Understanding Protein Function on a Genome-scale through the Analysis of Molecular Networks

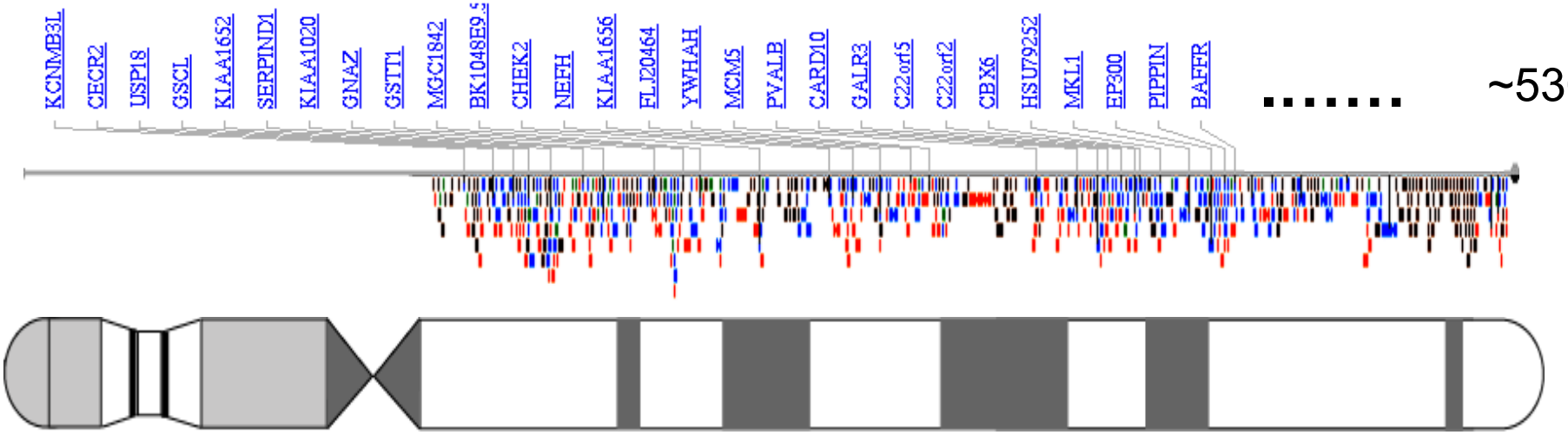
Mark B Gerstein
Yale

Slides at

Lectures.GersteinLab.org

(See Last Slide for References & More Info.)

The problem: Grappling with Function on a Genome Scale?



- 250 of ~530
  originally characterized on chr. 22
  [Dunham et al. Nature (1999)]

- >25K Proteins in Entire Human Genome
  (with alt. splicing)

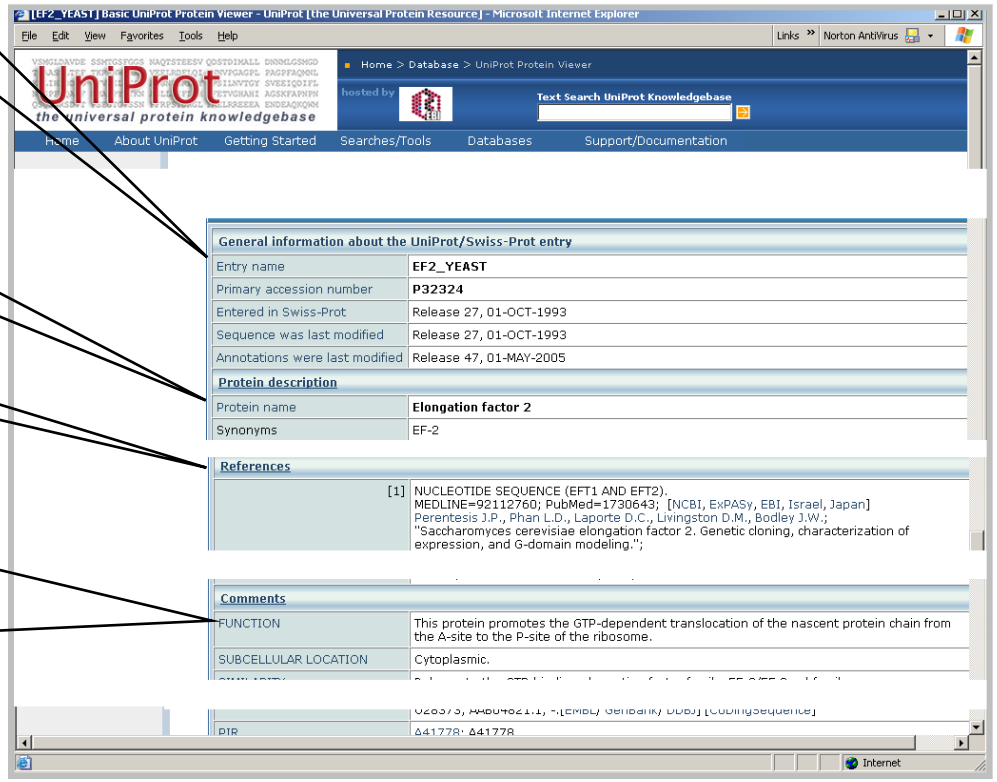
Traditional single molecule way to integrate evidence & describe function

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.

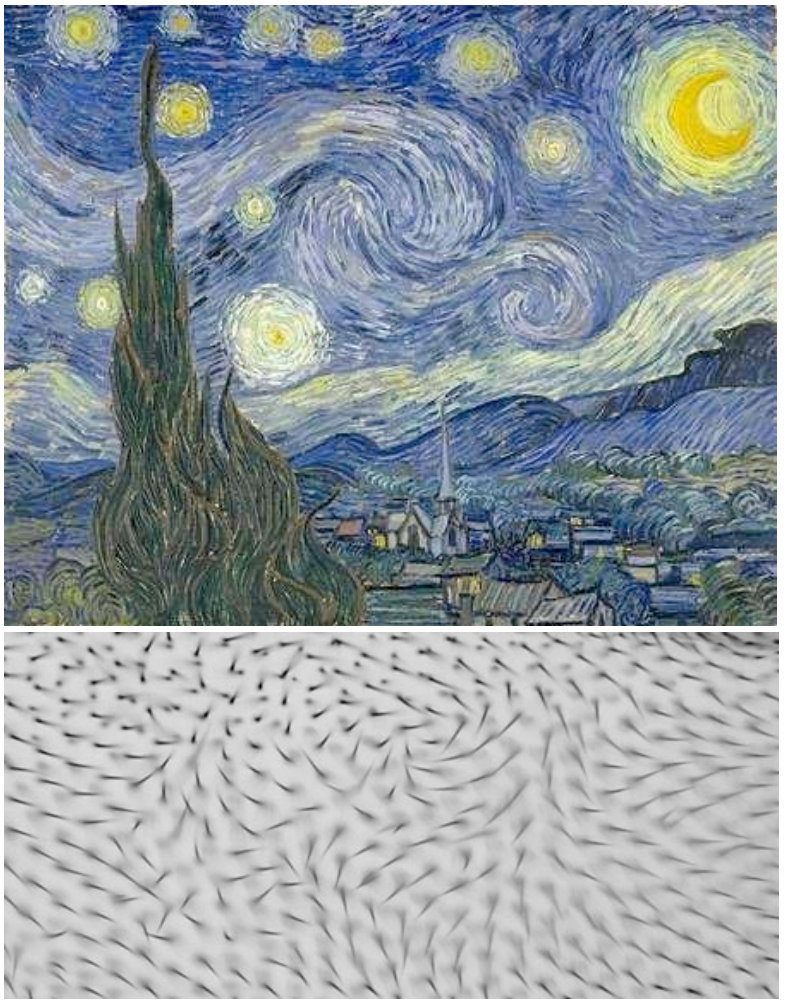


Some obvious issues in scaling single molecule definition to a genomic scale

- Fundamental complexities
  - ◊ Often >2 proteins/function
  - ◊ Multi-functionality:
    2 functions/protein
  - ◊ Role Conflation:
    molecular, cellular, phenotypic

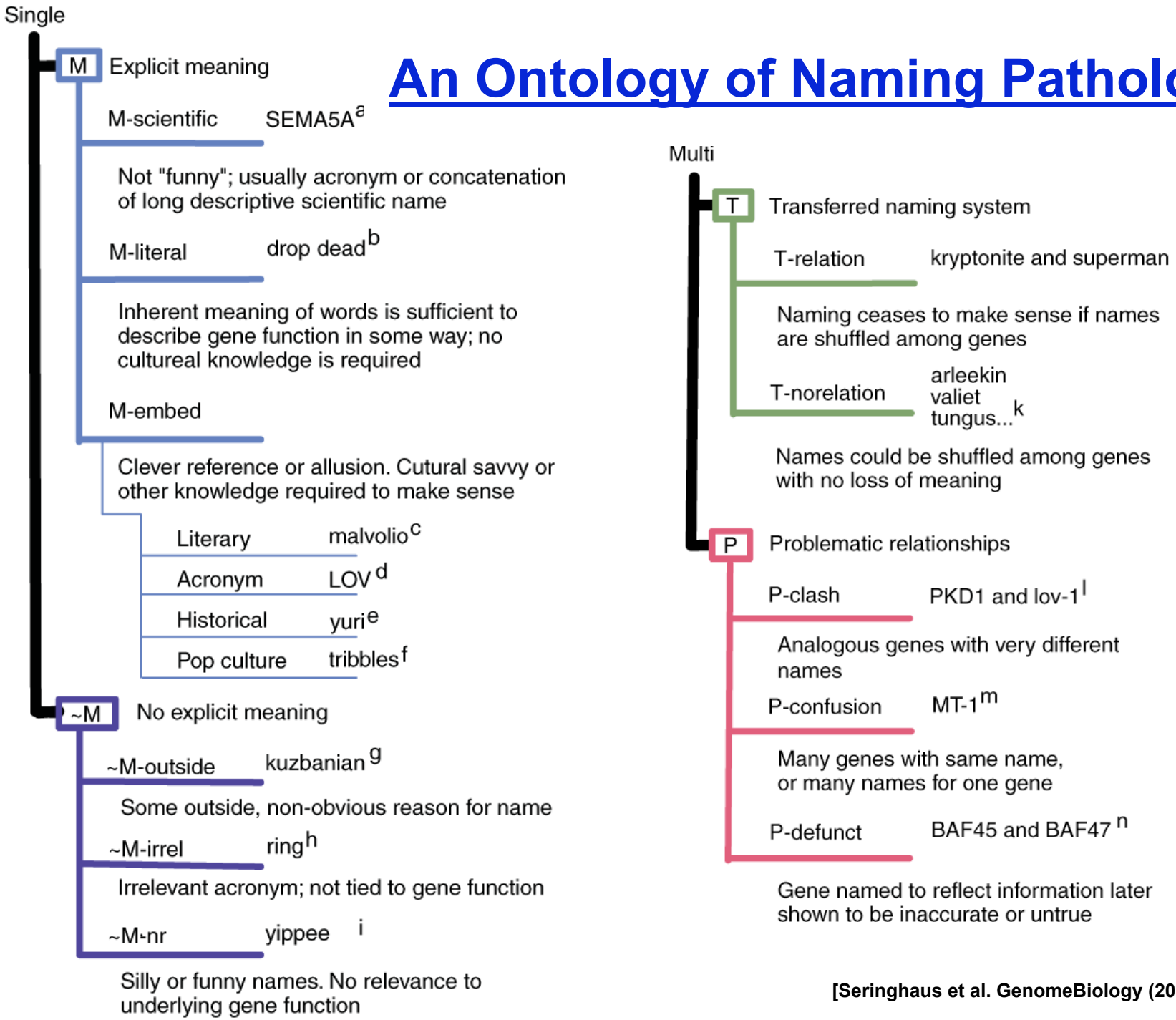
Some obvious issues in scaling single molecule definition to a genomic scale

- Fundamental complexities
  - ◊ Often >2 proteins/function
  - ◊ Multi-functionality:
    2 functions/protein
  - ◊ Role Conflation:
    molecular, cellular, phenotypic
- Fun terms… but do they scale?....
  - ◊ **Starry night** (P Adler, '94)



