networks



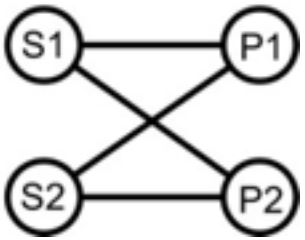
Metabolic networks

(a)

(b)



(c)



Organism	Genes in Genome	Genes in Model	Reactions	Metabolites	Date of reconstruction	Reference
Haemophilus influenzae	1,775	296	488	343	June 1999	[3]
Escherichia coli	4,405	660	627	438	May 2000	[5]
Saccharomyces cerevisiae	6,183	708	1,175	584	February 2003	[6]
Mus musculus	28,287	473	1220	872	January 2005	[7]
Homo sapiens	21,090[8]	3,623	3,673	--	January 2007	[9]
Mycobacterium tuberculosis	4,402	661	939	828	June 2007	[10]
Bacillus subtilis	4,114	844	1,020	988	September 2007	[11]
Synechocystis sp. PCC6803	3,221	633	831	704	October 2008	[12]
Salmonella typhimurium	4,489	1,083	1,087	774	April 2009	[13]
Arabidopsis thaliana	27,379	1,419,	1,567	1,748	February 2010	[14]

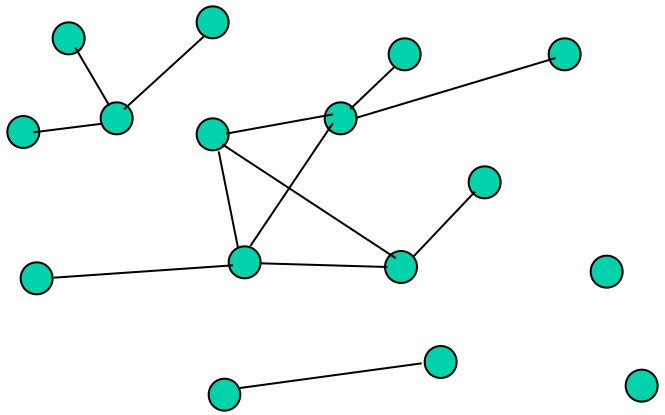
Outline

1. Learning biological networks: experiments
2. **Statistical properties of the networks**
3. Understanding networks structure: motifs, modules, etc
4. From structure to function
5. Compare/align networks
6. Dynamics of networks

Graph theory

Nodes: proteins

Edges: interactions

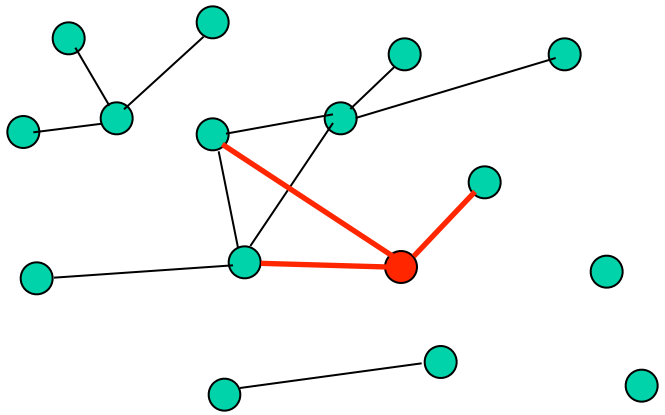


- Node degree
- Components
- Shortest path
- Diameter=Longest Shortest Path

Graph theory

Nodes: proteins

Edges: interactions

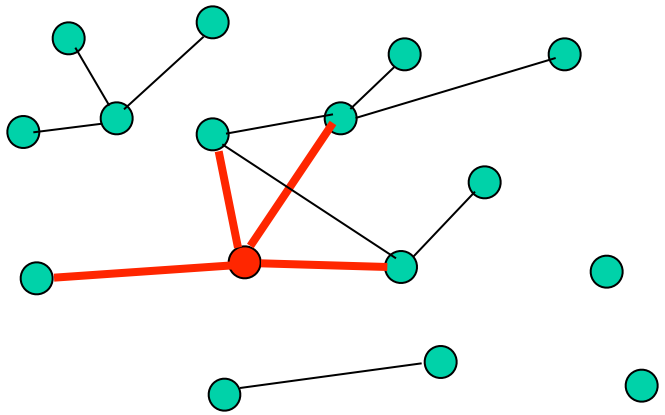


- Node degree d(k)
- Components
- Shortest path
- Diameter=Longest Shortest Path

Graph theory

Nodes: proteins

Edges: interactions

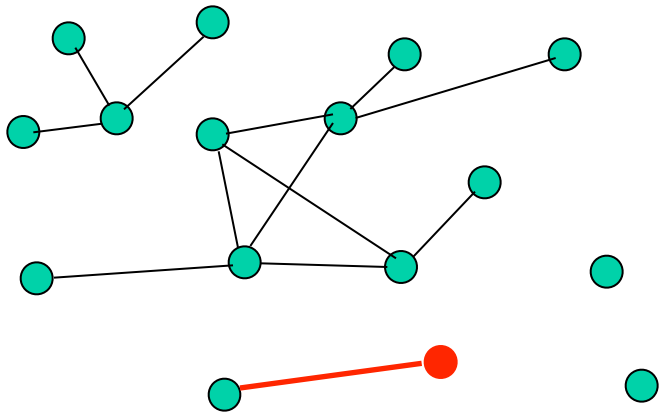


- Node degree d(k)
- Components
- Shortest path
- Diameter=Longest Shortest Path

Graph theory

Nodes: proteins

Edges: interactions

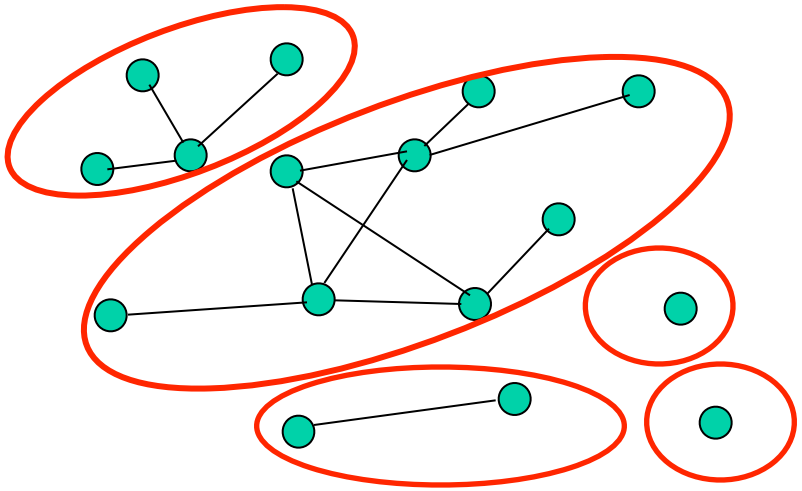


- Node degree d(k)
- Components
- Shortest path
- Diameter=Longest Shortest Path

Graph theory

Nodes: proteins

Edges: interactions

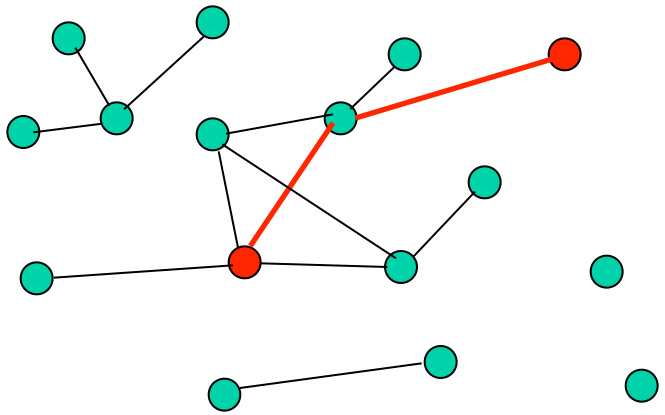


- Node degree
- Components
- Shortest path
- Diameter=Longest Shortest Path

Graph theory

Nodes: proteins

Edges: interactions

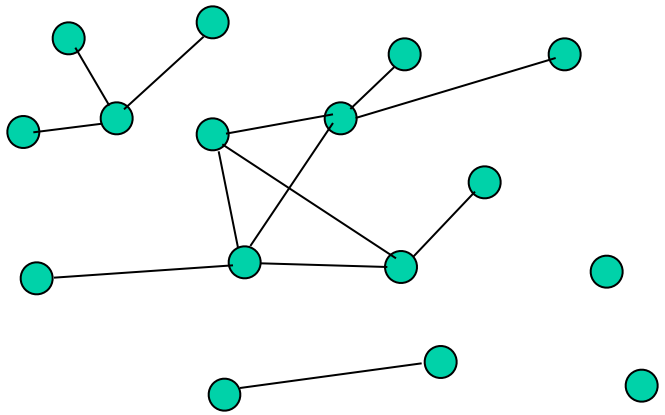


- Node degree
- Components
- Shortest path
- Diameter=Longest Shortest Path

Graph theory

Nodes: proteins

Edges: interactions

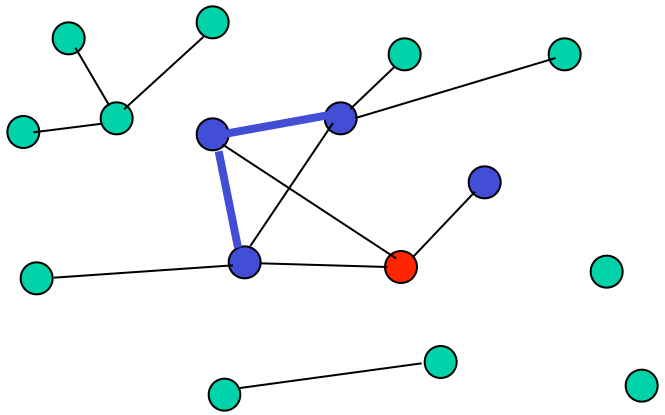


- Node degree
- Components
- Shortest path
- Diameter=Longest Shortest Path

Graph theory

Nodes: proteins

Edges: interactions

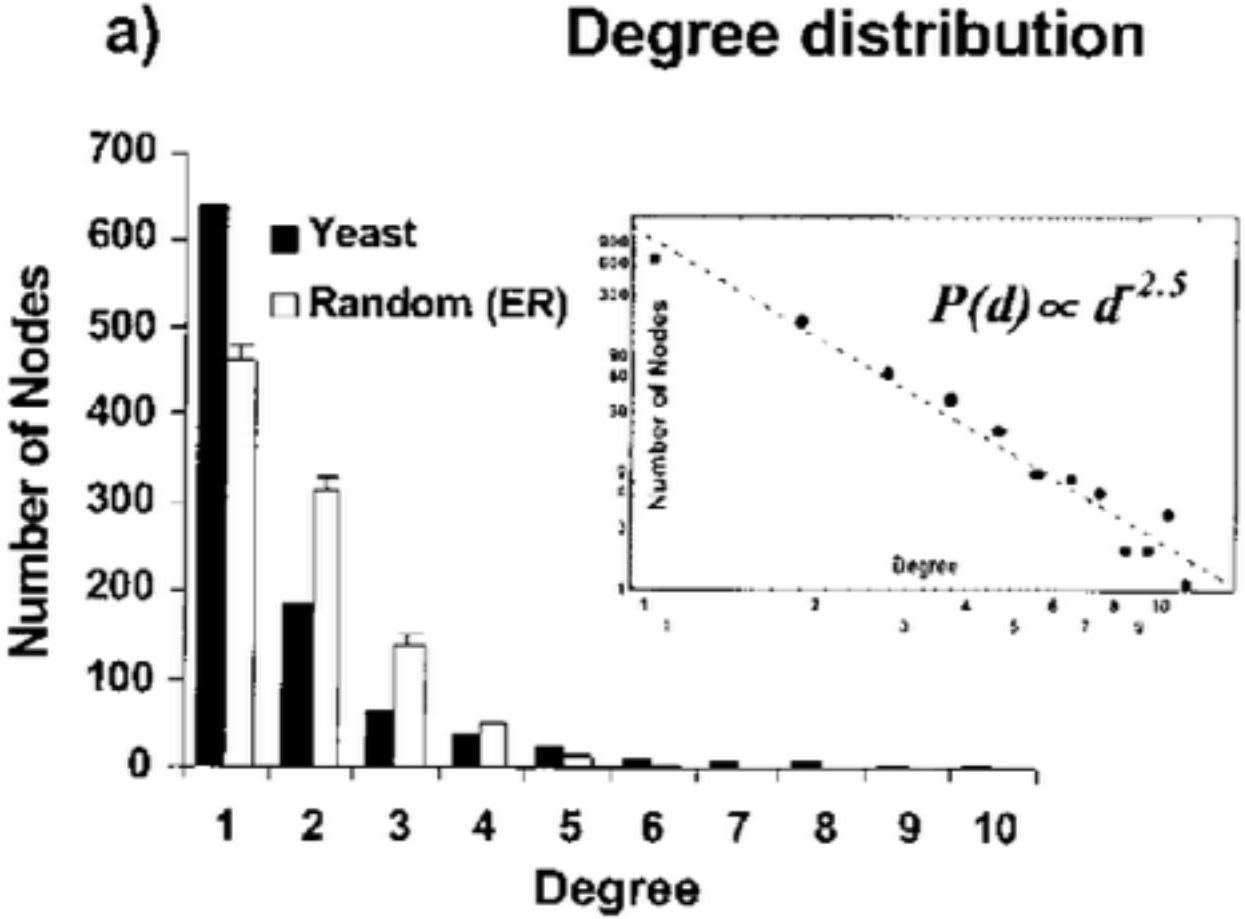


- Node degree
- Components
- Shortest path
- Diameter=<Shortest Path>
- Clustering coefficient

$$c_i = \Pr\{\Delta_{jk} = 1 | \Delta_{ij} = \Delta_{ik} = 1\}$$

it's not a random graph!

Andreas Wagner
Mol. Biol. Evol. 18(7):1283–1292. 2001



it's not a random graph!

Table 1
Comparison of Statistical Features Between Random Graphs and the Yeast Protein Interaction Network

		Random Graphs	
	Yeast	ER	PL ($\tau = 2.5$)
Whole graph			
Nodes	985	984.02 (10.39)	970.7 (81.57)
Degree	1.83	1.85 (0.98)	1.64 (1.76)
No. of components	163	108 (8)*	266.3 (30.6)*
Giant component			
Nodes	466	624.0 (38.7)*	336.9 (86)
Degree	2.3	2.07 (1.05)	2.50 (2.6)
Clustering coefficient ($\times 10^{-3}$)	22	0.59 (0.9)*	4.02 (2.3)*
Characteristic path length	7.14	15.88 (1.76)*	6.01 (1.14)

Wagner MBE 2000

it’s not a random graph!

