Biological Network Analysis

Mark B Gerstein

Yale

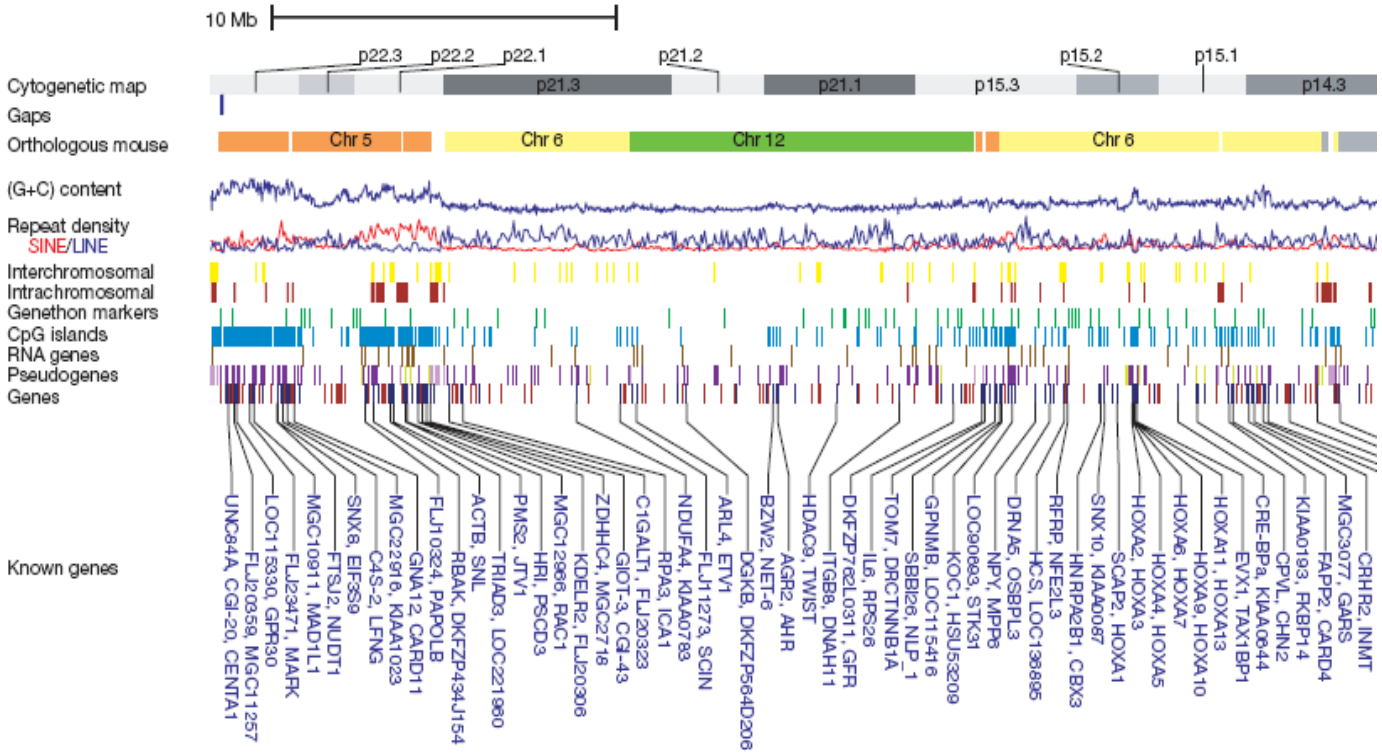
slides at

Lectures.GersteinLab.org

(See Last Slide for References & More Info.)

The problem: Grappling with Function on a Genome Scale?

sequence of human chr. 7



~1,200 protein-coding genes

(~950 pseudogenes)

[Hillier et al, Nature, 424, 157]

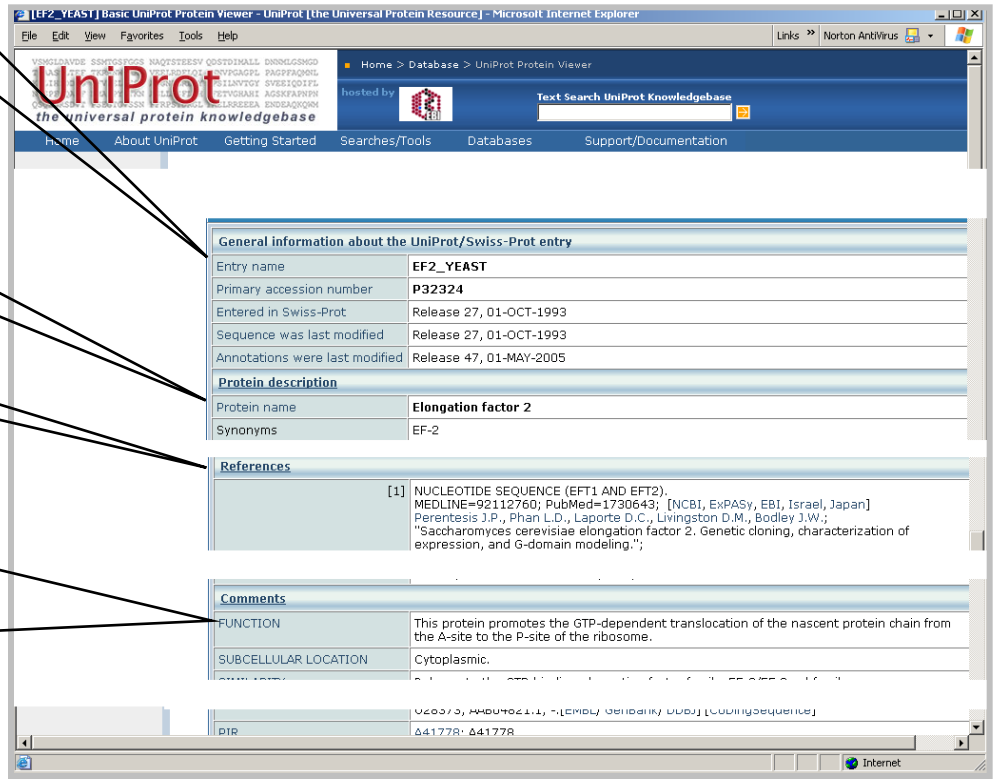
Traditional single molecule way to integrate evidence & describe function

EF2_YEAST

Descriptive Name:
Elongation Factor 2

Lots of references
to papers

Summary sentence describing function:
This protein promotes the GTP-dependent translocation of the nascent protein chain from the A-site to the P-site of the ribosome.



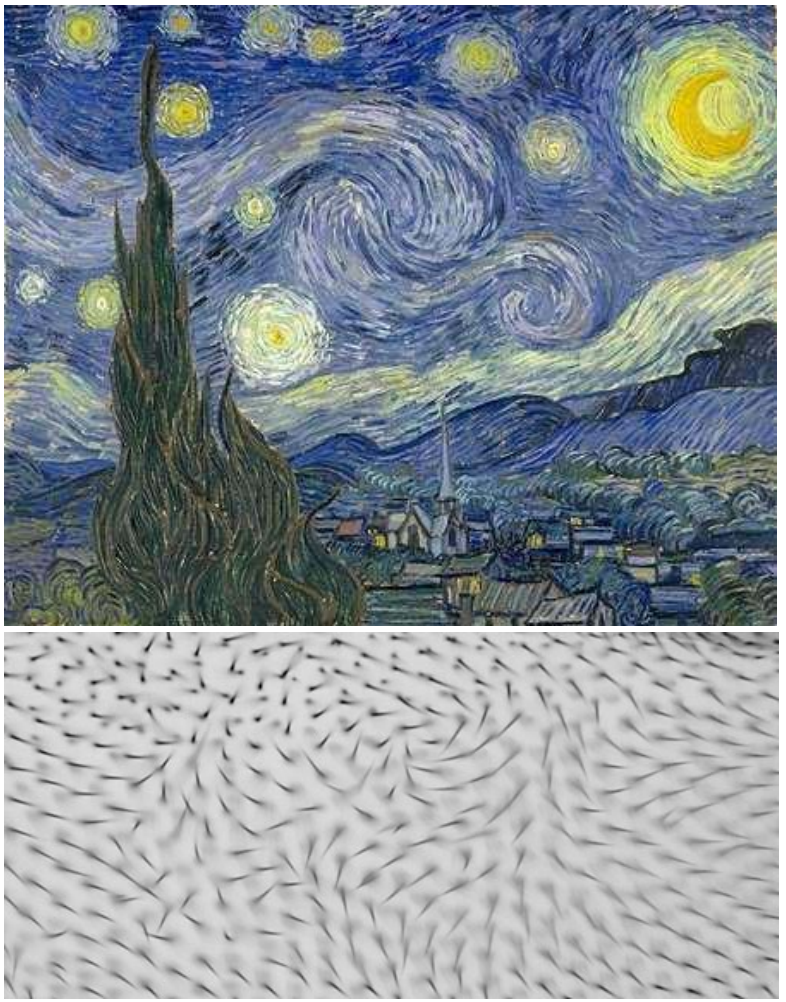
Some obvious issues in scaling single molecule definition to a genomic scale

- Fundamental complexities
  - Role Conflation:
    molecular, cellular, phenotypic
  - Often >2 proteins/function
  - Also Multi-functionality:
    2 functions/protein
    - phenotypically – e.g. Pleiotropic effects such as human PKU being involved in retardation & eczema
    - cellular role – e.g. Depending on the molecule it interacts with HSP70 is involved with protein folding, translocation of proteins into mitochondia, biogenesis of certain subunits..

[HSP from Craig et al, Rev Physiol Biochem Pharmacol (2006) 156:1 ; Terms from Seringhaus et al. GenomeBiology (2008)]

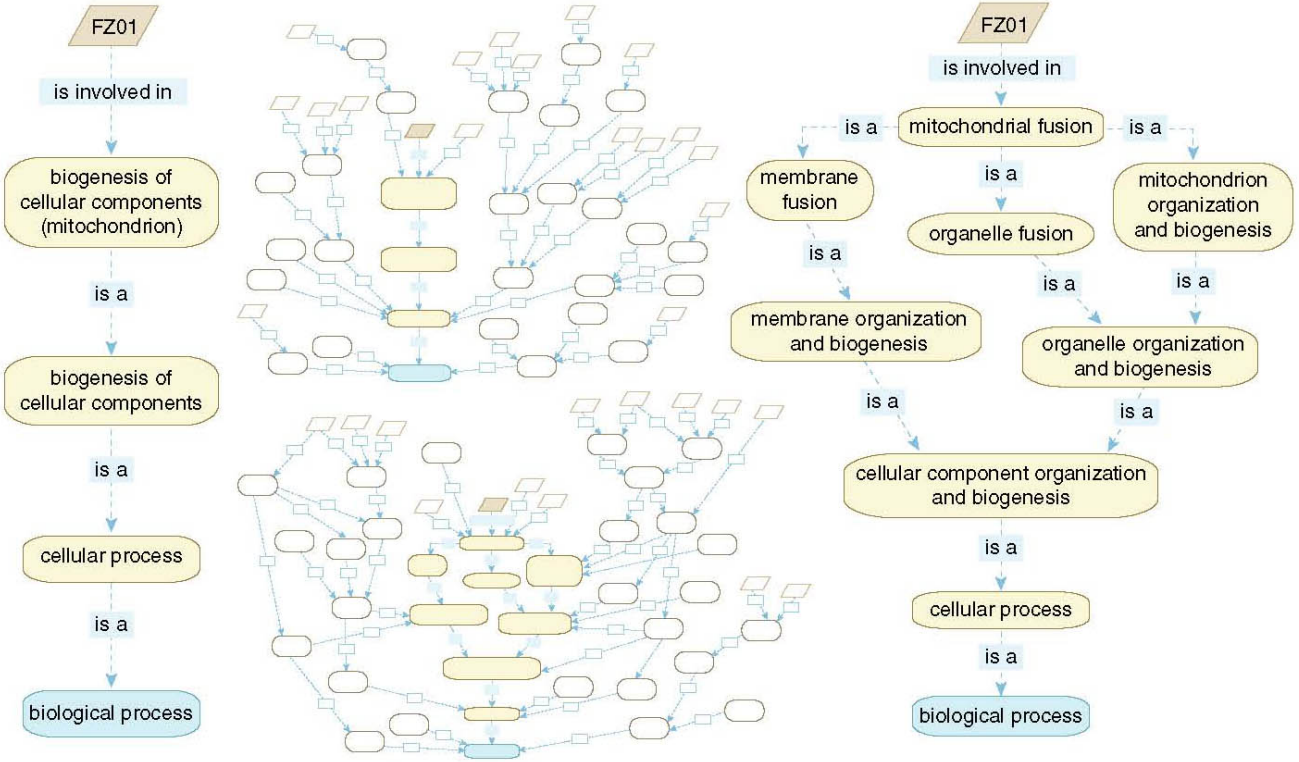
Some obvious issues in scaling single molecule definition to a genomic scale

- Fundamental complexities
  - Role Conflation:
    molecular, cellular, phenotypic
  - Often >2 proteins/function
  - Also Multi-functionality:
    2 functions/protein
    - phenotypically – e.g. Pleiotropic effects such as human PKU being involved in retardation & eczema
    - cellular role – e.g. Depending on the molecule it interacts with HSP70 is involved with protein folding, translocation of proteins into mitochondia, biogenesis of certain subunits..
- Fun terms… but do they scale?....
  - **Starry night** (P Adler, '94)



[HSP from Craig et al, Rev Physiol Biochem Pharmacol (2006) 156:1 ; Terms from Seringhaus et al. GenomeBiology (2008)]

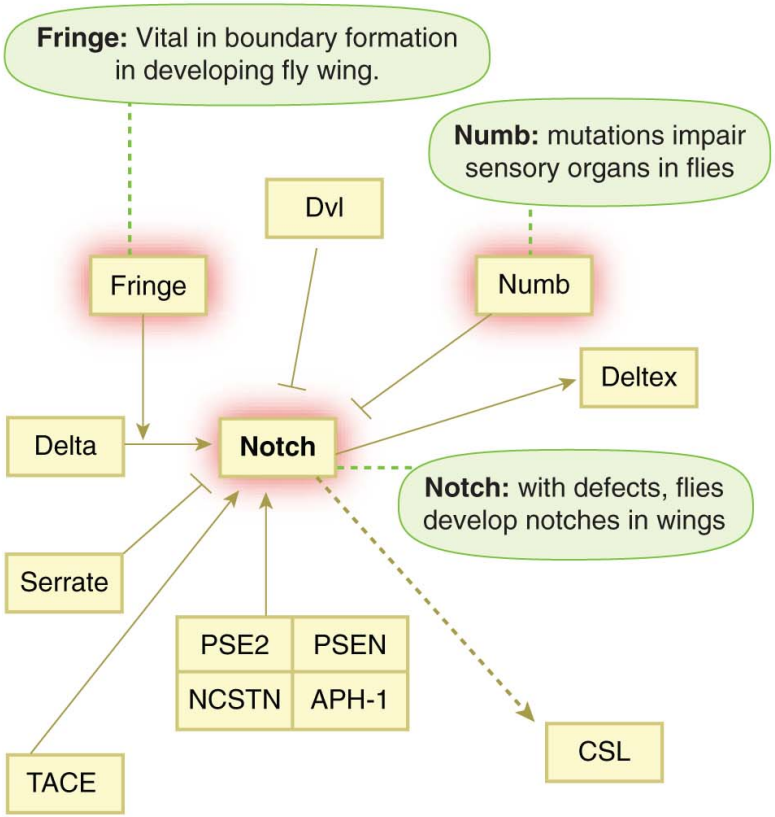
Hierarchies & DAGs of controlled-vocab terms but still have issues...



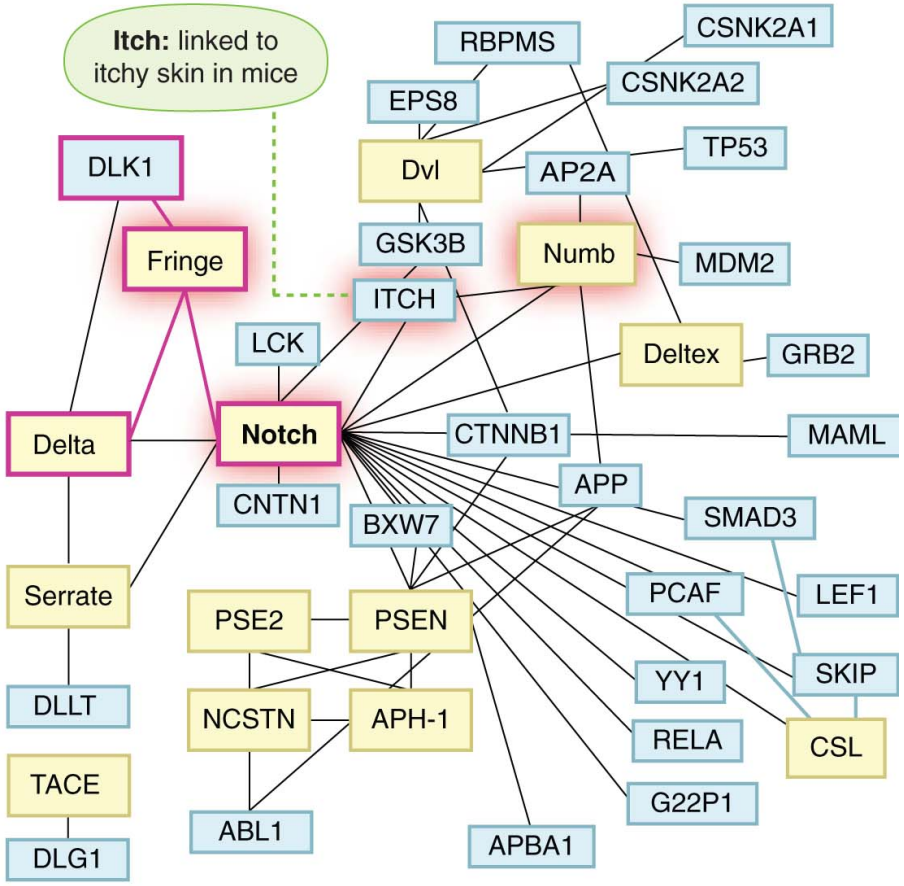
MIPS (Mewes et al.) **GO (Ashburner et al.)**

[Seringhaus & Gerstein, Am. Sci. '08]

Networks (Old & New)



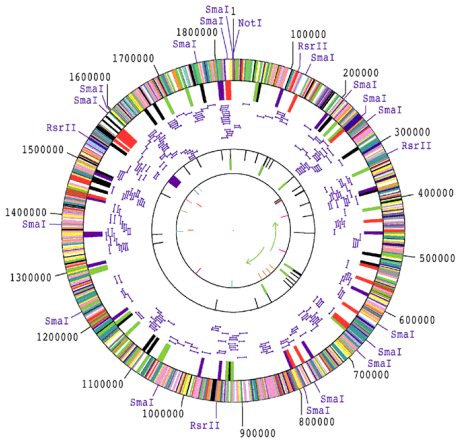
Classical KEGG pathway



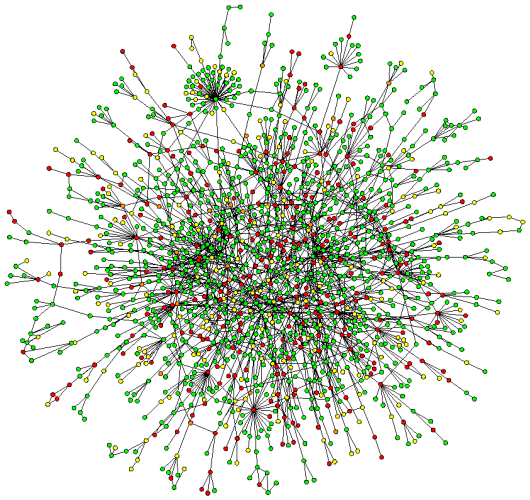
Same Genes in High-throughput Network

[Seringhaus & Gerstein, Am. Sci. '08]

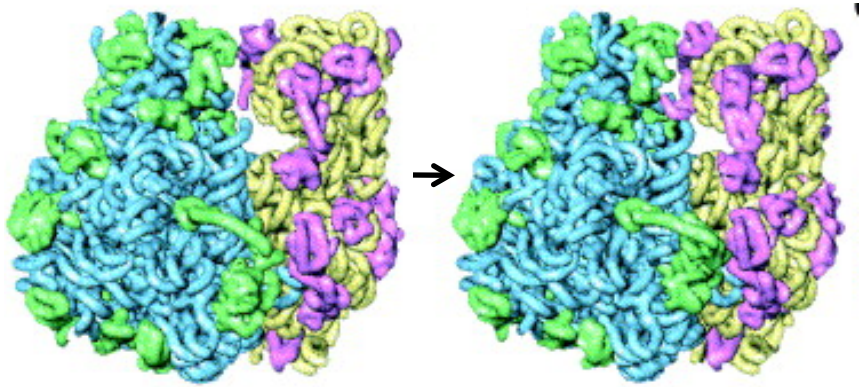
Networks occupy a midway point in terms of level of understanding



1D: Complete Genetic Partslist



~2D: Bio-molecular Network Wiring Diagram



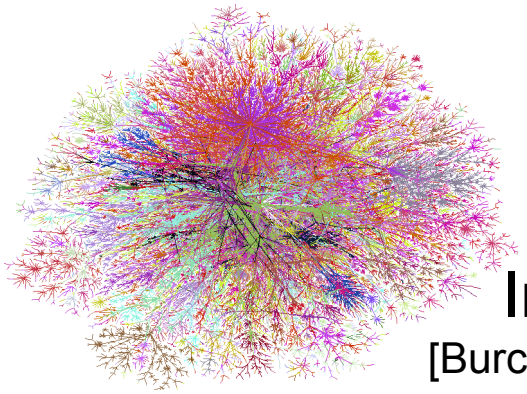
3D and 4D: Detailed structural understanding of cellular machinery (e.g. ribosome in different functional states)

[Fleischmann et al., Science, 269 :496]

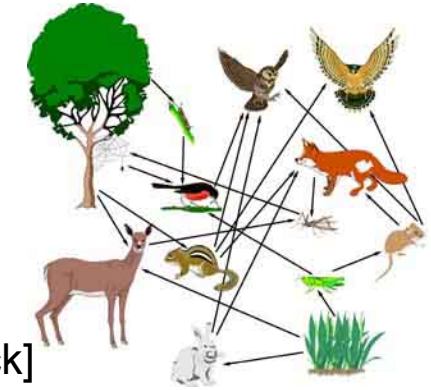
[Jeong et al. Nature, 41:411]

[Chiu et al. Trends in Cell Biol, 16:144]

Networks as a universal language

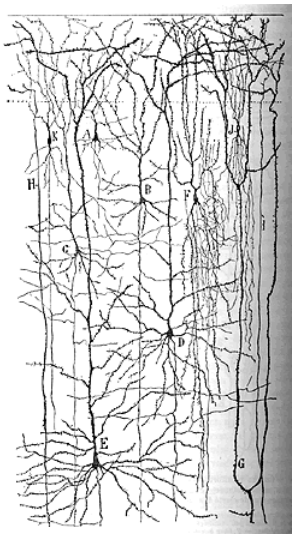


Internet
[Burch & Cheswick]

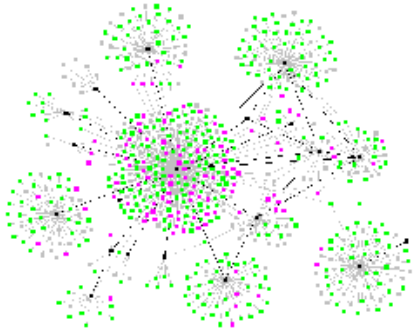


Food Web

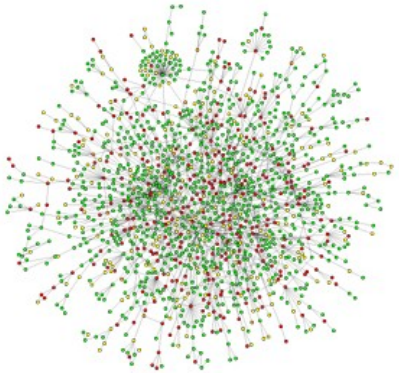
Electronic Circuit



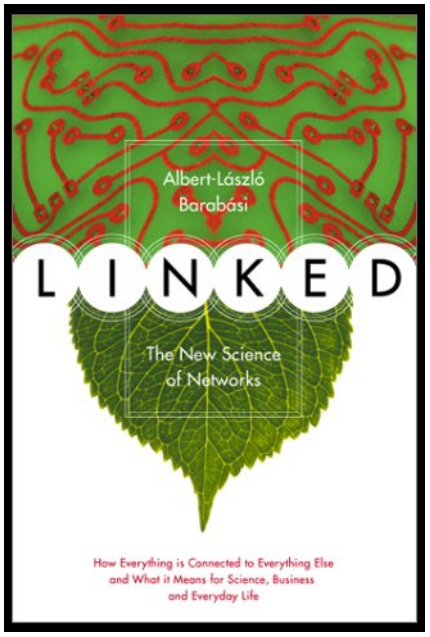
Neural Network
[Cajal]

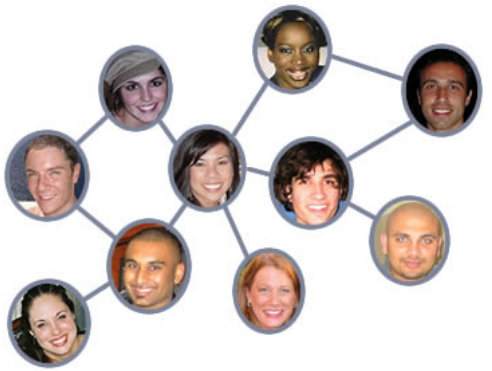


Disease Spread
[Krebs]



Protein Interactions
[Barabasi]