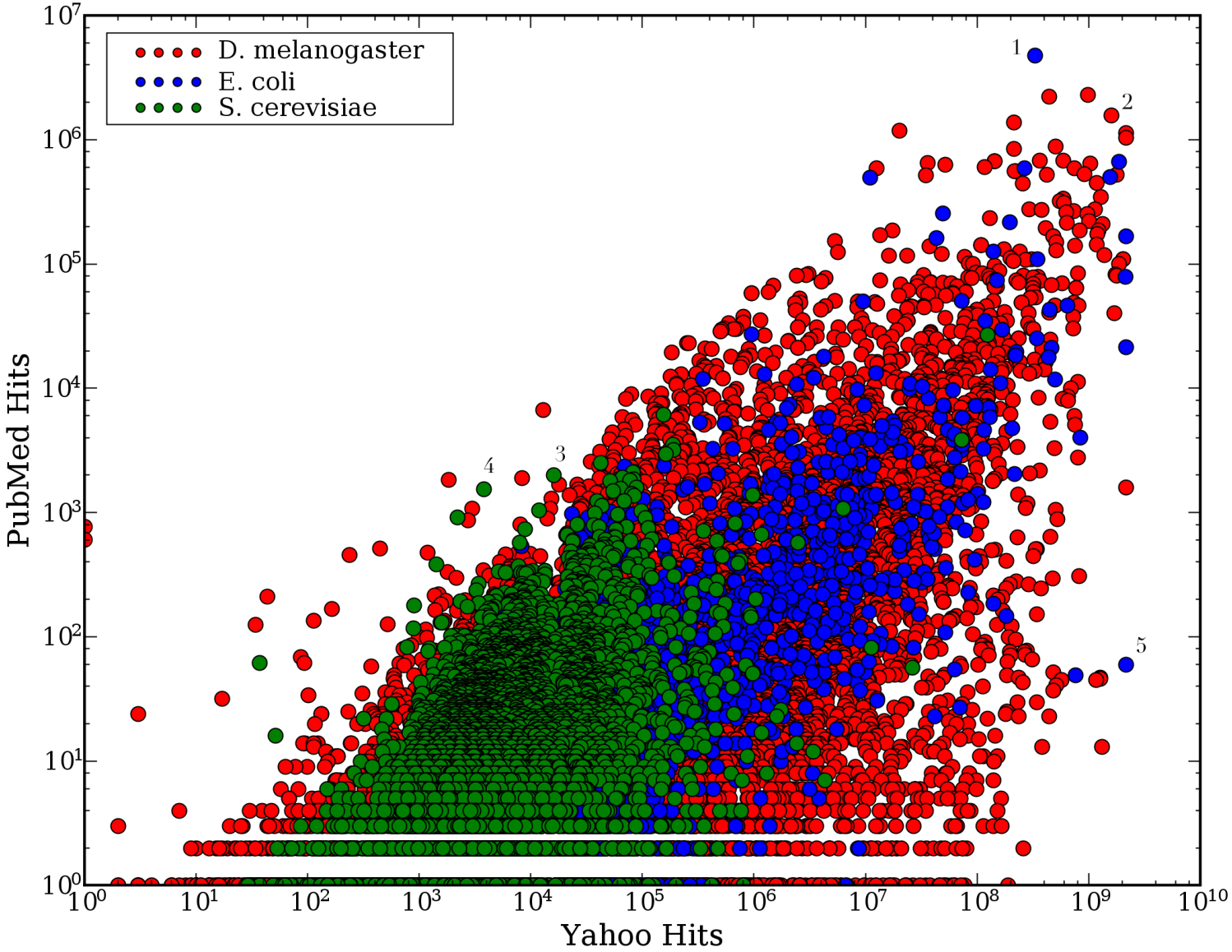
[Seringhaus et al. GenomeBiology (2008)]

An Ontology of Naming Pathologies



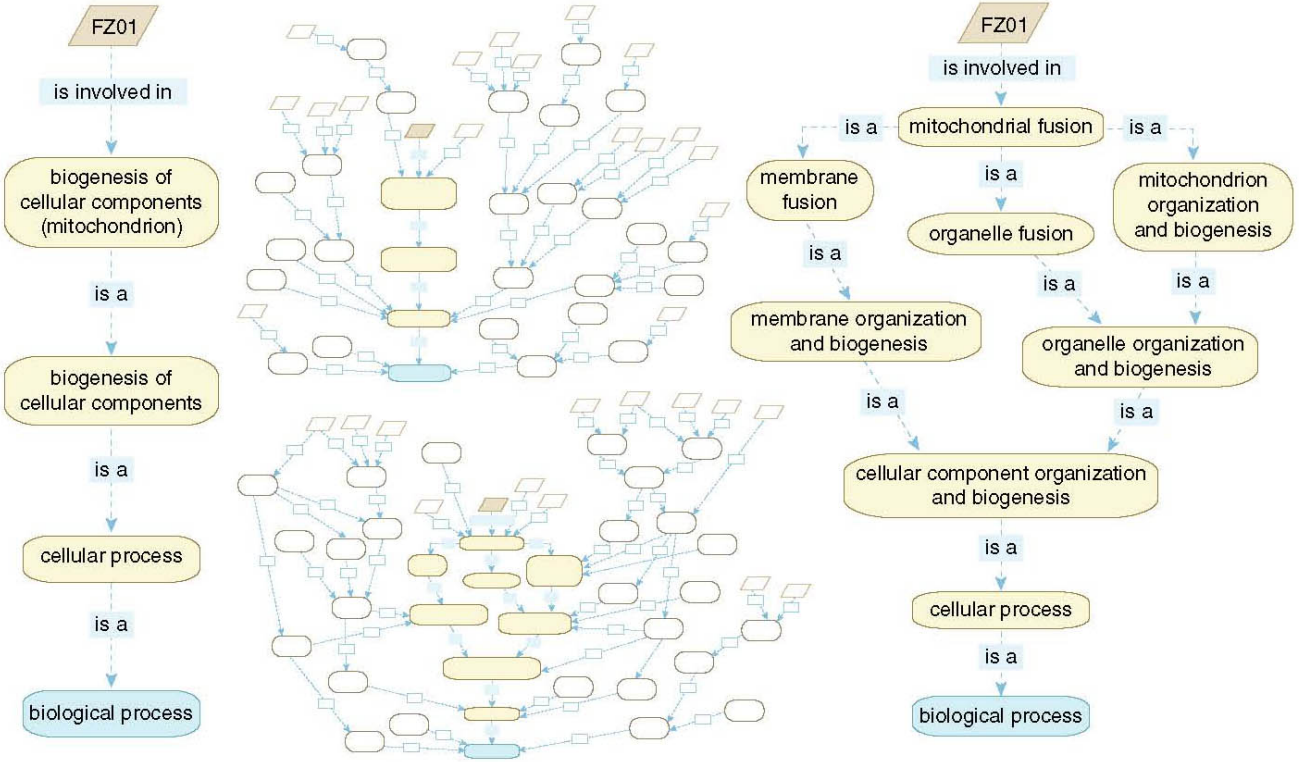
[Seringhaus et al. GenomeBiology (2008)]

Gene Name Skew



[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]

Towards Developing Standardized Descriptions of Function

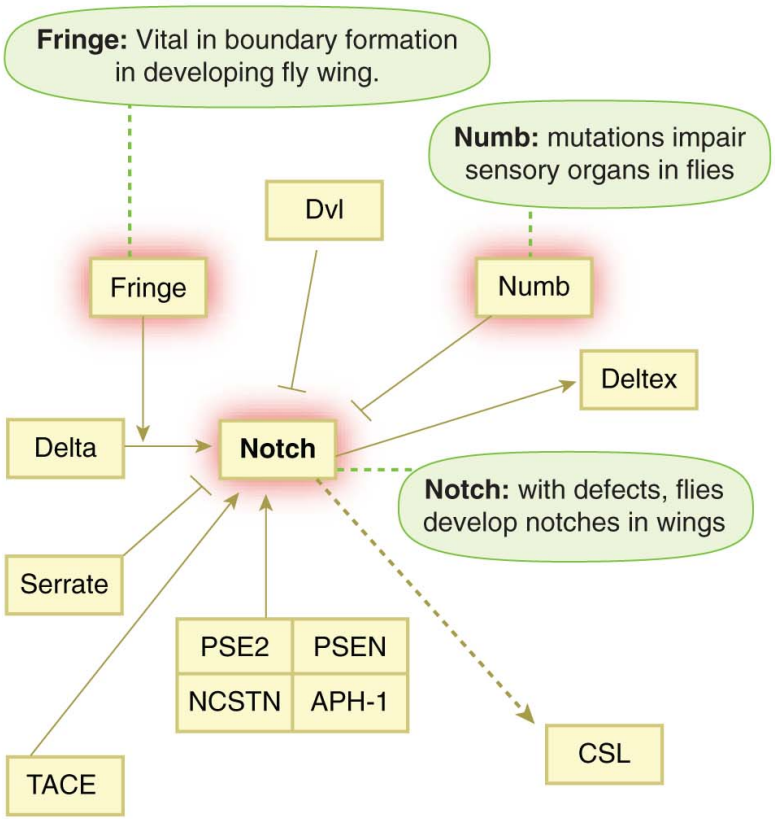
- Subjecting each gene to standardized expt. and cataloging effect
  - ◊ KOs of each gene in a variety of std. conditions => phenotypes
  - ◊ Std. binding expts for each gene (e.g. prot. chip)

- Function as a vector

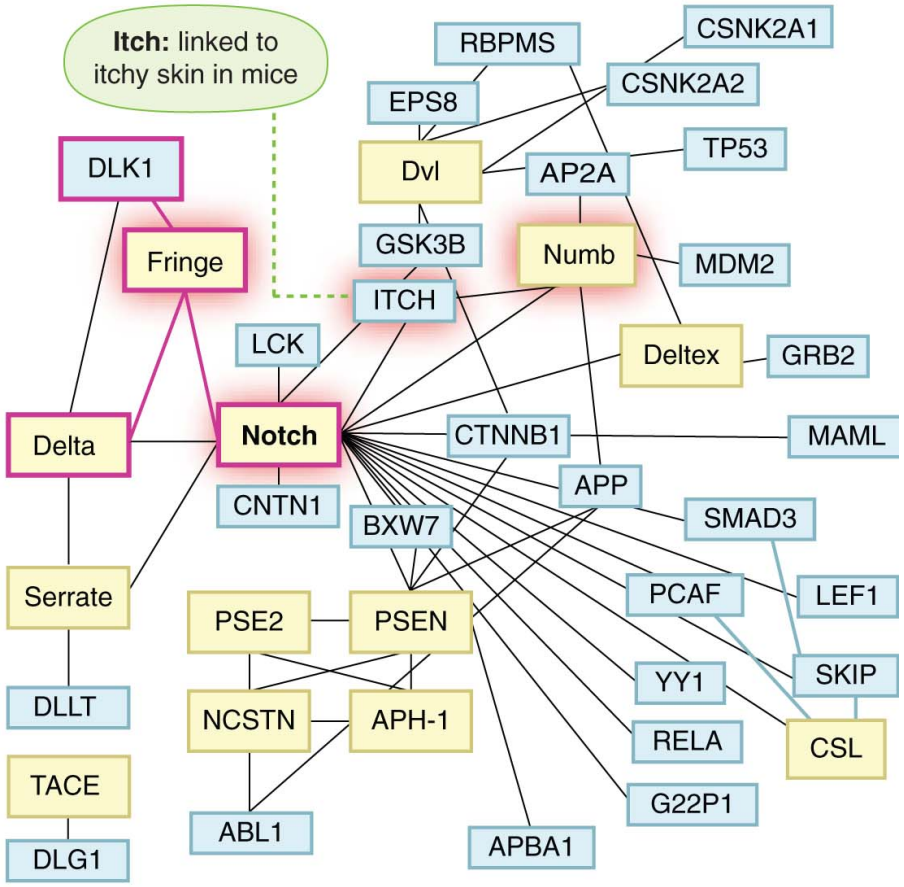
	nucleic acids		small molecules					proteins			
	DNA	RNA	ATP	Metal	CoA	NAD	.....	G protein	CDC28	Calmodulin	.....
protein 1	1.0	0	0	0	0	0	.....	0	0	0	.....
protein 2	0	0.9	0	0	0	0	.....	0	0	0	.....
protein 3	1.0	0	1.0	0	0	0	.....	0	0	0	.....
protein 4	0	0	0	0	0.8	0	.....	0	0	1.0	.....
protein 5	1.0	0	0	0	0	0	.....	0	0.9	0	.....
protein 6	0.9	0					.....				.....
protein 7	0	0.8					.....				.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

Interaction Vectors [Lan et al, IEEE 90:1848]

Networks (Old & New)