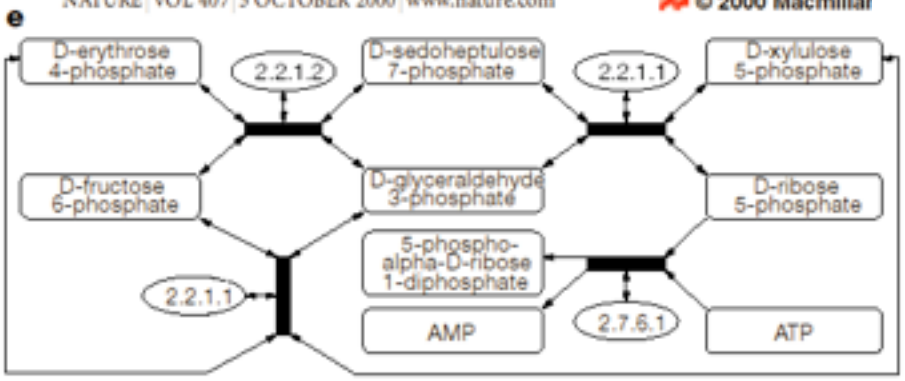
The large-scale organization of metabolic networks

H. Jeong*, B. Tombor†, R. Albert*, Z. N. Oltvai† & A.-L. Barabási*

* *Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA*
† *Department of Pathology, Northwestern University Medical School, Chicago, Illinois 60611, USA*

In a cell or microorganism, the processes that generate mass, energy, information transfer and cell-fate specification are seamlessly integrated through a complex network of cellular constituents and reactions[1]. However, despite the key role of these networks in sustaining cellular functions, their large-scale structure is essentially unknown. Here we present a systematic comparative mathematical analysis of the metabolic networks of

NATURE | VOL 407 | 5 OCTOBER 2000 | www.nature.com



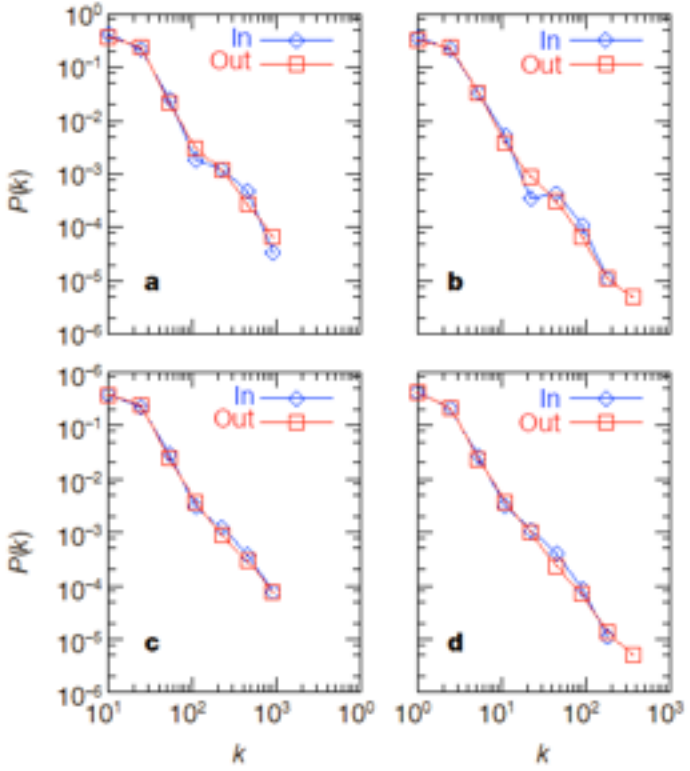
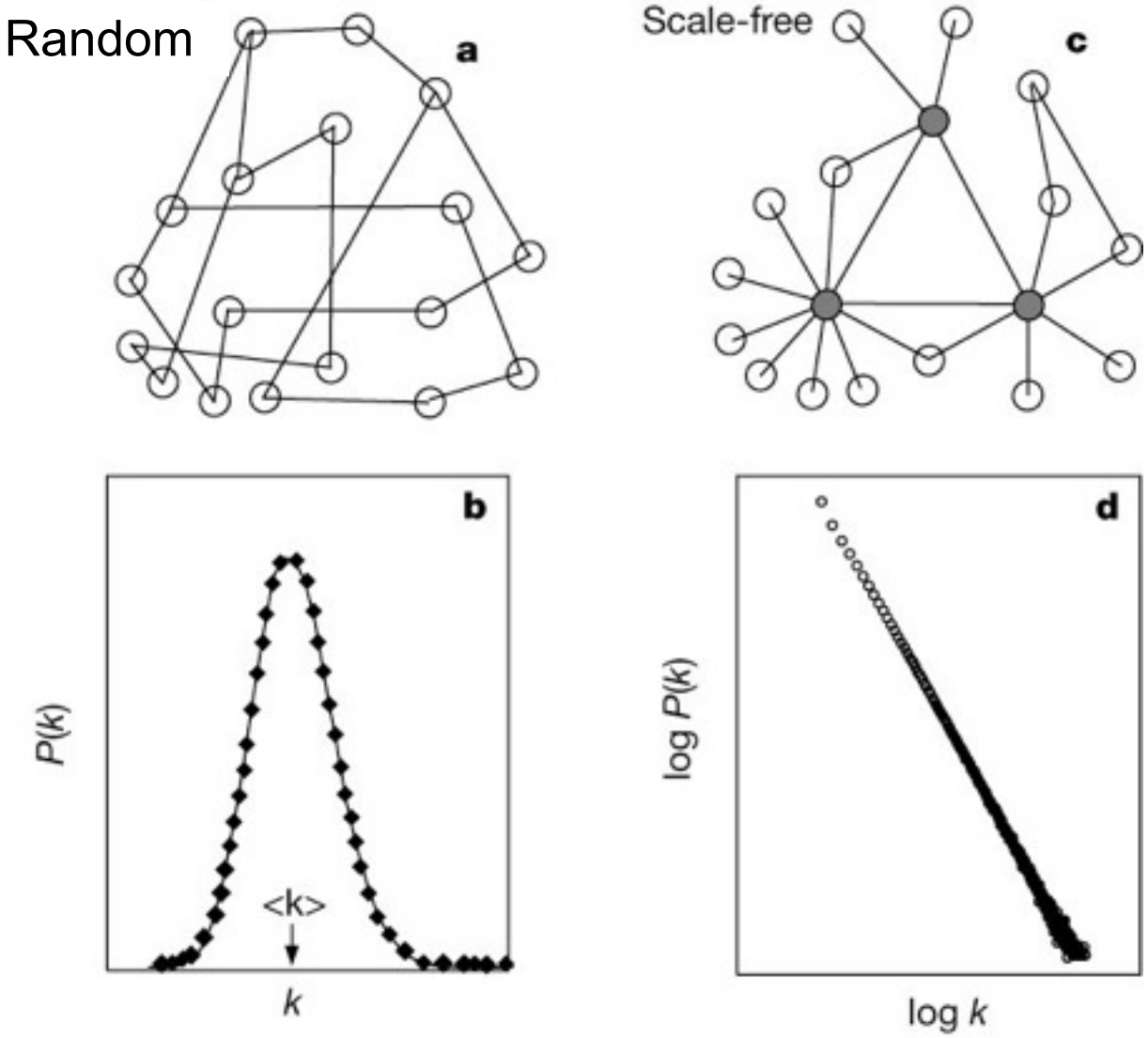


Figure 2 Connectivity distributions $P(k)$ for substrates. **a**, *Archaeoglobus fulgidus* (archae); **b**, *E. coli* (bacterium); **c**, *Caenorhabditis elegans* (eukaryote), shown on a log–log plot, counting separately the incoming (In) and outgoing links (Out) for each substrate. k_{in} (k_{out}) corresponds to the number of reactions in which a substrate participates as a product (educt). The characteristics of the three organisms shown in **a**–**c** and the exponents γ_{in} and γ_{out} for all organisms are given in Table 1 of the Supplementary Information. **d**, The connectivity distribution averaged over all 43 organisms.

IT’S ALMOST SCALE-FREE (=POWER-LAW) GRAPH

Random vs power-law

Barabasi A et.al. Nature:411(2001)
Wagner A Mol Biol Evol:18(2001)



Other power-law networks

Barabasi A et.al. Nature:411(2001)

Other power-law networks:

- Metabolic network
- Network of social interactions:

  scientific collaborations, actors in films
- The Internet:

  links, physical connections

The web of human sexual contacts

Promiscuous individuals are the vulnerable nodes to target in safe-sex campaigns.

Unlike clearly defined 'real-world' networks[1], social networks tend to be subjective to some extent[2,3] because the perception of what constitutes a social link may differ between individuals. One unambiguous type of co
is sexual contact, and h
sexual behaviour of a r
individuals[4] to reveal the
tures of a sexual-contact
that the cumulative di
number of different sexu
year decays as a scale-fr
has a similar exponen
females. The scale-free n
human sexual contacts i
egic safe-sex campaigns
most efficient way to pr
sexually transmitted dise

Many real-world net

naires. The response rate was 59%, which corresponds to 2,810 respondents. Two independent analyses of non-response error revealed that elderly people, particularly women, are under-represented in the sam-

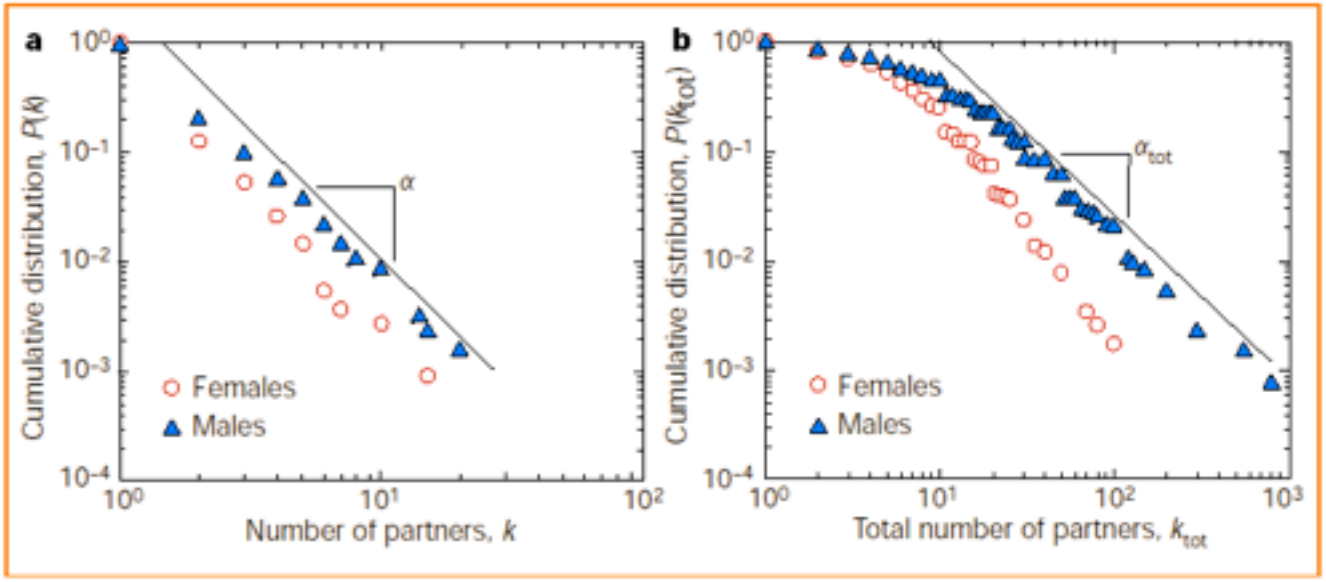


Figure 2 Scale-free distribution of the number of sexual partners for females and males. **a,** Distribution of number of partners, k, in the previous 12 months. Note the larger average number of partners for male respondents: this difference may be due to 'measurement bias' — social expectations may lead males to inflate their reported number of sexual partners. Note that the distributions are both linear, indicating scale-free power-law behaviour. Moreover, the two curves are roughly parallel, indicating similar scaling exponents. For females, $\alpha = 2.54 \pm 0.2$ in the range $k > 4$, and for males, $\alpha = 2.31 \pm 0.2$ in the range $k > 5$. **b,** Distribution of the total number of partners k_{tot} over respondents' entire lifetimes. For females, $\alpha_{tot} = 2.1 \pm 0.3$ in the range $k_{tot} > 20$, and for males, $\alpha_{tot} = 1.6 \pm 0.3$ in the range $20 < k_{tot} < 400$. Estimates for females and males agree within statistical uncertainty.

Random vs power-law

Barabasi A et.al. Nature:411(2001)
Wagner A Mol Biol Evol:18(2001)

The network of protein-protein interactions (and other molecular biological networks) are scale-free networks!

WHY?

- Scale-free networks are "better"…

  OR/AND

- Biological networks became scale-free due to evolution.

Random

- Removal of a randomly picked node significantly increases the average path.
- All nodes are of equal "importance".

Power-law

- Removal of a random node slightly increases the average path.
- Removal of a highly-connected node leads to drastic increase of the average path!

Let's check it

POWER LAW NETWORKS

- Tolerant to random "attacks",
- But more sensitive to targeted attacks!

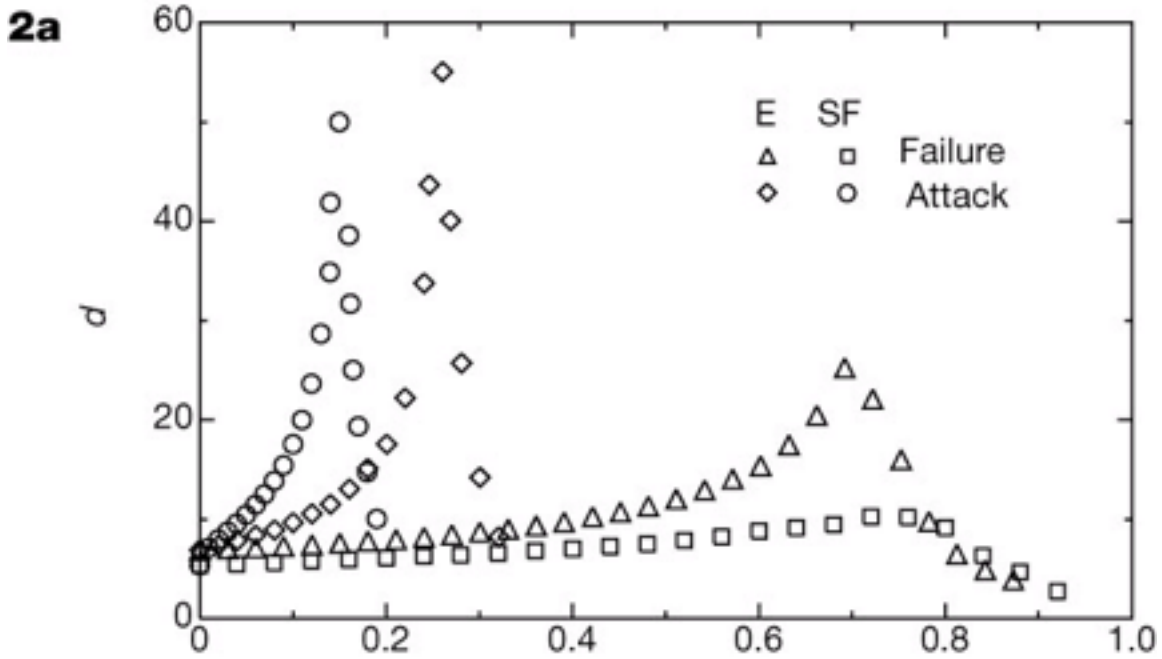
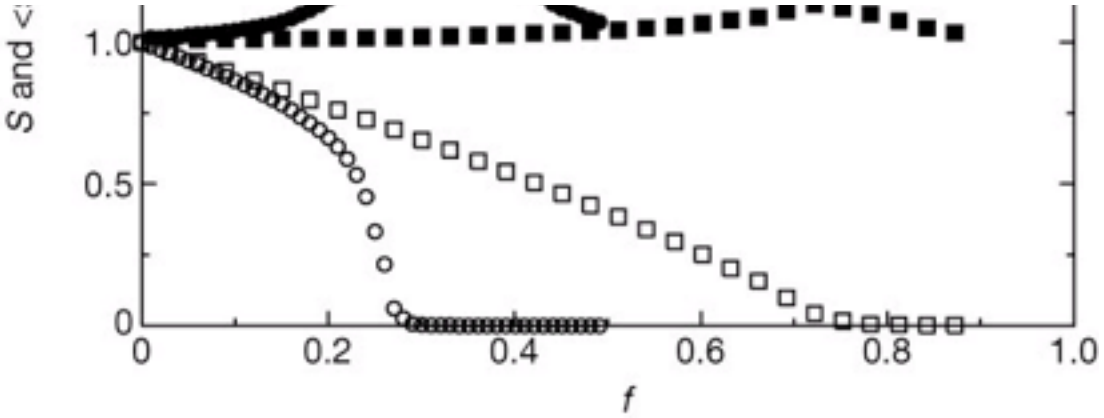


Figure 2 Changes in the diameter *d* of the network as a function of the fraction *f* of the removed nodes. **a**, Comparison between the exponential (E) and scale-free (SF) network models, each containing *N* = 10,000 nodes and 20,000 links (that is, *k* = 4).