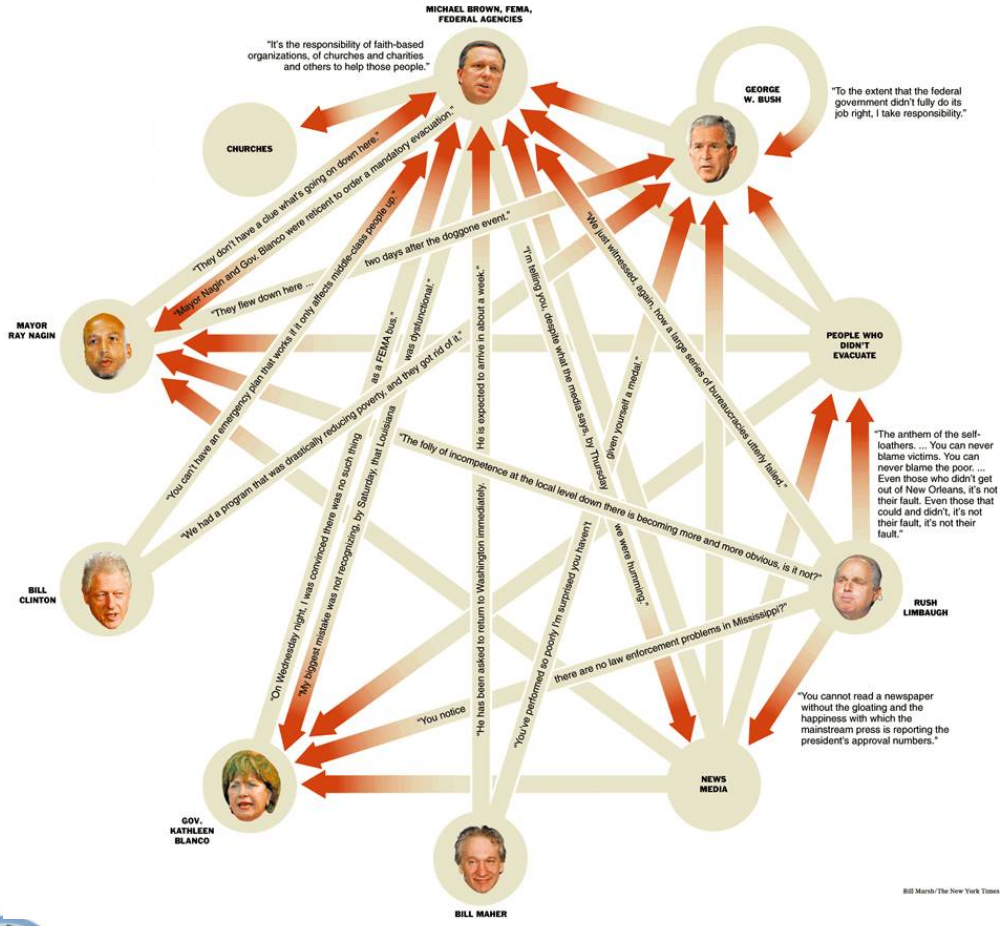


Social Network

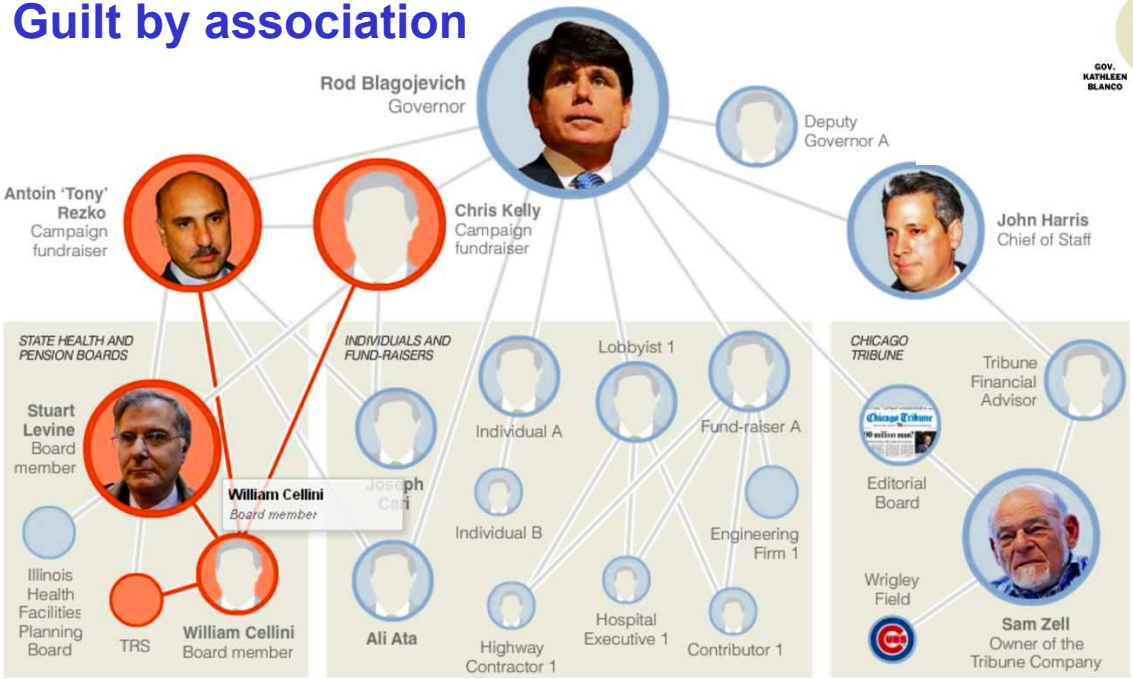
Key comparisons

- Electronic circuits share similar design principles compared to biological networks (synthetic biology)
- Social Systems provide useful intuition
- Biological systems and computer operating systems (The genome has often been called the OS for a living organism)
  - ◊ Execute information processing tasks
  - ◊ adaptive systems shaped by changing environments (Natural vs man-made systems)

Using the position in networks to describe function



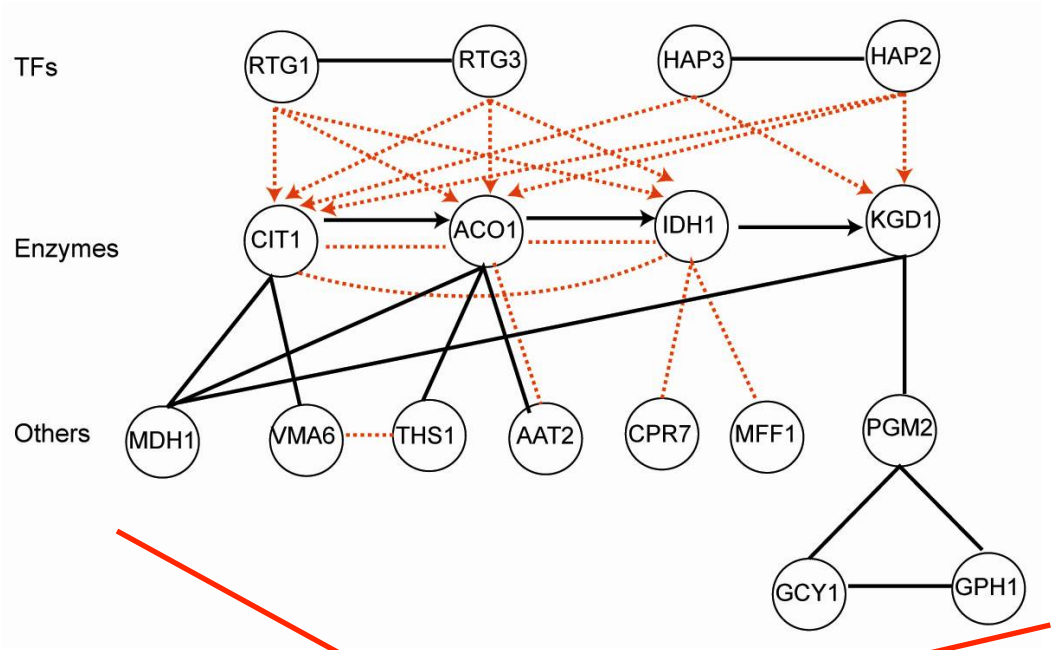
Guilt by association

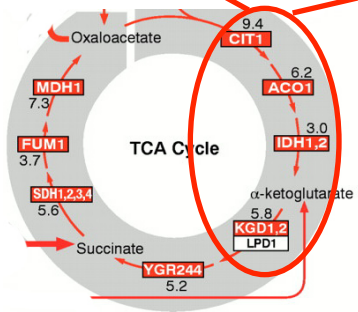


Finding the causal regulator (the "Blame Game")

[NY Times, 2-Oct-05, 9-Dec-08]

Combining networks forms an ideal way of integrating diverse information





Part of the TCA cycle

Metabolic pathway

Transcriptional regulatory network

Physical protein-protein Interaction

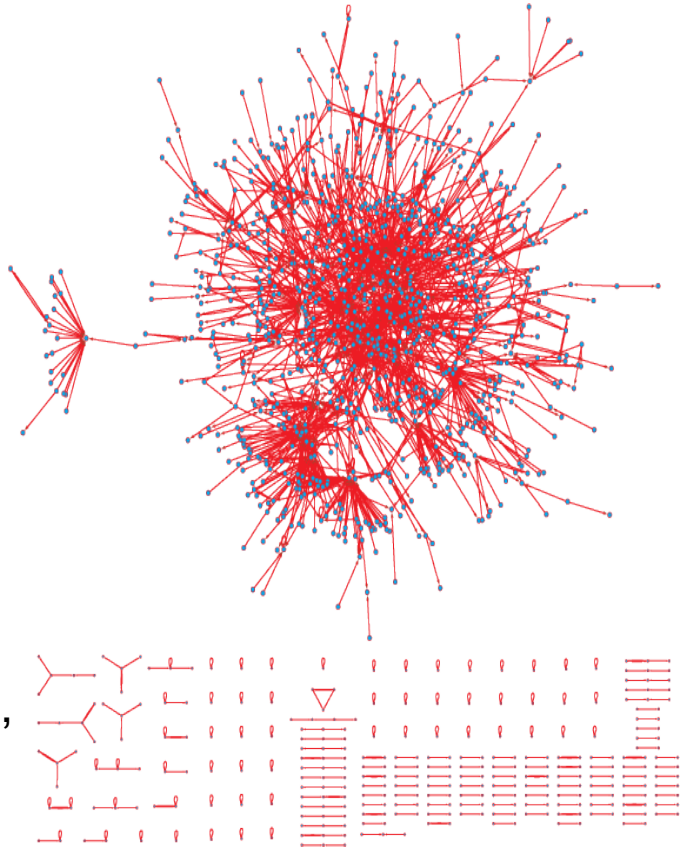
Co-expression Relationship

Genetic interaction (synthetic lethal)
Signaling pathways

Outline: Molecular Networks

- Why Networks?
- Generating Networks
  - Scanning for Targets of Modular Domains
  - Propagating Known Information
    (yeast ppi)
- Central Network Points
  - Hubs & Bottlenecks
    (yeast ppi & reg. net)
- Networks & Variation
  (human ppi & miRNA-targ. net)
- Social Network Comparisons
  (reg. net. in many organisms)
  - in rel. to social hierarchy
  - scaling in rel. to partnerships
- Computer OS Comparisons
  (E. coli reg. net)

Example: yeast PPI network



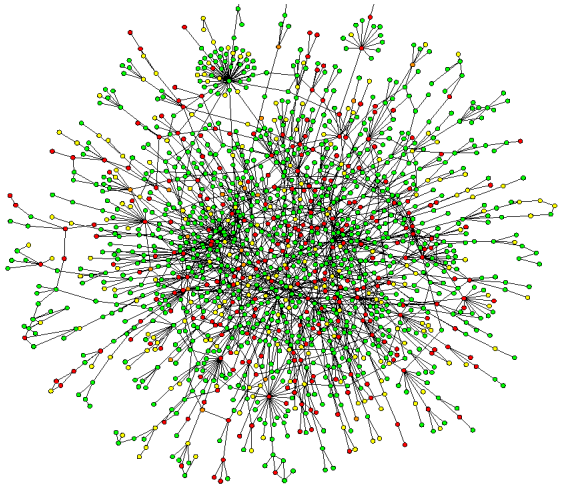
Actual size:

- ~6,000 nodes
  → Computational cost: ~18M pairs
- Estimated ~15,000 edges
  → Sparseness: 0.08% of all pairs (Yu et al., 2008)

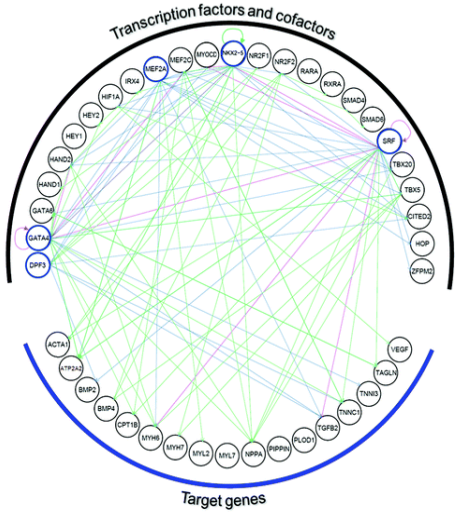
Known interactions:

- Small-scale experiments: accurate but few
  → Overfitting: ~5,000 in BioGRID, involving ~2,300 proteins
- Large-scale experiments: abundant but noisy
  → Noise: false +ve/-ve for yeast two-hybrid data up to
  45% and 90% (Huang et al., 2007)

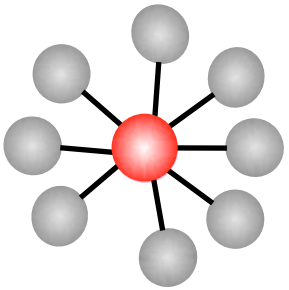
Different Types of Molecular Networks



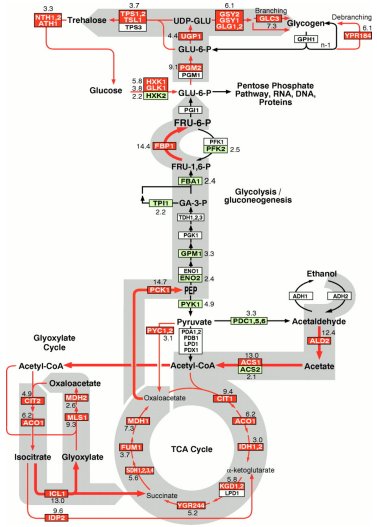
Protein-protein Interaction networks



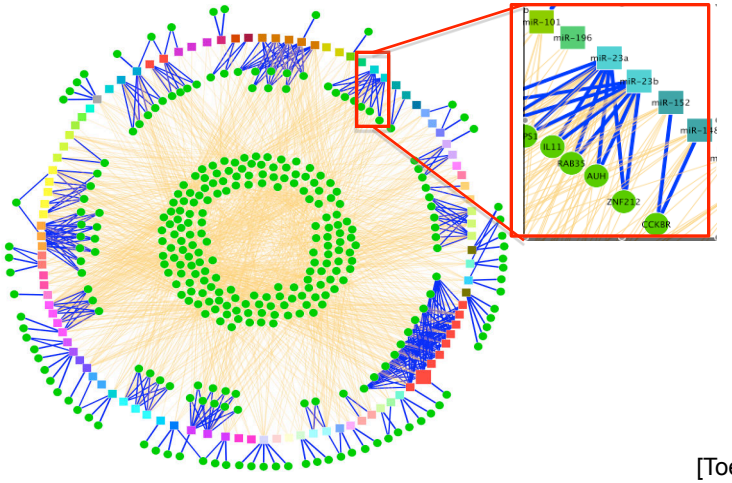
TF-target-gene Regulatory networks



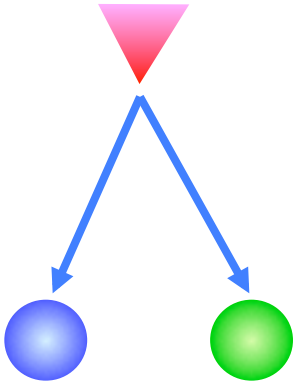
Undirected



Metabolic pathway networks



miRNA-target networks

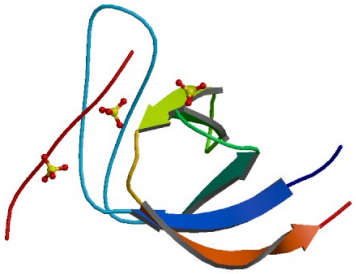


Directed

[Toenjes, *et al*, *Mol. BioSyst.* (2008); Jeong *et al*, *Nature* (2001); [Horak, et al, Genes & Development, 16:3017-3033; DeRisi, Iyer, and Brown, Science, 278:680-686]

Generating Networks

How do we construct large molecular networks.
From connecting sequence patterns for modular interaction domains to matching sequences



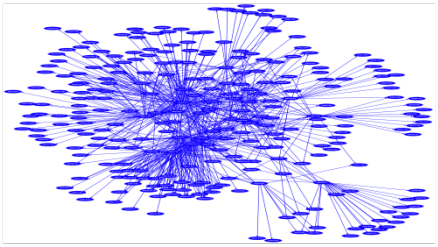
Profiling for interaction specificity

Scanning for target candidates

Proteome
with pre-computed features

Predicting interaction targets

Reconstructing networks/pathways



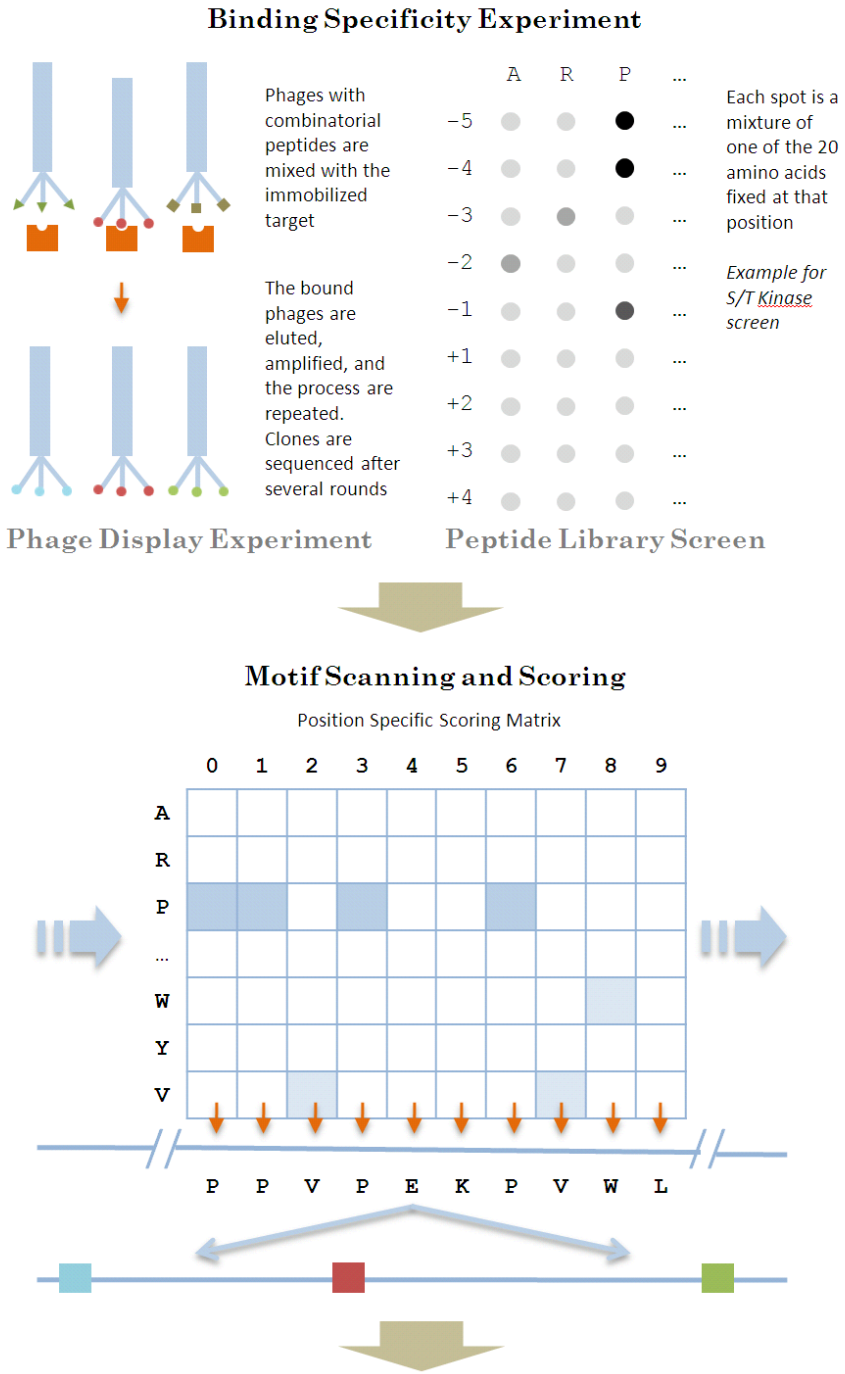
Modular protein domains interact with peptides containing a specific motif

Search for these sequences in a target proteome to find potential interactions

Motips

Determining the interaction specificity

- Experimental methods to identify domain interaction specificity
  - ◊ phage display experiment (e.g. SH3)
  - ◊ the peptide library screen (e.g. S/T Kinase)
- Scan the target proteome with normalized PWMs

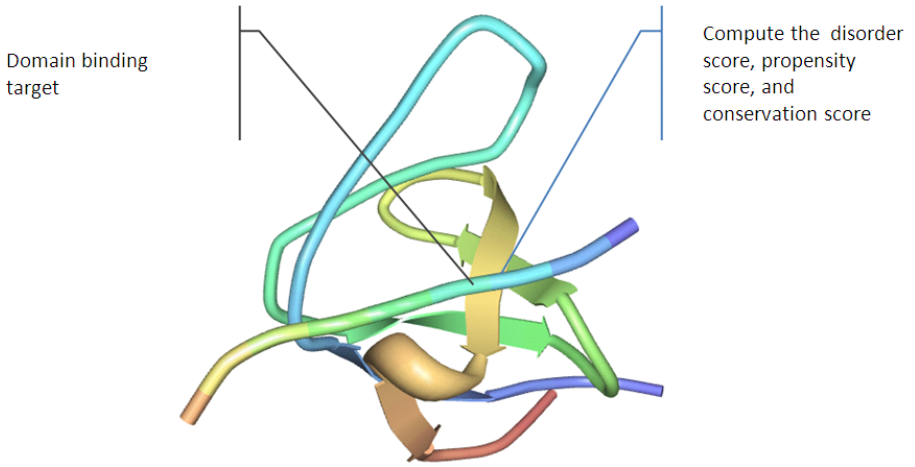


[Lam et al., BMC Bioinfo. ('10, press)]

Integrating structural and conservation features

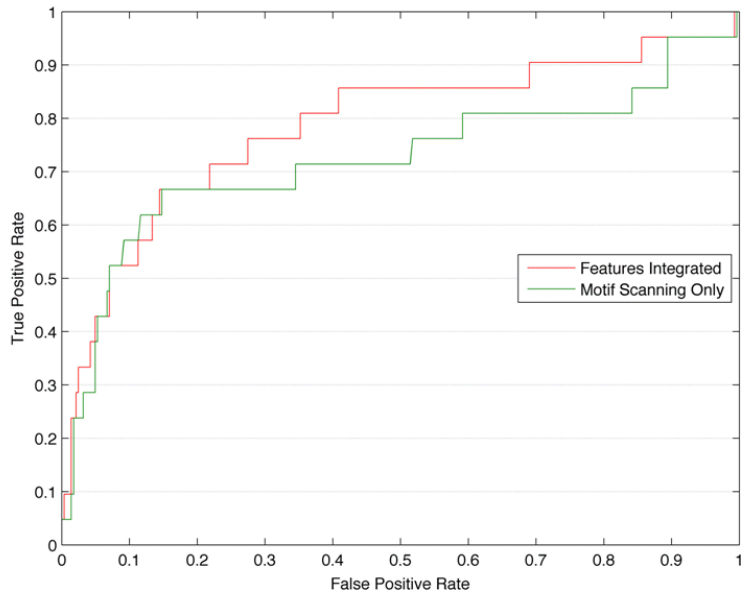
- Calculate features for potential interaction sites:
  - ◊ Surface accessibility
  - ◊ Protein disorder
  - ◊ Sequence Conservation
- Build a naïve Bayes classifier based on a validated training set to integrate the features

Features Scoring and Integration



***Prk1* Kinase AUC:**
79% vs 72%

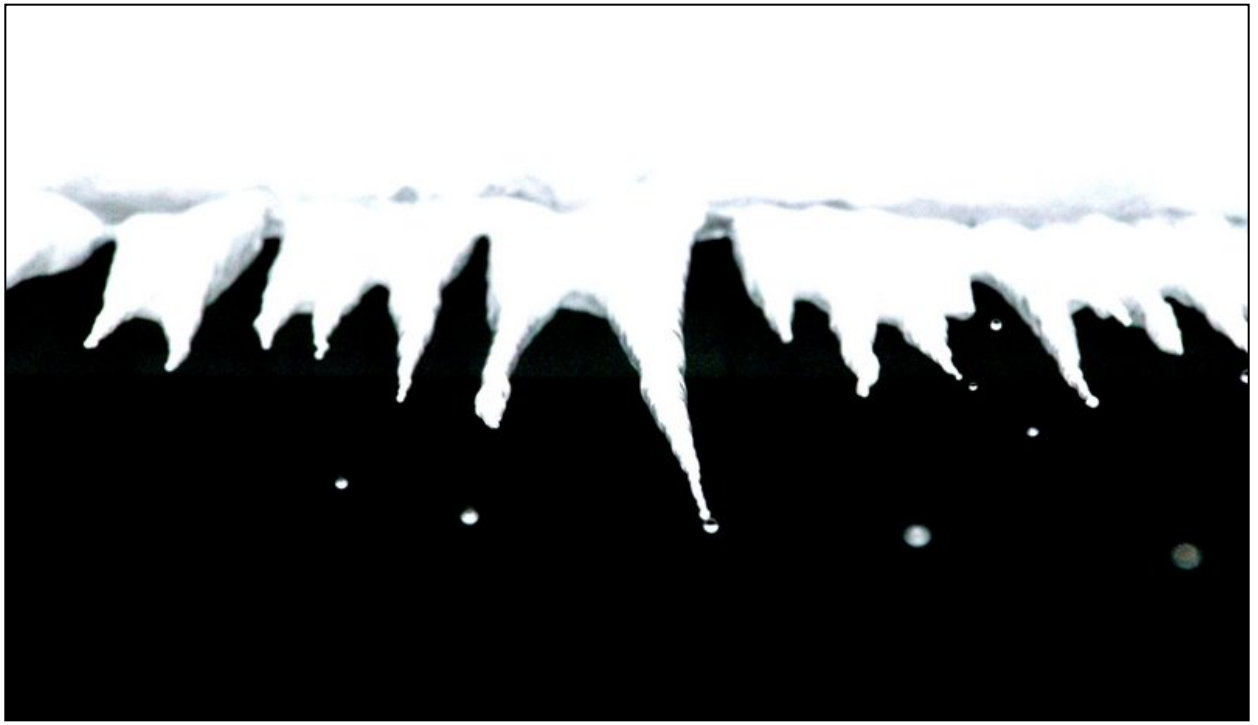
Domain Target Prediction



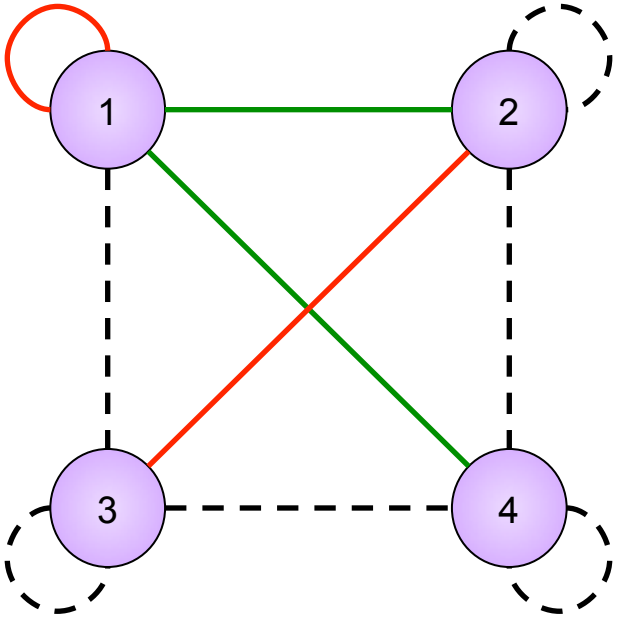
[Lam et al., BMC Bioinfo. ('10, press)]

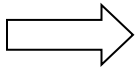
Generating Networks #2

How do we construct large molecular networks?
From extrapolating correlations between functional genomics data with fairly small sets of known interactions, making best use of the known training data.