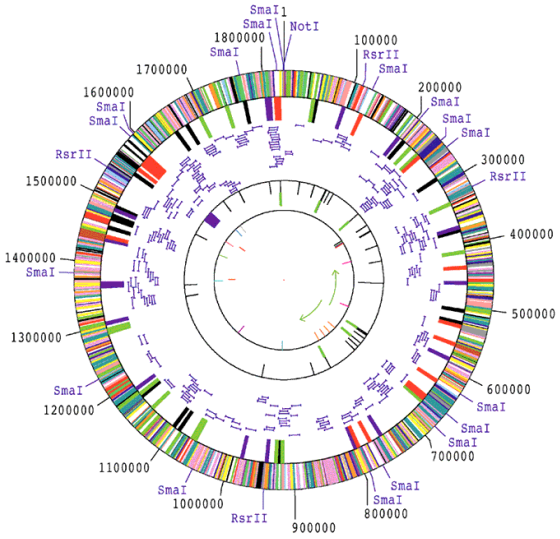
Classical KEGG pathway

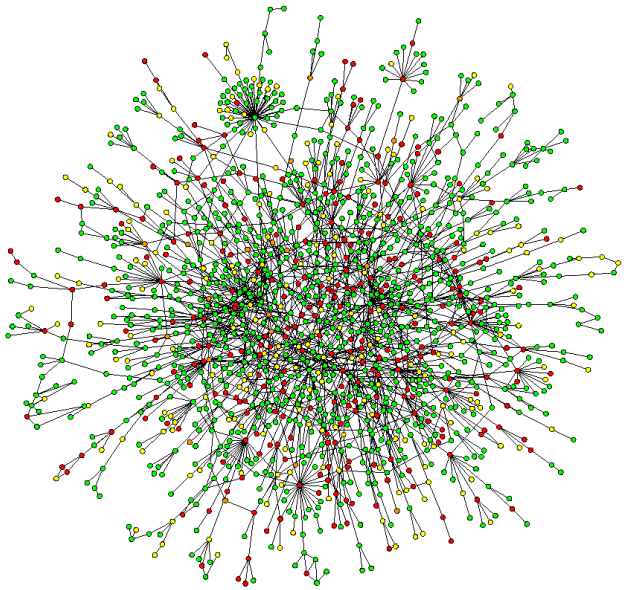


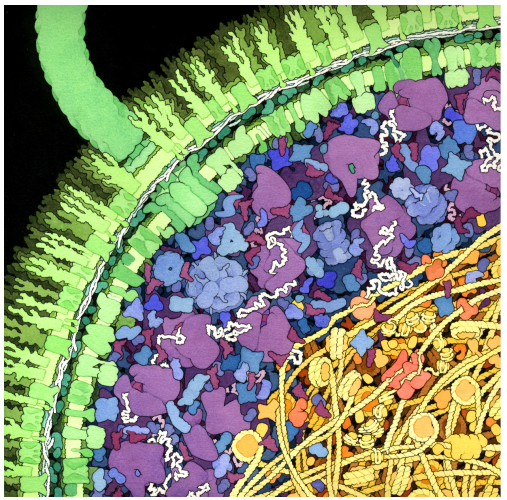
Same Genes in High-throughput Network

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

Networks occupy a midway point in terms of level of understanding







1D: Complete Genetic Partslist

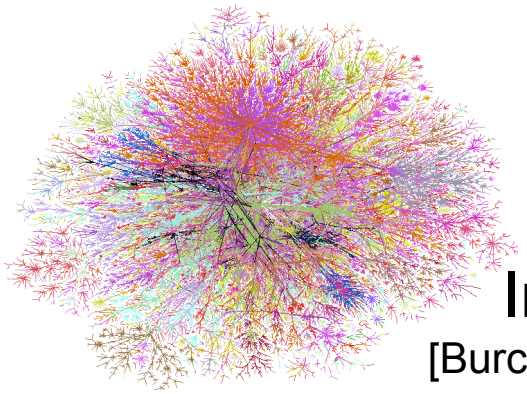
~2D: Bio-molecular Network Wiring Diagram

3D: Detailed structural understanding of cellular machinery

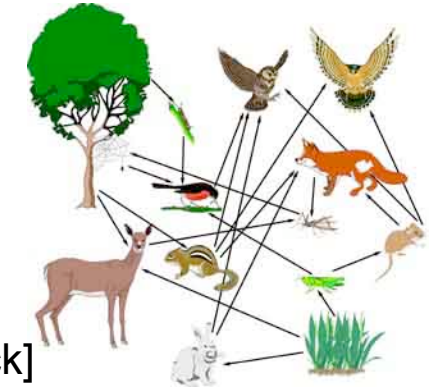
[Fleischmann et al., Science, 269 :496]

[Jeong et al. Nature, 41:411]

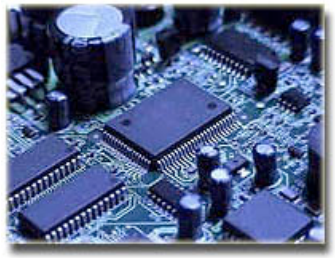
Networks as a universal language



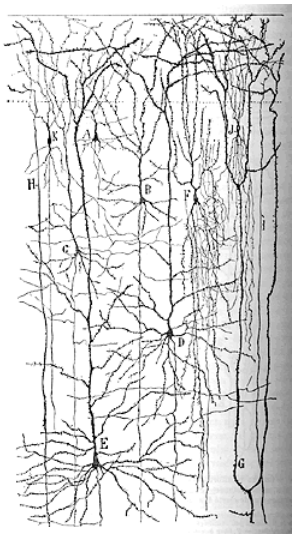
Internet
[Burch & Cheswick]



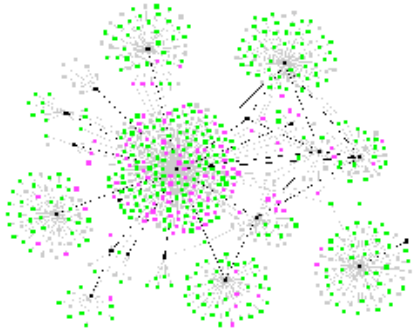
Food Web



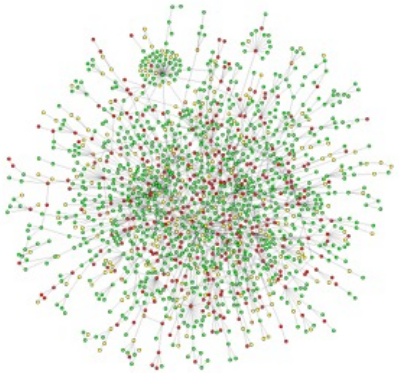
Electronic Circuit



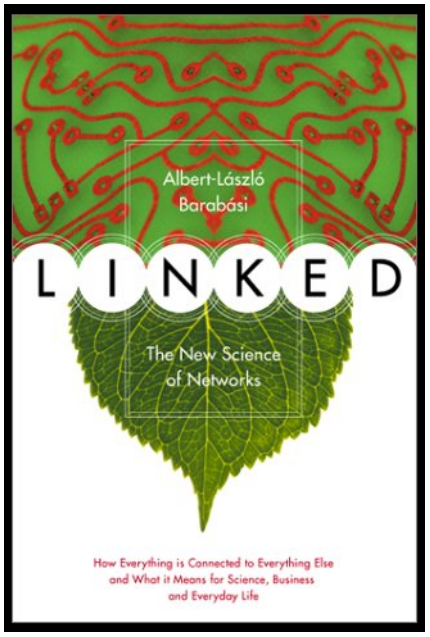
Neural Network
[Cajal]

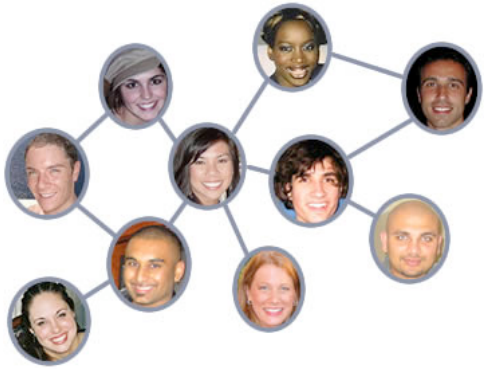


Disease Spread
[Krebs]



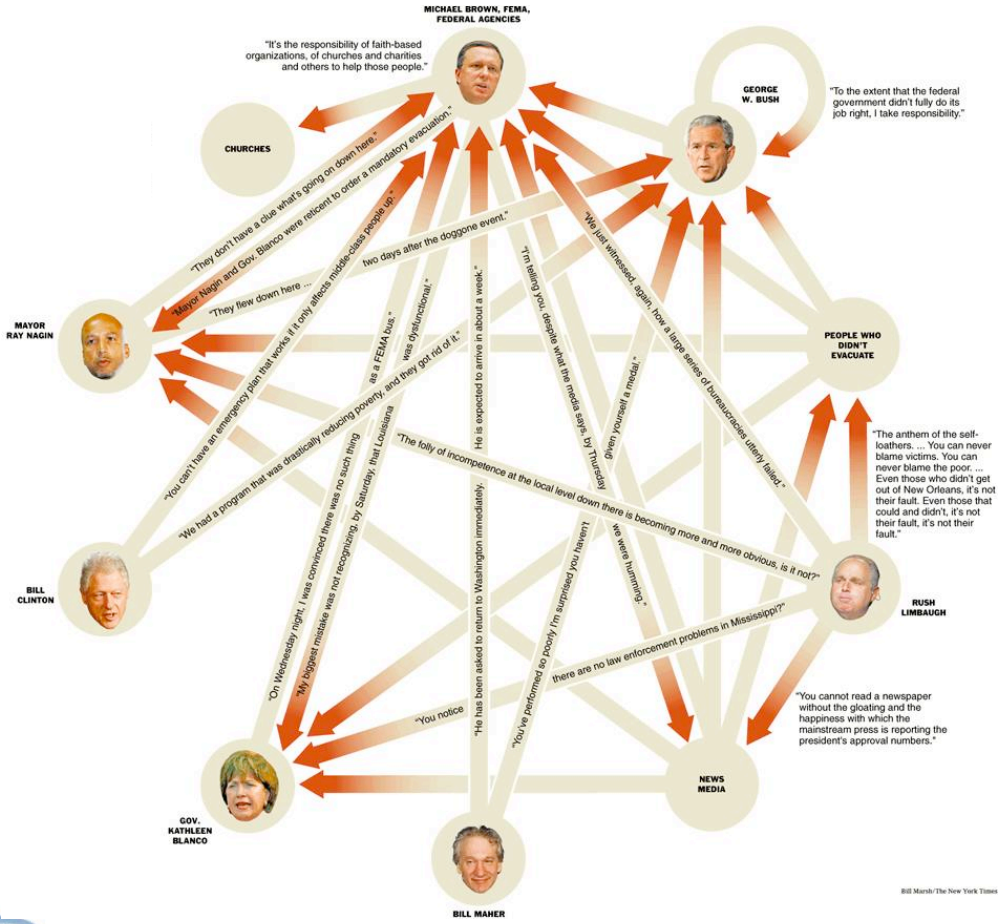
Protein Interactions
[Barabasi]