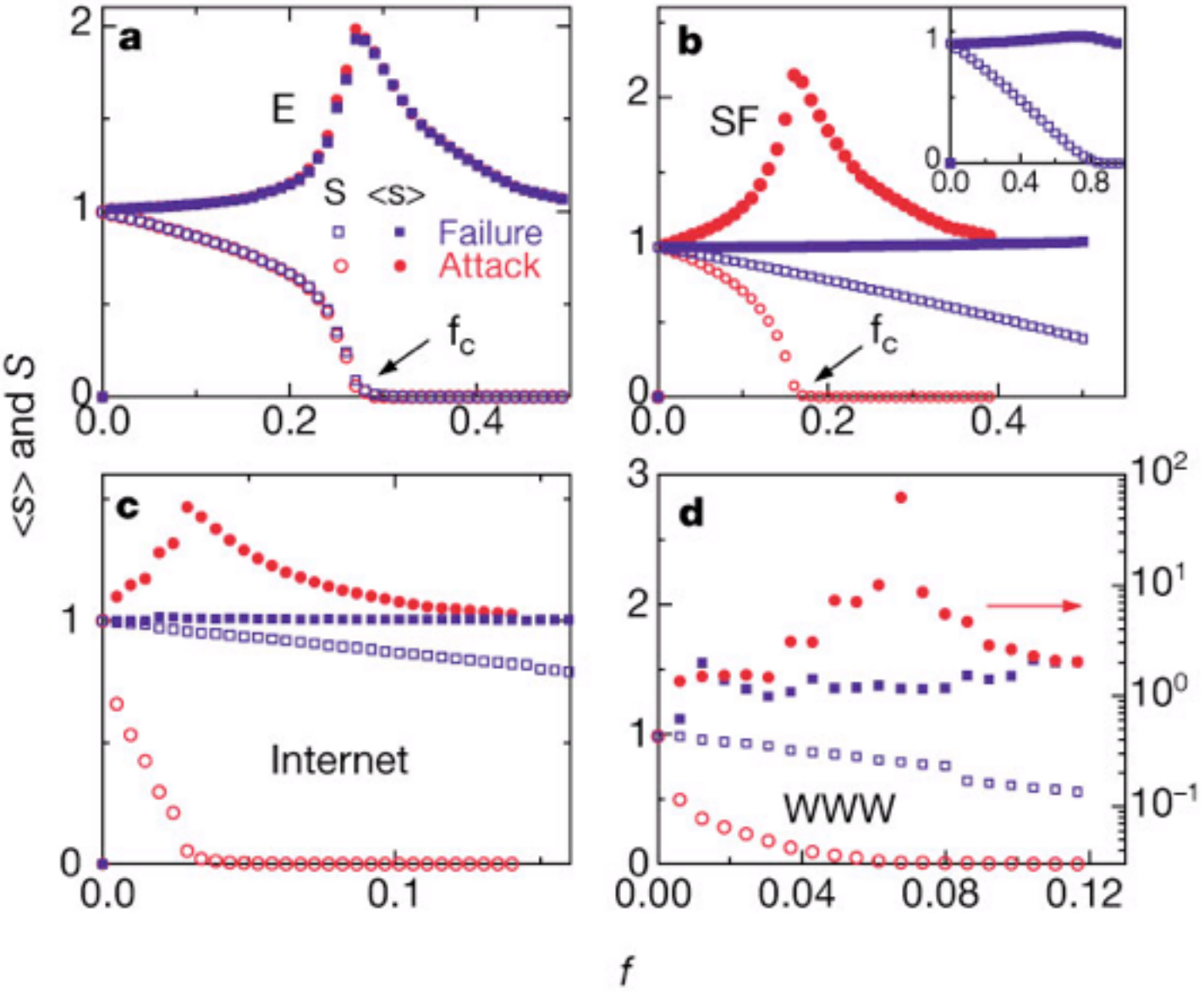
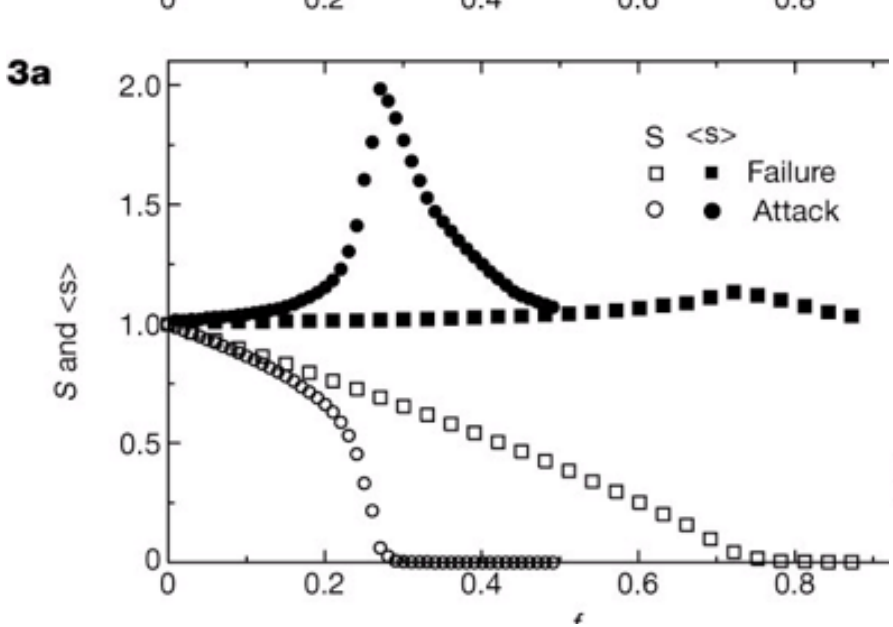


Figure 3 Network fragmentation under random failures and attacks. The relative size of the largest cluster S (open symbols) and the average size of the isolated clusters s (filled symbols) as a function of the fraction of removed nodes f for the same systems as in Fig. 2. The size S is defined as the fraction of nodes contained in the largest cluster (that is, $S = 1$



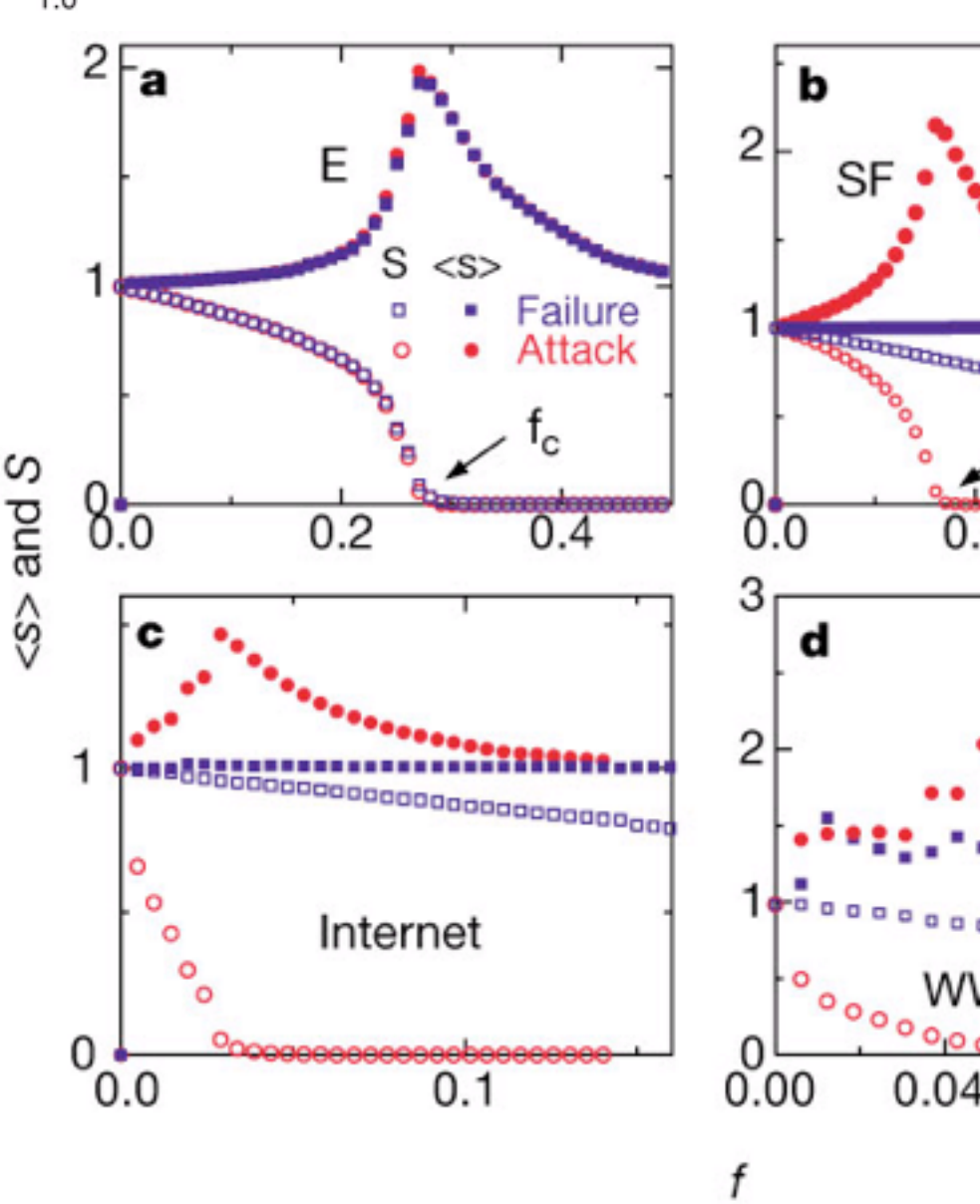


Figure 3 Network fragmentation under rando
the largest cluster *S* (open symbols) and the av
symbols) as a function of the fraction of remo
The size *S* is defined as the fraction of nodes
for $f = 0$). **a**, Fragmentation of the exponential
and attacks (circles). **b**, Fragmentation of the
(blue squares) and attacks (red circles). The in
whole range of *f*, indicating that the main clus
completely deflated

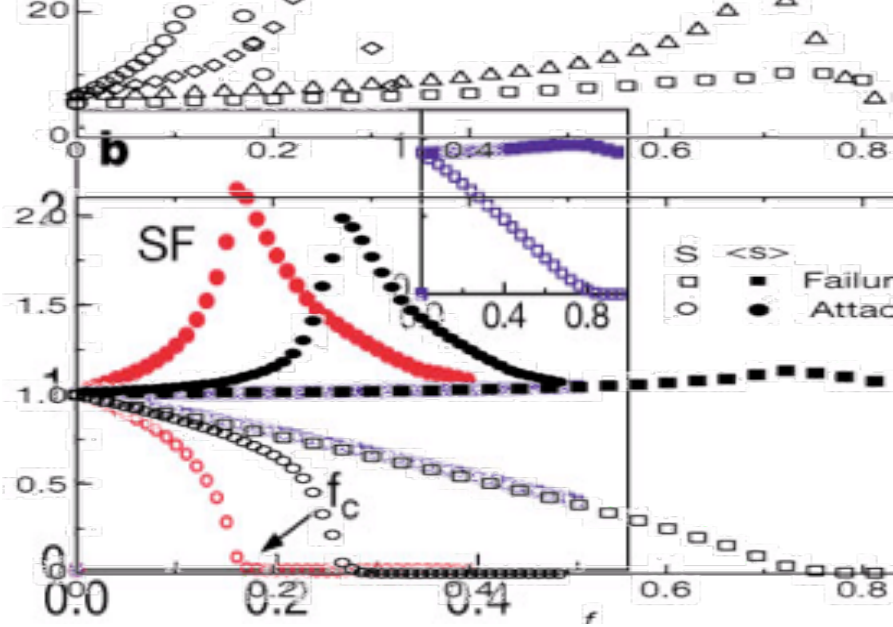


Figure 3 Network fragmentation under random failures and attacks. The relative size of the largest cluster S (open symbols) and the average size of the isolated clusters s (filled symbols) as a function of the fraction of removed nodes f for the same systems as in Fig. 2. The size S is defined as the fraction of nodes contained in the largest cluster (that is, $S = 1$ for $f = 0$). **a**, Fragmentation of the exponential network under random failures (squares) and attacks (circles). **b**, Fragmentation of the scale-free network under random failures (blue squares) and attacks (red circles). The inset shows the error tolerance curves for the whole range of f, indicating that the main cluster falls apart only after it has been completely deflated

0.0 0.1 0.00 0.04 0.08 0.12

f

Random | Power-law

Equally stable to random failures

More sensitive to attacks

POWER-LAW NETWORKS

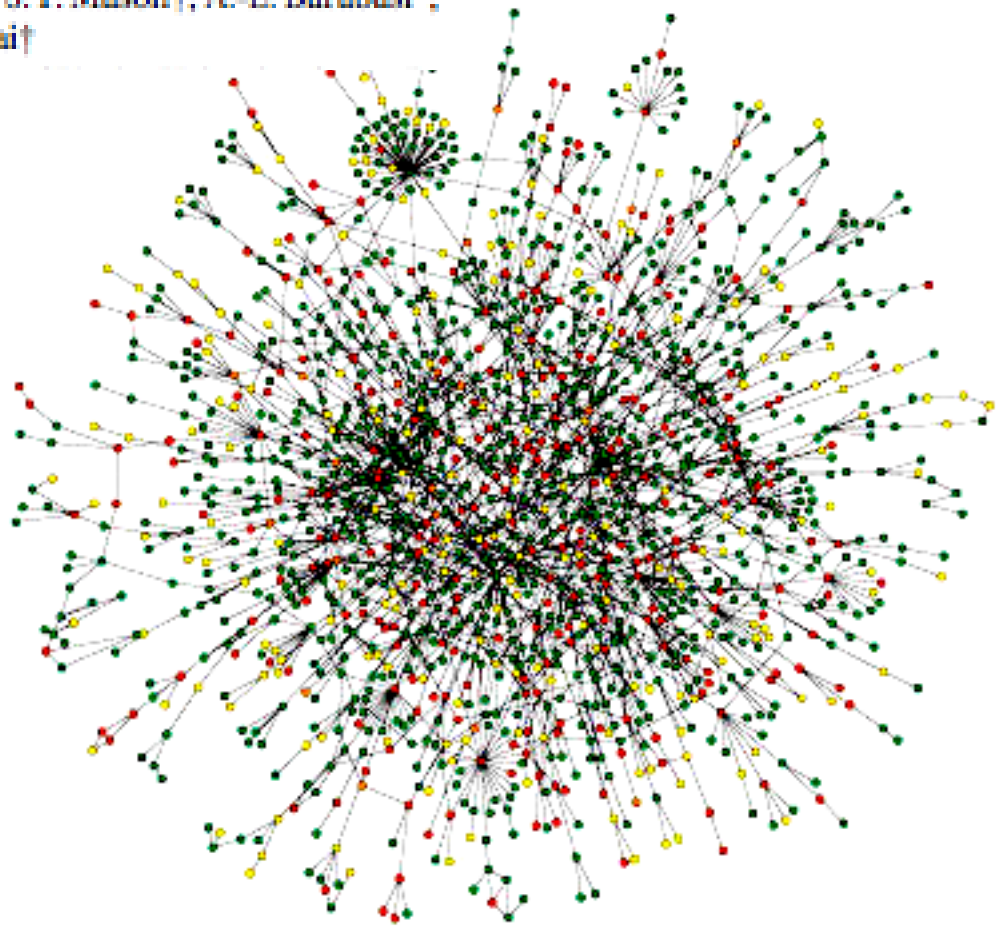
- Tolerant to random "attacks",
- But more sensitive to targeted attacks!