Training sets





	1	2	3	4
1	0	1	?	1
2	1	?	0	?
3	?	0	?	?
4	1	?	?	?

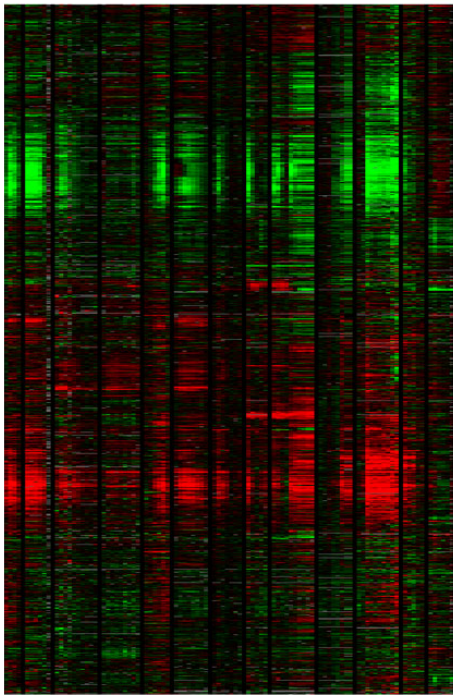
Known interactions

Known non-interactions

Unknown

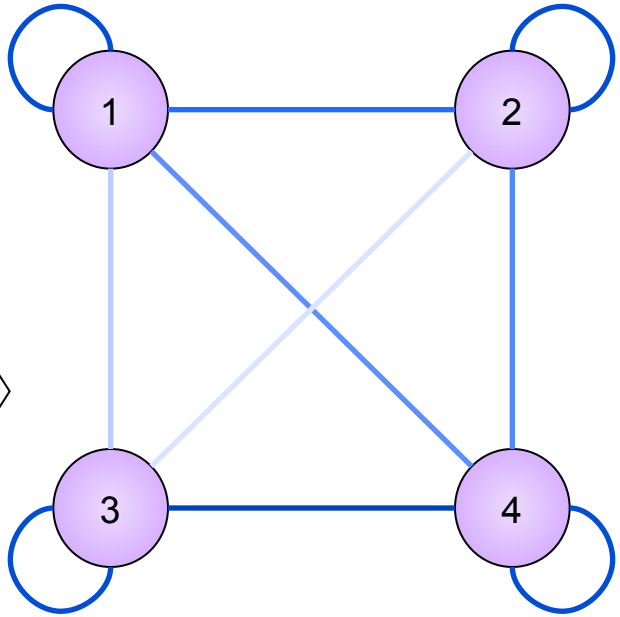
Network prediction: features

- Example 1: gene expression



Gasch et al., 2000

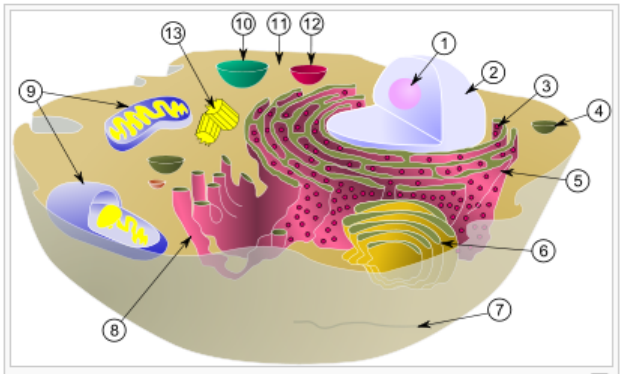
$x_1 = (0.2, 2.4, 1.5, \ldots)$
$x_2 = (0.8, 2.2, 1.5, \ldots)$
$x_3 = (4.3, 0.1, 7.5, \ldots)$
…
$sim(x_1, x_2) = 0.62$
$sim(x_1, x_3) = -0.58$
…



Similarity scale:

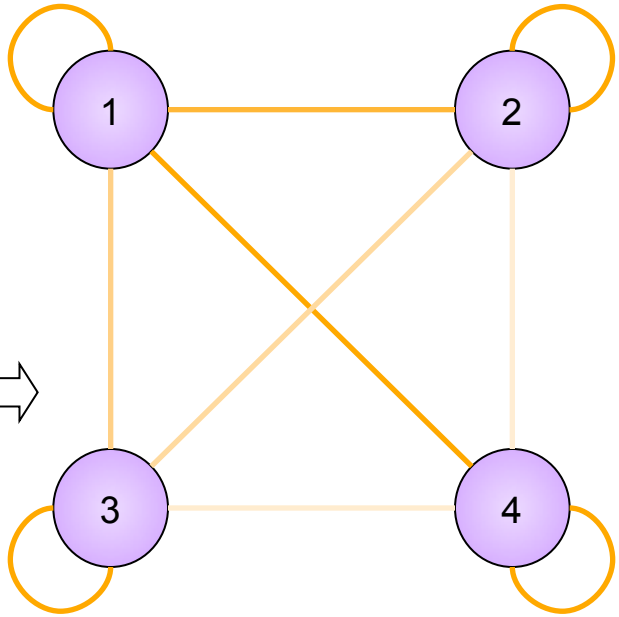
Network prediction: features

- Example 2: sub-cellular localization



http://www.scq.ubc.ca/wp-content/yeasttwohybridtranscript.gif

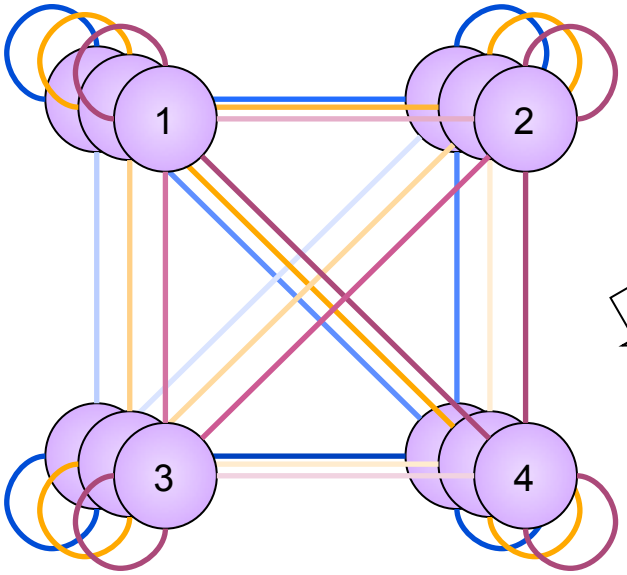
$x_1 = (1, 1, 0, 0, \ldots)$
$x_2 = (1, 1, 1, 0, \ldots)$
$x_3 = (1, 0, 1, 0, \ldots)$
…
$sim(x_1, x_2) = 0.81$
$sim(x_1, x_3) = 0.12$
…

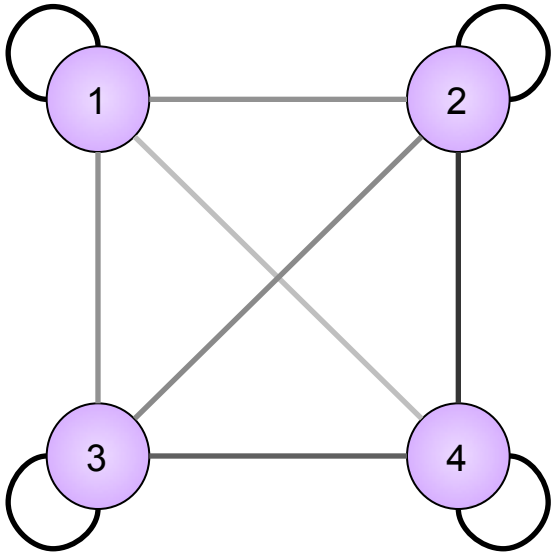


Similarity scale:

1 **-1**

Data integration & Similarity Matrix





	1	2	3	4
1	1.00	0.57	0.55	0.40
2	0.57	1.00	0.66	0.89
3	0.55	0.66	1.00	0.79
4	0.40	0.89	0.79	1.00

Learning methods

An endless list:

- Docking (e.g. Schoichet and Kuntz 1991)
- Evolutionary (e.g. Ramani and Marcotte, 2003)
- Topological (e.g. Yu et al., 2006)
- Bayesian (e.g. Jansen et al., 2003)
- **Kernel methods**
  - ◊ Global modeling:
    - em (Tsuda et al., 2003)
    - kCCA (Yamanishi et al., 2004)
    - kML (Vert and Yamanishi, 2005)
    - Pairwise kernel (Pkernel) (Ben-Hur and Noble, 2005)
  - ◊ Local modeling:
    - Local modeling (Bleakley et al., 2007)

Let's compare in a public challenge!
(DREAM: Dialogue for Reverse Engineering Assessment and Methods)

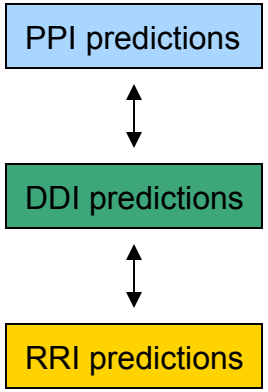
Our work: efficiently propagating known information

Training set expansion

- Motivation: lack of training examples
- Expand training sets horizontally

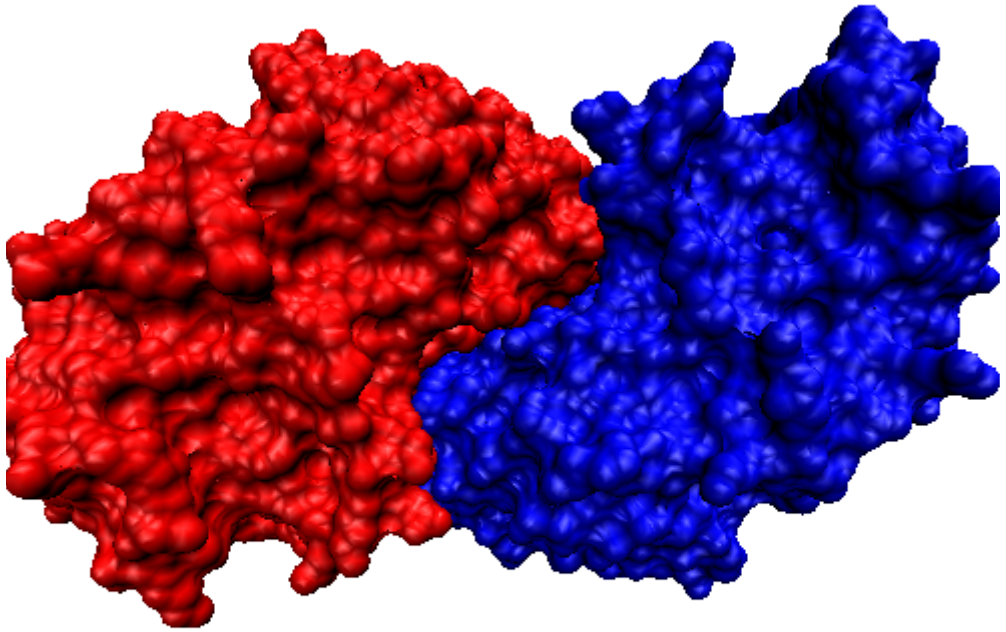
Multi-level learning

- Motivation: hierarchical nature of interaction
- Expand training sets vertically



DREAM3 *in silico* regulatory network reconstruction challenge

Protein interaction

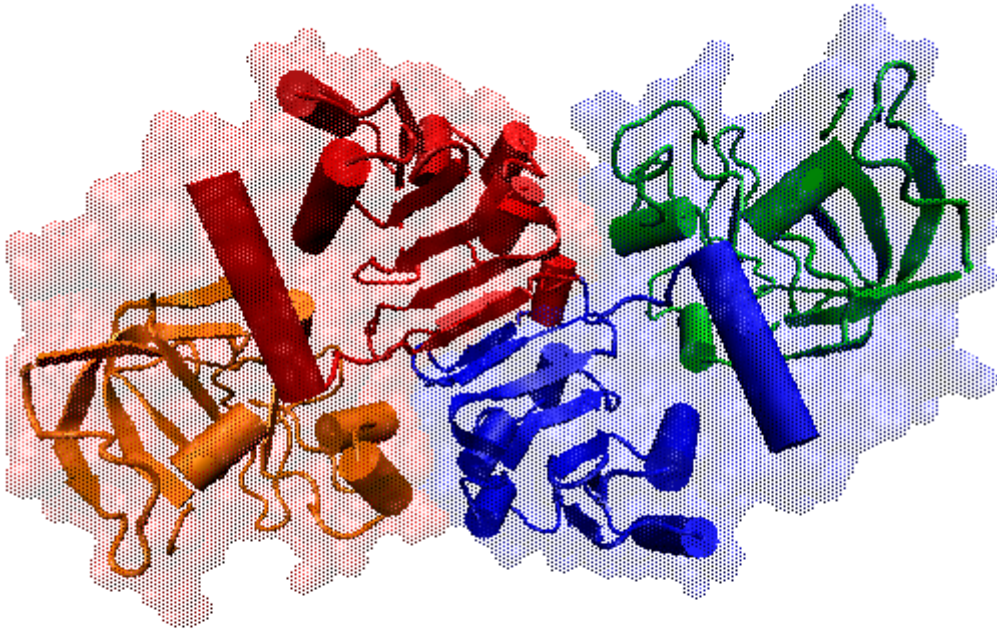


Yeast NADP-dependent alcohol dehydrogenase 6 (PDB: 1piw)

Protein-level features for interaction prediction: functional genomic information

[Yip and Gerstein, BMC Bioinfo. ('09, press)]

Domain interaction

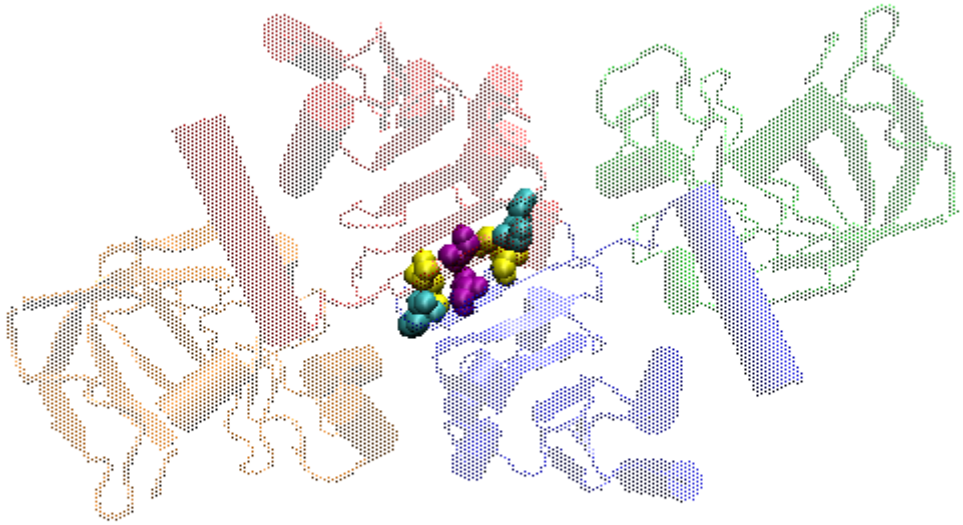


Pfam domains: PF00107 (inner) and PF08240 (outer)

Domain-level features for interaction prediction: evolutionary information

[Yip and Gerstein, BMC Bioinfo. ('09, press)]

Residue interaction

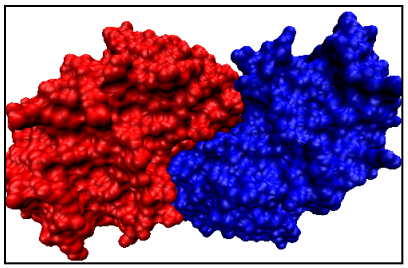


Interacting residues: 283 (yellow) with 287 (cyan), and 285 (purple) with 285

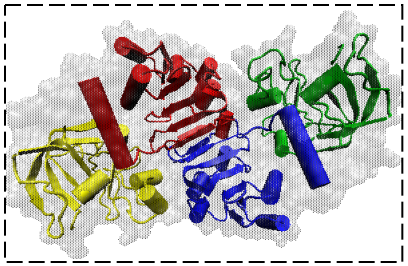
Residue-level features for interaction prediction: physical-chemical information

[Yip and Gerstein, BMC Bioinfo. ('09, press)]

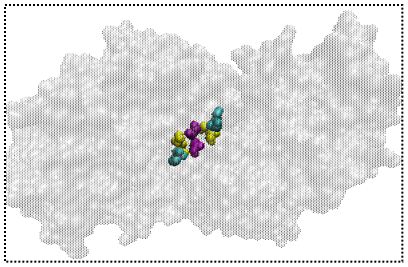
Combining the three problems



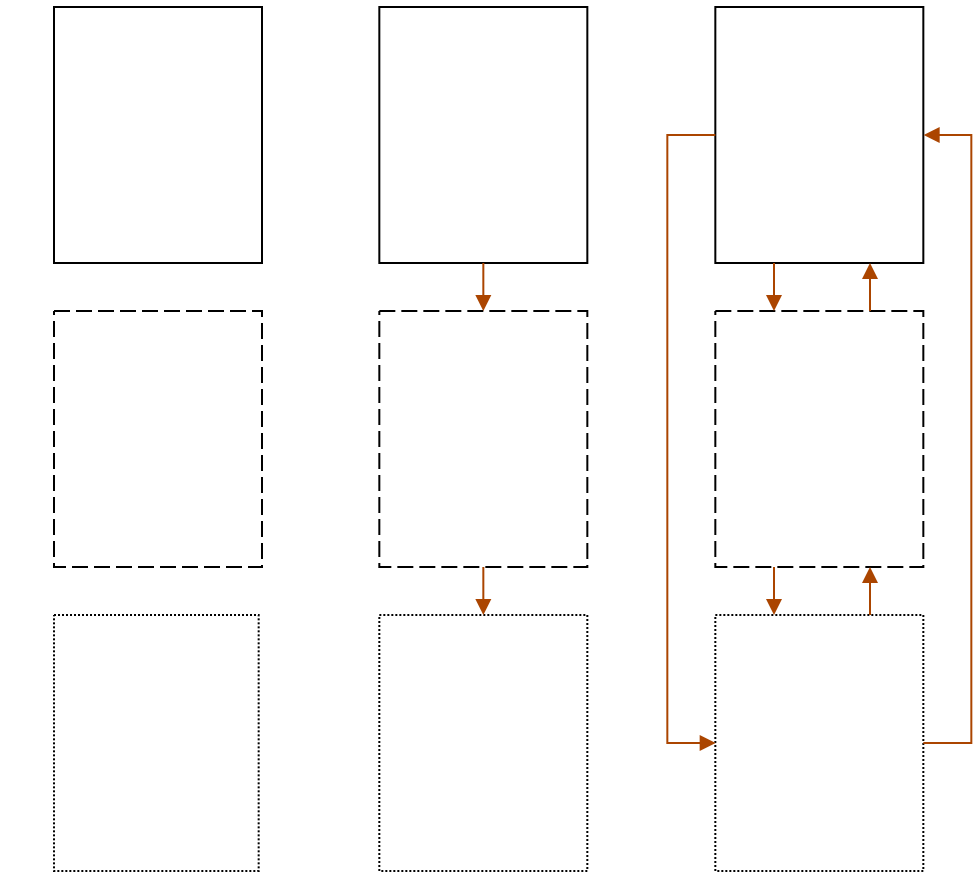
Protein interactions



Domain interactions



Residue interactions

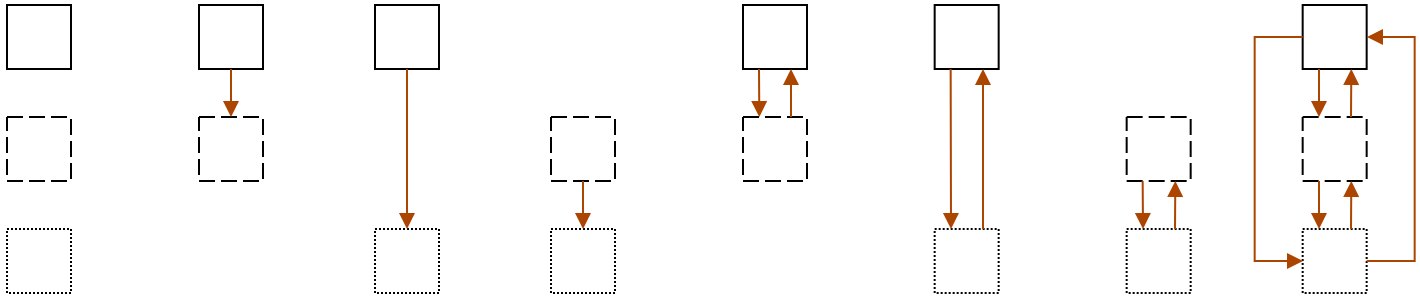


i. Independent levels ii. Unidirectional flow iii. Bidirectional flow

[Yip and Gerstein, BMC Bioinfo. ('09, press)]

Empirical results (AUCs)

	Ind. levels	Unidirectional flow			Bidirectional flow			
Level		PD	PR	DR	PD	PR	DR	PDR
Proteins	71.68				72.23	72.50		**72.82**
Domains	53.18	61.51			**71.71**		68.94	71.20
Residues	57.36		54.89	53.81		72.26	63.16	**77.86**



- Highest accuracy by bidirectional flow
- Additive effect: 2 vs. 3 levels

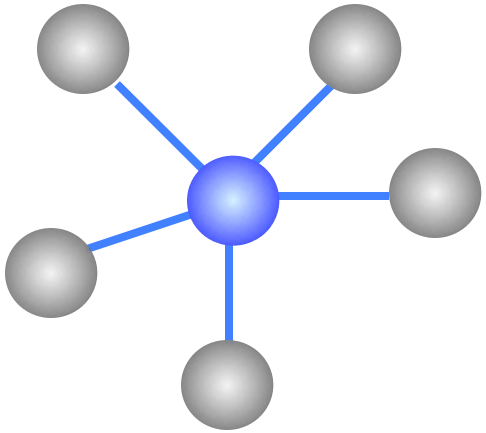
[Yip and Gerstein, BMC Bioinfo. ('09, press)]

Finding Central Points in Networks: Hubs & Bottlenecks

Where are key points networks ? How do we locate them ?

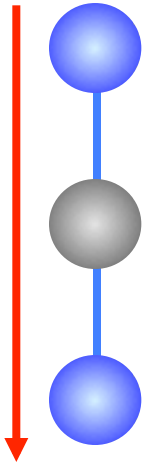
Global topological measures

Indicate the gross topological structure of the network



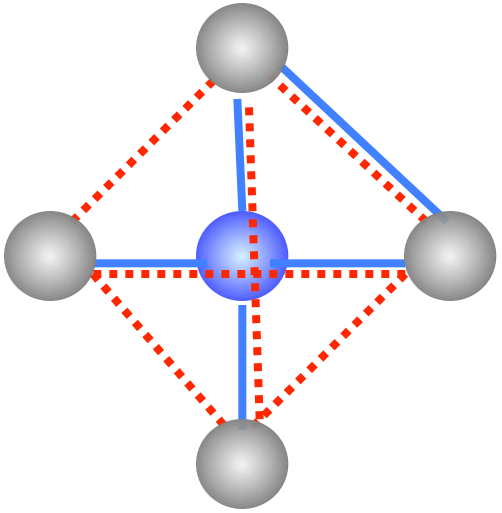
Degree (K)

5



Path length (L)

2



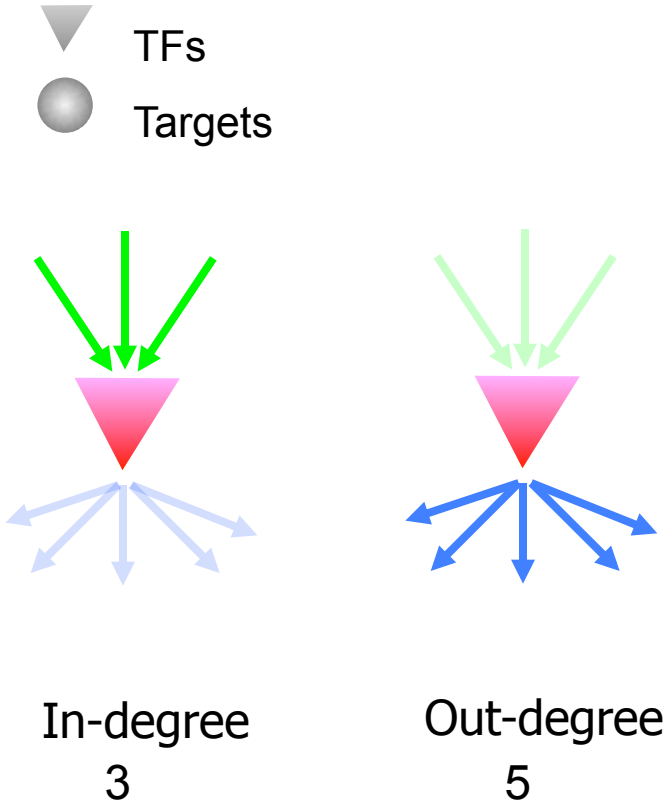
Clustering coefficient (C)

1/6

Interaction and expression networks are ***undirected***

[Barabasi]

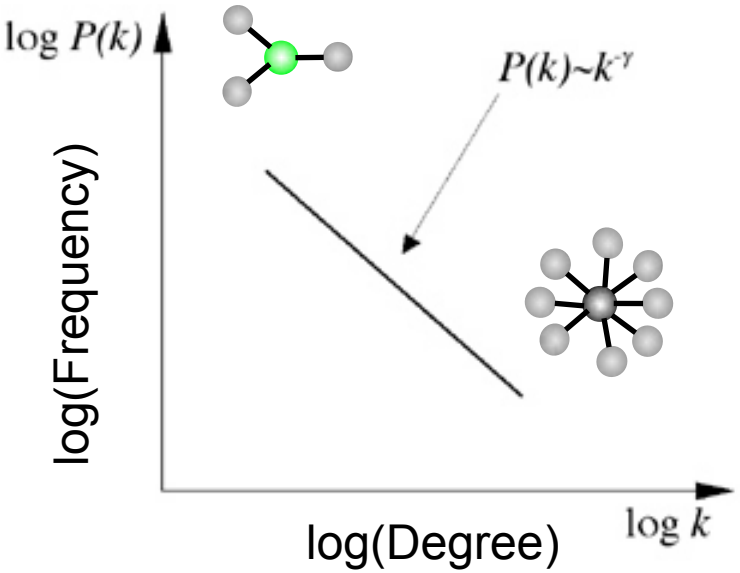
Global topological measures for directed networks