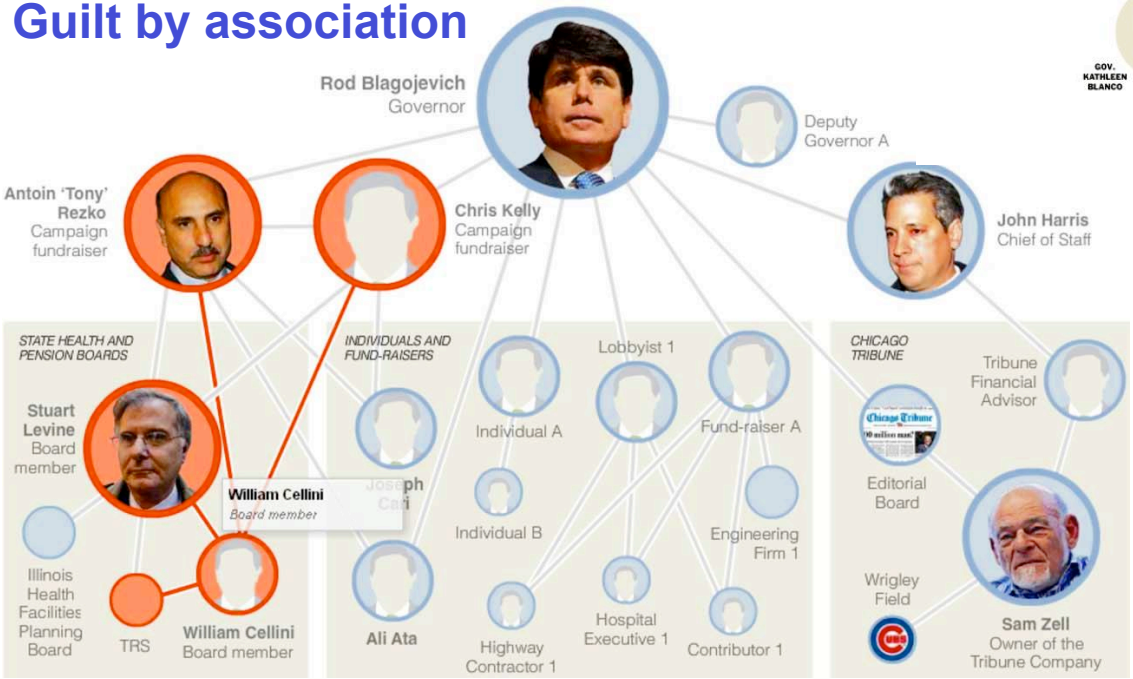


Social Network

Using the position in networks to describe function



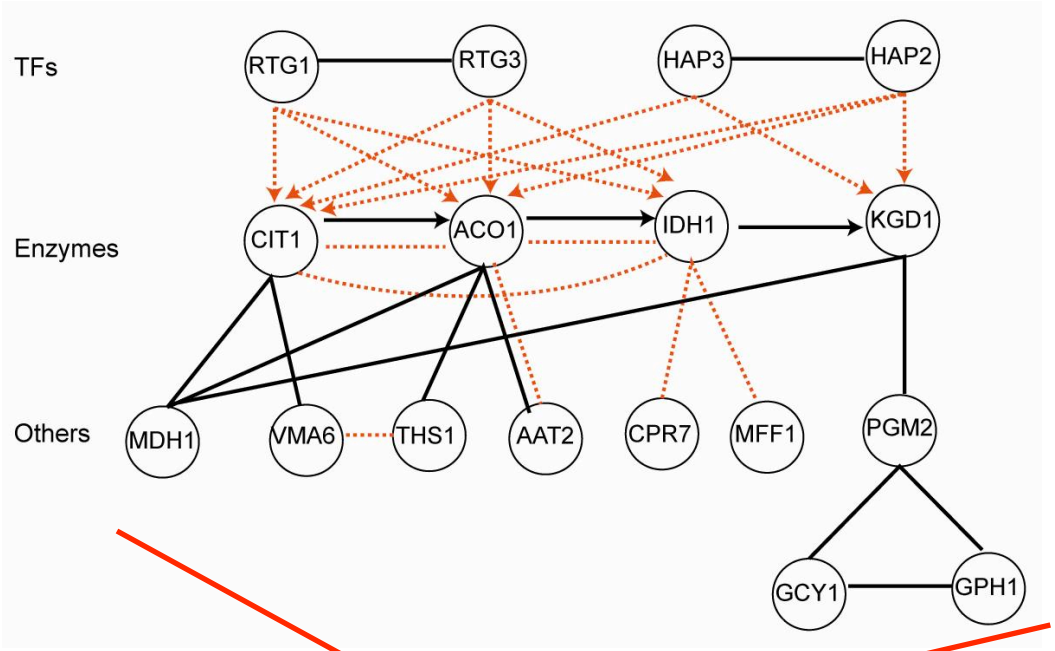
Guilt by association

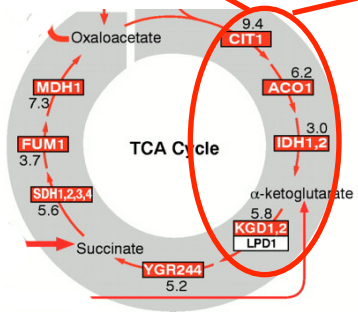


Finding the causal regulator (the "Blame Game")

[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

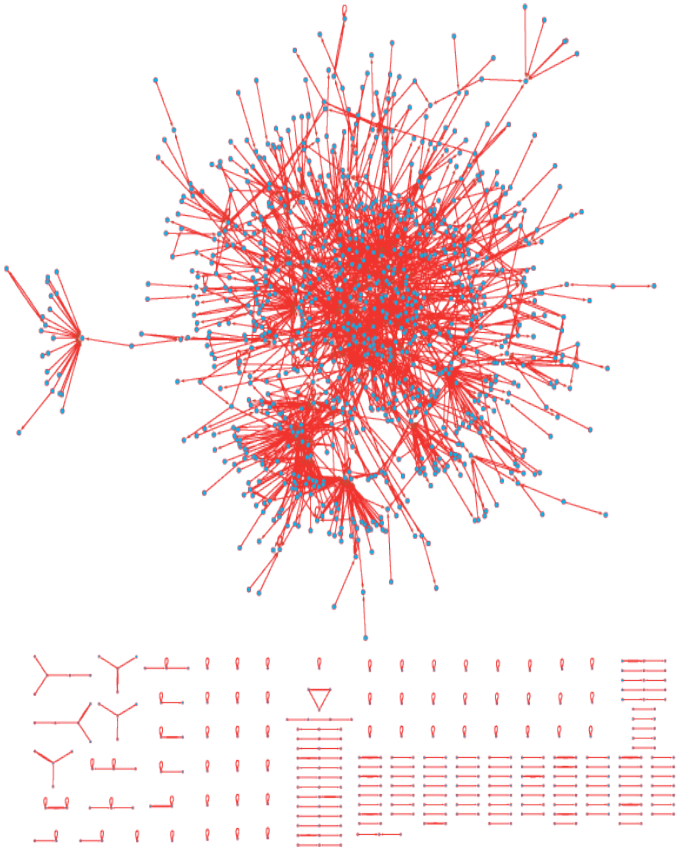
Co-expression Relationship

Genetic interaction (synthetic lethal)
Signaling pathways

Outline: Molecular Networks

- Why Networks?
- Predicting Networks (yeast)
  - ◊ Propagating known information
- Network Structure: Key Positions (yeast)
  - ◊ Hubs & Bottlenecks
  - ◊ Tops of Hierarchy
- Dynamics & Variation of Networks
  - ◊ Across cellular states (yeast)
  - ◊ Across environments (in prokaryotes)
- Protein Networks & Human Variation

Example: yeast PPI network



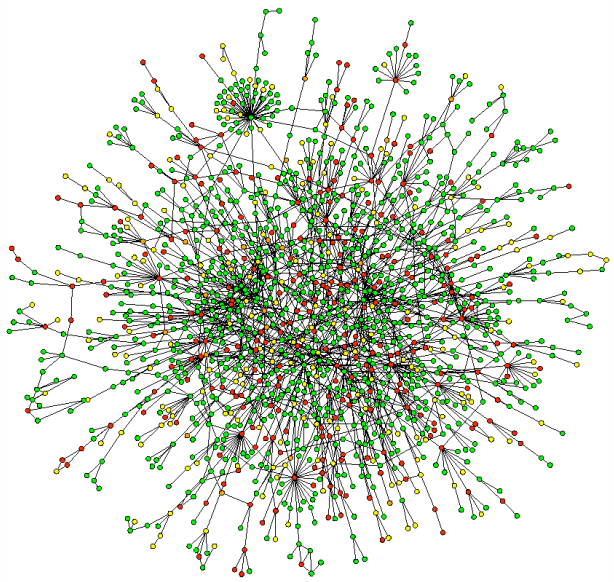
Actual size:

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

Known interactions:

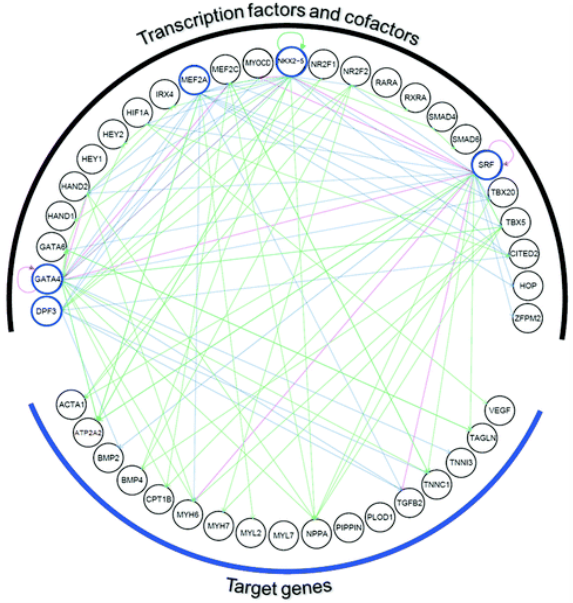
- Small-scale experiments: accurate but few
  → Overfitting: ~5,000 in BioGRID, involving ~2,300 proteins
- Large-scale experiments: abundant but noisy
  → Noise: false +ve/-ve for yeast two-hybrid data up to 45% and 90% (Huang et al., 2007)

Types of Networks

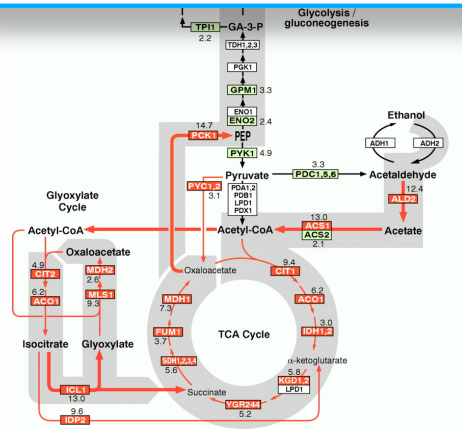


Interaction networks

Nodes: proteins or genes
Edges: interactions



Regulatory networks



Metabolic networks

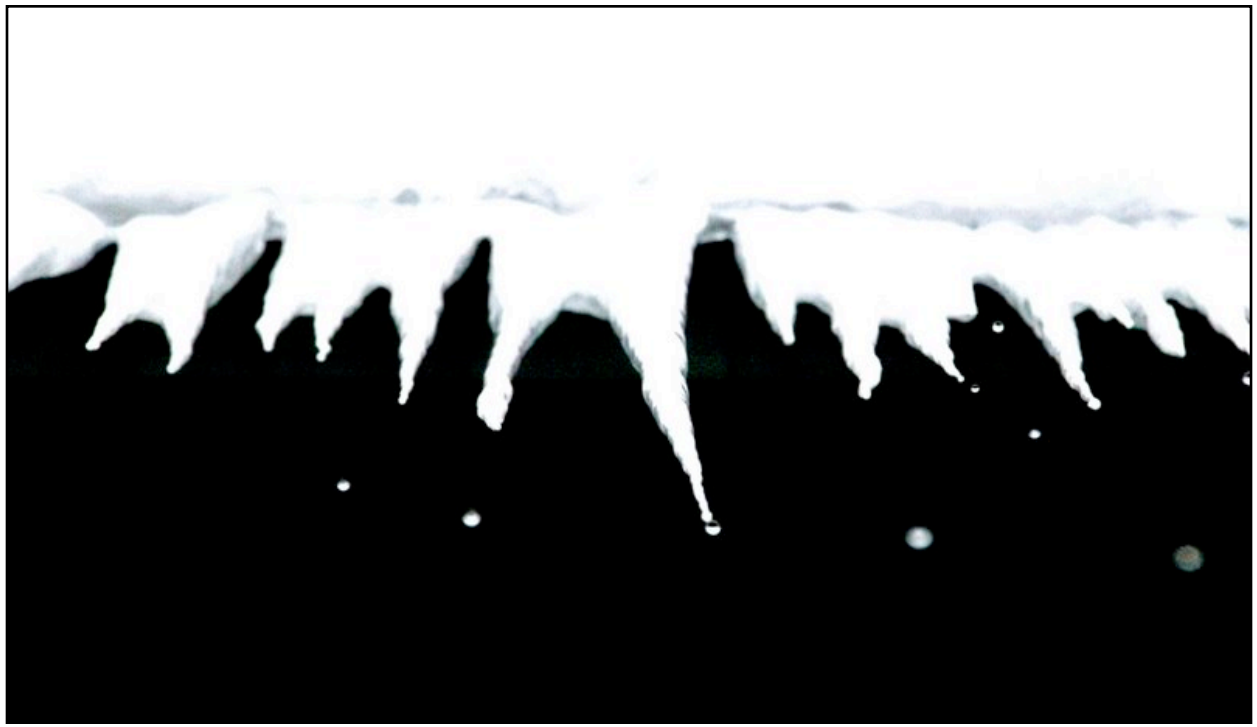
[Horak, et al, Genes & Development, 16:3017-3033]

[DeRisi, Iyer, and Brown, Science, 278:680-686]

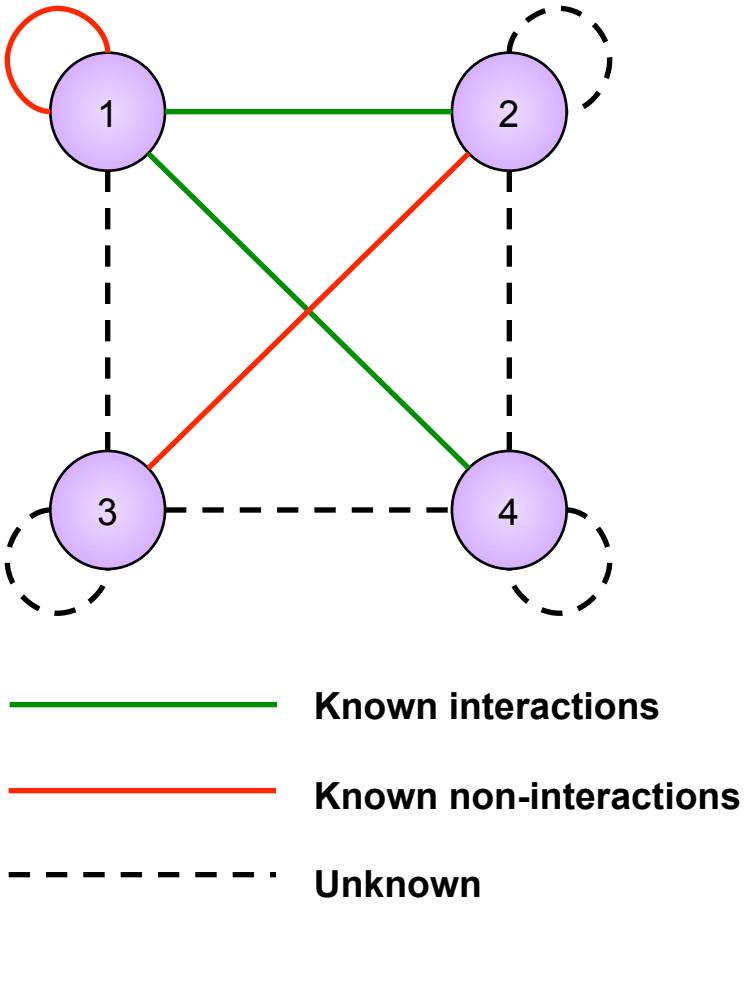
[Jeong et al, Nature, 41:411]

Predicting Networks

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.