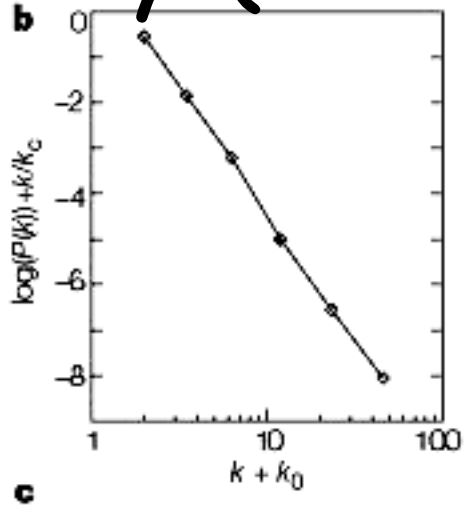
Outline

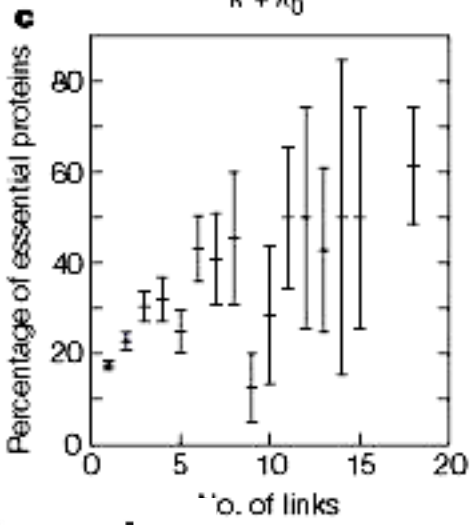
1. Learning biological networks: experiments
2. Statistical properties of the networks
3. **Understanding networks structure: motifs, modules, etc**
4. **From structure to function**
5. Compare/align networks
6. Dynamics of networks

Lethality and centrality (2001)

H. Jeong*, S. P. Mason†, A.-L. Barabási*,
Z. N. Oltvai†







Lethality and centrality in protein networks

The most highly connected proteins in the cell are the most important for its survival.

A simple physical model for scaling in protein–protein interaction networks

Eric J. Deeds*, Orr Ashenberg†, and Eugene I. Shakhnovich‡§

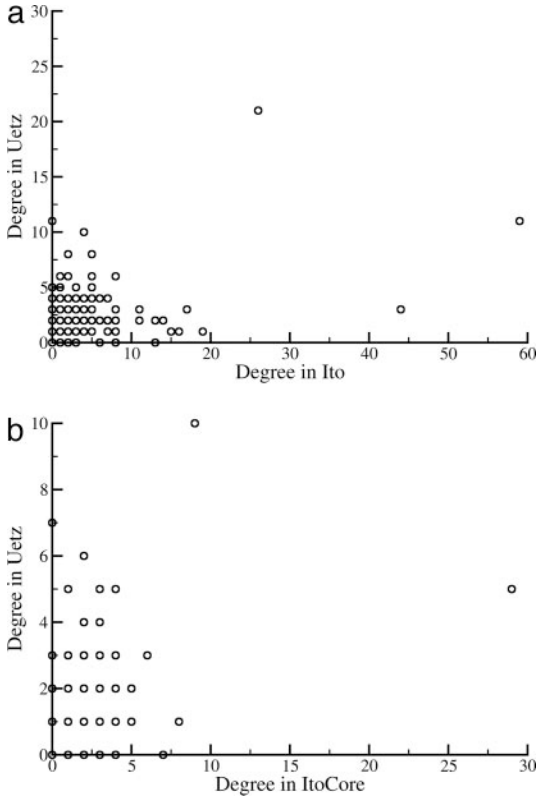


Fig. 1. Correlation between PPI networks. (*a*) The correlation between the network degree of a given protein in the Ito (13) and Uetz (12) data sets. Each point corresponds to a particular protein that exhibited interactions in both experiments. (*b*) A plot similar to *a* but comparing the ItoCore data set with Uetz.

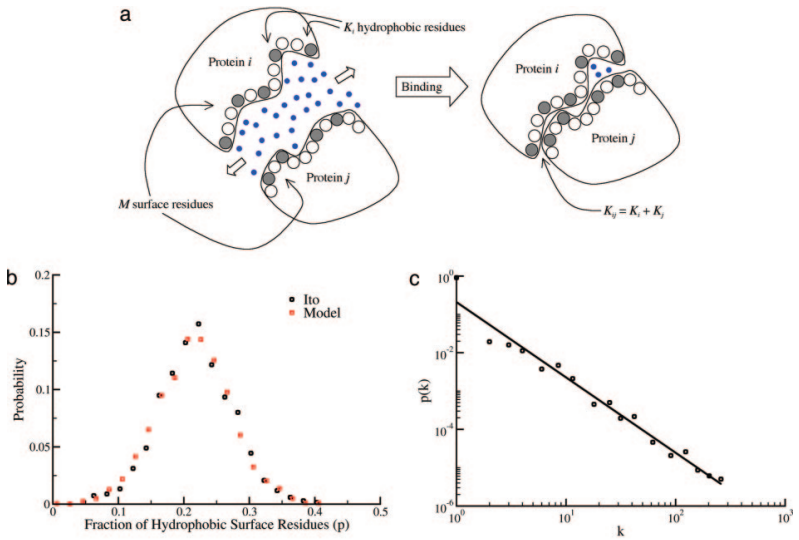
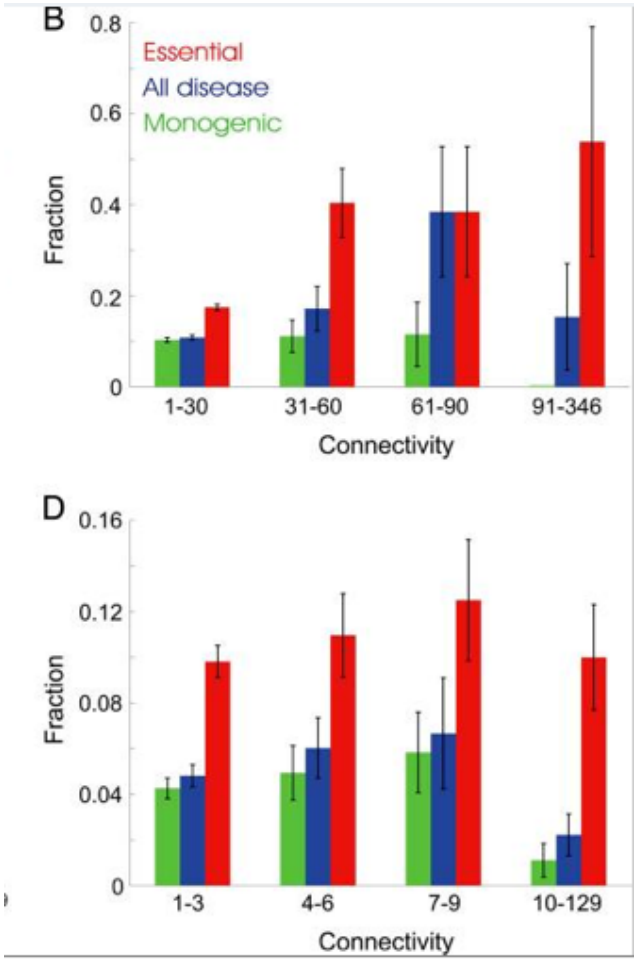


Fig. 2. A physical model for PPI measurements. (*a*) A schematic of the model described in the text. Association free energies are largely the result of desolvation of the two protein surfaces. The overall burial of hydrophobic groups is represented by the sum of the contributions from each protein. (*b*) The distribution of surface hydrophobicities in yeast proteins. The fraction of surface residues that are hydrophobic (defined as residues AVILMFYW) is calculated according to the description in the supporting information. This distribution is taken from proteins in the Ito experiment (13). The red squares represent the model hydrophobicities sampled from a Gaussian distribution with the same mean and standard deviation as the Ito proteins themselves. (*c*) A degree distribution for the realization of the model used in *b*. The cutoff was chosen such that the power-law fit gives an exponent of approximately −2.0, close to that of Ito graph. The degrees in this plot are shifted by +1 to allow for orphans (nodes of degree 0) to be displayed on a log–log plot. Note that the fraction of orphans in the graph is very high.

Lethality and centrality (2008)



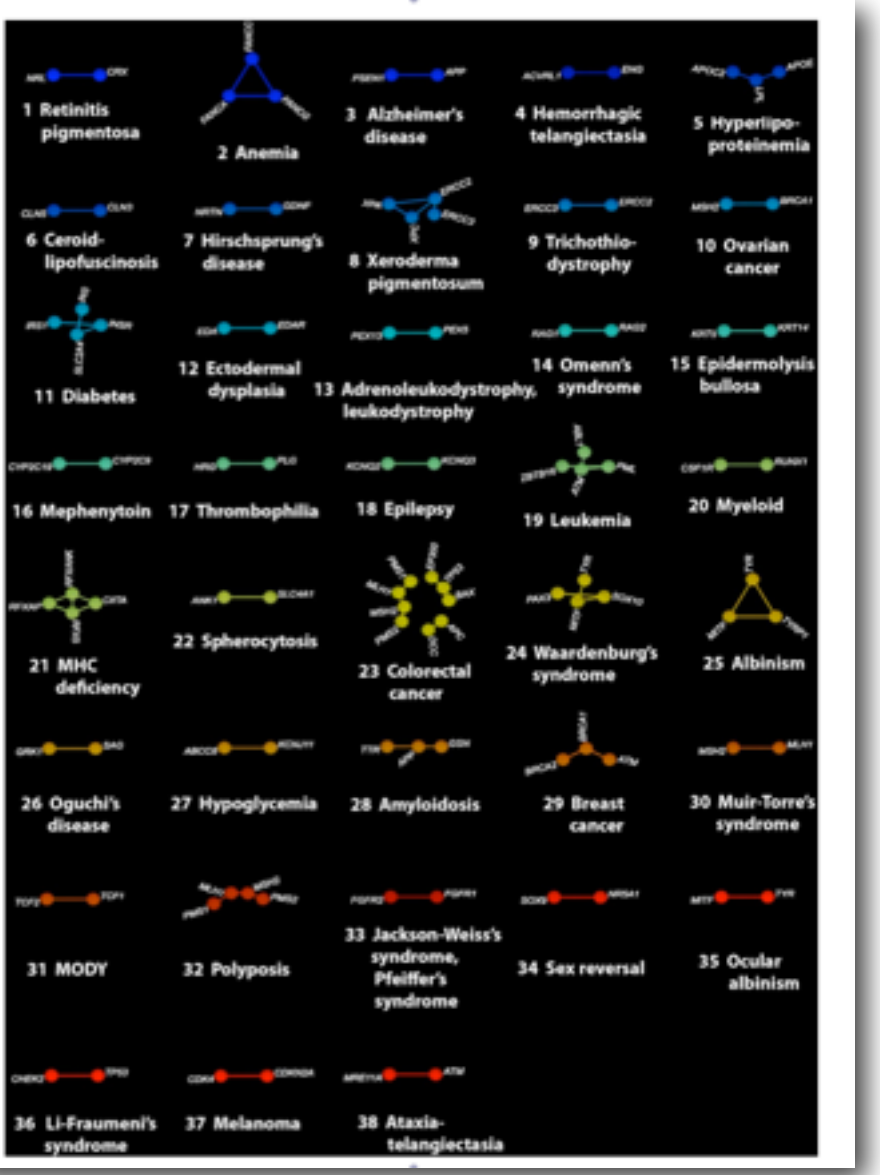
GeneWays

Y2H

Network properties of genes harboring inherited disease mutations

Igor Feldman*, Andrey Rzhetsky*†‡, and Dennis Vitkup*‡

PNAS | March 18, 2008



Network properties of genes harboring inherited disease mutations

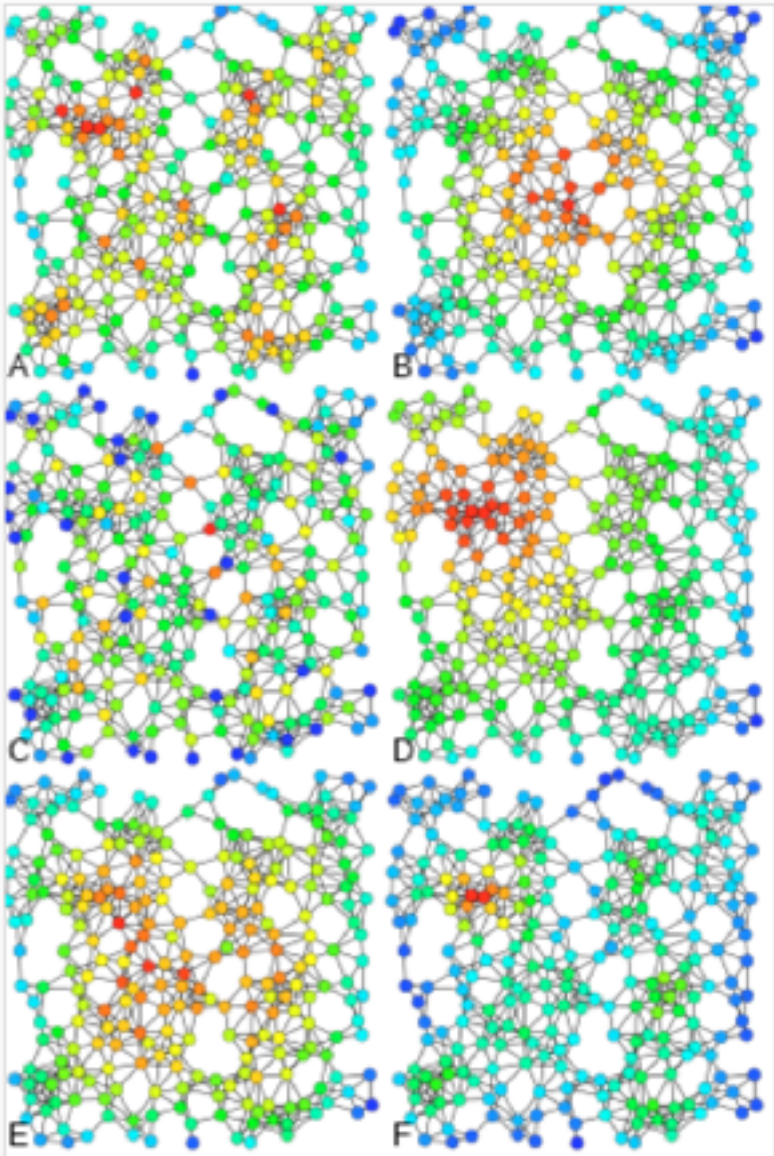
Igor Feldman*, Andrey Rzhetsky*†‡, and Dennis Vitkup*‡

PNAS | March 18, 2008

degree centrality = degree

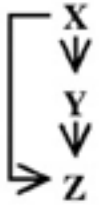
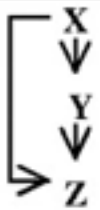
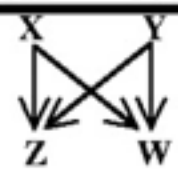
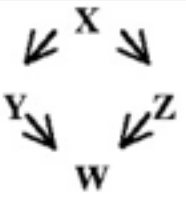
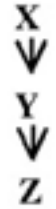
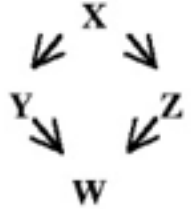
Closeness centrality (farness) =
sum of its distances to all other nodes

Betweenness centrality of i = the number of shortest path
between all other nodes which go through i

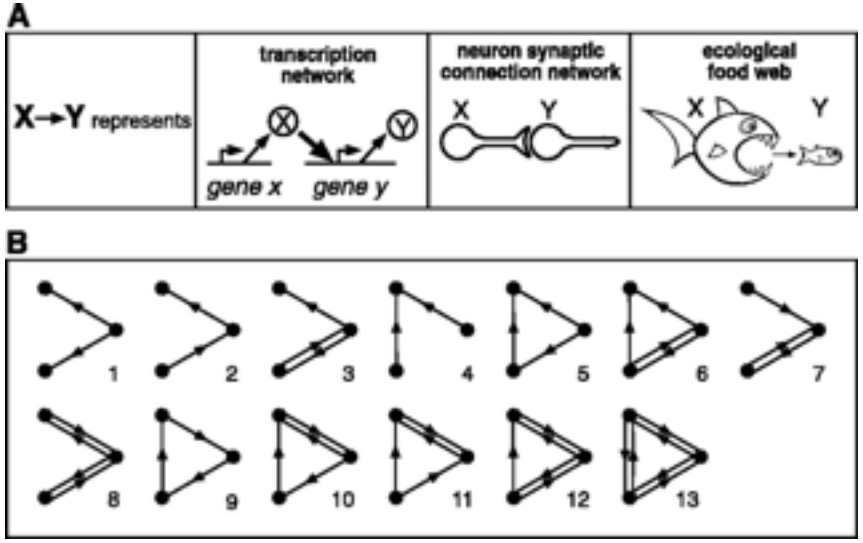


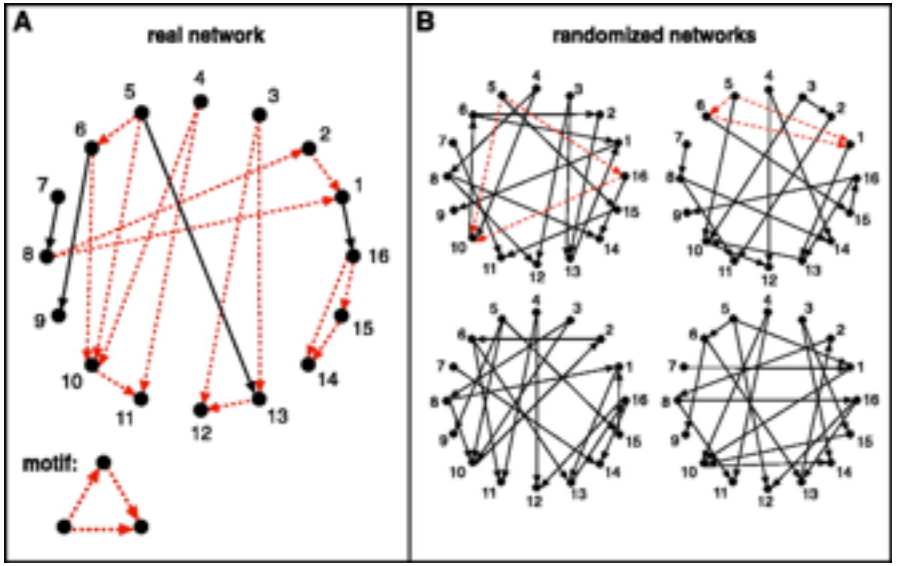
Examples of A) Degree centrality, B) Closeness centrality, C) Betweenness centrality, D) Eigenvector centrality, E) Katz centrality and F) Alpha centrality of the same graph.