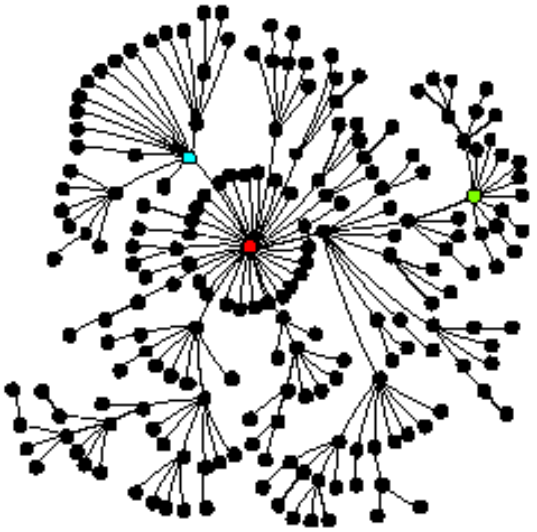


Regulatory and metabolic networks are ***directed***

Scale-free networks

Power-law distribution





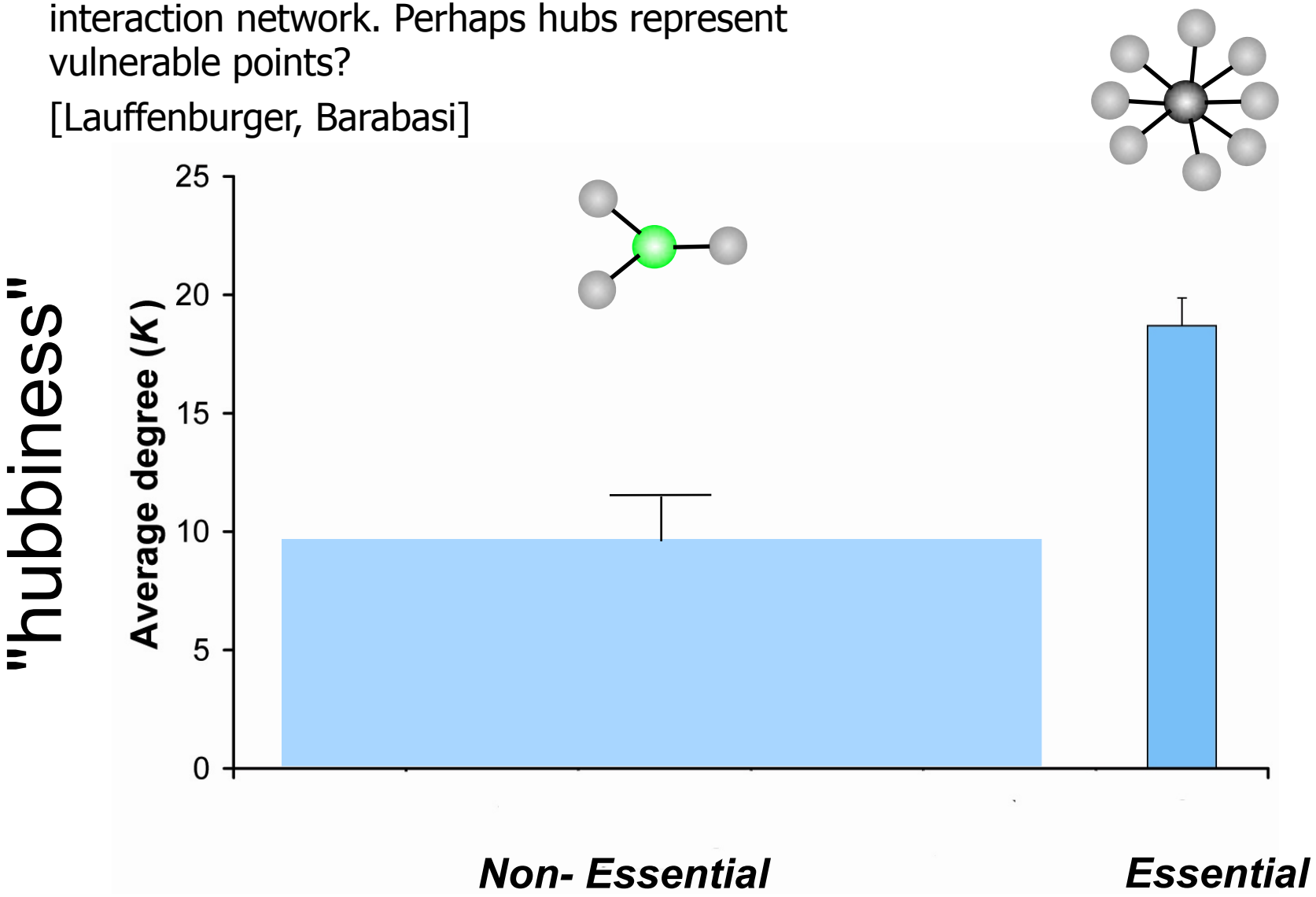
Hubs dictate the structure of the network

[Barabasi]

Hubs tend to be Essential

Integrate gene essentiality data with protein interaction network. Perhaps hubs represent vulnerable points?

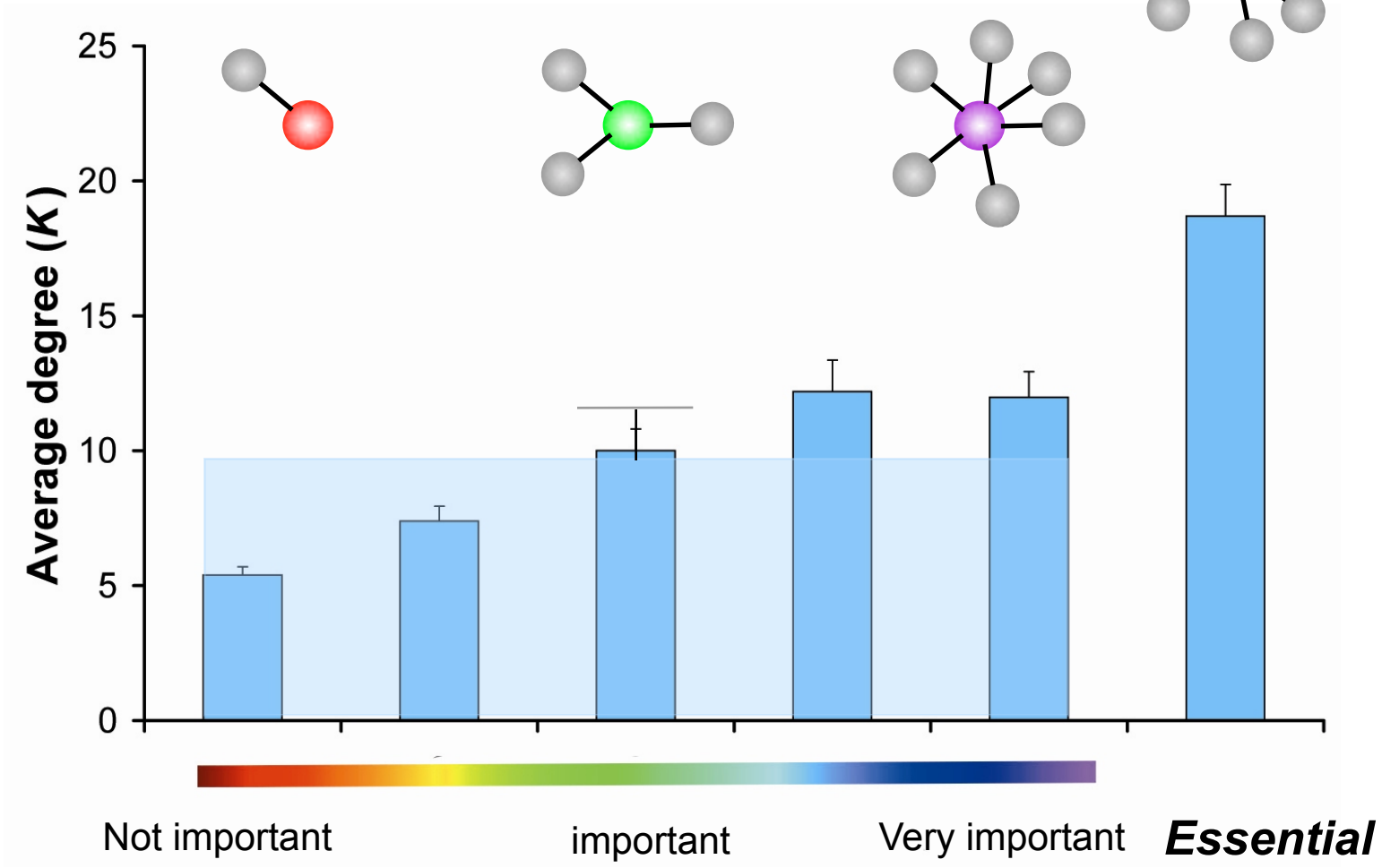
[Lauffenburger, Barabasi]



[Yu et al., 2003, TIG]

Relationships extends to "Marginal Essentiality"

Marginal essentiality measures relative importance of each gene (e.g. in growth-rate and condition-specific essentiality experiments) and scales continuously with "hubbiness"



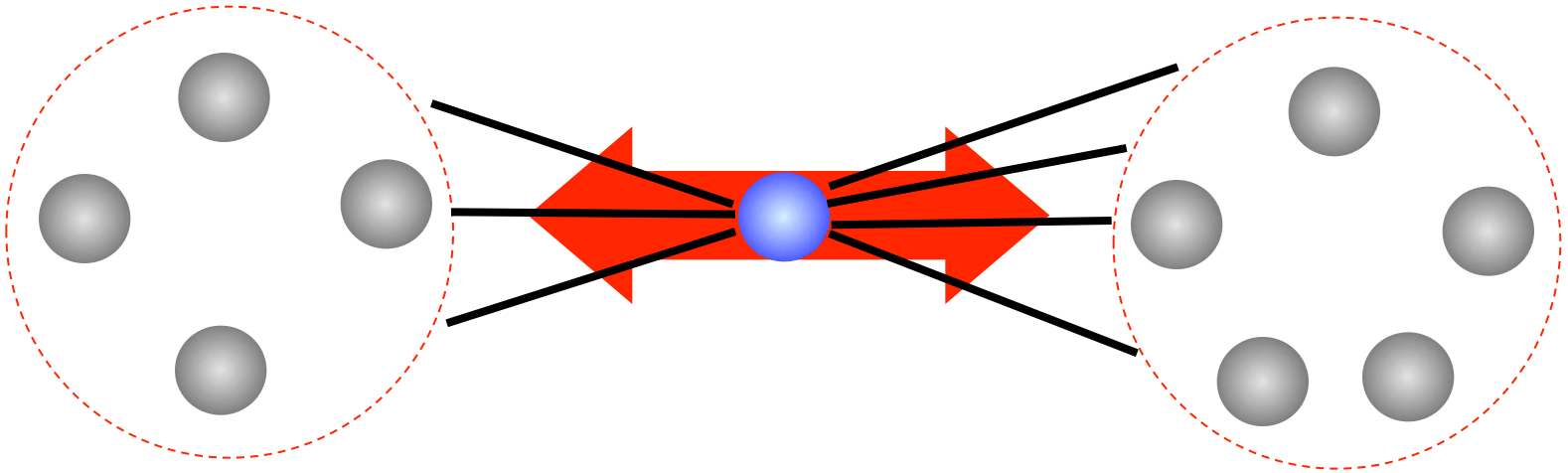
[Yu et al., 2003, TIG]

Another measure of Centrality: Betweenness centrality

Betweenness of a node is the number of shortest paths of pairs of vertices that run through it -- a measure of information flow.

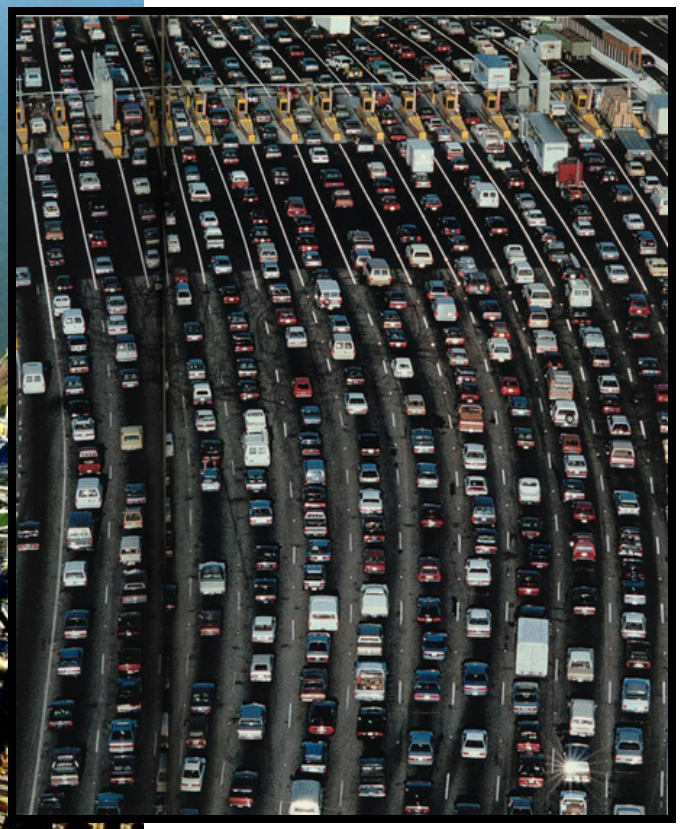
Freeman LC (1977) Set of measures of centrality based on betweenness. Sociometry 40: 35–41.

Girvan & Newman (2002) PNAS 99: 7821.

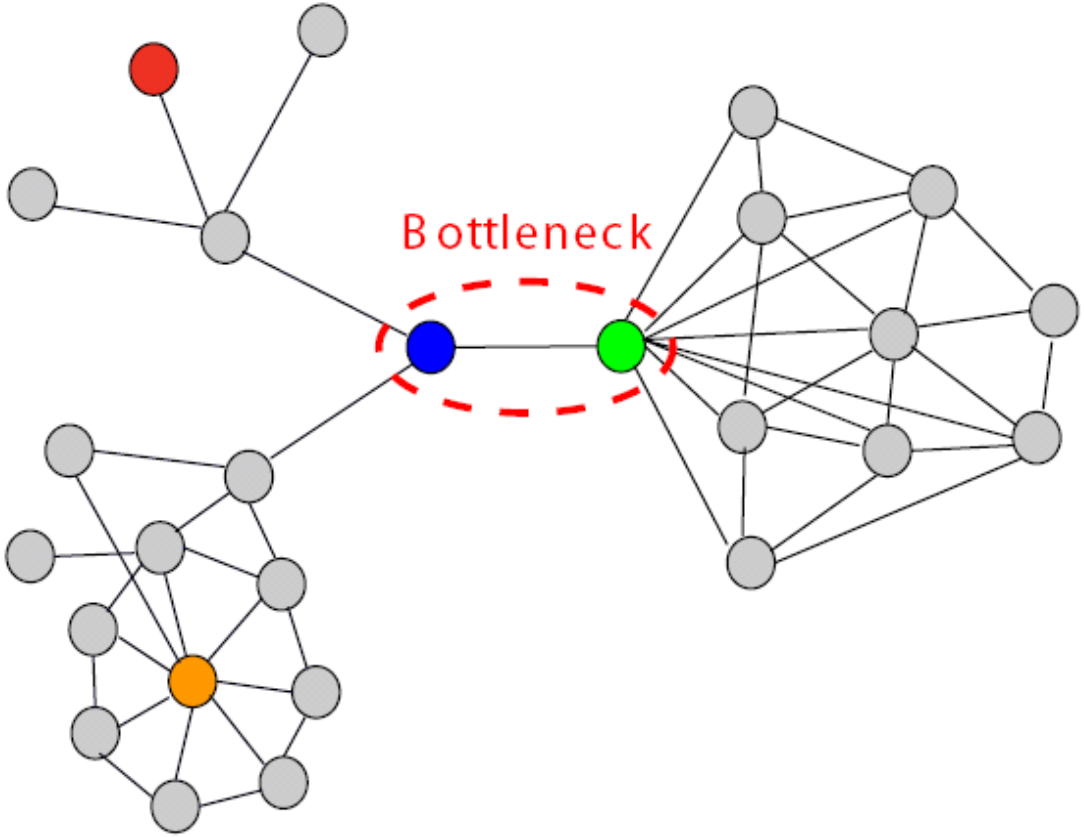


Betweenness centrality -- Bottlenecks

Proteins with high betweenness are defined as *Bottlenecks* (top 20%), in analogy to the traffic system



Bottlenecks & Hubs



Hub-bottleneck **node**

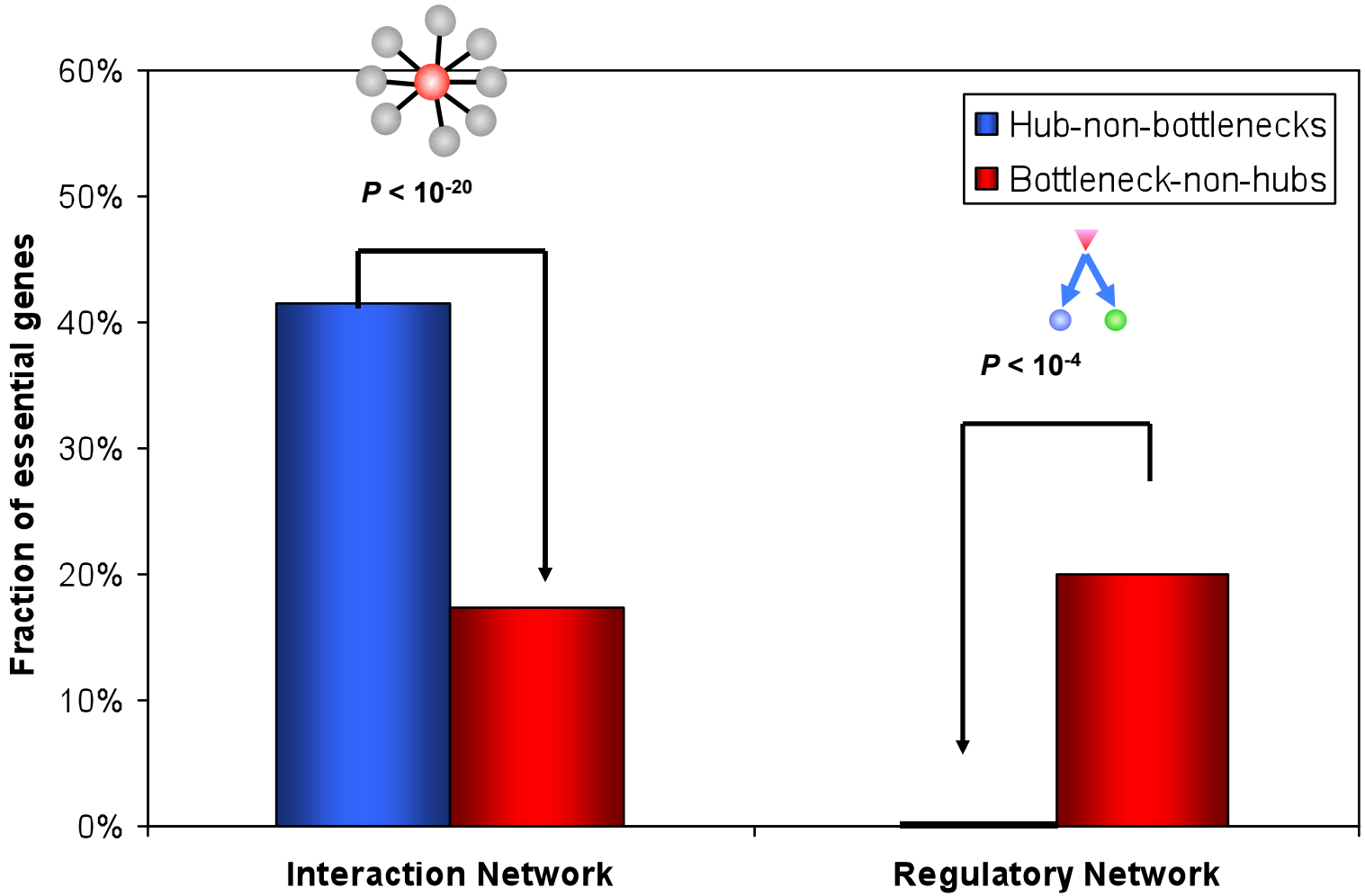
Non-hub-bottleneck **node**

Hub-non-bottleneck **node**

Non-hub-non-bottleneck **node**

[Yu et al., PLOS CB (2007)]

Bottlenecks are what matters in regulatory networks



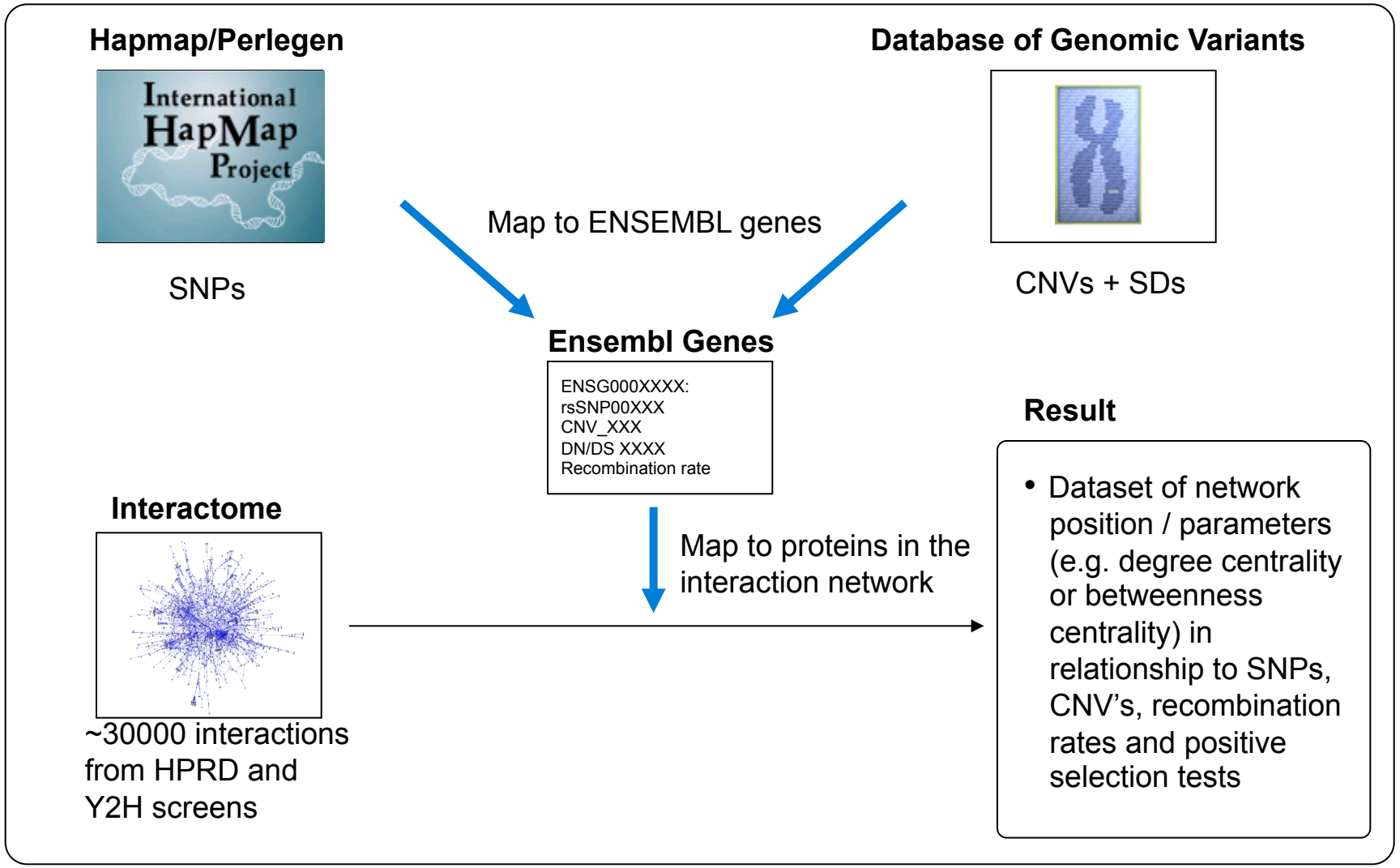
[Yu *et al.*, *PLoS Comput Biol* (2007)]

Networks & Variation

Which parts of the network vary most in sequence?
Which are under selection, either positive or negative?