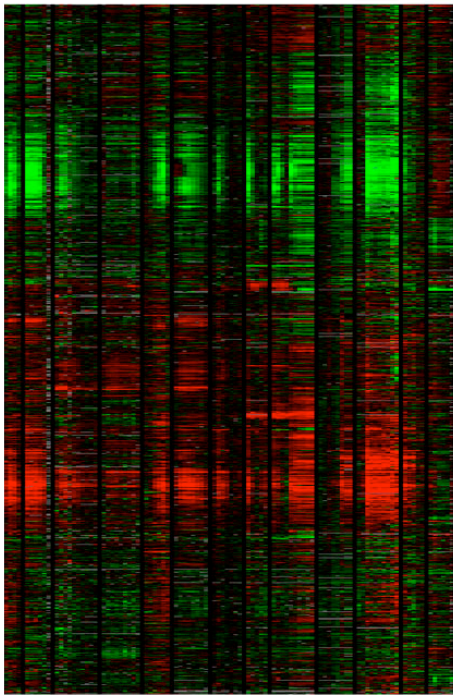


Network prediction: known information



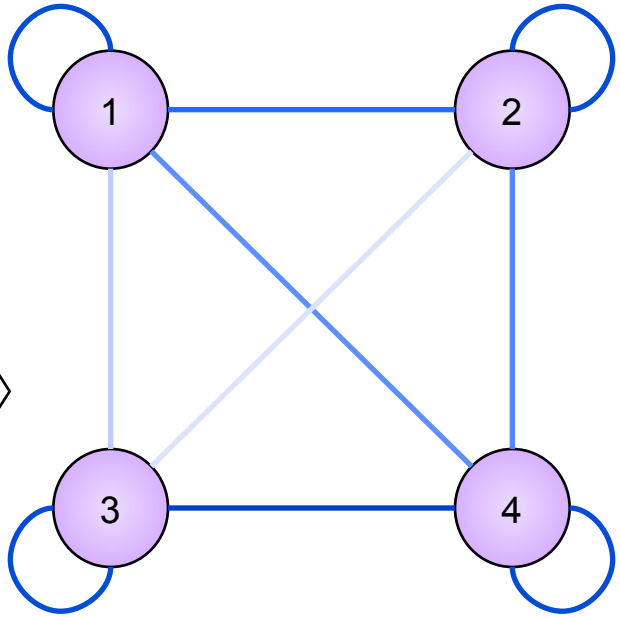
Network prediction: features

- Example 1: gene expression



Gasch et al., 2000

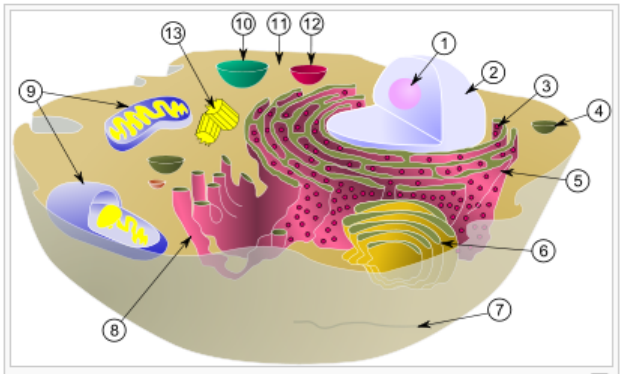
$x_1 = (0.2, 2.4, 1.5, \ldots)$
$x_2 = (0.8, 2.2, 1.5, \ldots)$
$x_3 = (4.3, 0.1, 7.5, \ldots)$
…
$sim(x_1, x_2) = 0.62$
$sim(x_1, x_3) = -0.58$
…



Similarity scale:

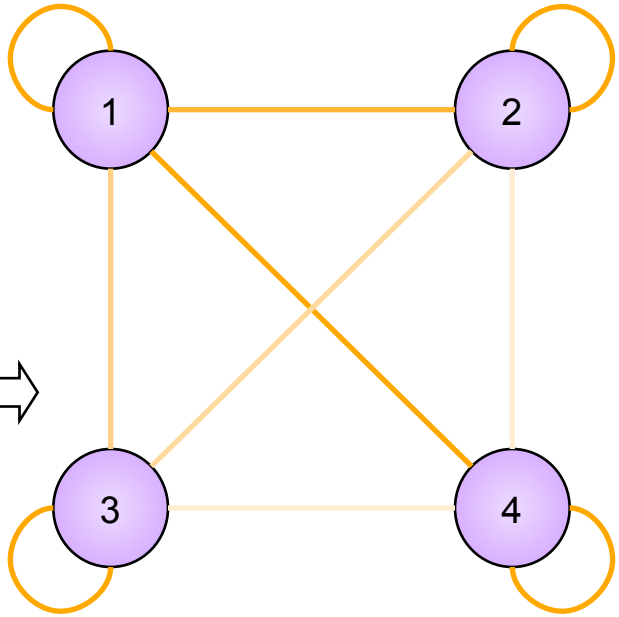
Network prediction: features

- Example 2: sub-cellular localization



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

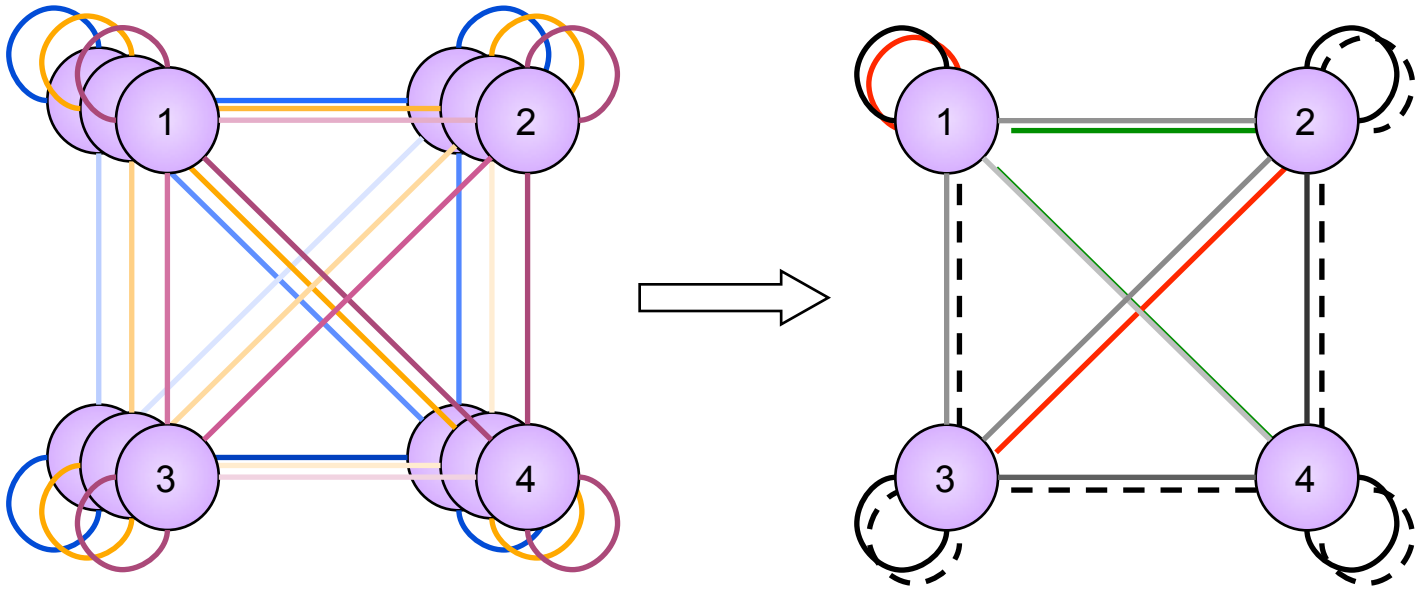
$x_1 = (1, 1, 0, 0, \ldots)$
$x_2 = (1, 1, 1, 0, \ldots)$
$x_3 = (1, 0, 1, 0, \ldots)$
…
$sim(x_1, x_2) = 0.81$
$sim(x_1, x_3) = 0.12$
…



Similarity scale:

1 **-1**

Network prediction: data integration



Learning methods

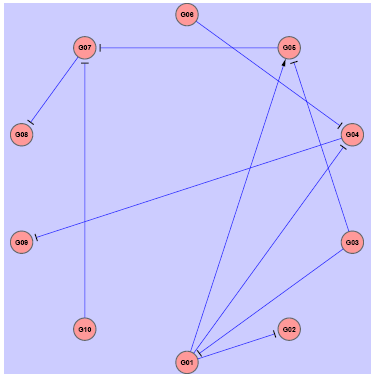
An endless list:

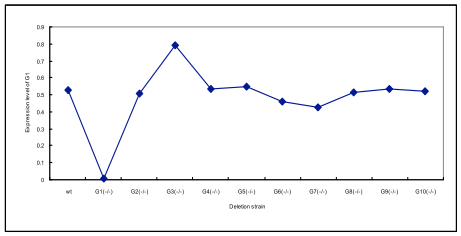
- Docking (e.g. Schoichet and Kuntz 1991)
- Evolutionary (e.g. Ramani and Marcotte, 2003)
- Topological (e.g. Yu et al., 2006)
- Bayesian (e.g. Jansen et al., 2003)
- **Kernel methods**
  - ◊ Global modeling:
    - em (Tsuda et al., 2003)
    - kCCA (Yamanishi et al., 2004)
    - kML (Vert and Yamanishi, 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)

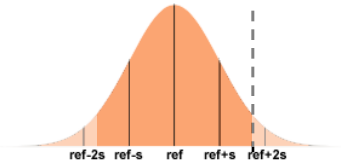
DREAM3: *in silico* regulatory network reconstruction

Actual network | **Expression data** | **Modeling** | **Predictions**



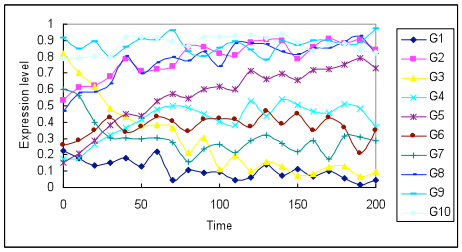


Deletion strains



Prob(signal|point)
= 2Φ((point – ref) / s) – 1

Noise models



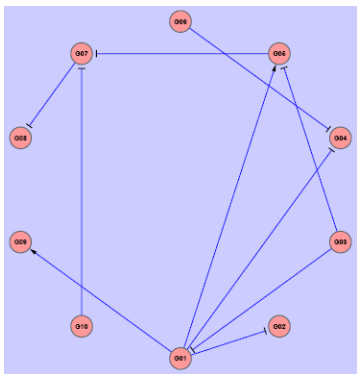
Time series after initial perturbation

$$\frac{dy_j}{dt} = a_{j0} - a_{jj}y_j + \sum_{k\in S_j} a_{jk}y_k$$

$$\frac{dy_j}{dt} = \frac{b_{j1}}{1+\exp\left(a_{j0} + \sum_{k\in S_j} a_{jk}y_k\right)} - b_{j2}y_j$$

$$\frac{dy_j}{dt} = a_{j0}\prod_{k_1\in S_1}\left(\frac{b_{jk_1}}{y_{k_1}^{c_{jk_1}} + b_{jk_1}}\right)\prod_{k_2\in S_2}\left(\frac{y_{k_2}^{c_{jk_2}}}{y_{k_2}^{c_{jk_2}} + b_{jk_2}}\right) - a_{j1}y_j$$

Expression rate models



[Yip et al., DREAM3]

Accuracy (AUC)	E. Coli 1	E. Coli 2	Yeast 1	Yeast 2	Yeast 3
Size-10	0.928	0.912	0.949	0.747	0.714
Size-50	0.930	0.924	0.917	0.792	0.805
Size-100	0.948	0.960	0.915	0.856	0.783

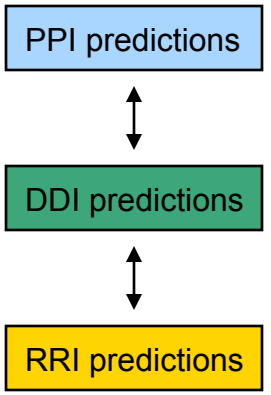
Our work: efficiently propagating known information