Motifs

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-pa**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			

Motifs





25 OCTOBER 2002 VOL 298 SCIENCE

Network Motifs: Simple Building Blocks of Complex Networks

R. Milo,[1] S. Shen-Orr,[1] S. Itzkovitz,[1] N. Kashtan,[1] D. Chklovskii,[2] U. Alon[1]*

Clusters in networks

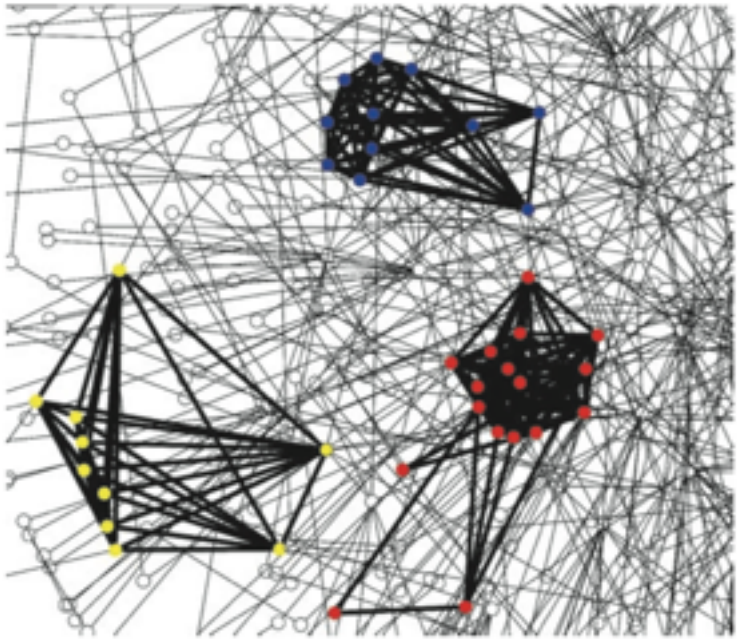


Fig. 2. Fragment of the protein network. Nodes and interactions in discovered clusters are shown in bold. Nodes are colored by functional categories in

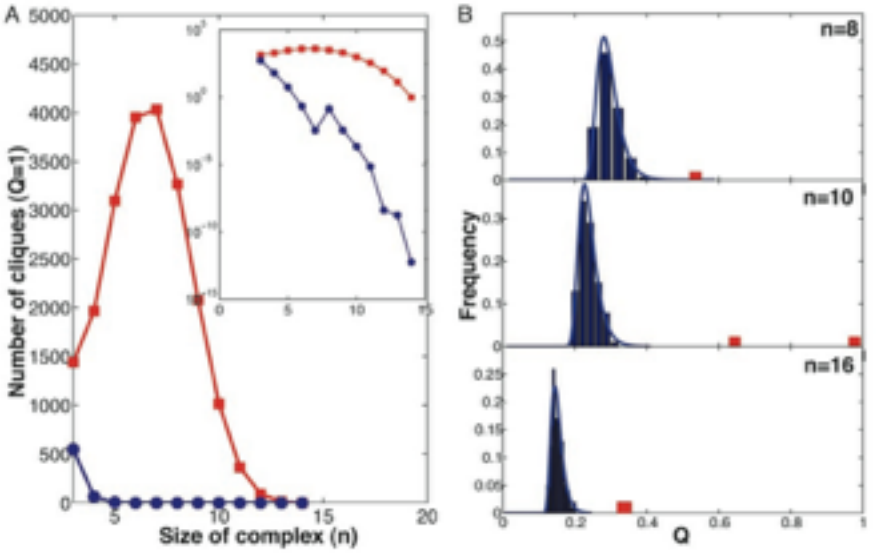


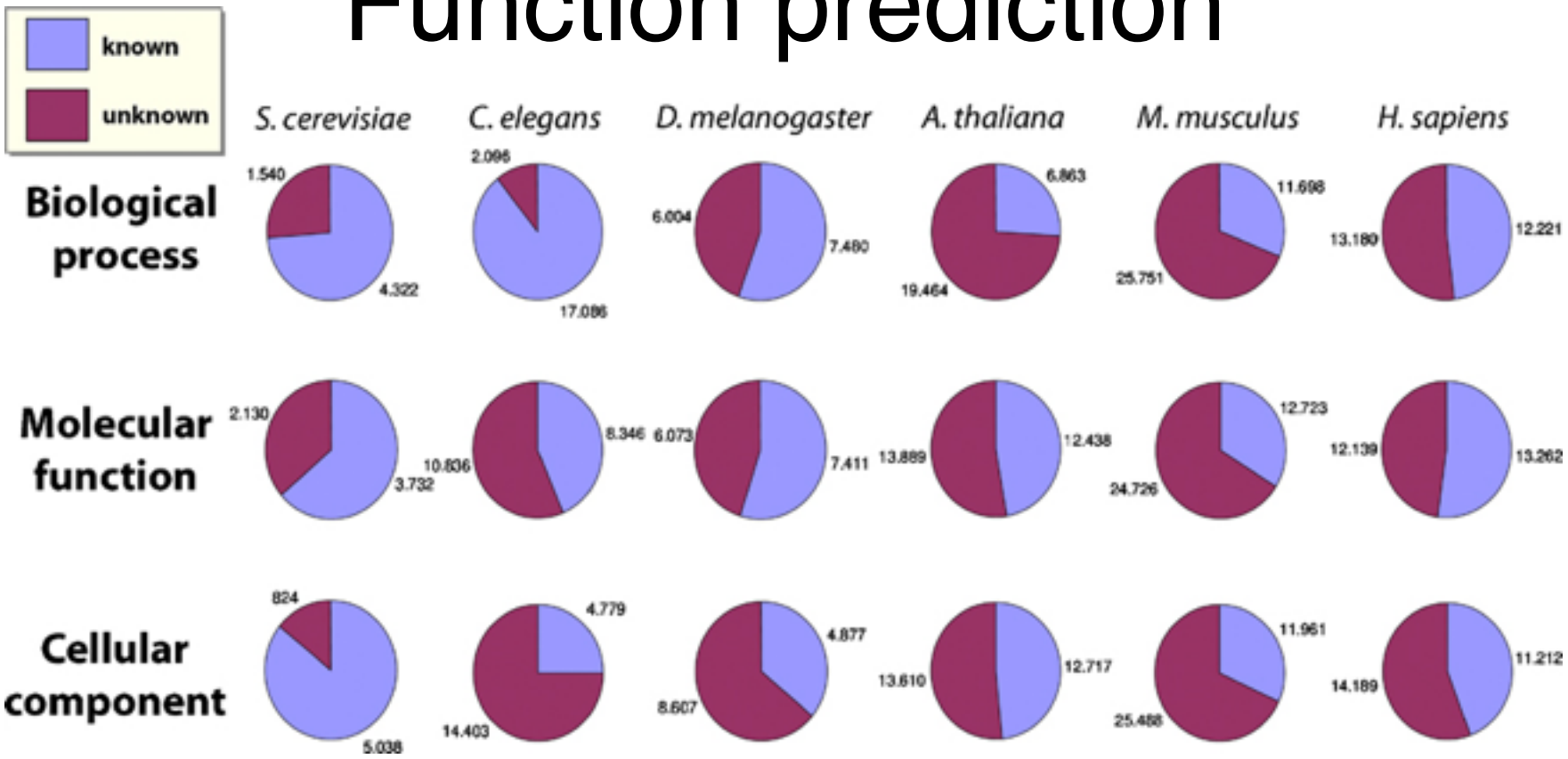
Fig. 1. Statistical significance of complexes and modules. (*A*) Number of complete cliques ($Q - 1$) as a function of clique size enumerated in the network of protein interactions (red) and in randomly rewired graphs (blue, averaged >1,000 graphs). Inset shows the same plot in log-normal scale. Note the dramatic enrichment in the number of cliques in the protein-interaction graph. Most of these cliques are parts of bigger complexes and modules. (*B*) Distribution of Q of clusters found by the MC search procedure in the randomly rewired graphs (blue bars). The blue line shows approximation of this distribution by the Fisher–Tippett extreme value distribution (EVD) with two fitted parameters. Red bars show complexes found in the original network of protein interactions. Sizes of the subgraphs are $n = 8$, 10, and 16. Note that real complexes have many more interactions than the tightest complexes found in randomly rewired graphs.

Protein complexes and functional modules in molecular networks

Victor Spirin and Leonid A. Mirny*

Harvard–MIT Division of Health Sciences and Technology, 16-343, Massachusetts Institute of Technology, 77 Massachusetts Avenue,

Function prediction



FROM:

Network-based prediction of protein function

Roded Sharan, Igor Ulitsky & Ron Shamir

Extent of annotation of proteins in model species. For each species, the charts give the fractions and numbers of annotated and unannotated proteins, according to the three ontologies of the GO annotation. The numbers are based on the Entrez Gene and the WormBase databases as of September 2006

Function prediction

Global protein function prediction from protein-protein interaction networks

Alexei Vazquez, Alessandro Flammini, Amos Maritan & Alessandro Vespignani

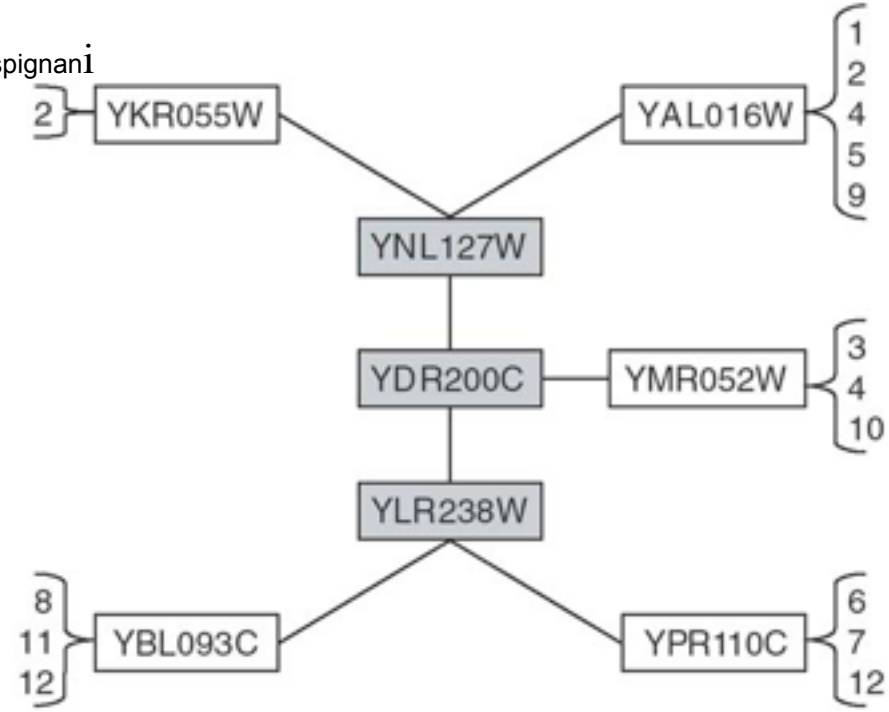


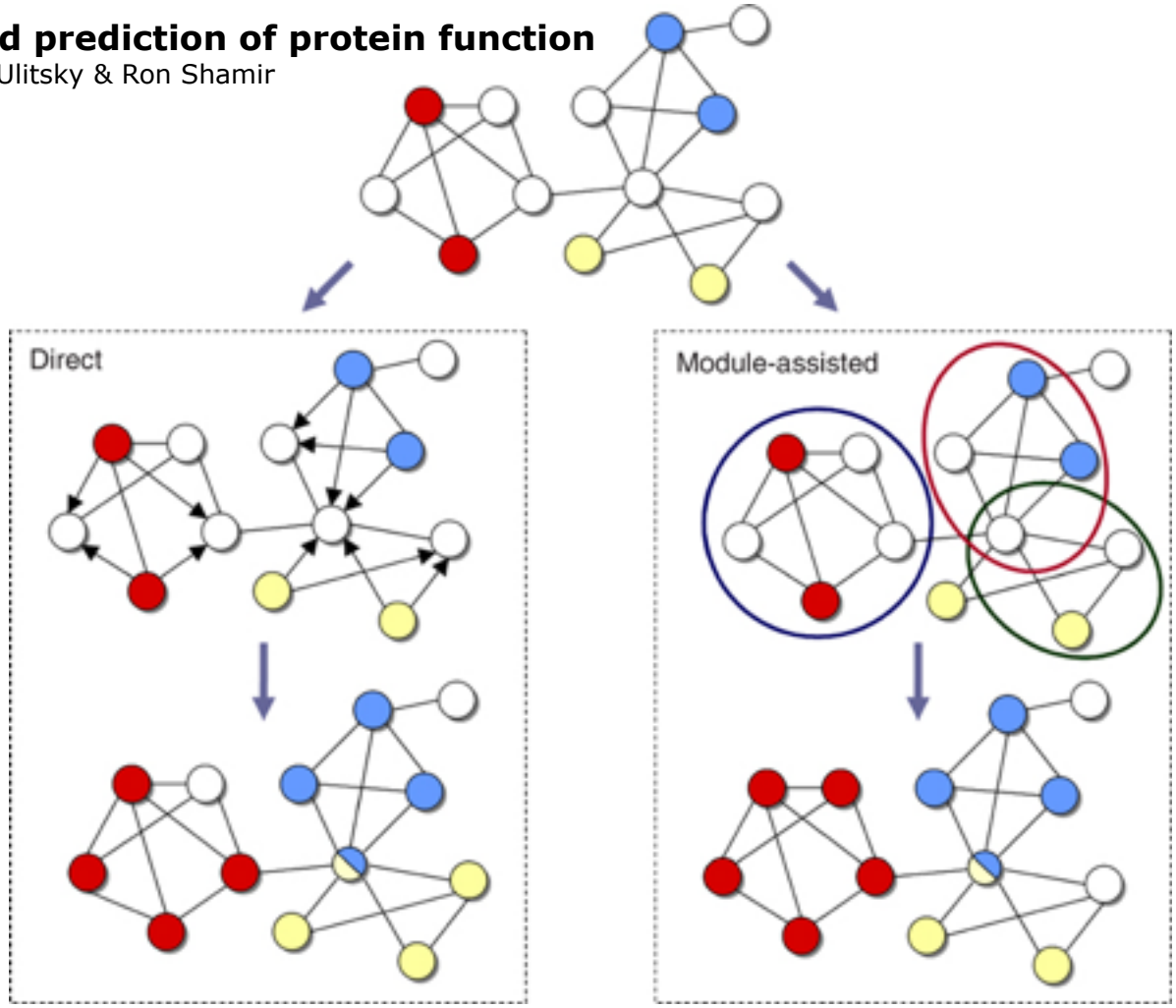
Figure 1. Illustration of the method.
Subgraph of the protein interaction network of the yeast *Saccharomyces cerevisiae*. Proteins in gray boxes are unclassified (unknown function); the others are classified proteins (functions in brackets) and are labeled according to the following criteria: 1, cell growth; 2, budding, cell polarity and filament formation; 3, pheromone response, mating-type determination, sex-specific proteins; 4, cell cycle checkpoint proteins; 5, cytokinesis; 6, rRNA synthesis; 7, tRNA synthesis; 8, transcriptional control; 9, other transcription activities; 10, other pheromone response activities; 11, stress response; 12, nuclear organization. Given one of these proteins of unknown function, if we take as a prediction the function that appears more often in the neighbor proteins of known function, then we obtain the following classification (from top to bottom) YNL127W (2), YDR200C (3,4,10) and YLR238W (12). Our method, however, considers also the interactions among unclassified proteins. If we iterate once more the 'majority rule' by taking into account the interactions among the three unclassified proteins, we obtain the following classification: YNL127W (2,4), YDR200C (3,4,10) and YLR238W (12). This way we determined another possible function for YNL127W.