METHODOLOGY: MAP SNP AND CNV DATA ONTO ENSEMBL GENES, AND THEN MAP ENSEMBL GENES TO THE KNOWN INTERACTOME

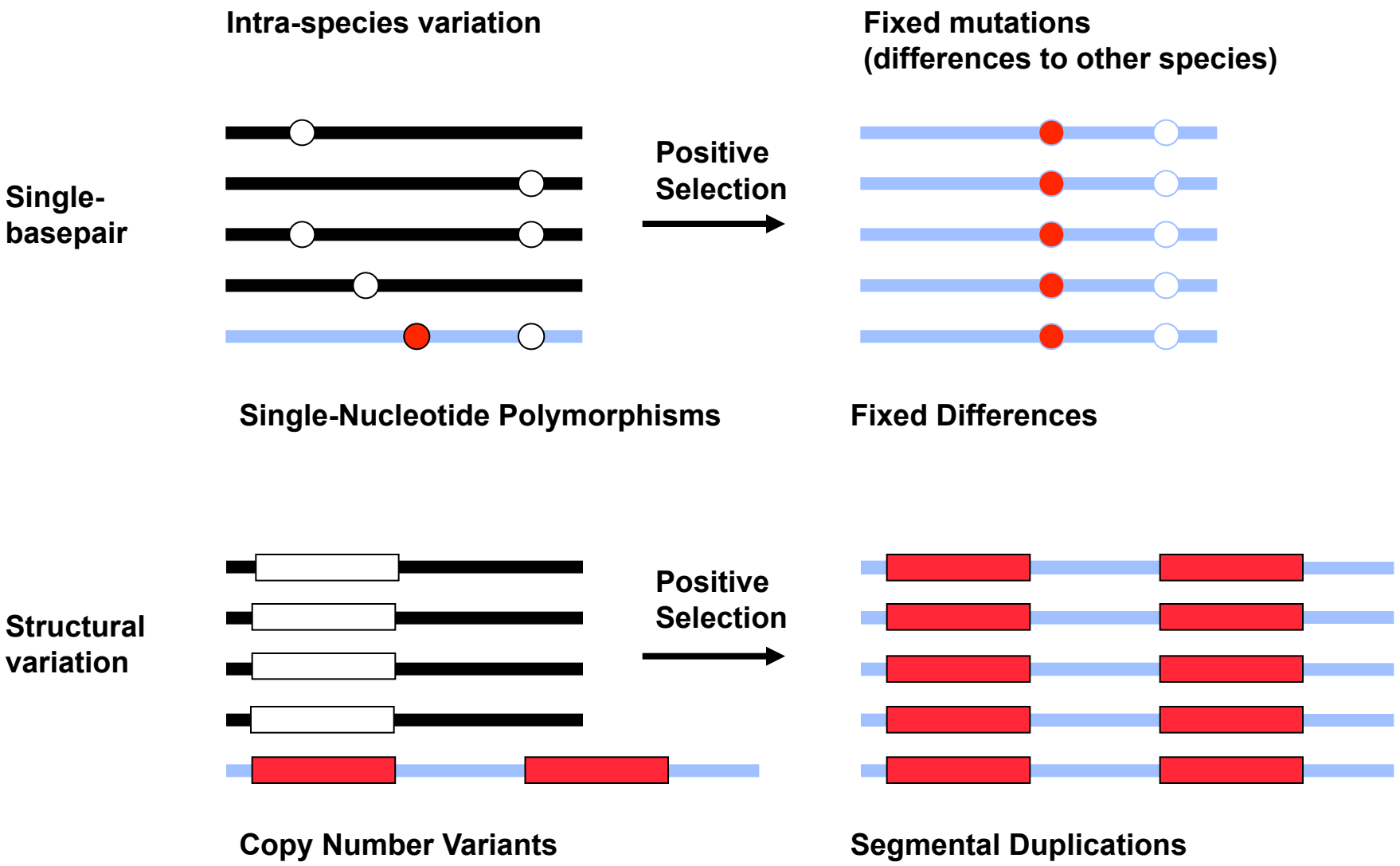
ILLUSTRATIVE



* From Nielsen et al. *PLoS Biol.* (2005) and Bustamante et al. *Nature* (2005)

Source: PMK

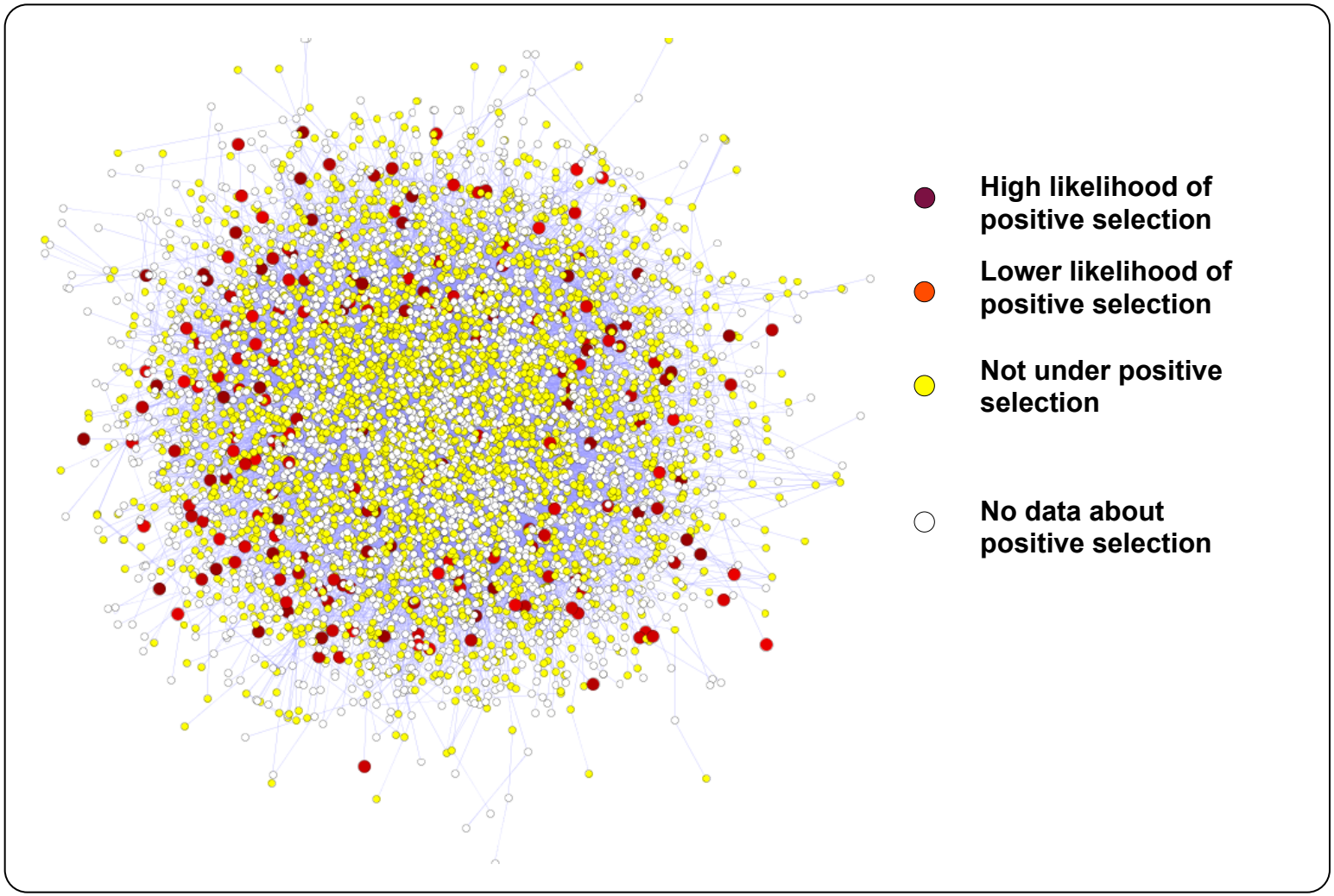
ADAPTIVE EVOLUTION CAN BE SEEN ON TWO DIFFERENT LEVELS



Source: PMK

POSITIVE SELECTION LARGELY TAKES PLACE AT THE NETWORK PERIPHERY

Positive selection in the human interactome

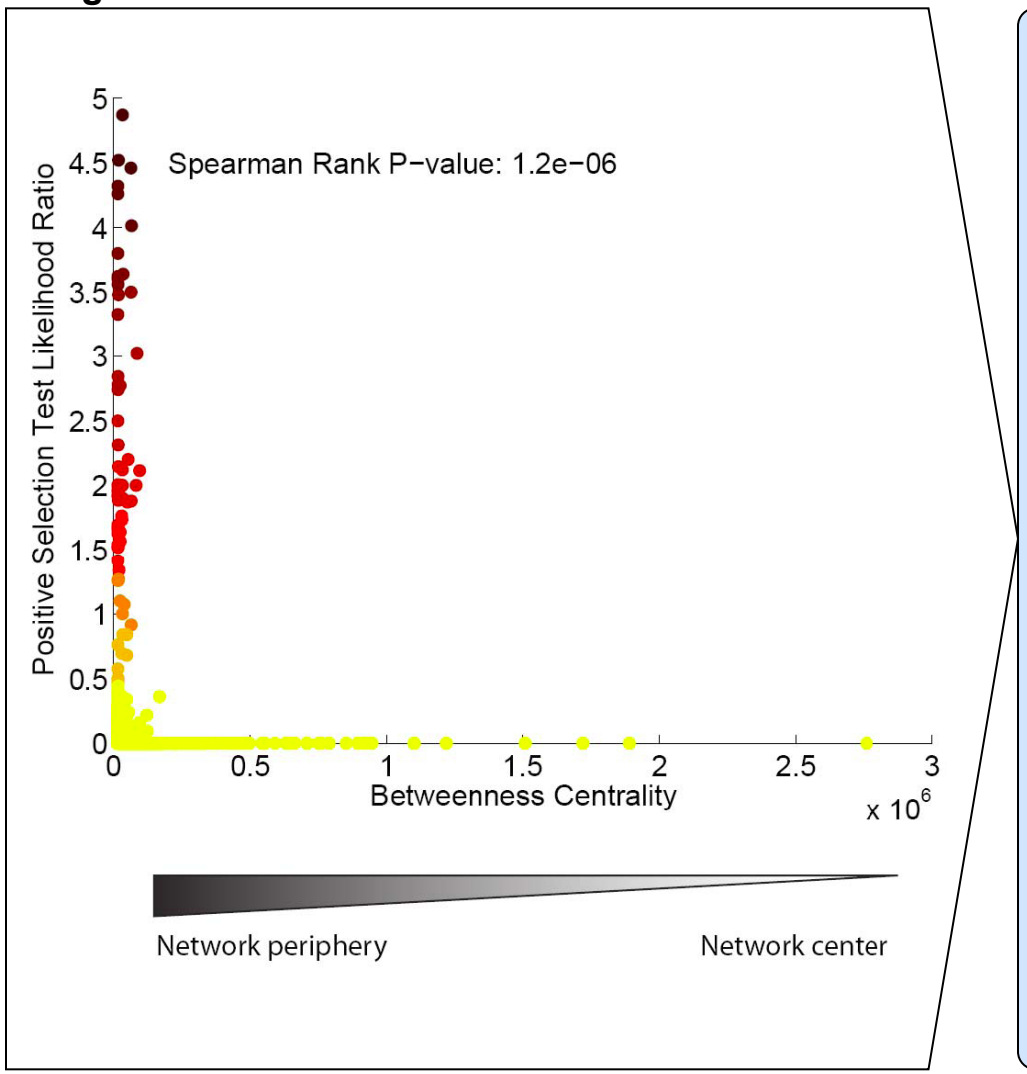


Source: Nielsen et al. *PLoS Biol.* (2005), HPRD, and Kim et al. PNAS (2007)

CENTRAL PROTEINS ARE LESS LIKELY TO BE UNDER POSITIVE SELECTION

Hubs

Degree vs. Positive Selection



Reasoning

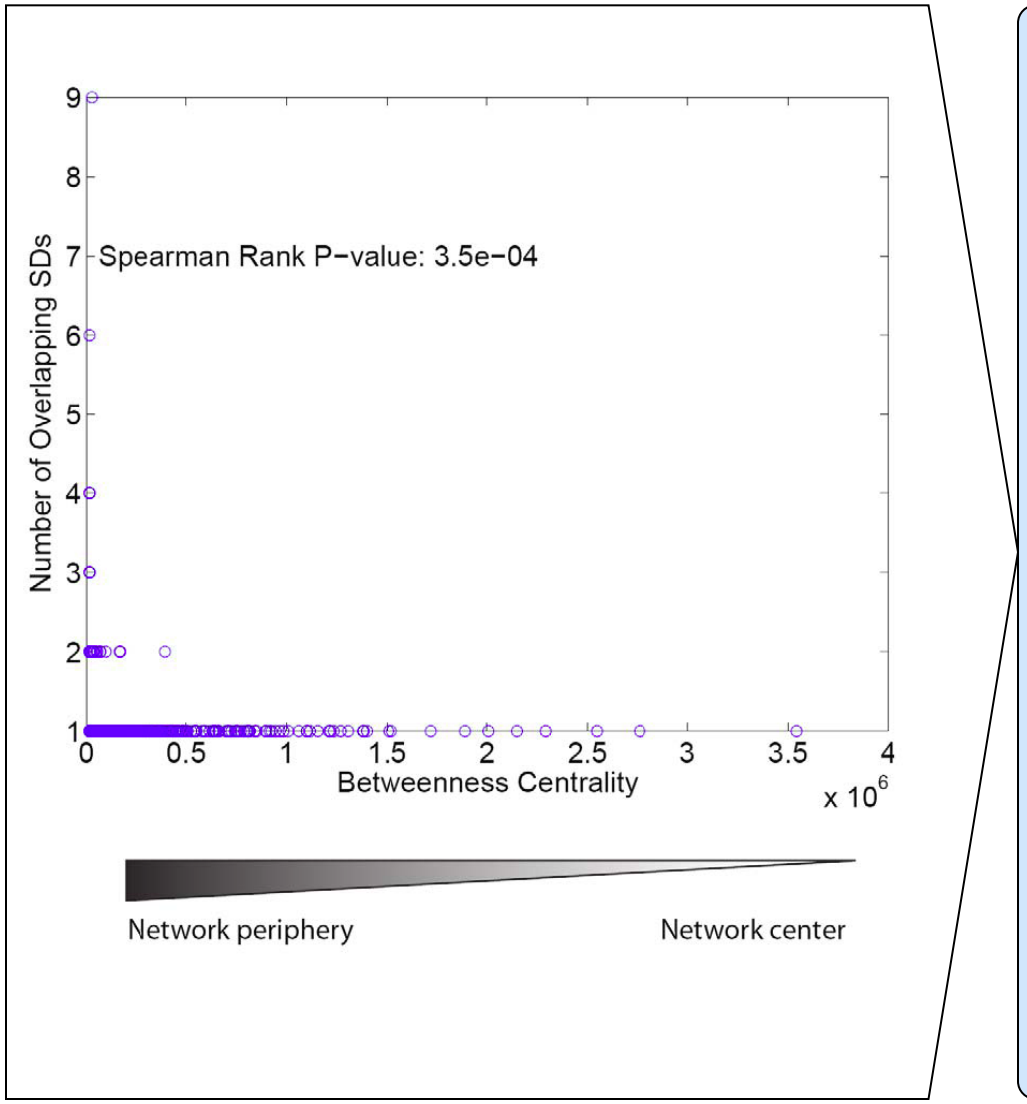
- Peripheral genes are likely to under positive selection, whereas hubs aren't
- This is likely due to the following reasons:
  - Hubs have stronger structural constraints, the network periphery doesn't
  - Most recently evolved functions (e.g. "environmental interaction genes" such as sensory perception genes etc.) would probably lie in the network periphery
- Effect is independent of any bias due to gene expression differences

* With a probability of over 80% to be positively selected as determined by Ka/Ks. Other tests of positive selection (McDonald Kreitmann and LDD) corroborate this result.

Source: Nielsen et al. *PLoS Biol.* (2005), Bustamante et al. *Nature* (2005), HPRD, Rual et al. *Nature* (2005), and Kim et al. PNAS (2007)

CENTRAL NODES ARE LESS LIKELY TO LIE INSIDE OF SDs

Centrality vs. SD occurrence



Reasoning

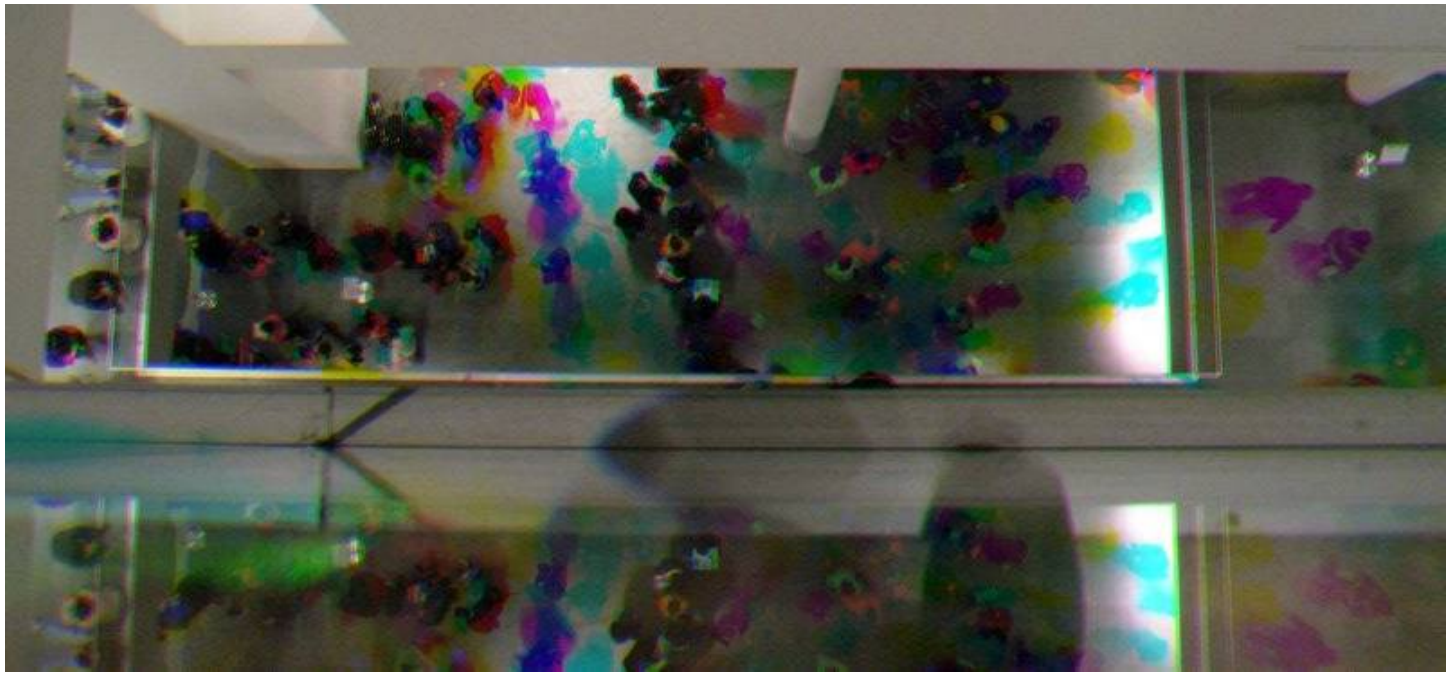
- This result also confirms our initial hypothesis – peripheral nodes tend to lie in regions rich in SDs.
- Since segmental duplications are a different mechanism of ongoing evolution, the less constrained peripheral proteins are enriched in them.
- Note that despite the small size of our dataset for known SD's we get significant correlations. It is to be expected that the correlations will get clearer as more data emerges*

* Specifically, a number of the SDs are likely not fixed, but rather common CNVs in the reference genome

Source: Database of genetic variation, HPRD, Rual et al. *Nature* (2005), and Kim et al. PNAS (2007)

Networks & Variation 2

Variation in the miRNA network

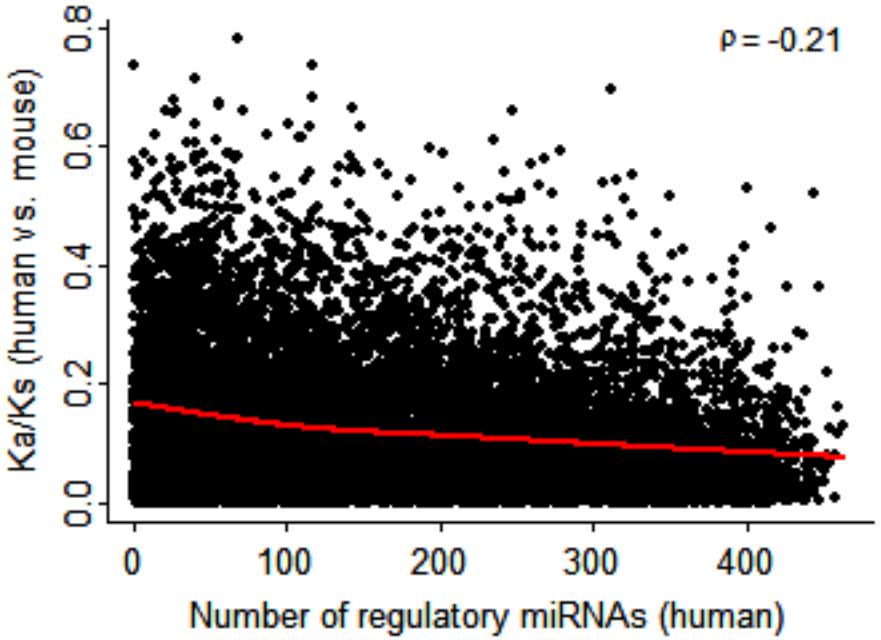


Analyze Regulation in microRNA-target Network

- Relationship between target in degree (number of micro-RNAs that regulate gene) & evolutionary rate of gene?
  - ◊ In deg. related 3' UTR size

- Expectation: more regulation, more constraint

Relationship between microRNA regulation and protein evolution



Human vs.	Number of genes	Correlation	P-value
chimpanzee	11326	-0.11	2.E-32
mouse	13280	-0.21	7.E-128
rat	12270	-0.20	4.E-107
cow	11683	-0.21	8.E-115
chicken	8061	-0.18	1.E-57

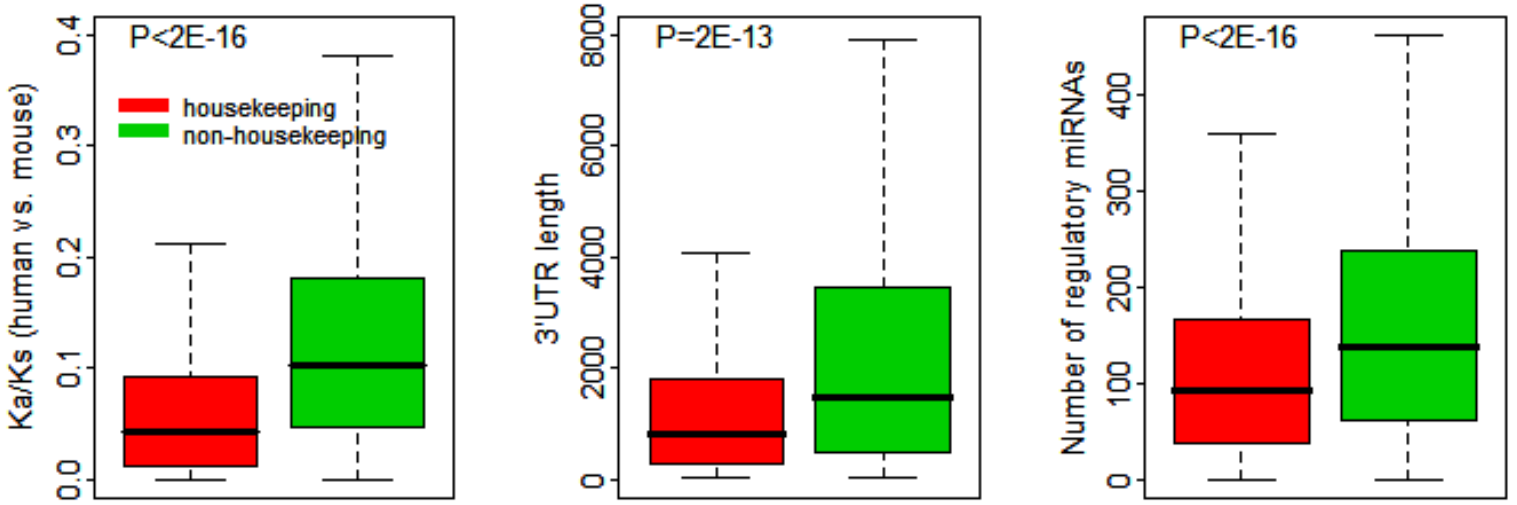
Important genes are regulated more intensively regulated by the microRNAs

[Cheng et al., BMC Genomics, 2009 (in press)]

MicroRNA regulation: a two-way strategy

For non-housekeeping genes, functionally critical genes are intensively regulated by miRNAs and prefer long 3'UTR.

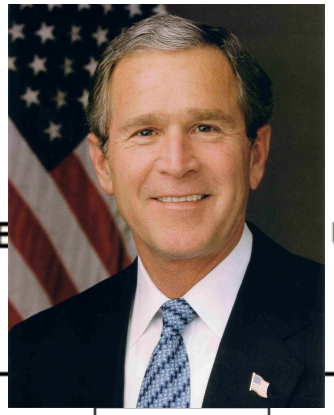
housekeeping genes, however conserved, are selected to have shorter 3'UTRs to avoid miRNA regulation.



[Cheng et al., BMC Genomics, 2009 (in press)]

Social Network Comparison #1 Comparing the Yeast Regulatory Network to a Governmental Hierarchy

Social Hierarchy



THE GOVE UNITED STATES

LEGISLATIVE BRANCH

THE CONGRESS

SENATE HOUSE

ARCHITECT OF THE CAPITOL
UNITED STATES BOTANIC GARDEN
GENERAL ACCOUNTING OFFICE
GOVERNMENT PRINTING OFFICE
LIBRARY OF CONGRESS
CONGRESSIONAL BUDGET OFFICE

EX SIDENT

WHITE HOUSE OFFICE
OFFICE OF THE VICE PRESIDE
COUNCIL OF ECONOMIC ADVIS
COUNCIL ON ENVIRONMENTAL
NATIONAL SECURITY COUNCIL
OFFICE OF ADMINISTRATION

NAGEMENT AND BUDGET
TIONAL DRUG CONTROL POLICY
LICY DEVELOPMENT
IENCE AND TECHNOLOGY POLICY
E U.S. TRADE REPRESENTATIVE

JUDICIAL BRANCH

THE SUPREME COURT OF THE UNITED STATES

UNITED STATES COURTS OF APPEALS
UNITED STATES DISTRICT COURTS
TERRITORIAL COURTS
UNITED STATES COURT OF INTERNATIONAL TRADE
UNITED STATES COURT OF FEDERAL CLAIMS
UNITED STATES COURT OF APPEALS FOR THE ARMED FORCES
UNITED STATES TAX COURT
UNITED STATES COURT OF APPEALS FOR VETERANS CLAIMS
ADMINISTRATIVE OFFICE OF THE UNITED STATES COURTS
FEDERAL JUDICIAL CENTER
UNITED STATES SENTENCING COMMISSION














INDEPENDENT ESTABLISHMENTS AND GOVERNMENT CORPORATIONS

AFRICAN DEVELOPMENT FOUNDATION
CENTRAL INTELLIGENCE AGENCY
COMMODITY FUTURES TRADING COMMISSION
CONSUMER PRODUCT SAFETY COMMISSION
CORPORATION FOR NATIONAL AND COMMUNITY SERVICE
DEFENSE NUCLEAR FACILITIES SAFETY BOARD
ENVIRONMENTAL PROTECTION AGENCY
EQUAL EMPLOYMENT OPPORTUNITY COMMISSION
EXPORT-IMPORT BANK OF THE U.S.
FARM CREDIT ADMINISTRATION
FEDERAL COMMUNICATIONS COMMISSION
FEDERAL DEPOSIT INSURANCE CORPORATION
FEDERAL ELECTION COMMISSION
FEDERAL HOUSING FINANCE BOARD

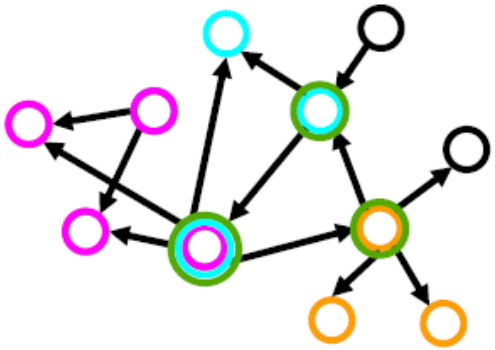
FEDERAL LABOR RELATIONS AUTHORITY
FEDERAL MARITIME COMMISSION
FEDERAL MEDIATION AND CONCILIATION SERVICE
FEDERAL MINE SAFETY AND HEALTH REVIEW COMMISSION
FEDERAL RESERVE SYSTEM
FEDERAL RETIREMENT THRIFT INVESTMENT BOARD
FEDERAL TRADE COMMISSION
GENERAL SERVICES ADMINISTRATION
INTER-AMERICAN FOUNDATION
MERIT SYSTEMS PROTECTION BOARD
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
NATIONAL ARCHIVES AND RECORDS ADMINISTRATION
NATIONAL CAPITAL PLANNING COMMISSION
NATIONAL CREDIT UNION ADMINISTRATION