Training set expansion

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

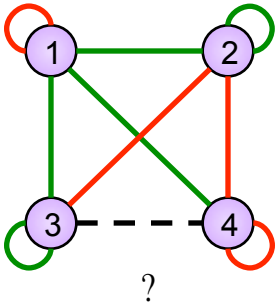
Multi-level learning

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



DREAM3 *in silico* regulatory network reconstruction challenge

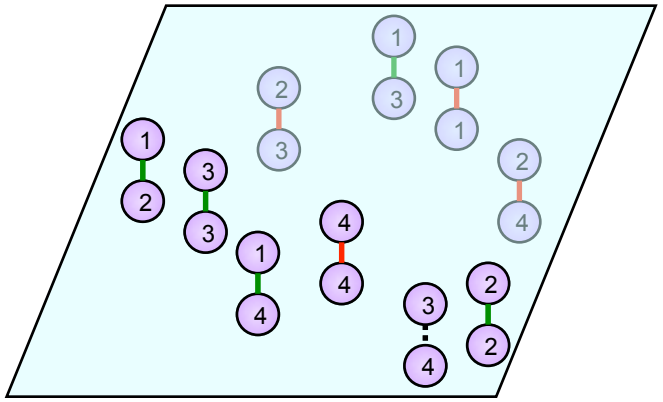
Global vs. local modeling



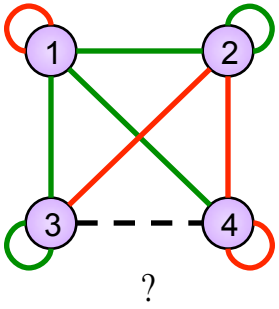
Global modeling: build one model for the whole network

Example - Pairwise kernel: consider object pairs instead of individual objects

Problem: $O(n^2)$ instances, $O(n^4)$ kernel elements

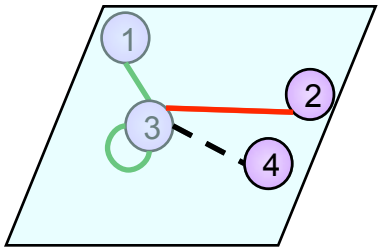


Global vs. local modeling



Local modeling: build one model for each node

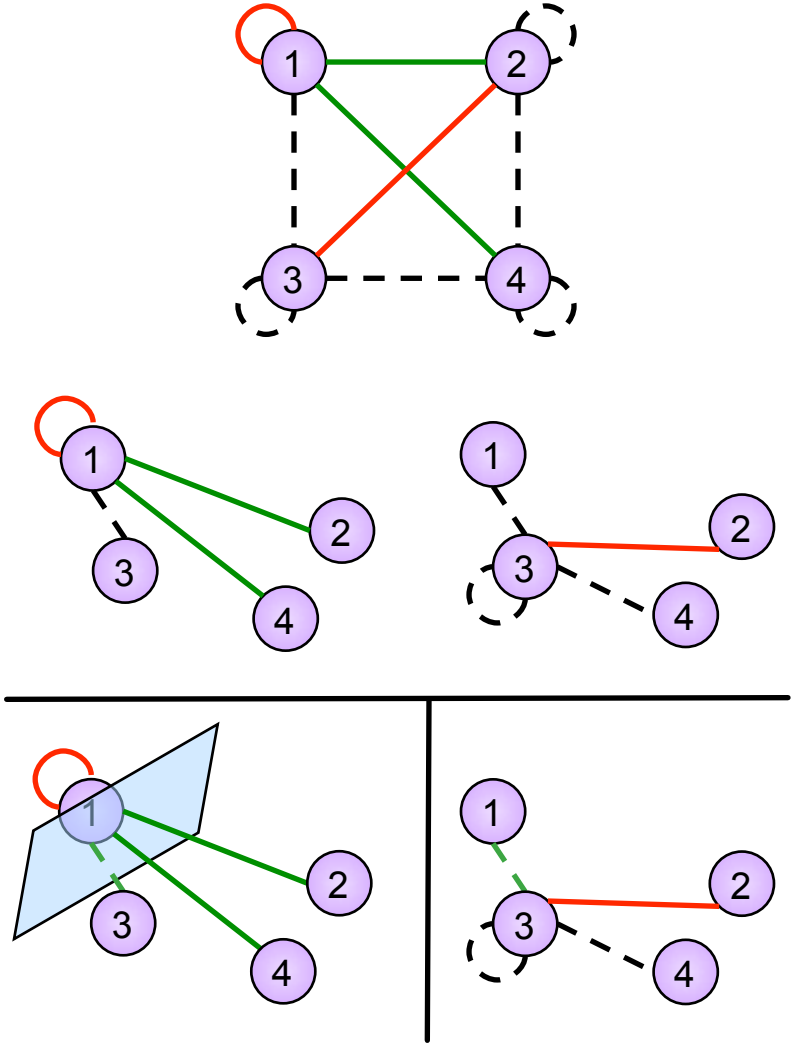
Model for node 3:



Problem: insufficient and unevenly distributed training data (what if node 3 has no known interactions at all?)

Prediction propagation

- Goal: keep the flexibility of local modeling, but tackle the data sparsity problem
- Motivation: some objects have more examples than others
- Our approach:
  - ◊ Learn models for objects with more examples first
  - ◊ Propagate the most confident predictions as auxiliary examples of other objects



[Yip and Gerstein, Bioinformatics ('09)]

Prediction accuracy (AUC)

	phy	loc	exp-gasch	exp-spellman	y2h-ito	y2h-uetz	tap-gavin	tap-krogan	int
Mode 1									
direct	58.04	66.55	64.61	57.41	51.52	52.13	59.37	61.62	70.91
kCCA	65.80	63.86	68.98	65.10	50.89	50.48	57.56	51.85	80.98
kML	63.87	68.10	69.67	68.99	52.76	53.85	60.86	57.69	73.47
em	71.22	75.14	67.53	64.96	55.90	53.13	63.74	68.20	81.65
local	71.67	71.41	72.66	70.63	67.27	67.27	64.60	67.48	75.65
local+pp	73.89	75.25	77.43	75.35	71.60	71.51	74.62	71.39	83.63
local+ki	71.68	71.42	75.89	70.96	69.40	69.05	70.53	72.03	81.74
local+pp+ki	72.40	75.19	77.41	73.81	70.44	70.57	73.59	72.64	83.59

Observations:

- Highest accuracy by training set expansion
- Over fitting of local modeling without training set expansion
- Prediction propagation theoretically related to co -training (Blum and Mitchell, 1998)
  - ◊ Semi-supervised (Similarity with PSI-BLAST)

[Yip and Gerstein, Bioinformatics ('09)]

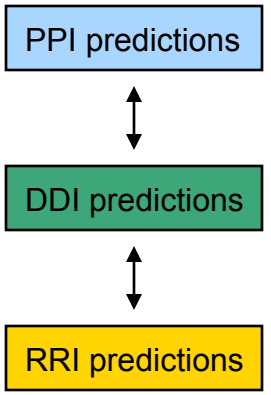
From horizontal to vertical

Training set expansion

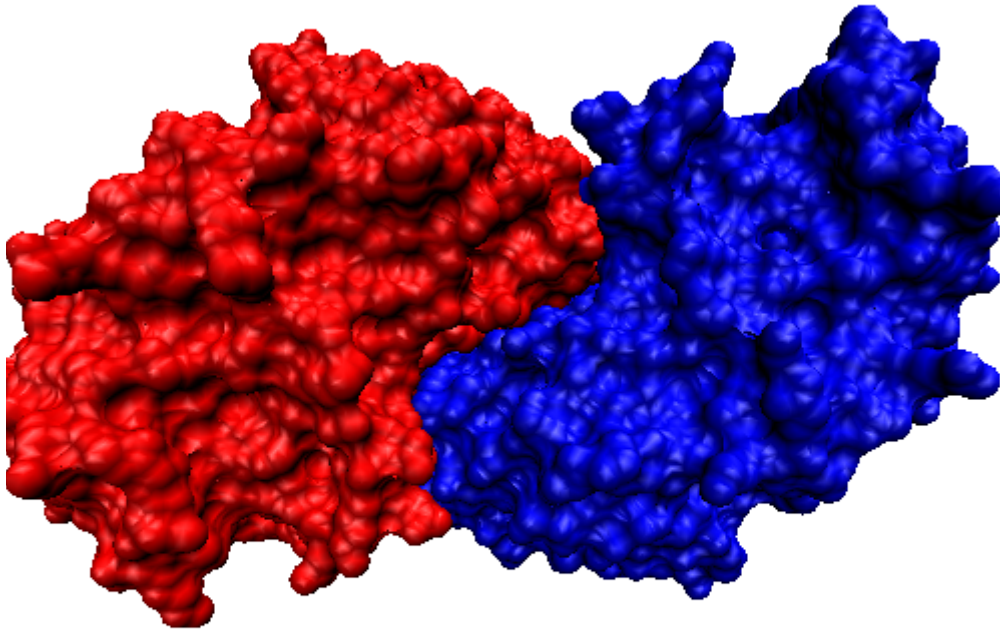
- Motivation: lack of training examples
- Expand training sets horizontally

Multi-level learning

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



Protein interaction

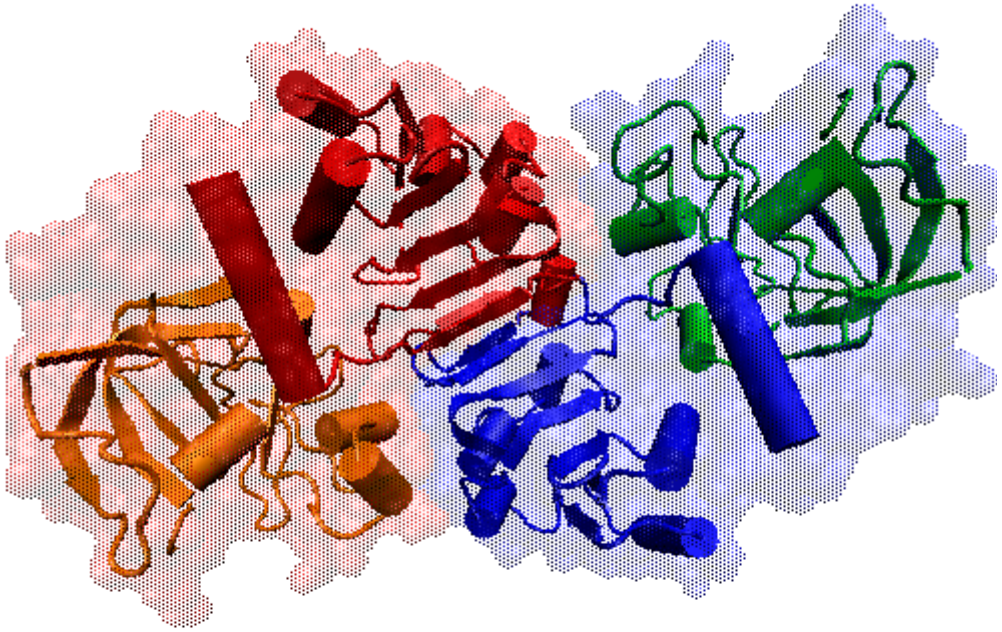


Yeast NADP-dependent alcohol dehydrogenase 6 (PDB: 1piw)

Protein-level features for interaction prediction: functional genomic information

[Yip and Gerstein, in revision]

Domain interaction

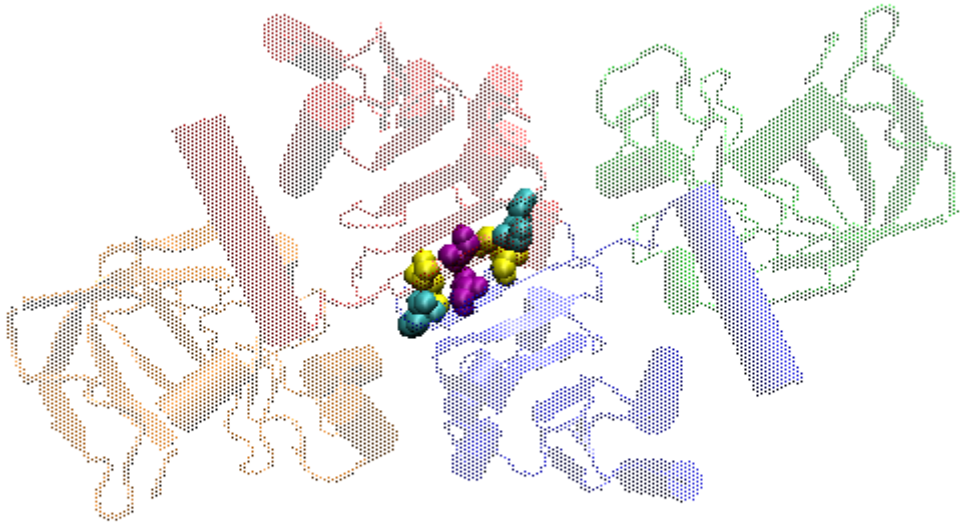


Pfam domains: PF00107 (inner) and PF08240 (outer)

Domain-level features for interaction prediction: evolutionary information

[Yip and Gerstein, in revision]

Residue interaction

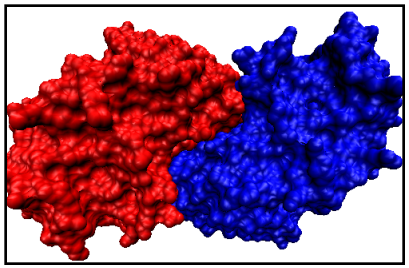


Interacting residues: 283 (yellow) with 287 (cyan), and 285 (purple) with 285

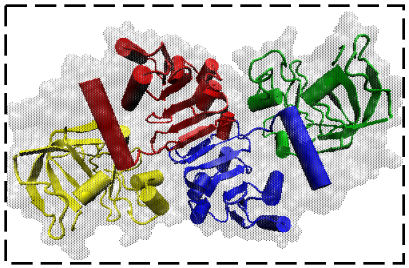
Residue-level features for interaction prediction: physical-chemical information