From network

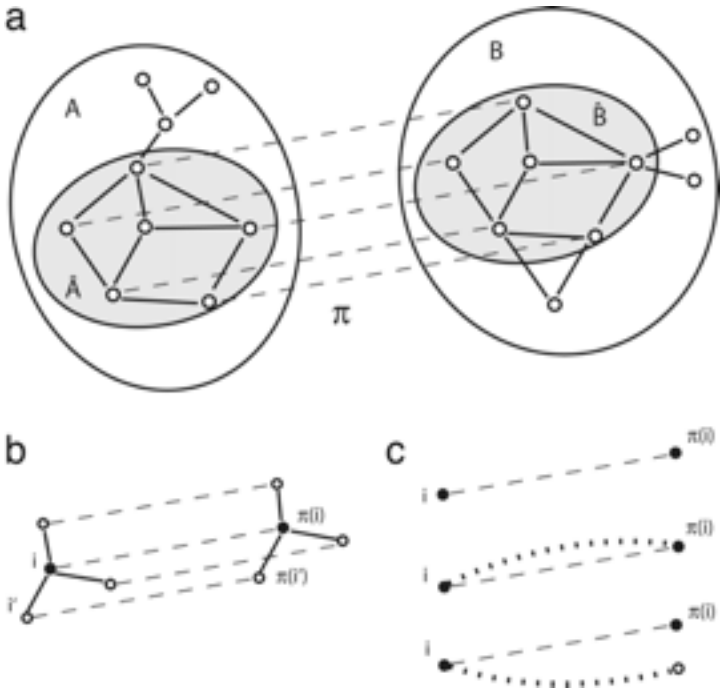
FROM:

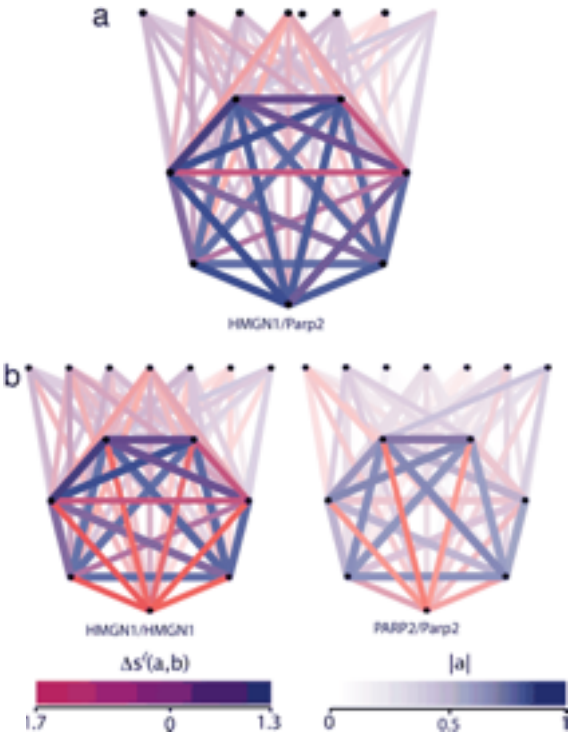
Network-based prediction of protein function

Roded Sharan, Igor Ulitsky & Ron Shamir



Alignment of networks





Cross-species analysis of biological networks by Bayesian alignment

Johannes Berg† and Michael Lässig

Institut für Theoretische Physik, Universität zu Köln, Zülpicherstrasse 77, 50937 Cologne, Germany

Outline

1. Learning biological networks: experiments
2. Statistical properties of the networks
3. Understanding networks structure: motifs, modules, etc
4. From structure to function
5. Compare/align networks
6. **Dynamics of networks**

Evolution of power-law graphs

1. Growth
2. Preferential attachment

~~Albert and Barabasi 2000~~

~~Herbert A. Simon 1955~~

Yule 1925

Table 1. Re-inventing Willis.

Phenomenon	Discovered	Re-discovered	
Yule's process	Yule (1925)	Fermi (1949)	Huberman and Adamic (1999)
Simon's process	Simon (1955)	Günter et al (1992)	Barabasi and Albert (1999)
Champernowne's process	Champernowne (1953)	Levy and Solomon (1996)	
Power law of word frequencies	Estoup (before 1916)	Condon (1928)	Zipf (1935)
Power law of scientific citing	Price (1965)	Silagadze (1997)	Redner (1998)

Evolution of graphs

- Growth
  1. start with m_0 nodes
  2. add a node with m edges
  3. connect these edges to existing nodes

  at timestep t : $t+m_0$ nodes, tm edges

Evolution of graphs

- Preferential attachment

  Probability Π of connection to node i depends on the degree k_i of this node.

E.g. $$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

"Rich gets richer"

Better evolution of graphs

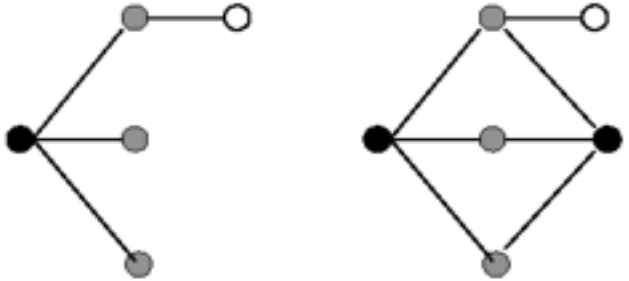
A. Wagner, M.Lassig, A.Maritan etc

- Gene duplication
- Mutations
- Preferential attachment

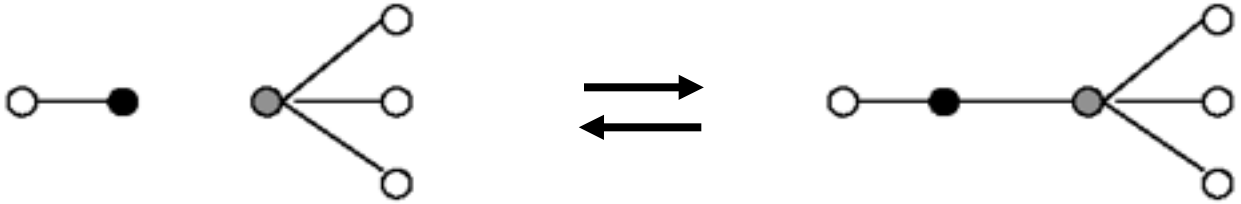
More biological neutral evolution of graphs

A.Wagner, M.Lassig, A.Maritan, S.Redner etc.

- Gene duplication



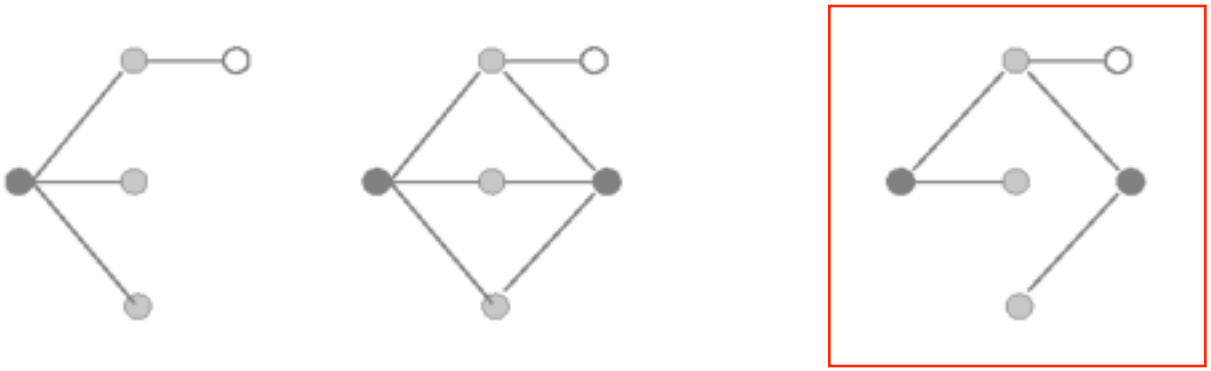
- Mutations



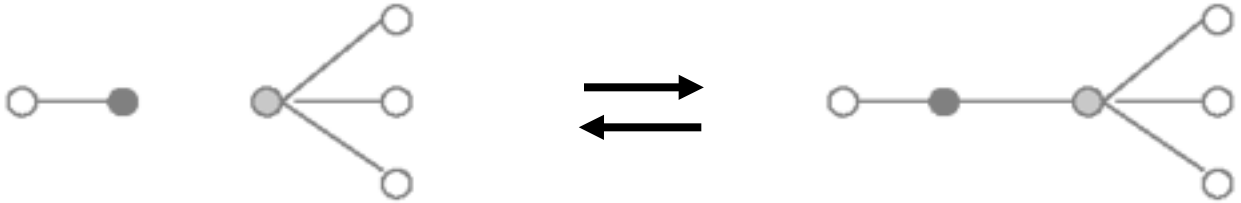
More biological neutral evolution of graphs

A.Wagner, M.Lassig, A.Maritan, S.Redner etc.

- Gene duplication



- Mutations (rich gets richer)



=> Broad (not power-law) distribution!

More biological neutral evolution of graphs

- Gene duplication and re-wiring

Infinite-Order Percolation and Giant Fluctuations in a Protein Interaction Network

J. Kim[1], P. L. Krapivsky[2], B. Kahng[1], and S. Redner[2]

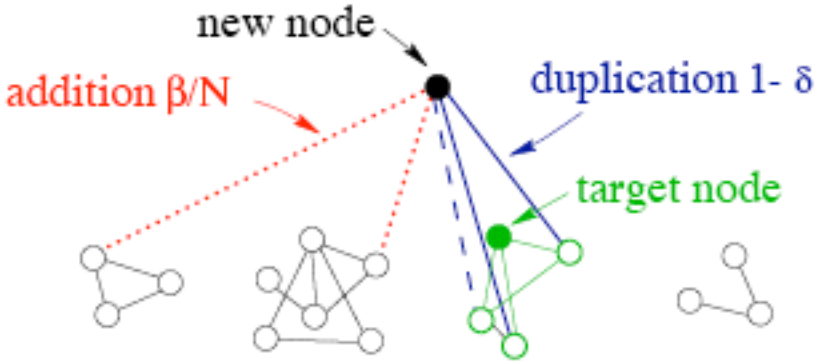
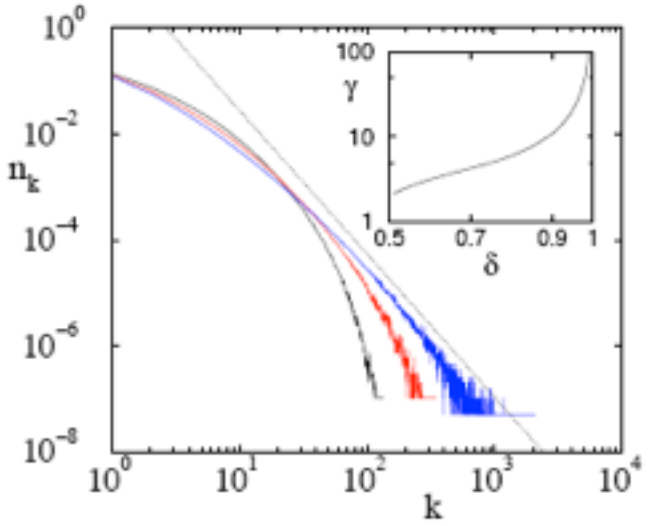


FIG. 1. Growth steps of the protein interaction network: The new node duplicates 2 out of the 3 links between the target node (shaded) and its neighbors. Each successful duplication occurs with probability $1-\delta$ (solid lines). The new node also attaches to any other network node with probability β/N (dotted lines). Thus 3 previously disconnected clusters are joined by the complete event.



Evidence for Network Evolution in an *Arabidopsis* Interactome Map

Arabidopsis Interactome Mapping Consortium*†

Plants have unique features that evolved in response to their environments and ecosystems. A full account of the complex cellular networks that underlie plant-specific functions is still missing. We describe a proteome-wide binary protein-protein interaction map for the interactome network of the plant *Arabidopsis thaliana* containing about 6200 highly reliable interactions between about 2700 proteins. A global organization of plant biological processes emerges from community analyses of the resulting network, together with large numbers of novel hypothetical functional links between proteins and pathways. We observe a dynami gene duplication events, providing evidence for a model of networks. This and future plant interactome maps should fa understand plant biology and improve crops.

Classical genetic and molecular approaches have provided fundamental understanding of processes such as growth control or development and molecular descriptions of genotype-to-phenotype relationships for a varie-

ty of plan protein-c *dopsis th* functiona the biolo complex is lacking and S2), genotype ated at th

*All authors with their affiliations and contributions are listed at the end of the paper.
†To whom correspondence should be addressed. E-mail: marc_vidal@dfci.harvard.edu; ecker@salk.edu; pascal_braun@dfci.harvard.edu; david_hill@dfci.harvard.edu