NATIONAL FOUNDATION ON THE ARTS AND THE HUMANITIES
NATIONAL LABOR RELATIONS BOARD
NATIONAL MEDIATION BOARD
NATIONAL RAILROAD PASSENGER CORPORATION (AMTRAK)
NATIONAL SCIENCE FOUNDATION
NATIONAL TRANSPORTATION SAFETY BOARD
NUCLEAR REGULATORY COMMISSION
OCCUPATIONAL SAFETY AND HEALTH REVIEW COMMISSION
OFFICE OF GOVERNMENT ETHICS
OFFICE OF PERSONNEL MANAGEMENT
OFFICE OF SPECIAL COUNSEL
OVERSEAS PRIVATE INVESTMENT CORPORATION
PEACE CORPS
PENSION BENEFIT GUARANTY CORPORATION

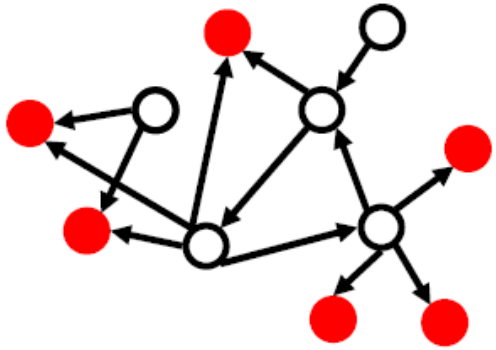
POSTAL RATE COMMISSION
RAILROAD RETIREMENT BOARD
SECURITIES AND EXCHANGE COMMISSION
SELECTIVE SERVICE SYSTEM
SMALL BUSINESS ADMINISTRATION
SOCIAL SECURITY ADMINISTRATION
TENNESSEE VALLEY AUTHORITY
TRADE AND DEVELOPMENT AGENCY
U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT
U.S. COMMISSION ON CIVIL RIGHTS
U.S. INTERNATIONAL TRADE COMMISSION
U.S. POSTAL SERVICE

Determination of "Level" in Regulatory Network Hierarchy with Breadth-first Search

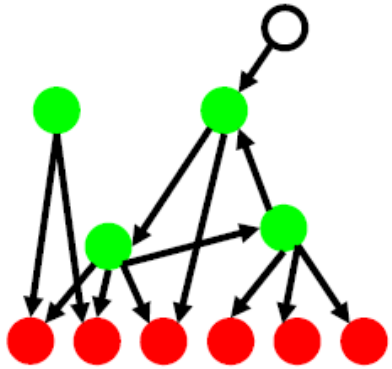
I. Example network with all 4 motifs



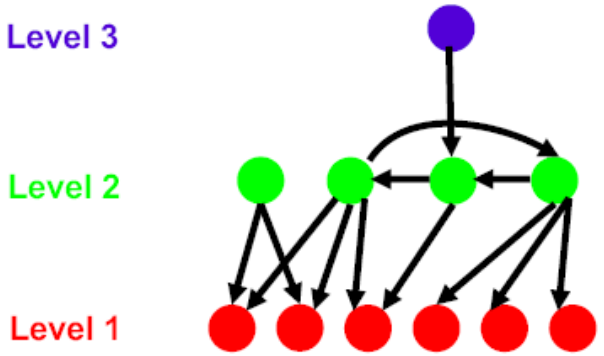
II. Finding terminal nodes (Red)



III. Finding mid-level nodes (Green)

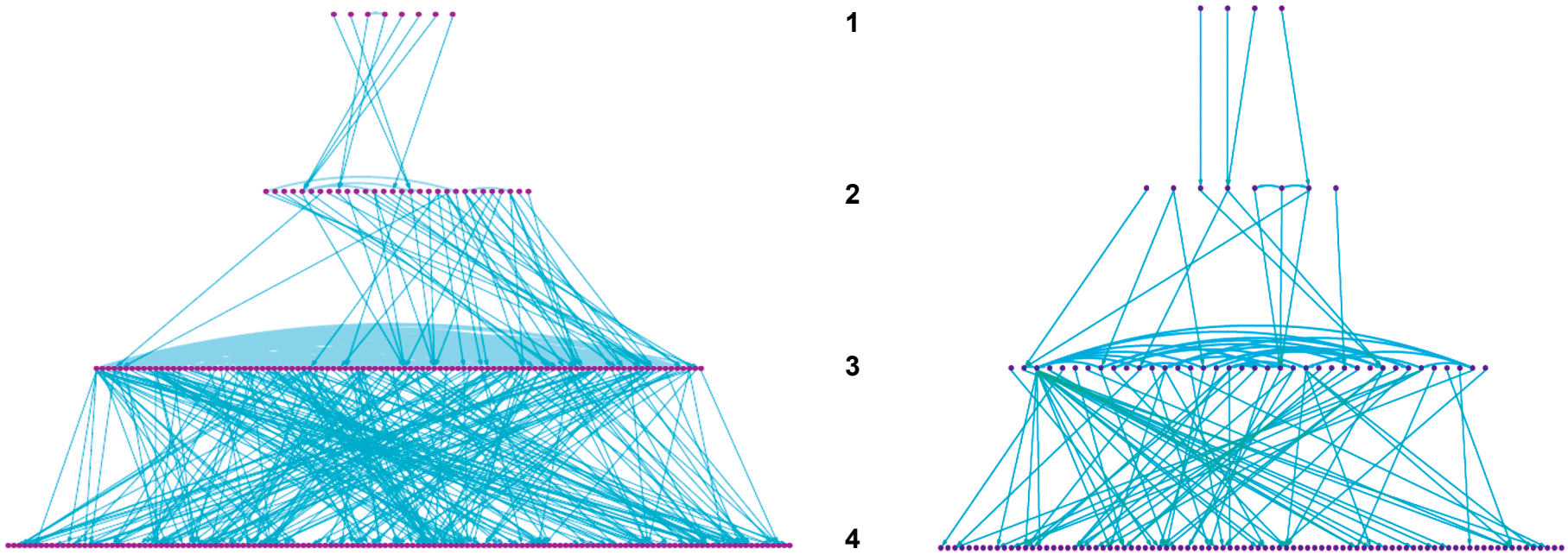


IV. Finding top-most nodes (Blue)



[Yu et al., PNAS (2006)]

Regulatory Networks have similar hierarchical structures

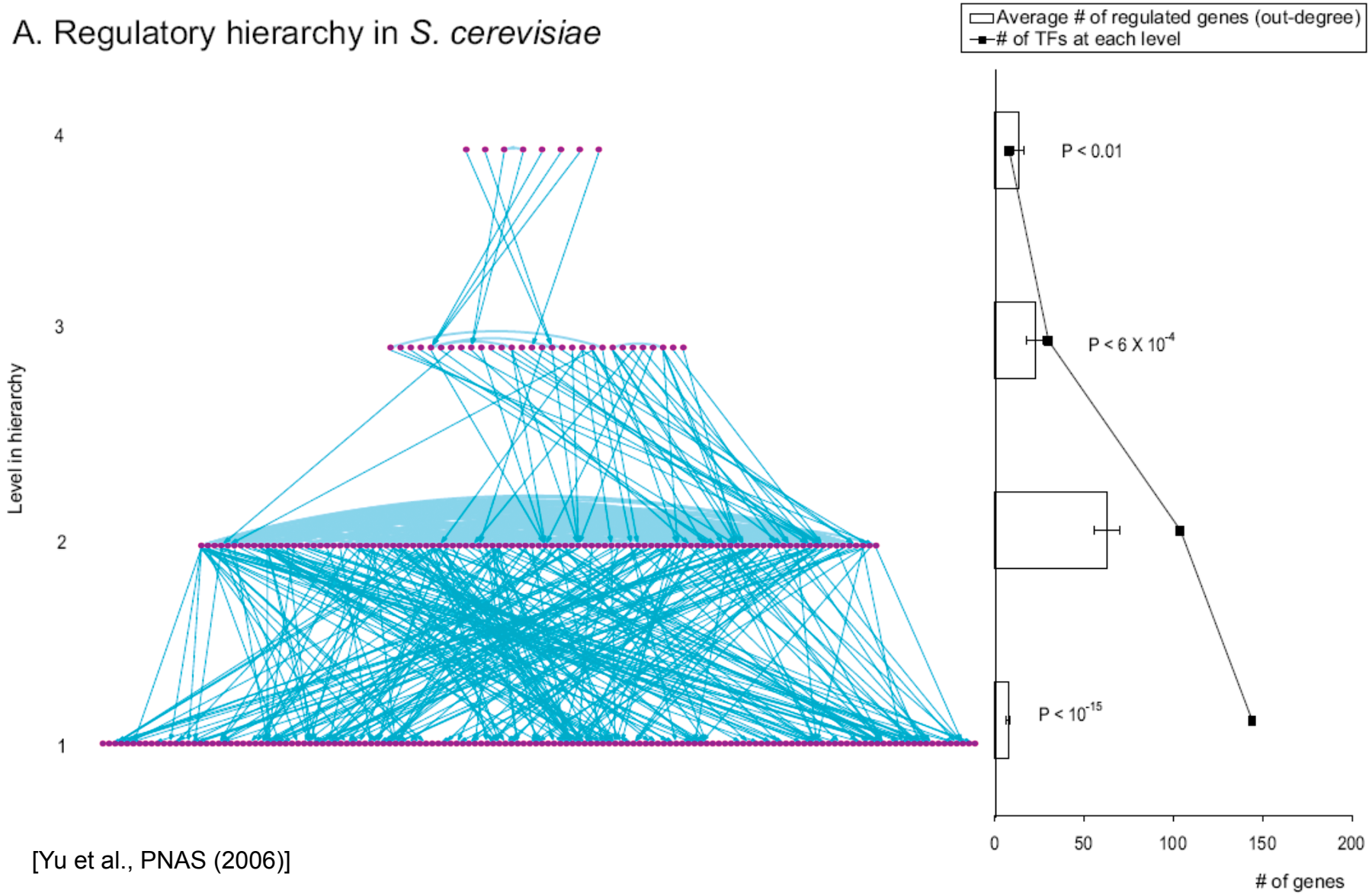


S. cerevisiae

E. coli

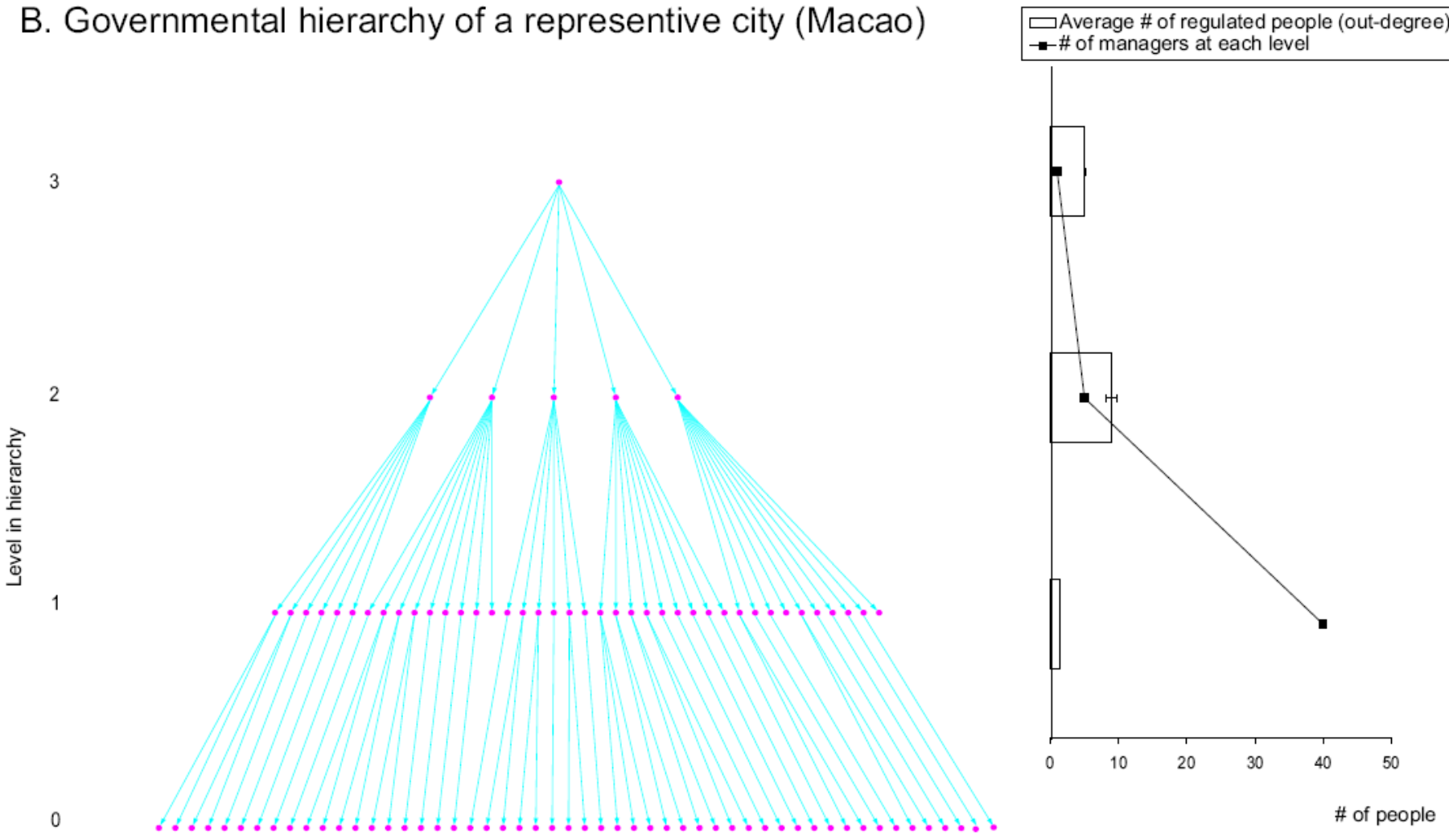
[Yu *et al.*, *Proc Natl Acad Sci U S A* (2006)]

Yeast Regulatory Hierarchy: the Middle-managers Rule

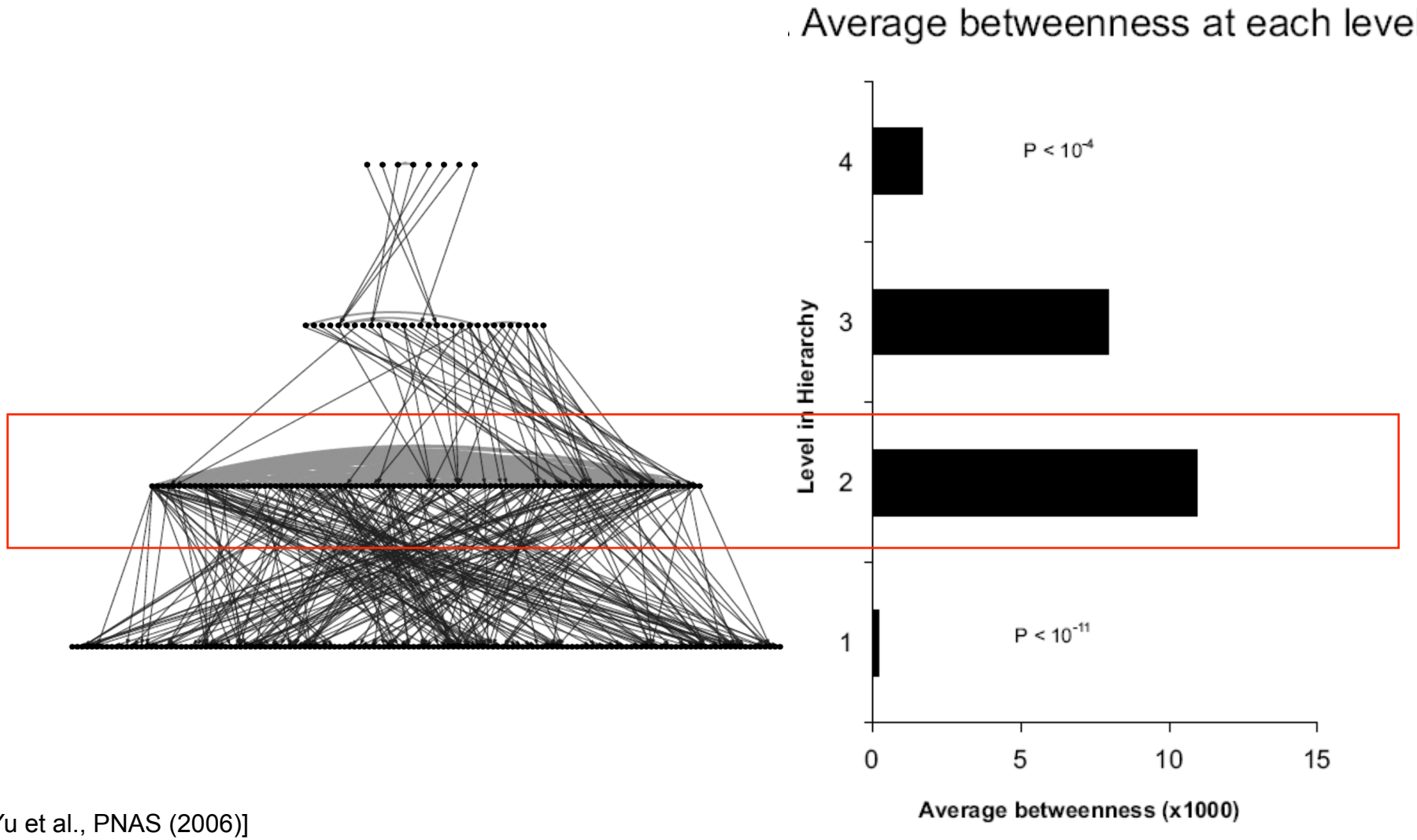


[Yu et al., PNAS (2006)]

Yeast Network Similar in Structure to Government Hierarchy with Respect to Middle-managers



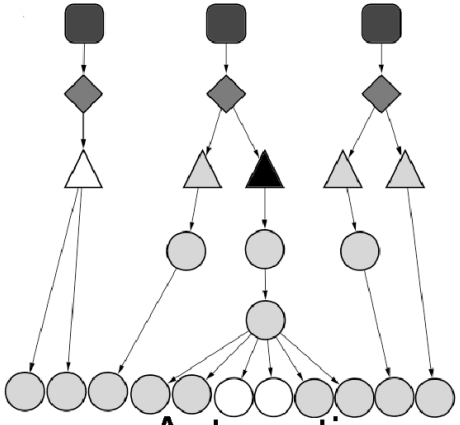
Characteristics of Regulatory Hierarchy: Middle Managers are Information Flow Bottlenecks



[Yu et al., PNAS (2006)]

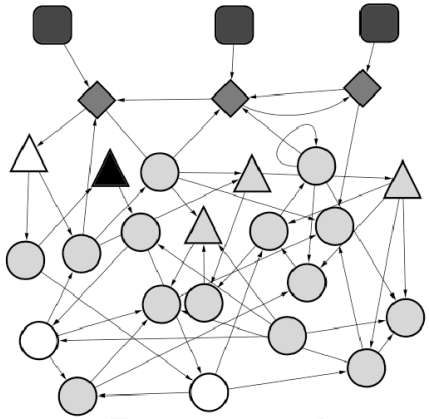
Social Network Comparison #2
Broadening the comparison to different types of hierarchies & different types of biological networks

Different kinds of Hierarchies



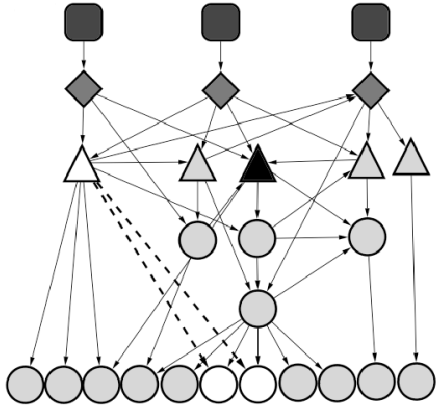
Autocratic

- Well-defined levels and a clear chain of command
- A military hierarchy



Democratic

- Without well-defined levels & with more co-regulatory partnerships
- A club or a scientific collaboration network



Intermediate

- High degree of co-regulation and can be organized into hierarchies
- A law firm

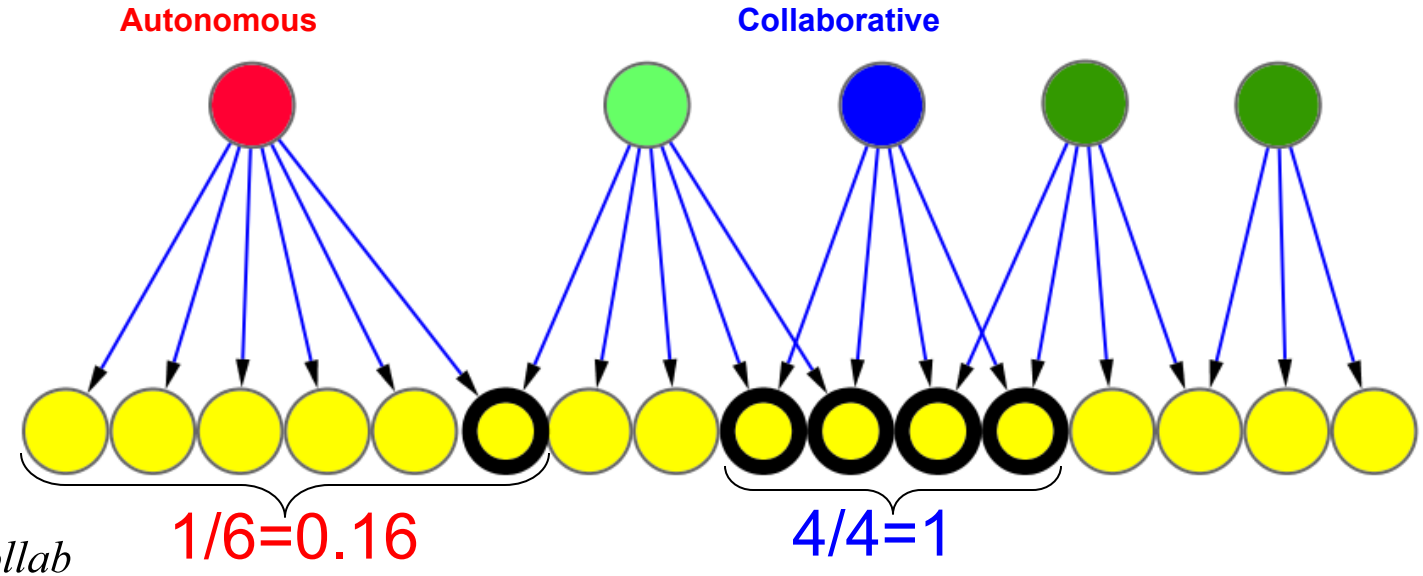
	Autocratic	Democratic	Intermediate
Betweenness (△)	1.03	3.6	3.3
Betweenness (▲)	4.1	1.08	3.4
Var. Betw. (triangles)	2.1	0.58	1.74
Var. Betw. (all)	2.9	1.4	1.9
$D_{Net\text{-}collab}$	0	0.91	0.71

Decreases σ^2 of 'stress'

Increases 'collaboration'

[Bhardwaj et al., PNAS (2010), in press]

Collaborative Nature of the Nodes

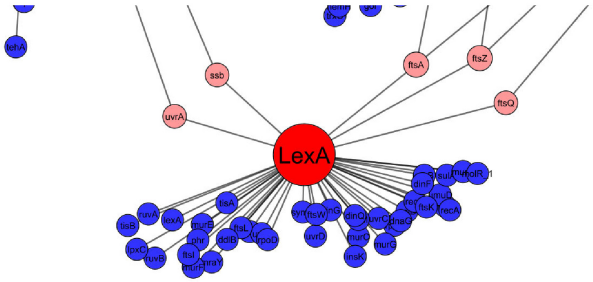


Degree of Collaboration D^i_{collab}

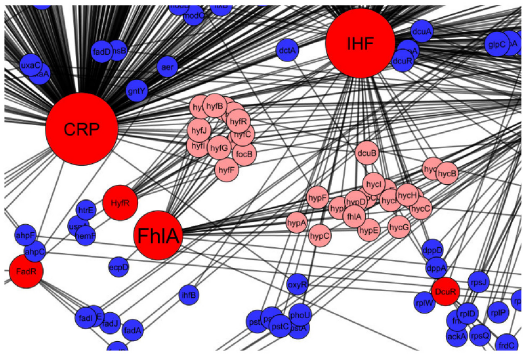
More Collaborative: Democratic

More Autonomous: Autocratic

Autonomous

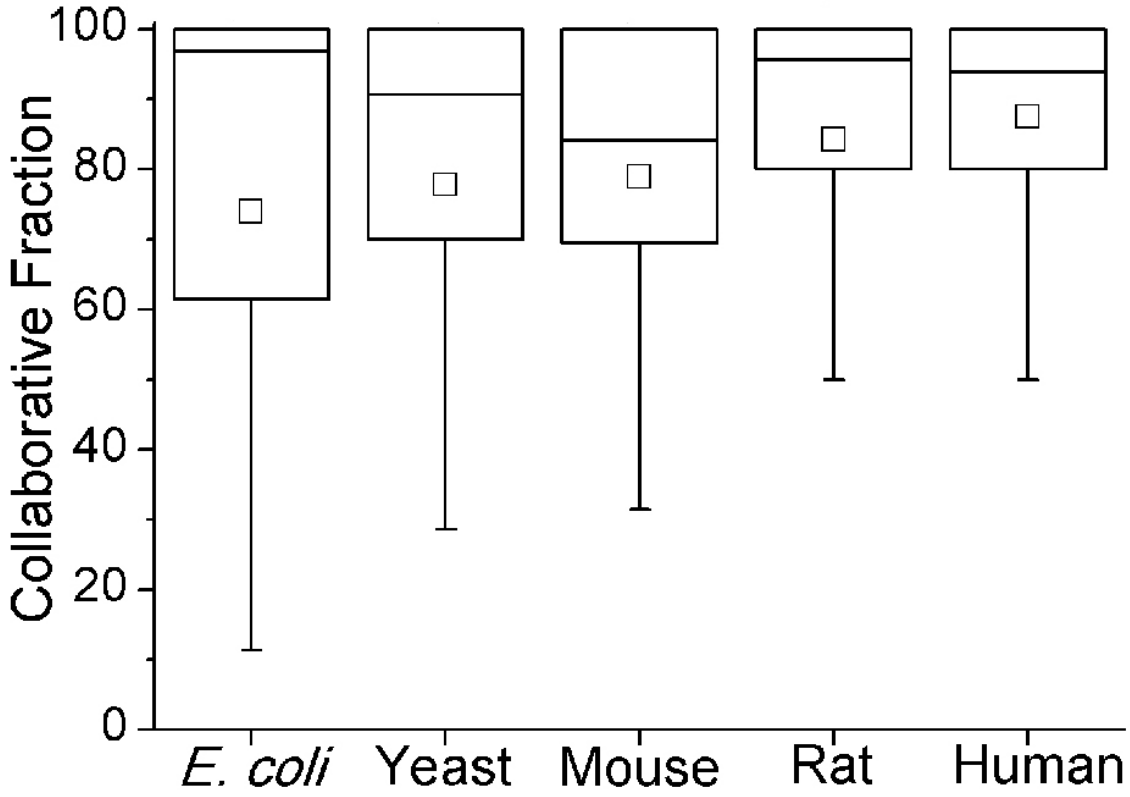


Collaborative



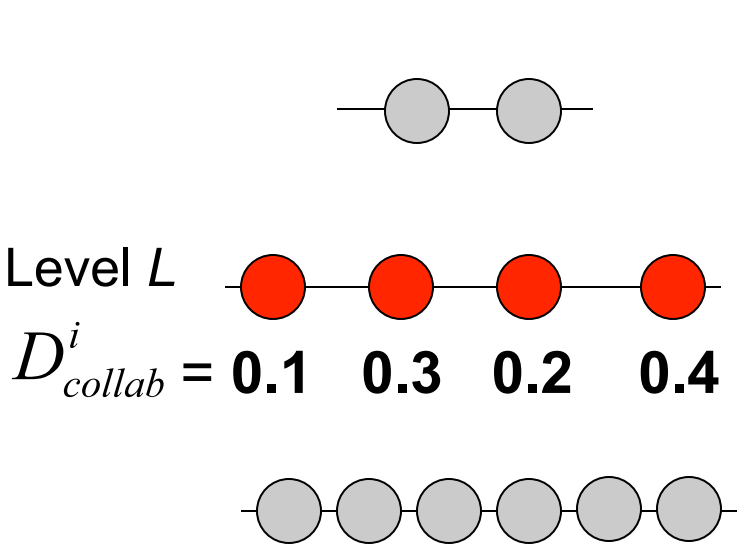
[Bhardwaj et al., PNAS (2010), in press]