[Yip and Gerstein, in revision]

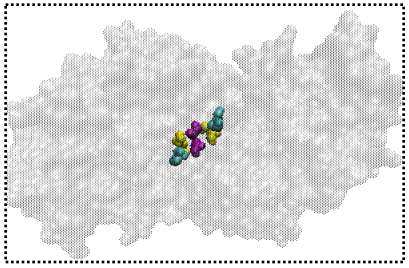
Combining the three problems



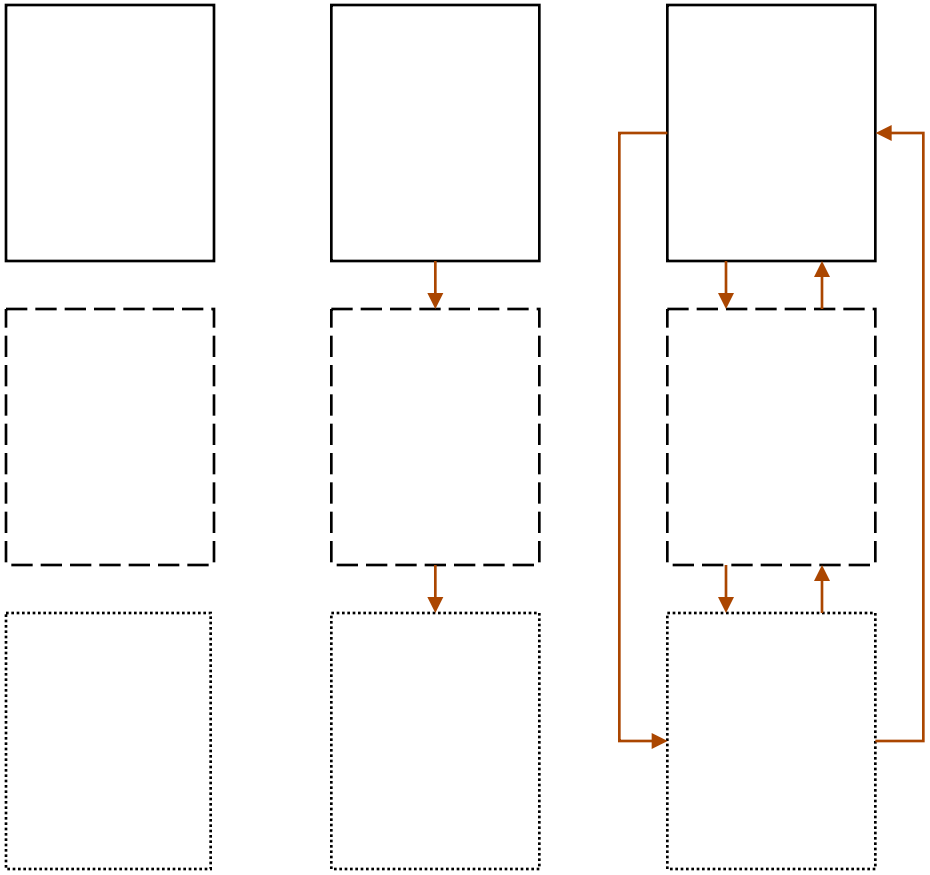
Protein interactions



Domain interactions



Residue interactions

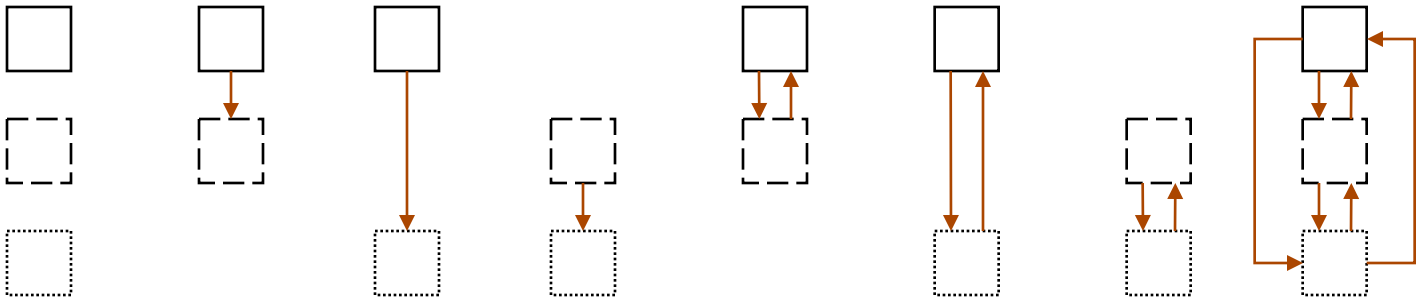


i. Independent levels ii. Unidirectional flow iii. Bidirectional flow

[Yip and Gerstein, in revision]

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

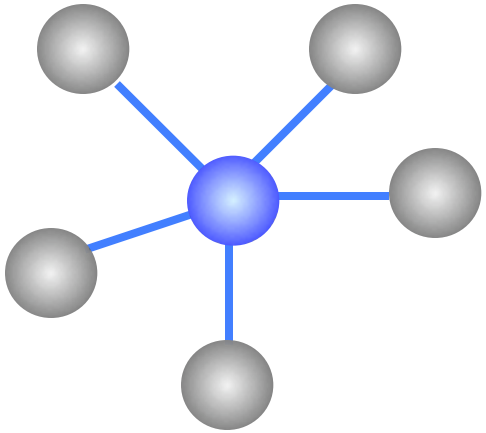
[Yip and Gerstein, in revision]

Finding Central Points in Networks: Hubs & Bottlenecks

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

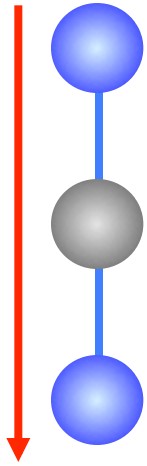
Global topological measures

Indicate the gross topological structure of the network



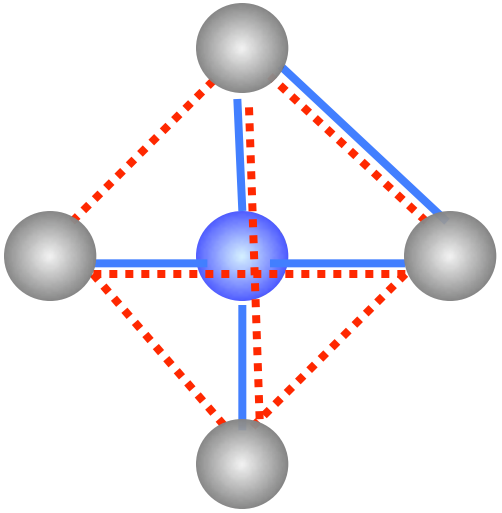
Degree (K)

5



Path length (L)

2



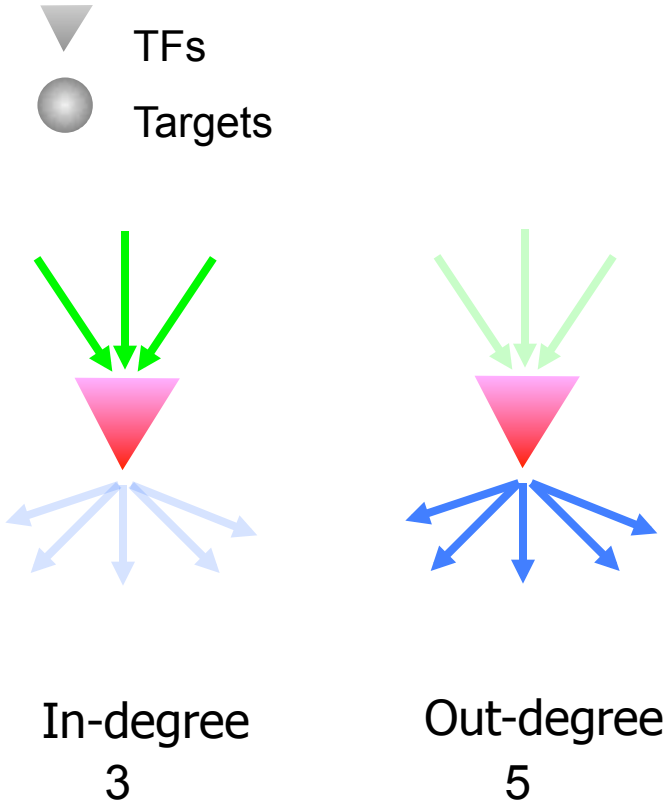
Clustering coefficient (C)

1/6

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

[Barabasi]

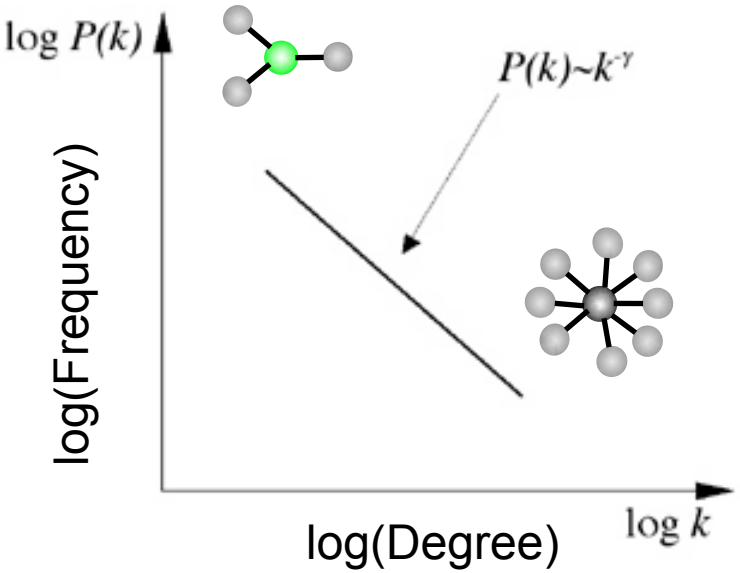
Global topological measures for directed networks

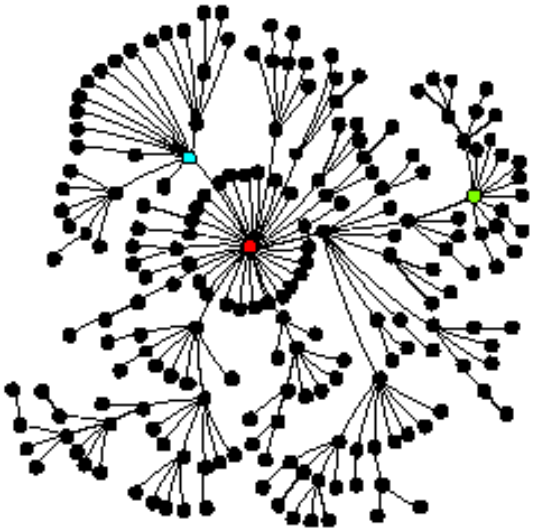


Regulatory and metabolic networks are ***directed***

Scale-free networks

Power-law distribution





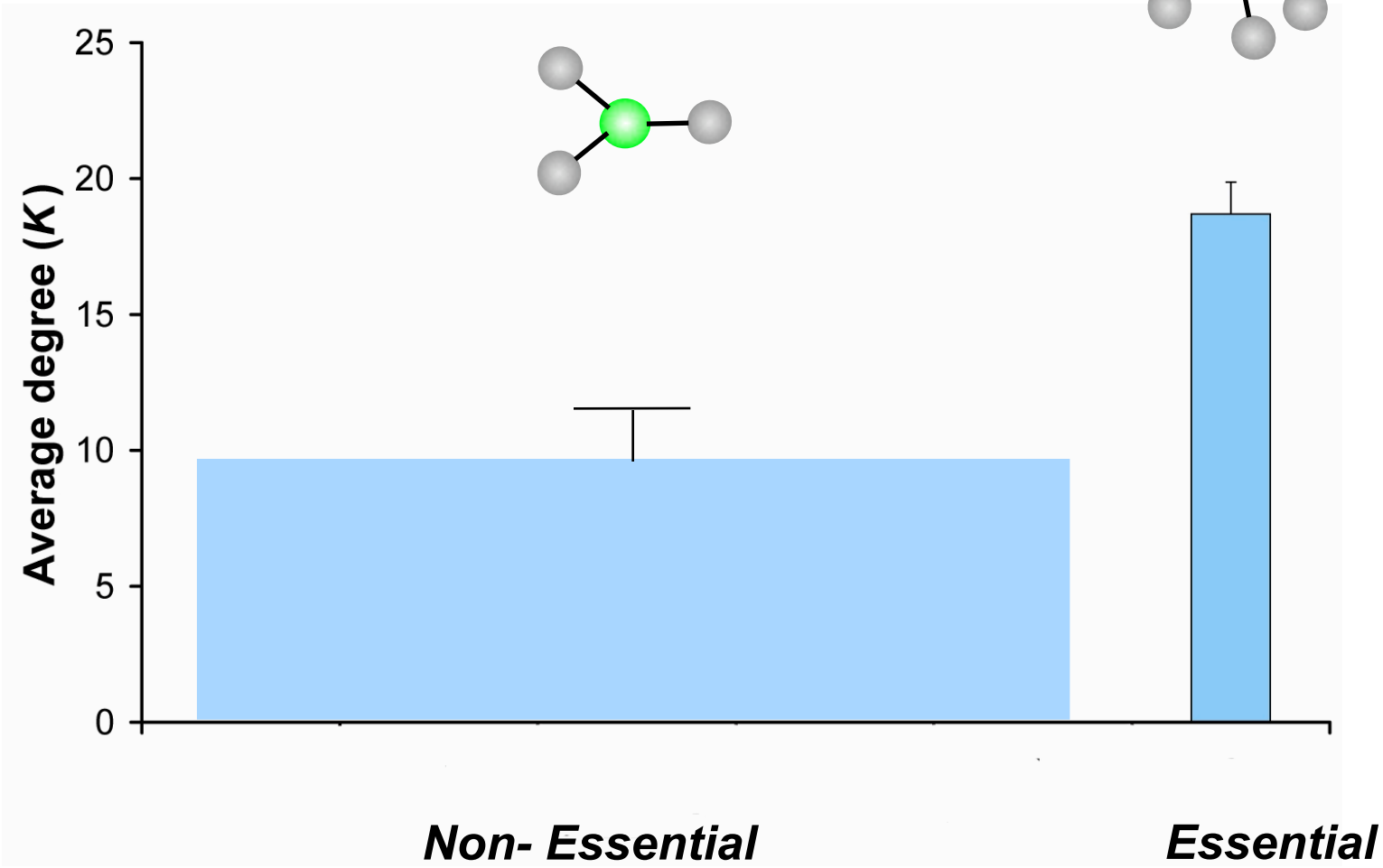
Hubs dictate the structure of the network