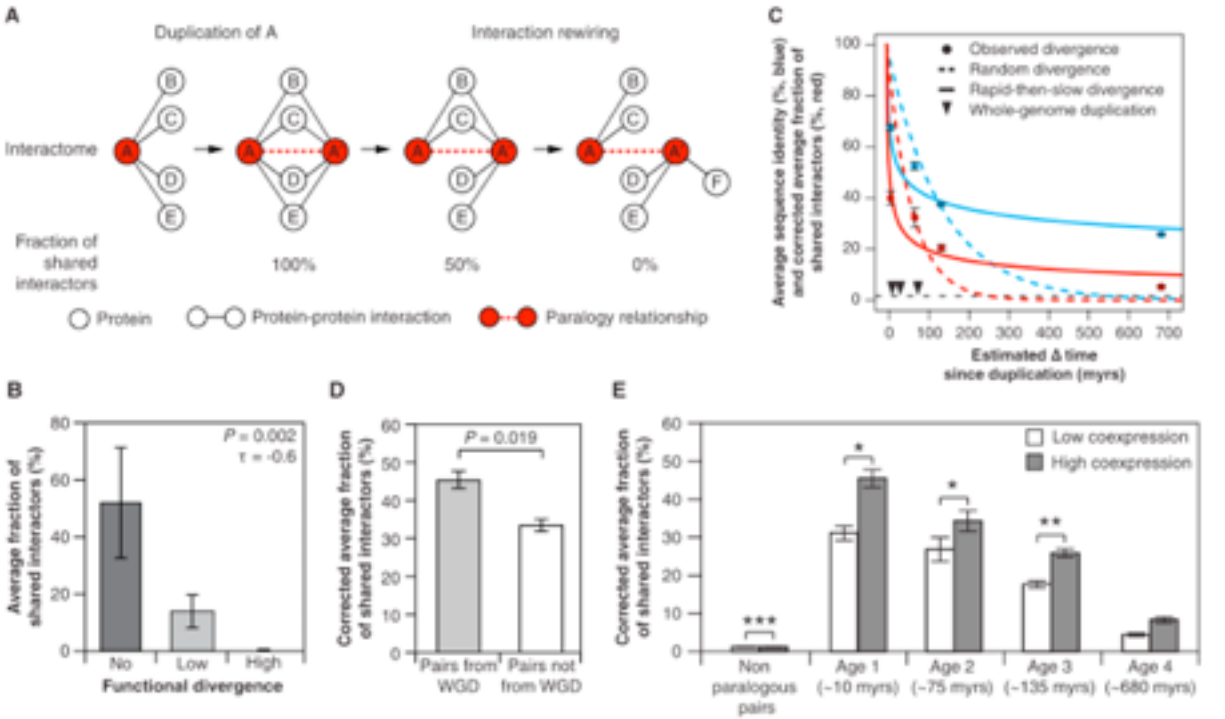
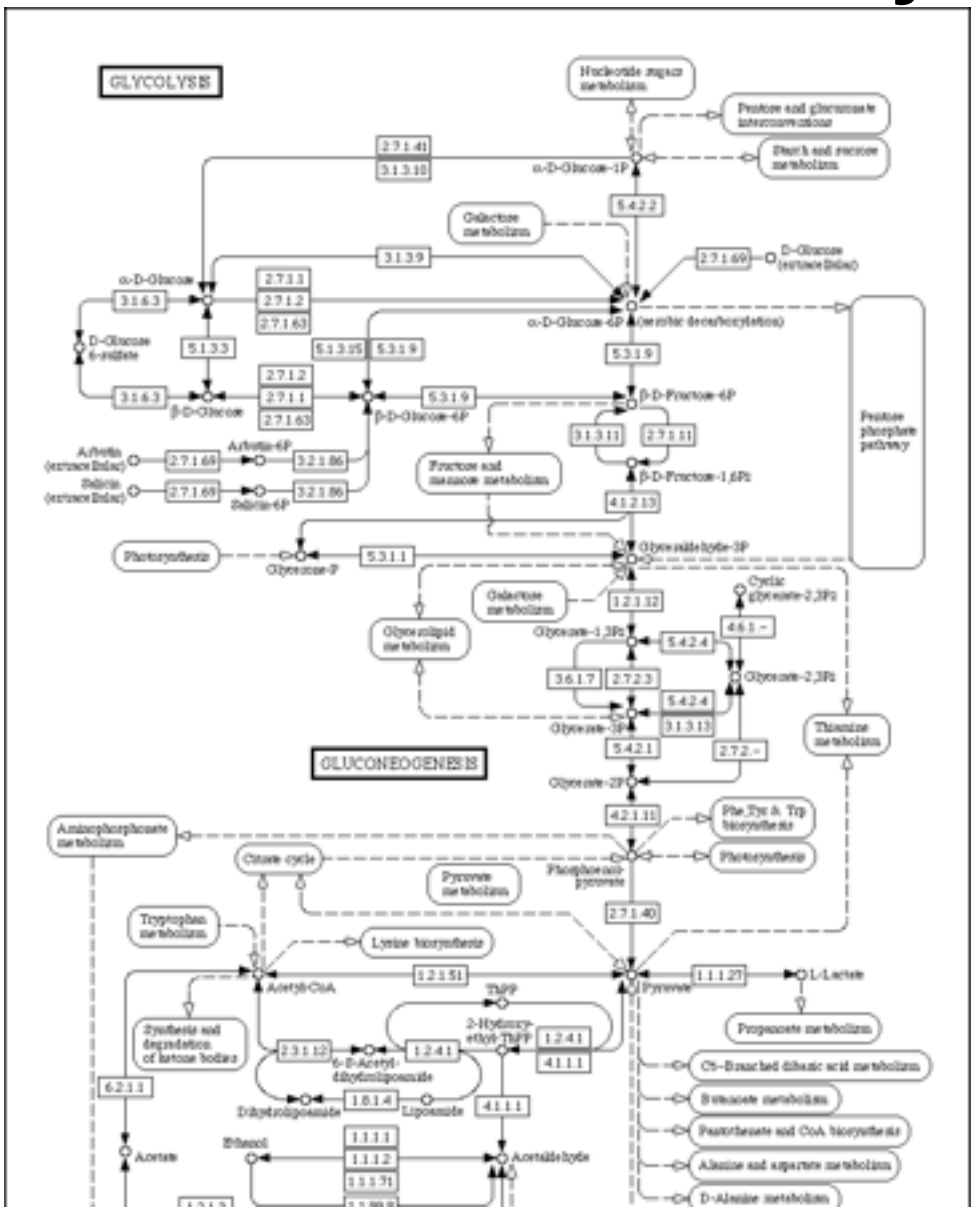


Fig. 4. Evidence for network evolution in AI-1_{MAIN}. **(A)** Interaction rewiring over time, according to the duplication-divergence model (*24*). **(B)** Average fraction of interactors shared between pairs of paralogous proteins with no (n = 4), low (n = 10), and high (n = 3) functional divergence (*28*). Error bars, mean ± SEM. P value, one-sided Kendall ranking correlation test (τ, association) (*3*). **(C)** Average fraction of shared interactors, corrected for low experimental coverage (*3*), and average protein sequence identity between pairs of paralogous proteins as a function of the estimated time elapsed since duplication. Error bars, mean ± SEM (*3*). Dashed black line, corrected average fraction of shared interactors of nonparalogous pairs; myrs, millio years. **(D)** Corrected average fraction of shared interactors (*3*) for pairs paralogous proteins originating from polyploidy events (n = 109), as com pared with other paralogous protein pairs of similar age (n = 147). Error bar mean ± SEM (*3*). P values, Mann-Whitney U test. **(E)** Corrected average frac tion of shared interactors (*3*) for pairs of paralogous proteins encoded by gen pairs with high or low coexpression correlation (top and bottom tertile, respe tively) as a function of phylogeny-based age group. Error bars, mean ± SE (*3*). *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Metabolic networks

Metabolic Pathways



KEGG database

Flux Balance Analysis

No accumulation of intermediates

# of molecules in = # of molecules out

Vin+Vout = 0

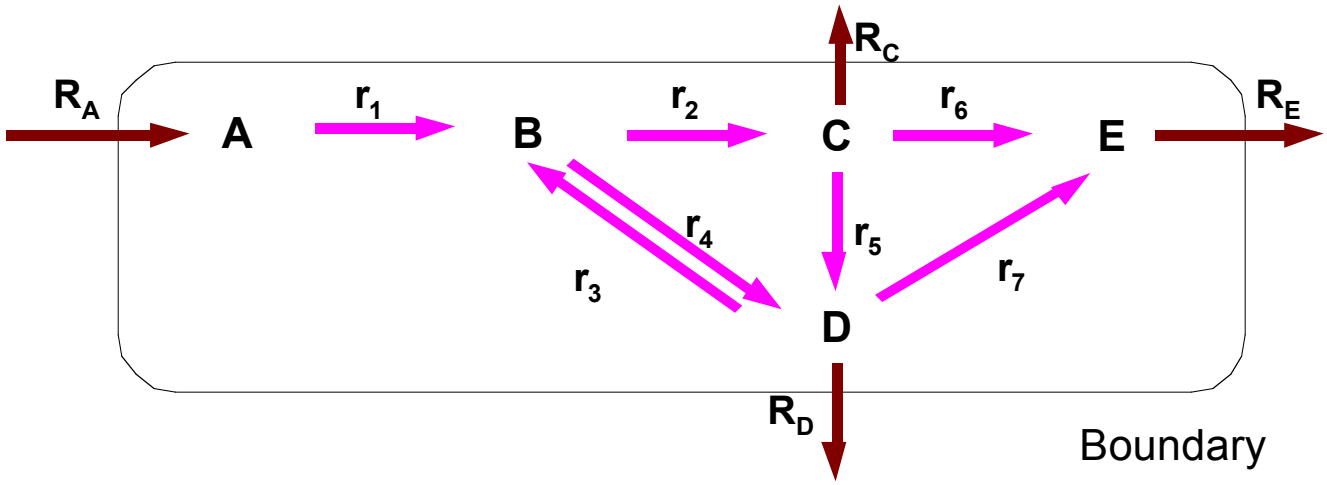
Example:

$2A+B \rightarrow 3D$

$D+C \rightarrow E$

$(2V_A+V_B)/3 = V_E$

Flux Balance Analysis



Steady state Mass Balance

A: $-r1=-R_A$
B: $-r1+r4-r2-r3=0$
C: $+r2-r5-r6=+R_C$
D: $+r3+r5-r4-r7=+R_D$
E: $+r6+r7=+R_E$

$S.v = b$

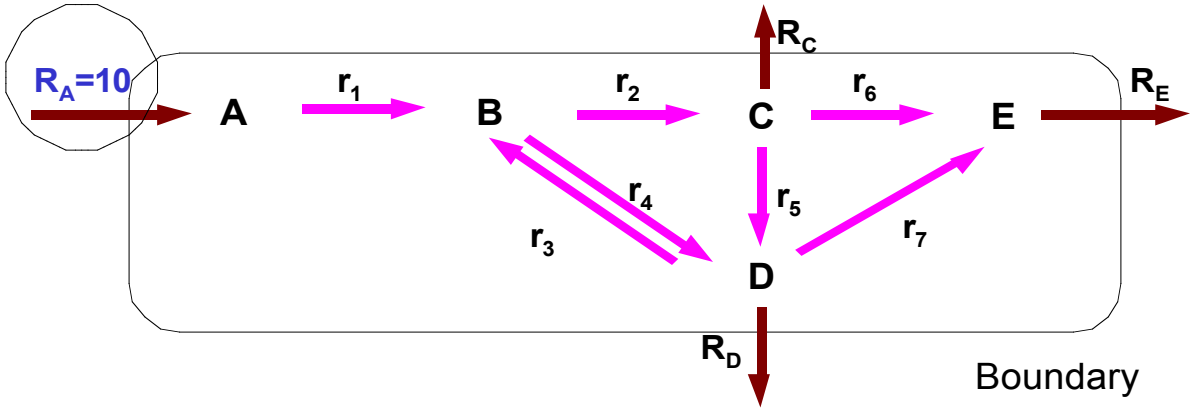
$$\begin{matrix} r_1 & r_2 & r_3 & r_4 & r_5 & r_6 & r_7 \end{matrix}$$

$$\begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} -R_A \\ 0 \\ R_C \\ R_D \\ R_E \end{bmatrix}$$

Internal fluxes

Transportation fluxes

Flux Balance Analysis



Steady state Mass Balance

A: $-r1 = -10$

B: $-r1 + r4 - r2 - r3 = 0$

C: $r2 - r5 - r6 - R_C = 0$

D: $r3 + r5 - r4 - r7 - R_D = 0$

E: $r6 + r7 - R_E = 0$

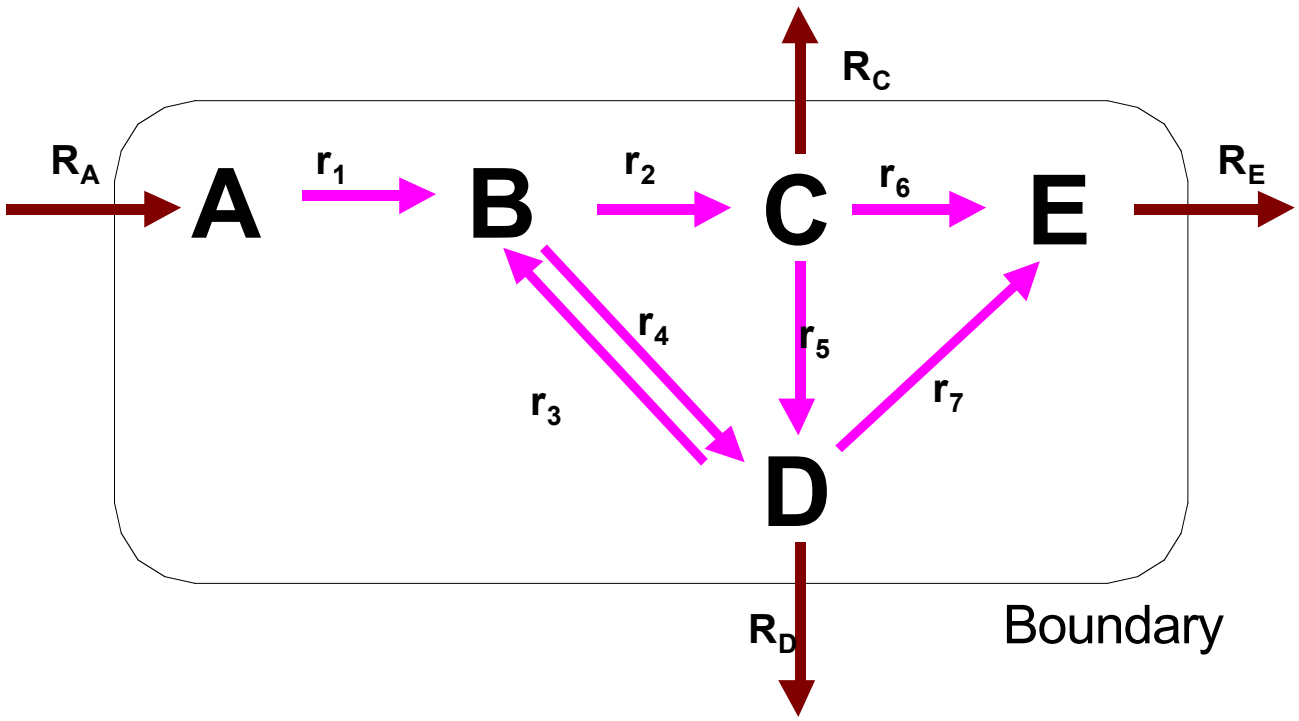
$$\begin{array}{c} \begin{bmatrix} r_1 & r_2 & r_3 & r_4 & r_5 & r_6 & r_7 & R_C & R_D & R_E \end{bmatrix} \\ \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & -1 \end{bmatrix} \end{array} = \begin{bmatrix} -10 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

10 Unknown fluxes

1 Known fluxes

$$S'.v' = b'$$

Flux Balance Analysis



Total Number of fluxes = 11
Total numer of known flux = 1
Total number of Metabolites = 5
Total number of d.f = 11-1+5=5
(i.e 5 possible solutions for this reaction network)

Flux Balance Analysis

- If cells optimize their growth rate then we need to find a solution that maximizes growth.
- Growth = biomass/time

```
+1BIOM-0.582GLY-0.0485MethylTHF-0.25GLN-     45.135ATP+44.96ADP+ 44.96Pi    -
0.25GLU-0.176PHE-0.131TYR-0.205SER-0.054TRP-0.229ASP-0.229ASN-0.326LYS-
0.087CYS-0.146MET-0.241THR-0.276ILE-0.21PRO-0.281ARG-0.488ALA-0.402VAL-
0.428LEU-0.09HIS-0.203GTP-0.136UTP-0.126CTP-0.0247dATP-0.0254dGTP-
0.0254dCTP-0.0247dTTP-0.00258PS -0.09675PE-0.02322PG-0.00645CL-
0.00785LPS-0.0276Pept-0.0341PTRSC-0.007SPRMD-0.154Glycogen;
```