Higher species are more show more collaborative nodes (more democratic)



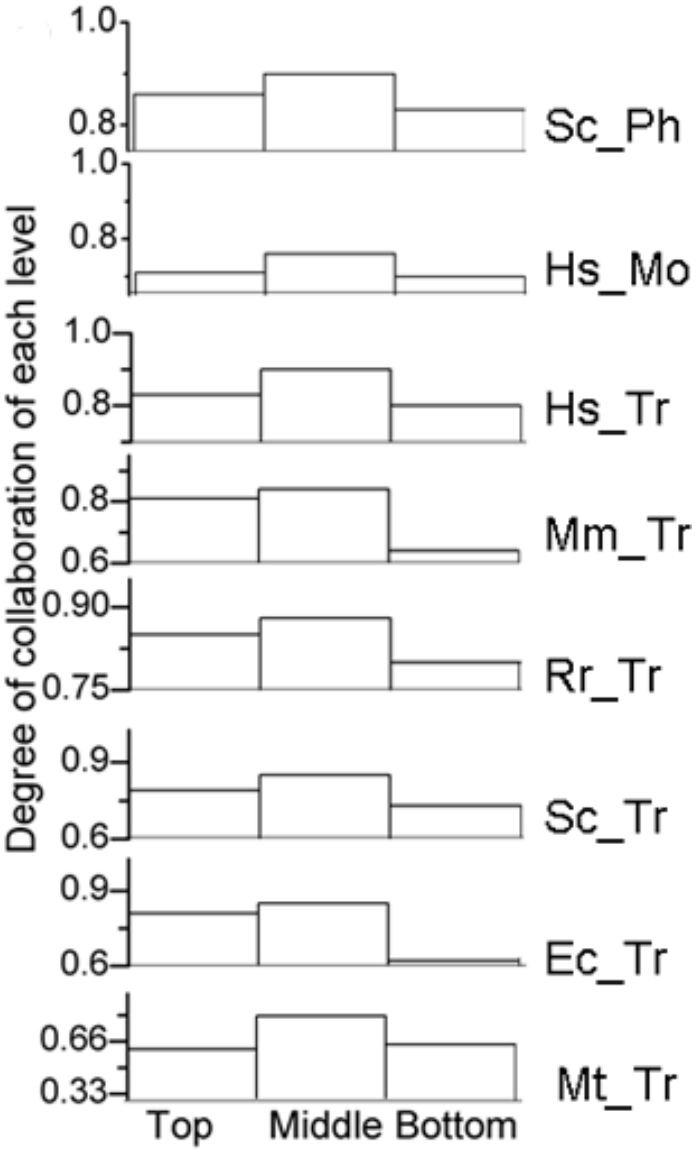
[Bhardwaj et al., PNAS (2010), in press]

Collaborative Nature of the Levels

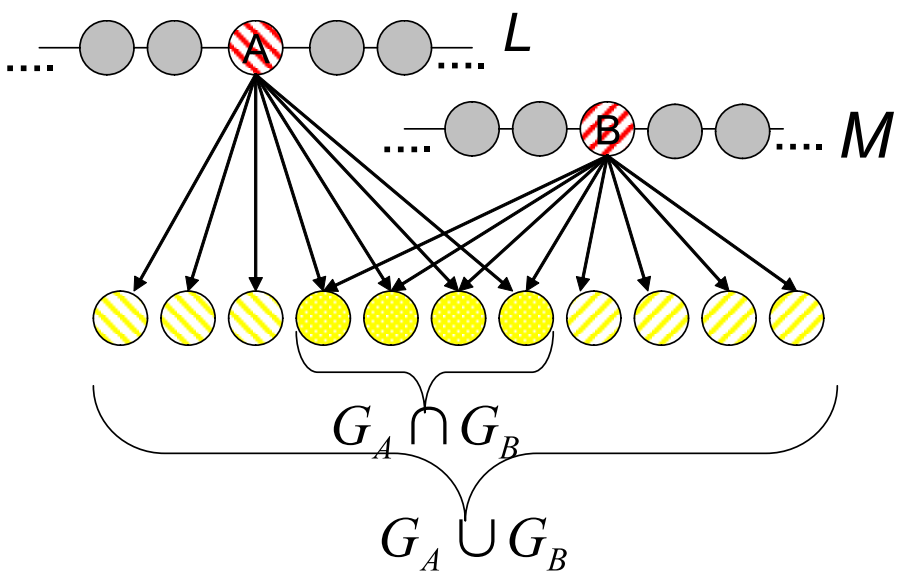


$$D^{L}_{Level-collab} = \frac{0.1+0.3+0.2+0.4}{4} = 0.25$$

[Bhardwaj et al., PNAS (2010), in press]

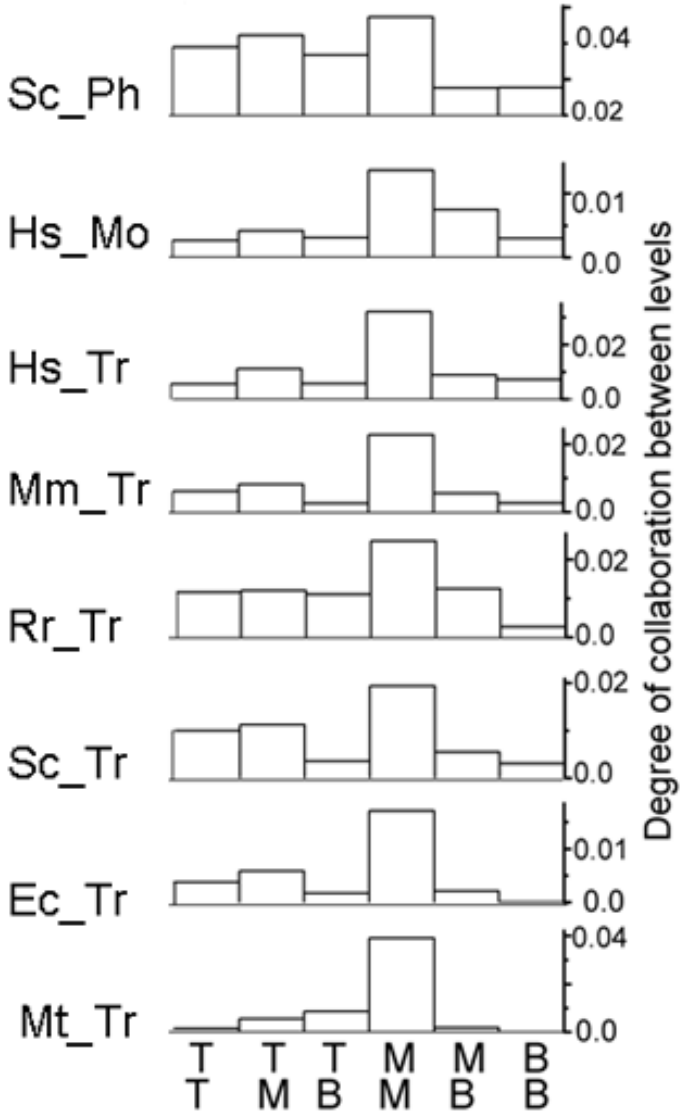


Collaboration Between Levels



$$D^{L,M}_{betw-level-collab} = \frac{\sum_{A \in L} \sum_{B \in M} \frac{G_A \cap G_B}{G_A \cup G_B}}{|L| \bullet |M|}$$

[Bhardwaj et al., PNAS (2010), in press]



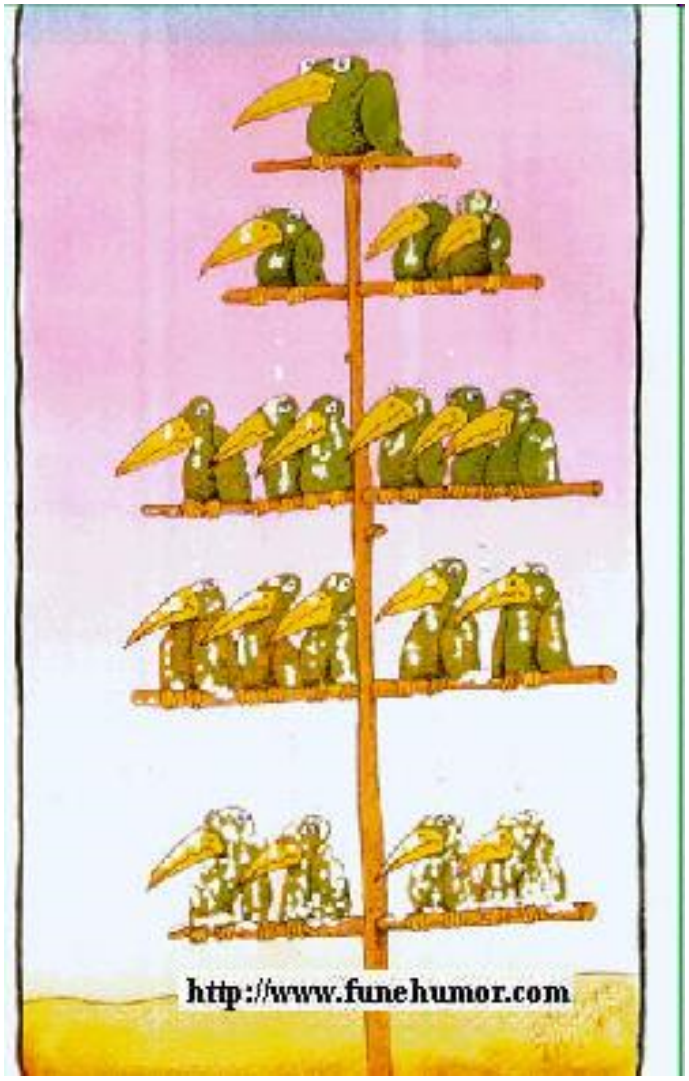
Middle Managers Interact the Most in Efficient Corporate Settings

- Floyd, S. W. et al (1992)
  Middle management involvement in strategy and its association with strategic type
  Strategic Management Journal 13, 153-167.

- Woodward, J. (1982) Industrial Organization: Theory and Practice (Oxford University Press, Oxford).

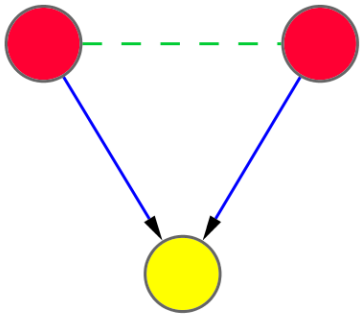
- Floyd, S. W. et al (1993)
  Dinosaurs or Dynamos? Recognizing Middle Management's Strategic Role
  The Academy of Management Executive 8, 47-57.

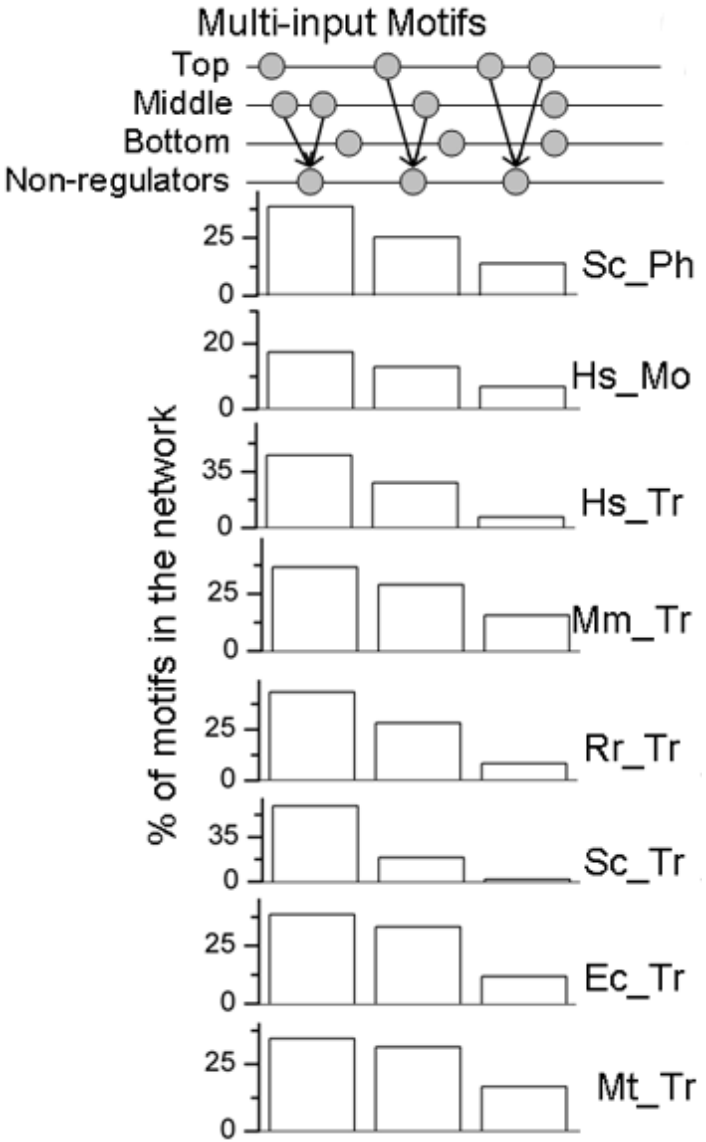
- Floyd, S. W. et al (1997)
  Middle management's strategic influence and organizational performance
  Journal of Management Studies 34, 465-485.



[Bhardwaj et al., PNAS (2010), in press]

Co-regulation Instantiates a Multi-Input Motif

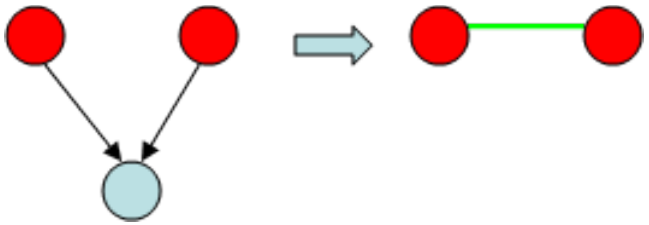




[Bhardwaj et al., PNAS (2010), in press]

Network Comparisons #3
Relating the size of co-regulation in partnership networks with the scale of the regulated

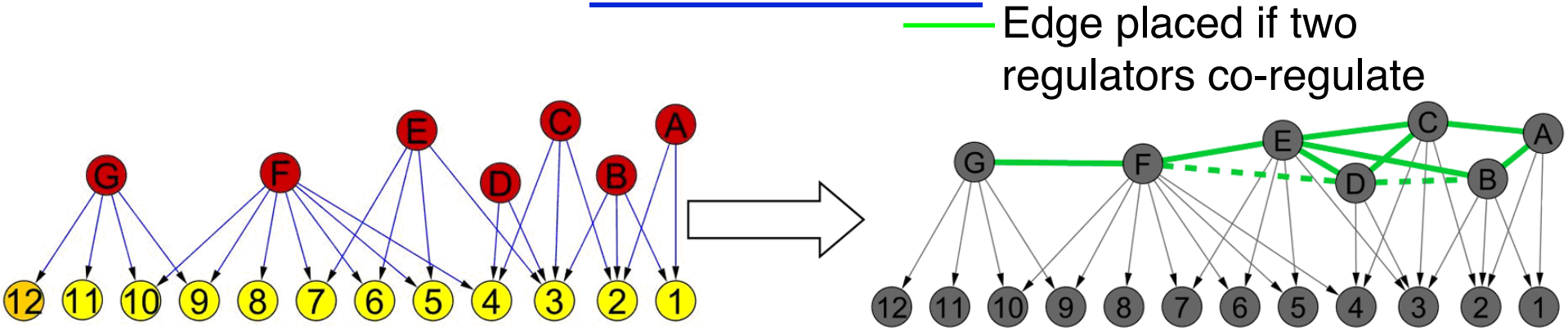
Co-regulation Partnerships



- Readily seen in many commonplace social contexts.
- An academic institution (say a high school), multiple teachers supervise the same set of students and have partnership interactions amongst themselves.

[Bhardwaj et al., PLoS Comp Biol (2010), in press]

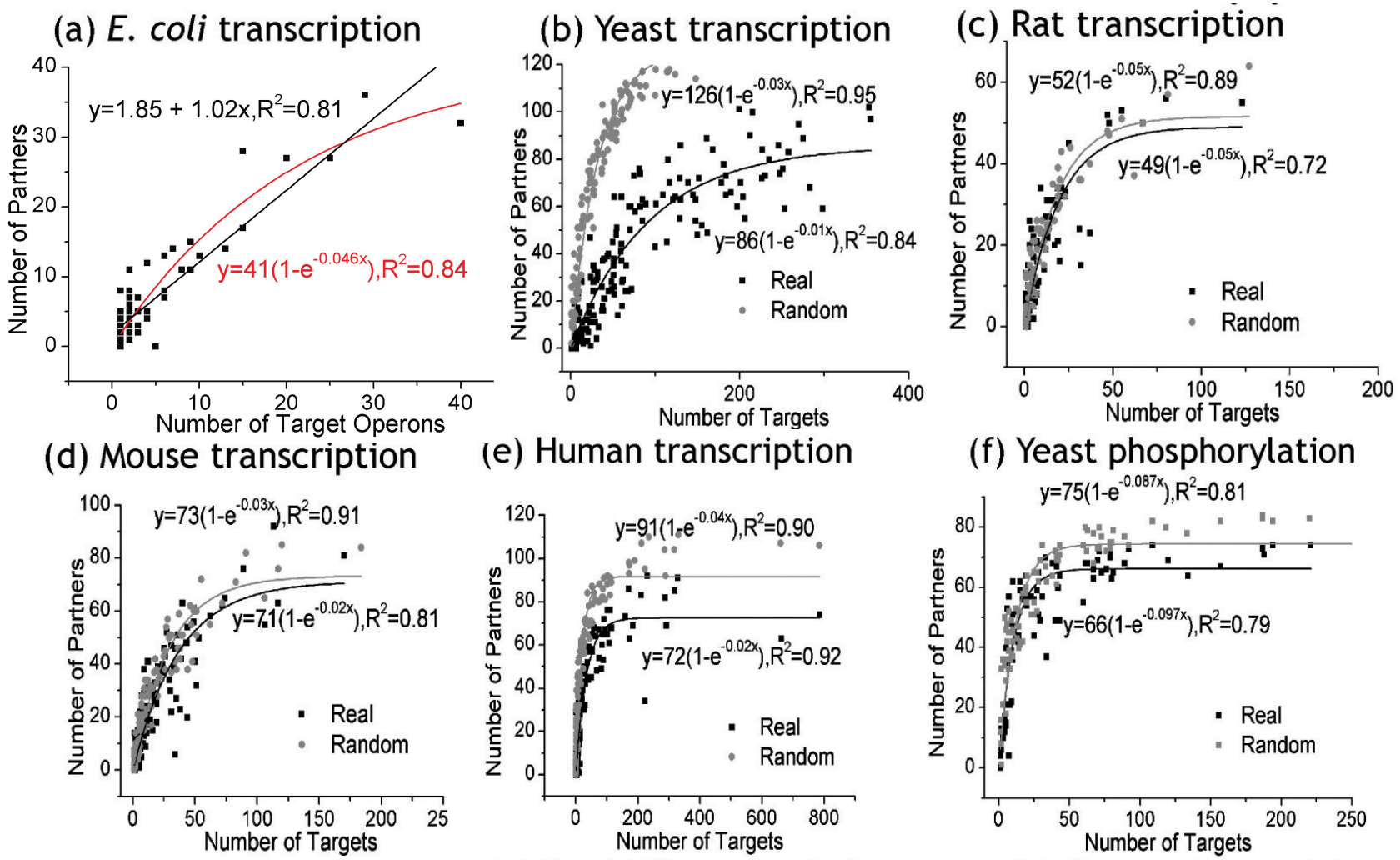
Building and Analysis of Networks



Network type	Species	Number of regulators	Number of targets	Number of interactions
Transcription	*E. coli*	160	1,420	3,123
Transcription	Yeast	157	4,410	12,873
Transcription	Mouse	144	1,092	2,403
Transcription	Rat	91	461	1,092
Transcription	Human	156	3,032	6,896
Phosphorylation	Yeast	87	1,337	4,083
Modification	Human	518	1,218	2,782

[Bhardwaj et al., PLoS Comp Biol (2010), in press]

Scaling of Regulators with Targets

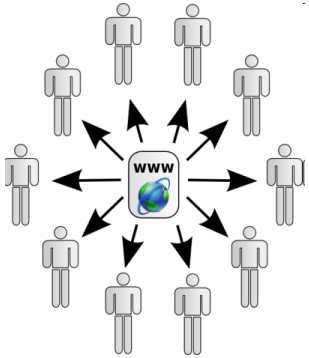


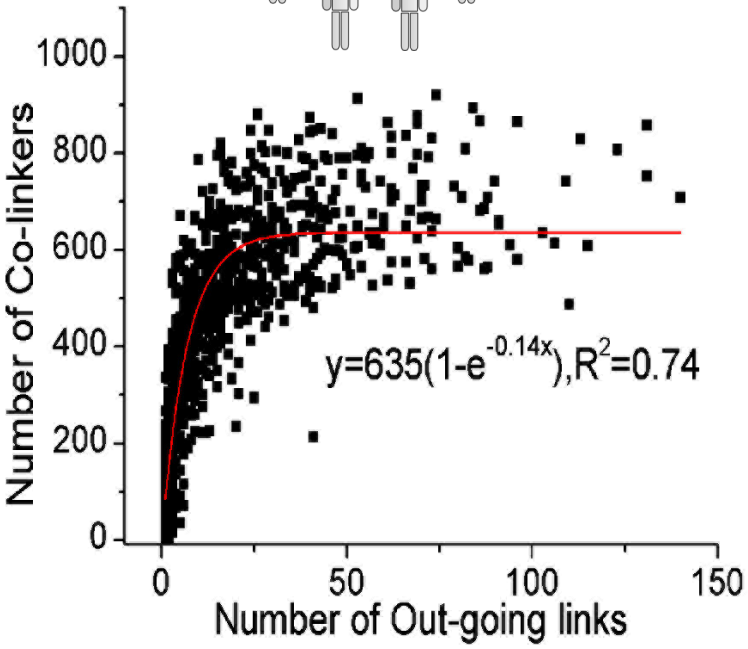
Linear in *E. coli* (Due to operons)
Exponential Saturation in others

[Bhardwaj et al., PLoS Comp Biol (2010), in press]

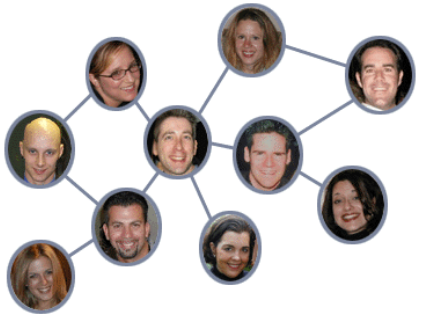
Comparison to Social Networks: Partnership networks effectively saturate with increasingly complex output

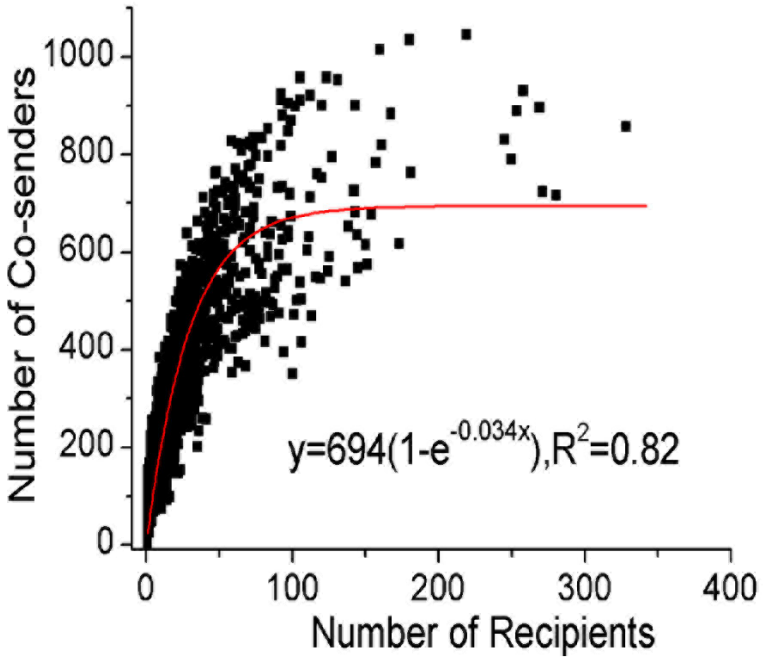
Blogs Network





Email Network



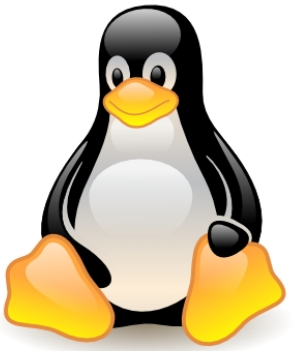


[Bhardwaj et al., PLoS Comp Biol (2010), in press]

Software Network Comparison
Comparing the structure and evolution of biological regulatory networks and software call graphs

E. Coli Transcriptional regulatory network vs Linux kernel call graph

		E. coli transcriptional regulatory network	Linux call graph
Basic properties of systems	Nodes	Genes (TFs & targets)	Functions (subroutines)
	Edges	Transcriptional regulation	Function calls
	External constraints	Natural environment	Hardware architecture, customer requirements
	Origin of evolutionary changes	Random mutation & natural selection	Designers' fine tuning



	E. coli transcriptional regulatory network	Linux call graph
Number of nodes	1378	12391
Number of persistent nodes	72* (5%)	5120 (41%)
Number of edges	2967	33553
Number of modules	64	3665
Number of comparative references	200 bacterial genomes	24 versions of kernels
Years of evolution	Billions years	20 years

[Yan et al., PNAS (2010), in press]

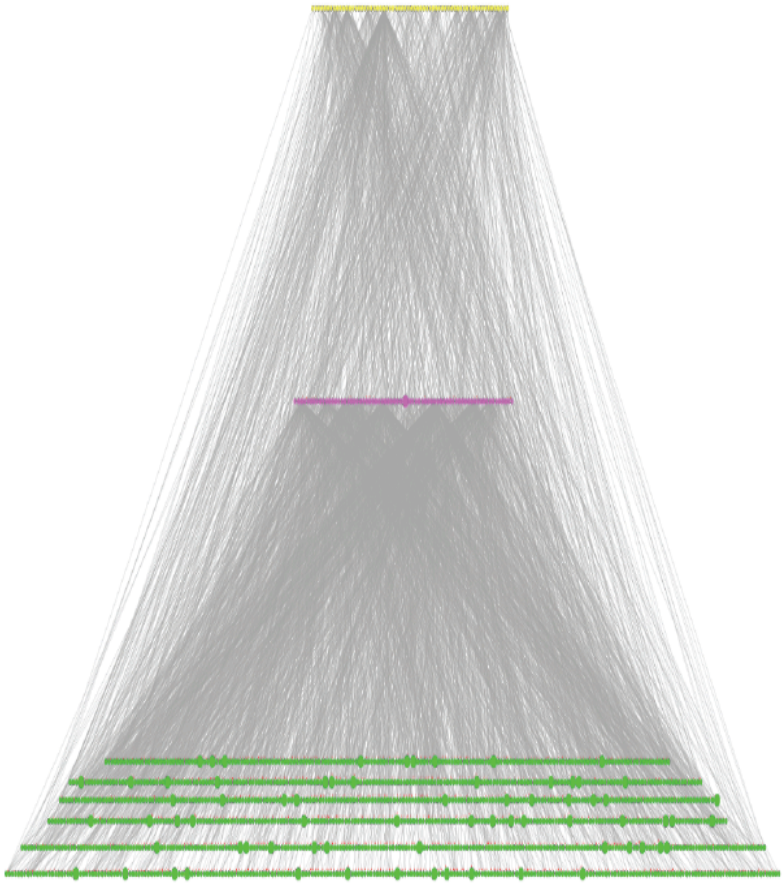
E. coli transcriptional regulatory network

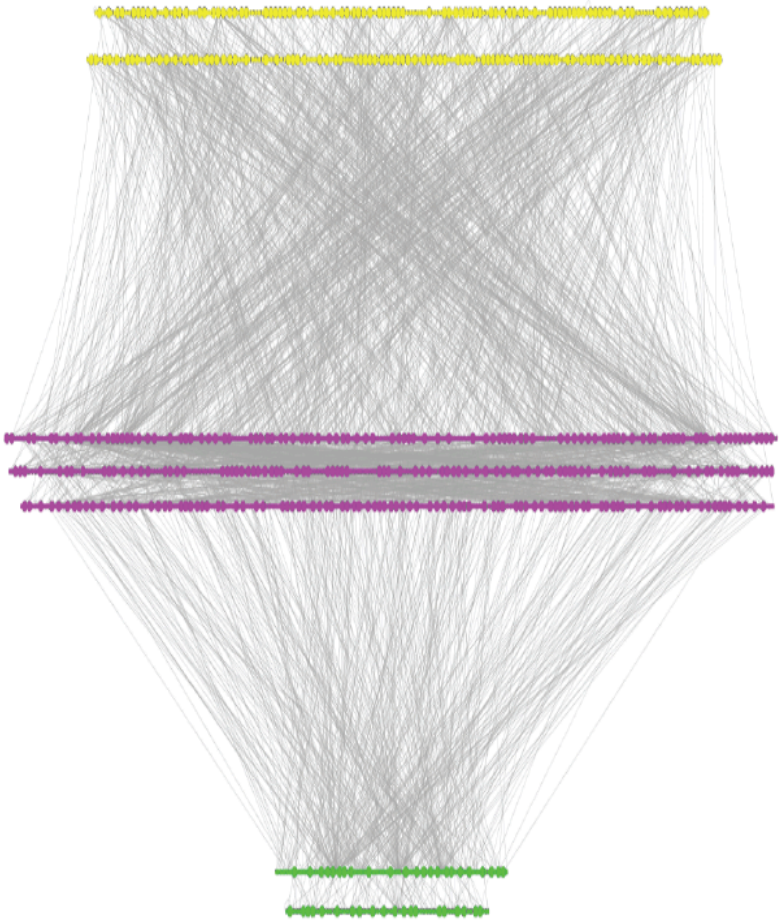
Linux call graph

master regulator

middle manager

workhorse

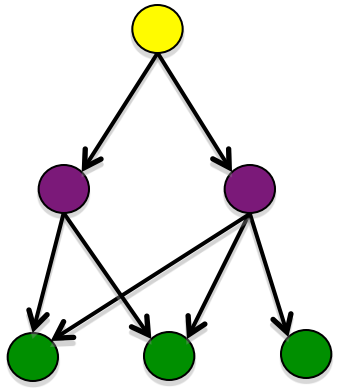




[Yan et al., PNAS (2010), in press]

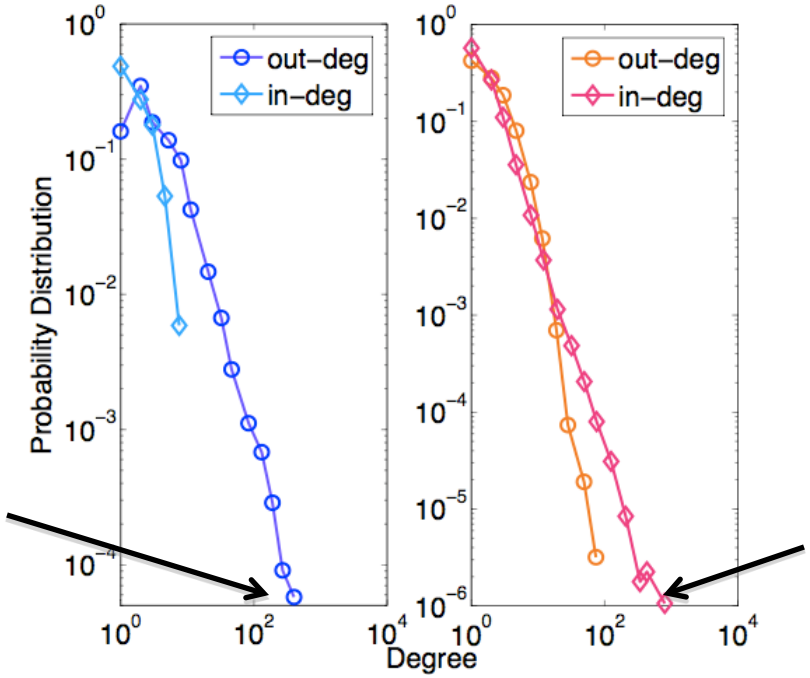
Comparison: hierarchical organization

Pyramidal vs
Top-heavy



	% in *E. coli* regulatory network	% in Linux call graph
master regulator	4.6	29.6
middle manager	5.1	58.2
workhorse	90.2	12.3

Degree distribution
Roles of hubs



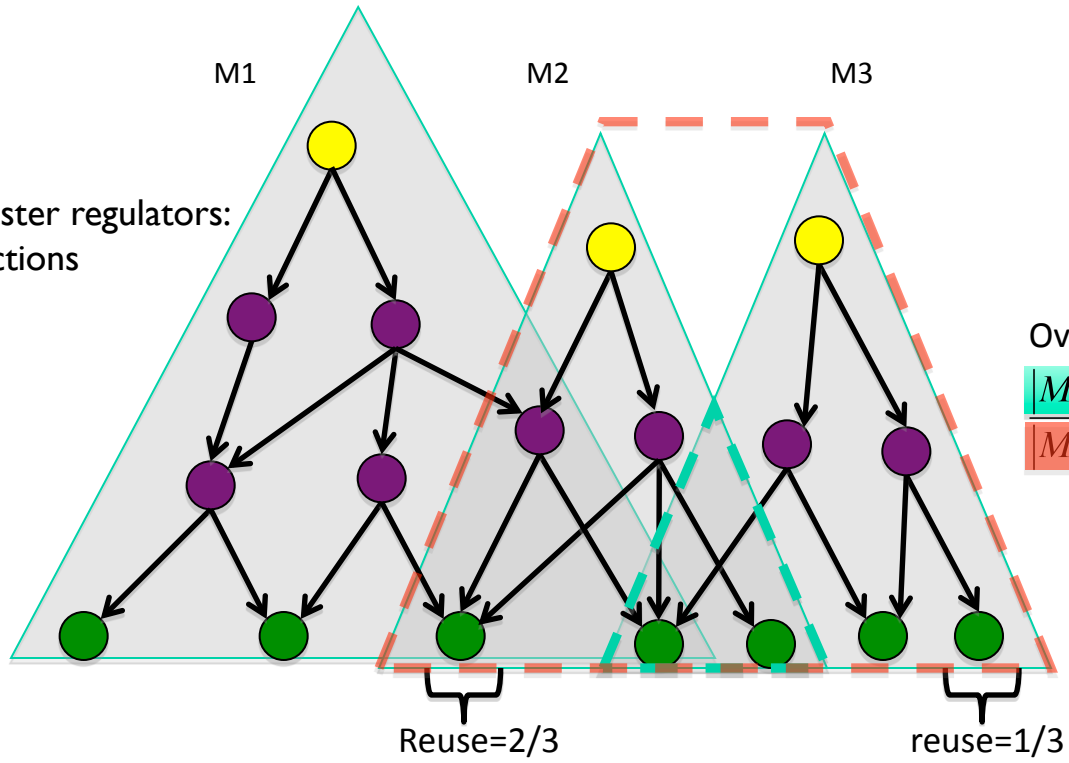
out-degree hubs
e.g. “crp”

in-degree hubs
e.g. “printk”

[Yan et al., PNAS (2010), in press]

Comparison: organization of modules

Modules are labeled by master regulators: TFs, high-level starting functions



Overlap(M2,M3)=

$$\frac{|M2 \cap M3|}{|M2 \cup M3|} = \frac{2}{11}$$

	E. Coli TRN	Linux call graph
# of modules	64	3665
Average overlap	4.3%	80.7%
Maximum node reuse	15.6%	87.5%
Average node reuse	3.5%	8.4%

TRN: modules overlap little, components are less generic: "ompF"

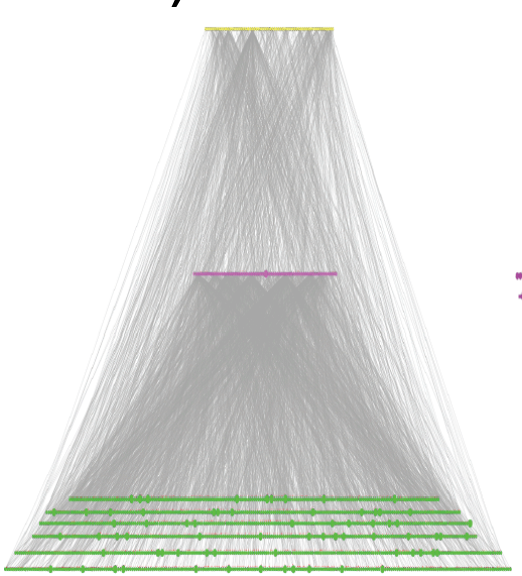
Call graph: modules overlap, Functions are highly reused (generic): "printk"

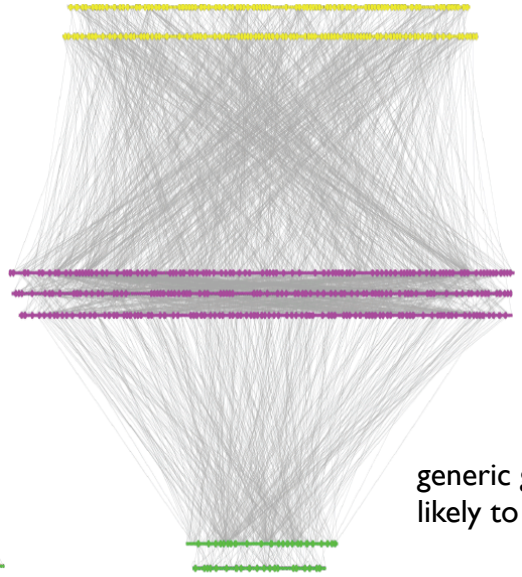
[Yan et al., PNAS (2010), in press]

Comparison of persistent components

- Persistent genes (preserve among different genomes) vs persistent functions (preserve among different releases)

specialized proteins
are preserved
across genomes



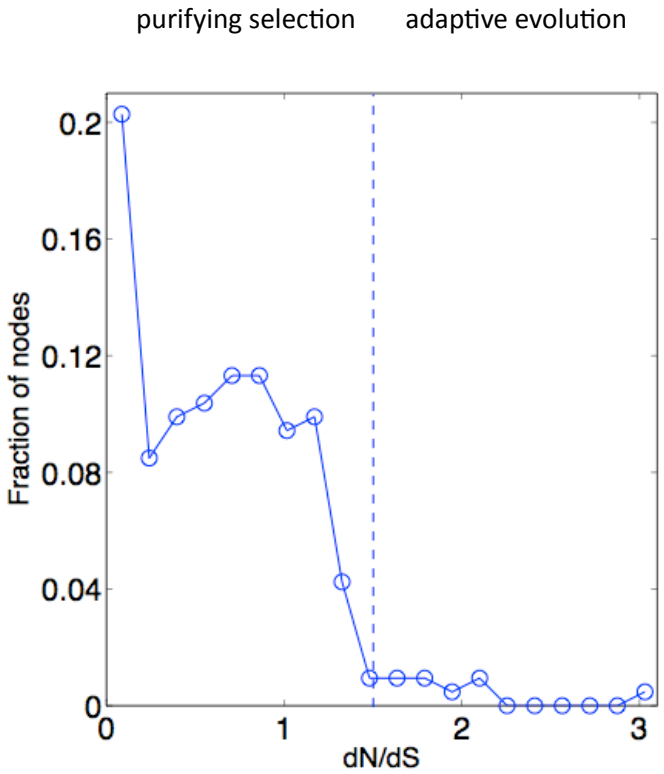


generic guys are more
likely to be persistent

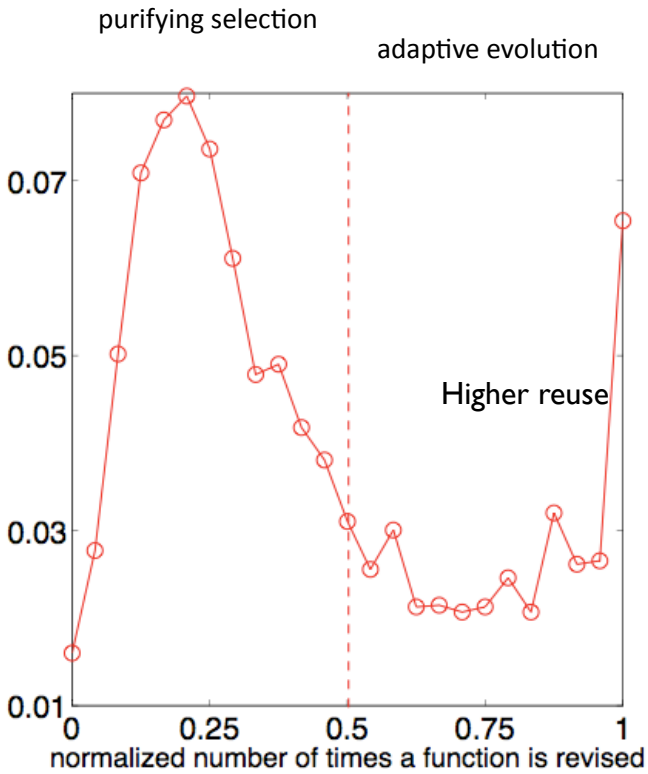
- Building of the hierarchy:
  - ◊ TRN: Bottom up. Regulatory changes are the main driving forces of evolution
  - ◊ Call graph: top down

[Yan et al., PNAS (2010), in press]

Evolutionary rate of persistent functions



Persistent genes evolve slowly

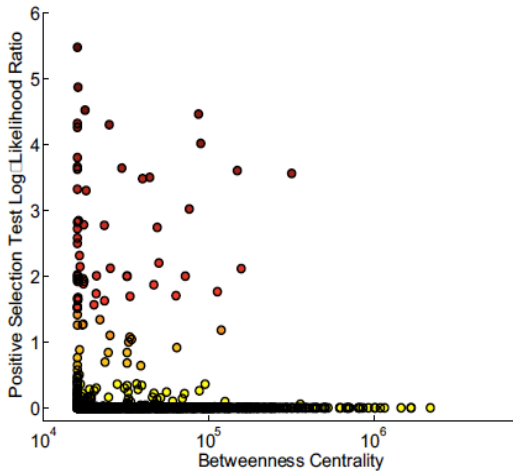


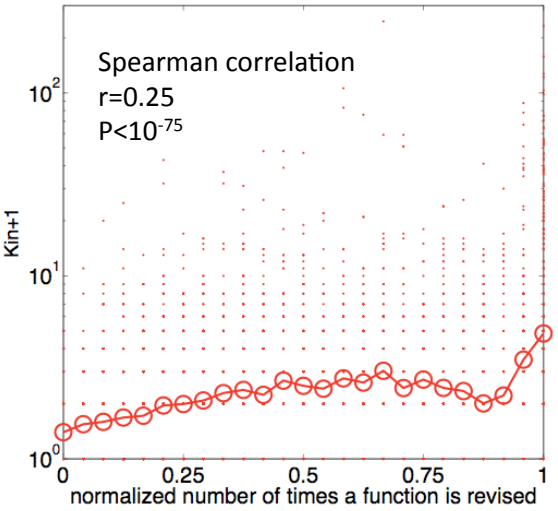
Two classes of persistent functions

[Yan et al., PNAS (2010), in press]

Why and so what?

The difference can be explained by the nature of hubs evolution: tinkering vs design





Kim et.al. PNAS 2007

- Independent modules:
  - robust
  - costly: the system needs a variety of tools for different tasks
- Overlap modules (reuse):
  - Less robust:
    - Breakdown of a generic component is harmful to the whole system
    - Fragile in the sense any change in a module may require compensating changes in a generic function
  - cost effective: components can be used by need to be fine-tuned

[Yan et al., PNAS (2010), in press]

Outline: Molecular Networks

- Why Networks?
- Generating Networks
  - Scanning for Targets of Modular Domains
  - Propagating Known Information
    (yeast ppi)
- Central Network Points
  - Hubs & Bottlenecks
    (yeast ppi & reg. net)
- Networks & Variation
  (human ppi & miRNA-targ. net)
- Social Network Comparisons
  (reg. net. in many organisms)
  - in rel. to social hierarchy
  - scaling in rel. to partnerships
- Computer OS Comparisons
  (E. coli reg. net)

Conclusions on Networks: Generation

- Predicting Networks
  - ◊ Scanning for sequence motifs recognized by modular protein domains (motips)
  - ◊ Extrapolating from the Training Set
  - ◊ Principled ways of using known information in the fullest possible fashion
    - Multi-level learning

Conclusions:
Analysis of Network Structure

- Centrality Measures in Protein Network
  - ◊ Hubs & Bottlenecks
  - ◊ Importance of later in regulatory networks

Conclusions: Connecting Networks & Variation

- Positive selection (adaptive evolution) at the network periphery
  - ◊ On a sequence level, it can be seen as positive selection of peripheral nodes
  - ◊ On a structural level, it can be seen as the pattern of SDs that display significantly higher allele frequencies in non-central genes
- miRNA network
  - ◊ More highly regulated genes are under more constraint in miRNA-target networks
  - ◊ Exception for housekeeping genes

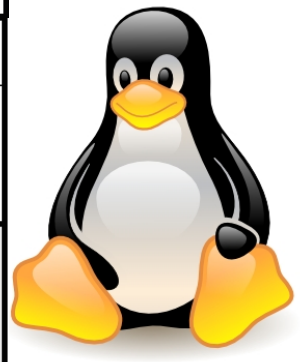
Conclusions: Comparison to Social Hierarchies

- Regulatory Network Hierarchies
  - ◊ Middle managers dominate, sitting at info. flow bottlenecks
  - ◊ Paradox of influence & essentiality

Conclusions: Comparison to Social and Regulatory Hierarchies

- Regulatory Network Hierarchies
  - ◊ Democratic v Autocratic
  - ◊ Collaborative (locally democratic) fraction of networks increases with organism complexity
  - ◊ Middle managers most collaborative
  - ◊ Most interaction occur between two middle managers (as seen in efficient corporate hierarchies)
- Number of collaborative partners saturates even while scale of targets governed increases
  - ◊ Also seen in social networks

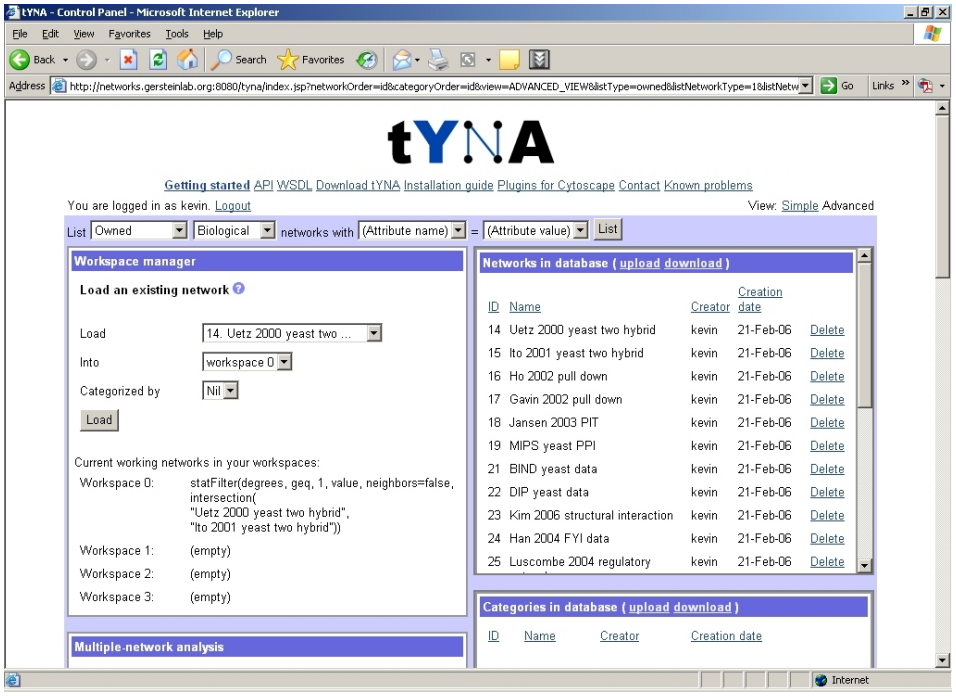
		E. coli transcriptional regulatory network	Linux call graph
Hierarchical organization	Structure	Pyramidal	Top-heavy
	Characteristic hubs	Upper-level TFs with high out-degree	Generic workhorse functions with high in-degree
Organization of modules	Downstream modules as labeled by	Master TFs responsible for sensing environmental signals	High-level starting functions which initiate execution for specific tasks
	Node reuse	Low	High
	Overlap between modules	Low	High
Persistent nodes	Characteristics	Specialized (non-generic) workhorses	Generic or reusable functions
	Location in hierarchy	Mostly bottom	Mostly top
	Evolutionary rate	Mostly conservative (e.g. dnaA)	Conservative (e.g. strlen) & adaptive (e.g. mempool_alloc)
Design principles	Building of hierarchy	Bottom up	Top down
	Optimal solution favors	Robustness	Cost-effectiveness (reuse of components)

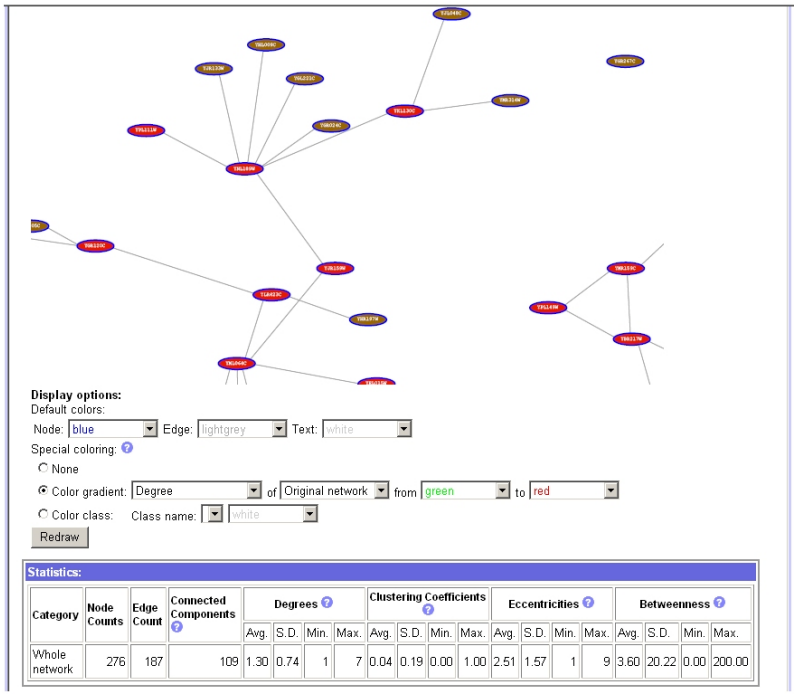


- an automated web tool

tYNA

(vers. 2 : "TopNet-like Yale Network Analyzer")





Normal website + Downloaded code (JAVA)
+ Web service (SOAP) with Cytoscape plugin

[Yu et al., NAR (2004); Yip et al. Bioinfo. (2006);
Similar tools include Cytoscape.org, Idekar, Sander et al]

Acknowledgements

H Yu
P Kim
K Yip
C Cheng
N Bhardwaj
K-K Yan
H Lam
R Alexander
G Fang
M Seringhaus
Y Xia
J Korbel
E Franzosa
B Turk
J Mok
M Snyder

Networks.GersteinLab.org

Job opportunities currently for postdocs & students

More Information on this Talk

SUBJECT: Networks

DESCRIPTION:
Brown Applied Math, Providence, RI; 2010.04.09, 16:00-17:00; [**I:BROWNMATH**] (Long networks talk, derived from [I:**MBINETS**], including **callgraph***, **coregscaling***, **reghier***, **& motips*** for 1st time. Whole talk took 2 hrs. with questions.)

(PPT works on mac & PC and has many photos. Paper references in the talk were mostly from Papers.GersteinLab.org. The above topic list can be easily cross-referenced against this website. Each topic abbrev. which is starred is actually a papers "ID" on the site. For instance, the topic **pubnet*** can be looked up at **http://papers.gersteinlab.org/papers/pubnet**)

.

PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as **kwpotppt** , that can be easily queried from flickr, viz: **http://www.flickr.com/photos/mbgmbg/tags/kwpotppt** .