Millimoles of metabolites present in 1 gm (dry wt.) of biomass

Flux Balance Analysis

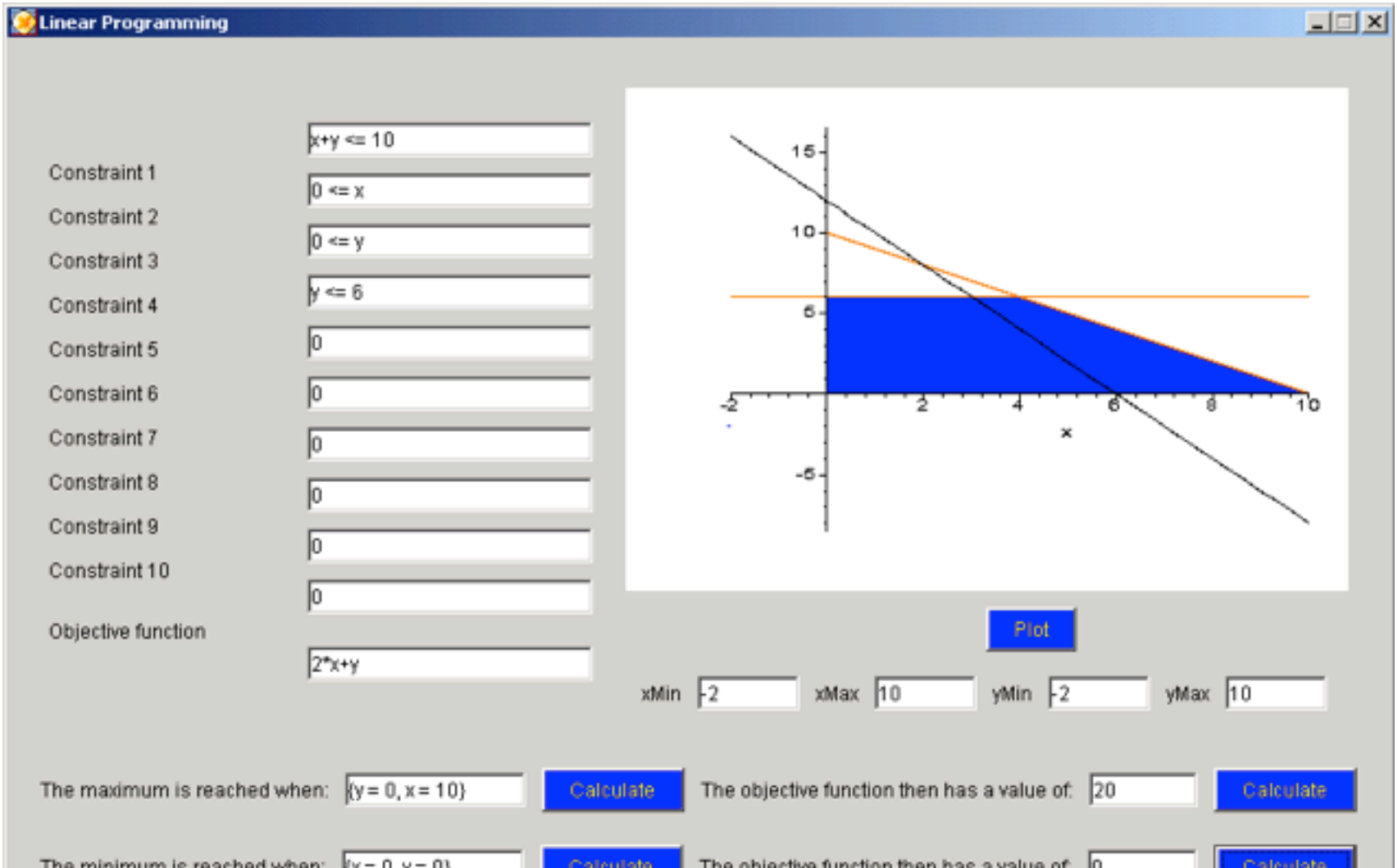
- Input: stoichiometric matrix
  optimization function (biomass)
- Constrains

$$\frac{d\mathbf{X}}{dt} = \mathbf{S} \bullet \mathbf{v} - \mathbf{b} = 0$$

Maximize: $Z = \sum c_i \cdot v_i = \mathbf{c} \bullet \mathbf{v}$

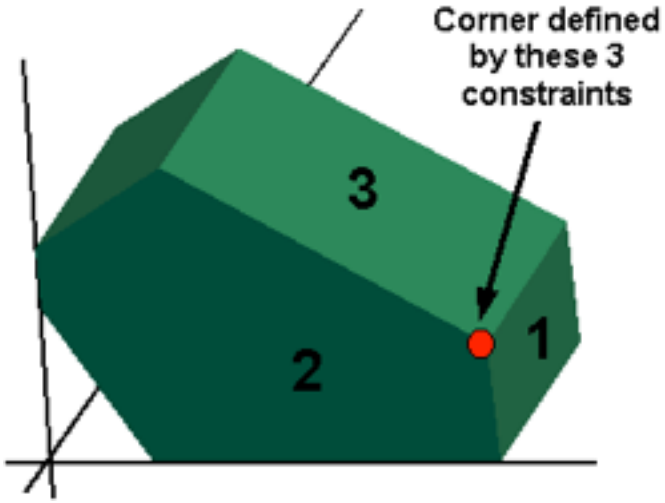
Flux Balance Analysis

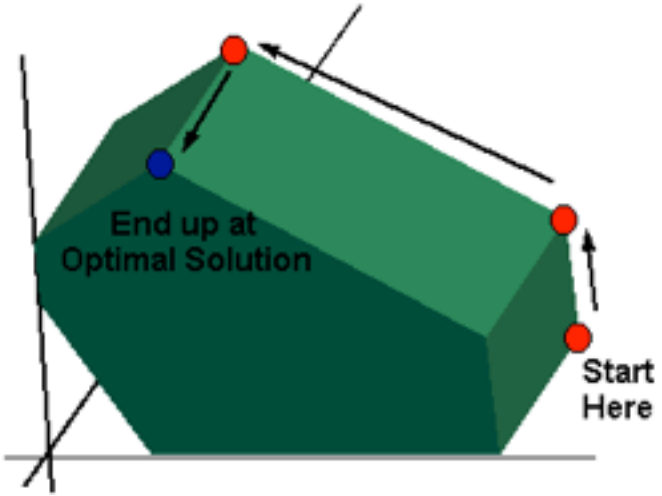
- Linear programming



Flux Balance Analysis

- Linear programming



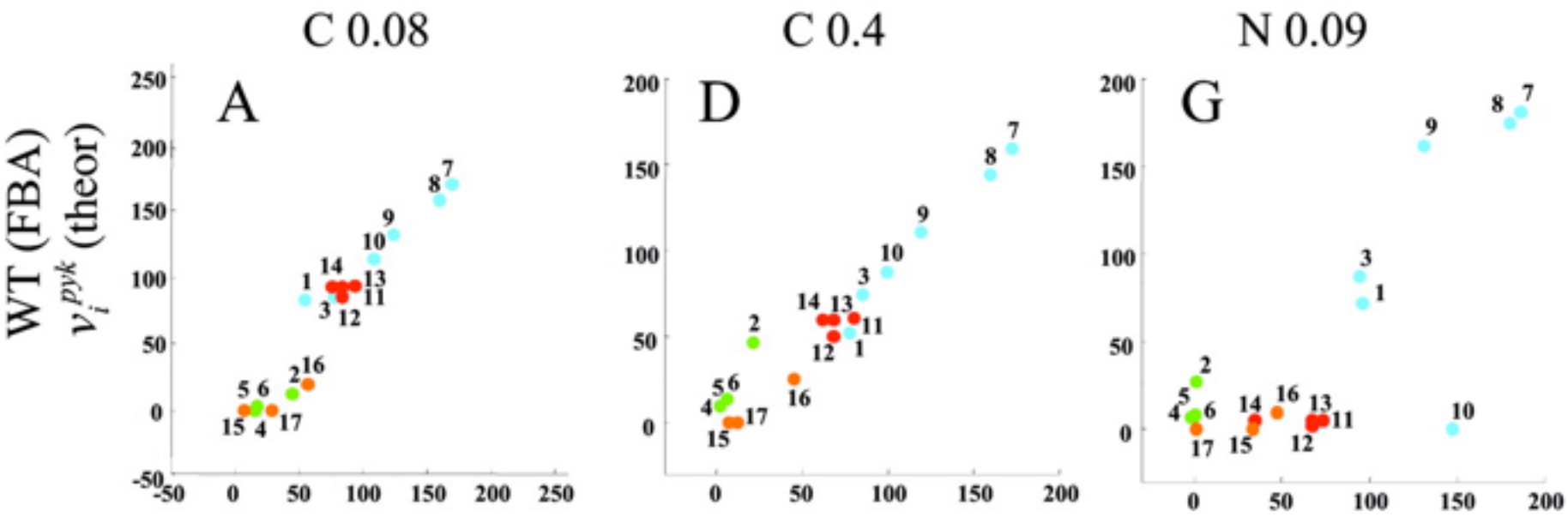


Flux Balance Analysis

- Effects of external conditions
- Effect of mutations
- Predictive cell physiology

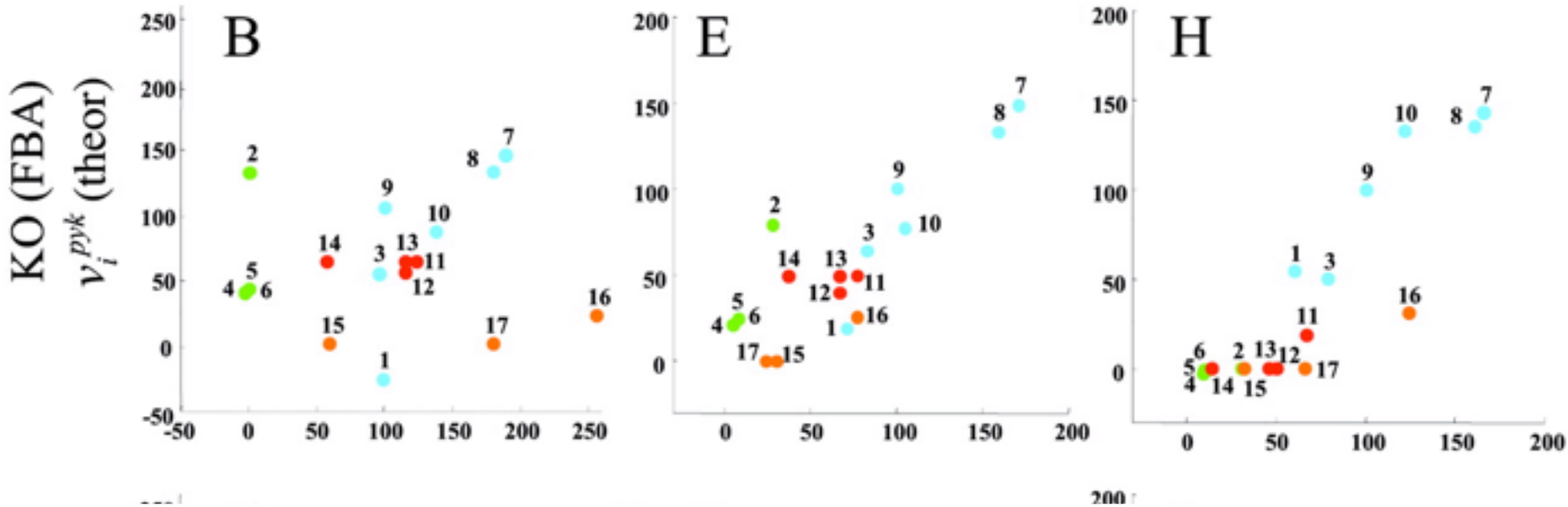
Flux Balance Analysis

- Effect of C and N starvation



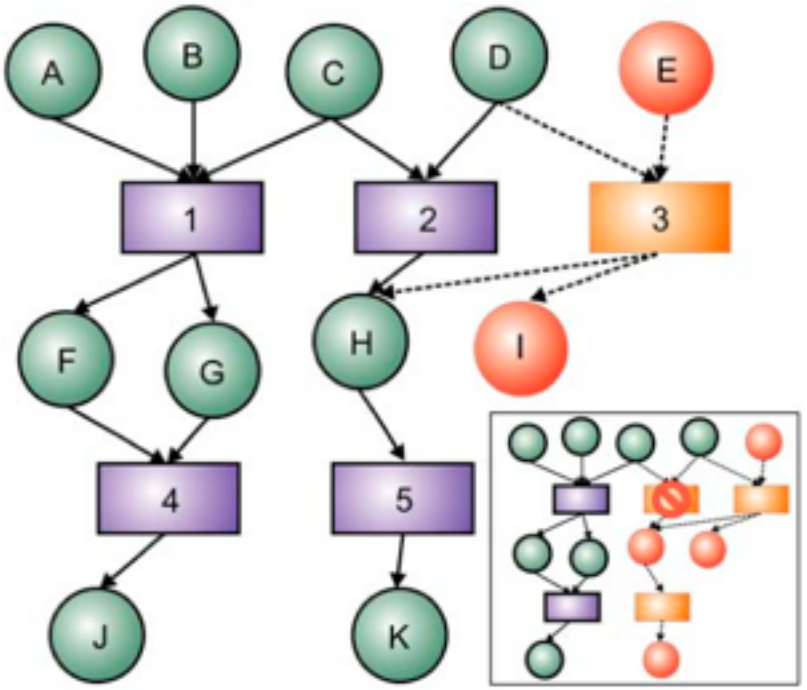
Flux Balance Analysis

- Effect of mutations and starvation



Predicting outcomes of knockouts

Topology-Based Metabolic Predictions

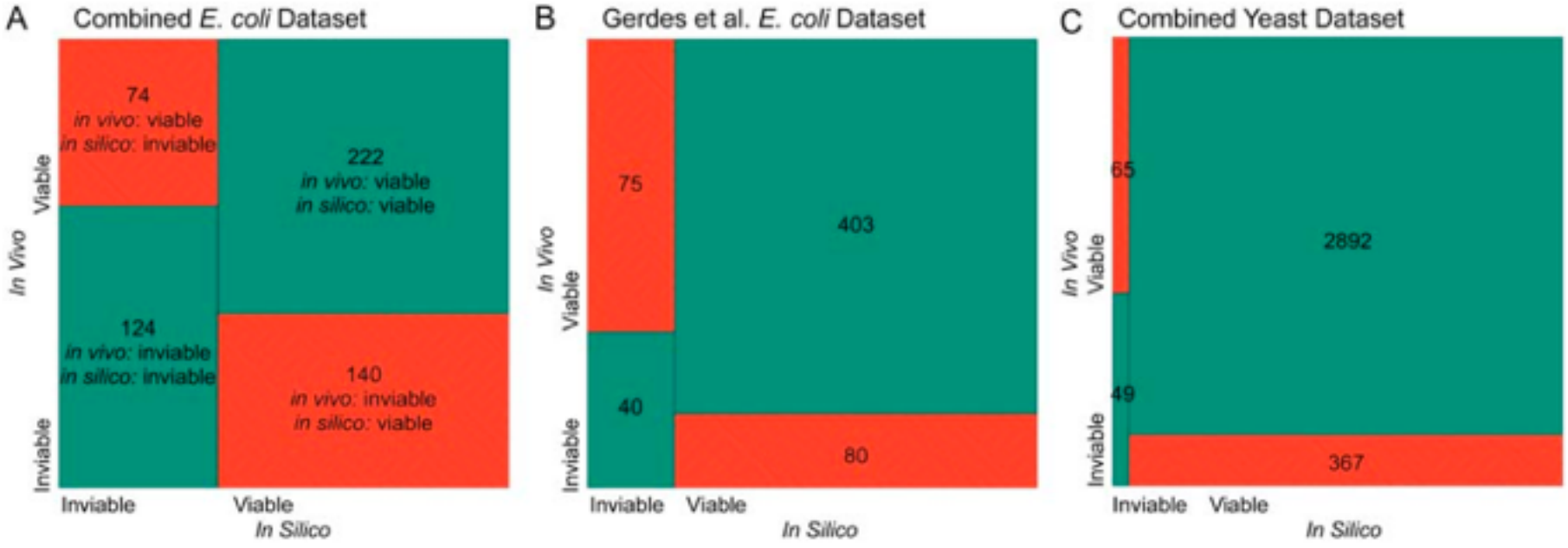


2304 Biophysical Journal Volume 91 September 2006 2304–2311

Using the Topology of Metabolic Networks to Predict Viability of Mutant Strains

Zeba Wunderlich* and Leonid A. Mirny†
*Biophysics Program, Harvard University, Cambridge, Massachusetts; and †Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts

Predicting outcomes of knockouts



2304 Biophysical Journal Volume 91 September 2006 2304–2311

Using the Topology of Metabolic Networks to Predict Viability of Mutant Strains

Zeba Wunderlich* and Leonid A. Mirny†
*Biophysics Program, Harvard University, Cambridge, Massachusetts; and †Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts

Networks

- Structure and dynamics of some biological network can be studied experimentally

- Networks don't look like random graphs, more like power-law graphs.
  - results of neutral evolution
  - results of selection