[Barabasi]

Hubs tend to be Essential

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

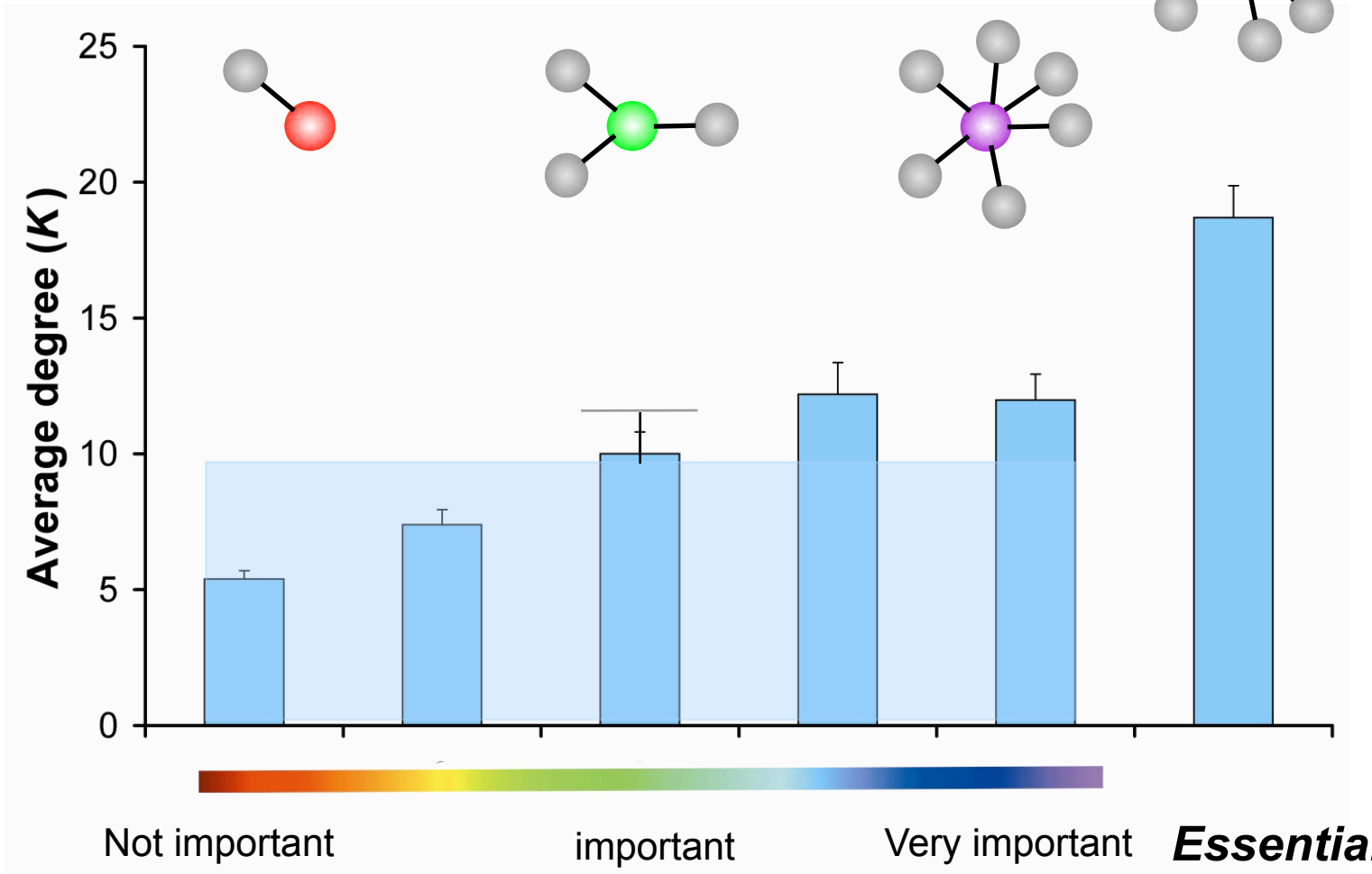
[Lauffenburger, Barabasi]



[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"



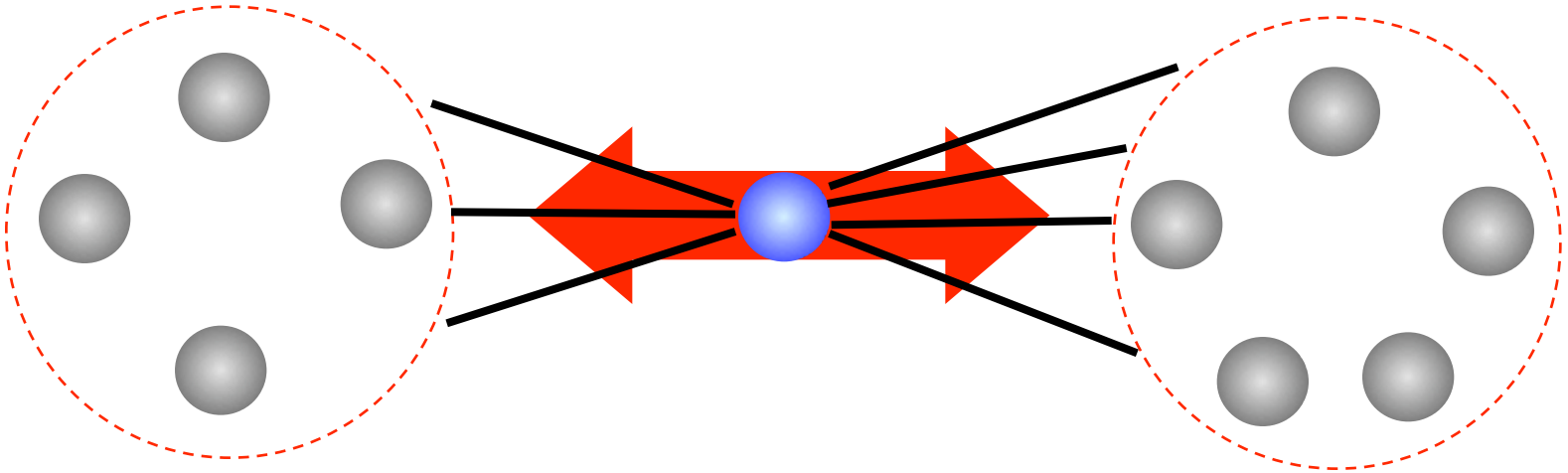
[Yu et al., 2003, TIG]

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.

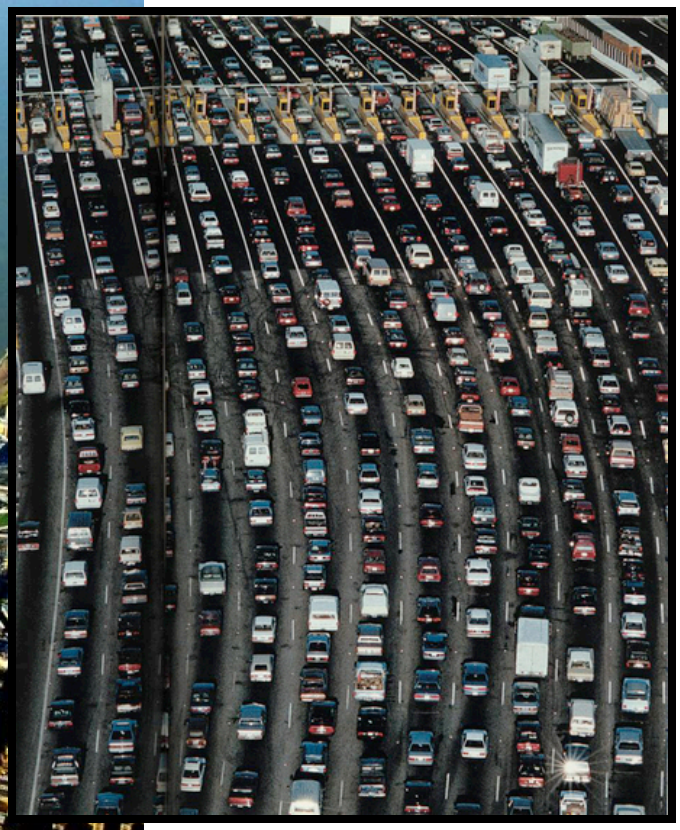
Freeman LC (1977) Set of measures of centrality based on betweenness. Sociometry 40: 35–41.

Girvan & Newman (2002) PNAS 99: 7821.

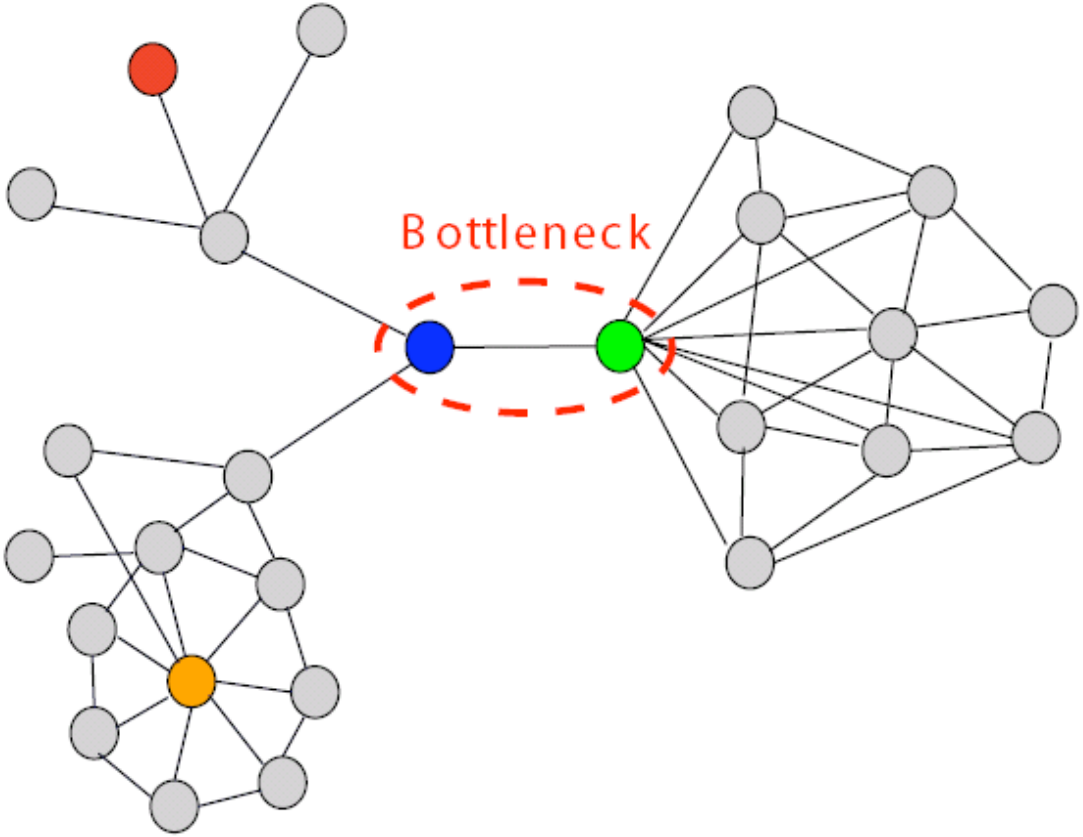


Betweenness centrality -- Bottlenecks

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



Bottlenecks & Hubs



Hub-bottleneck **node**

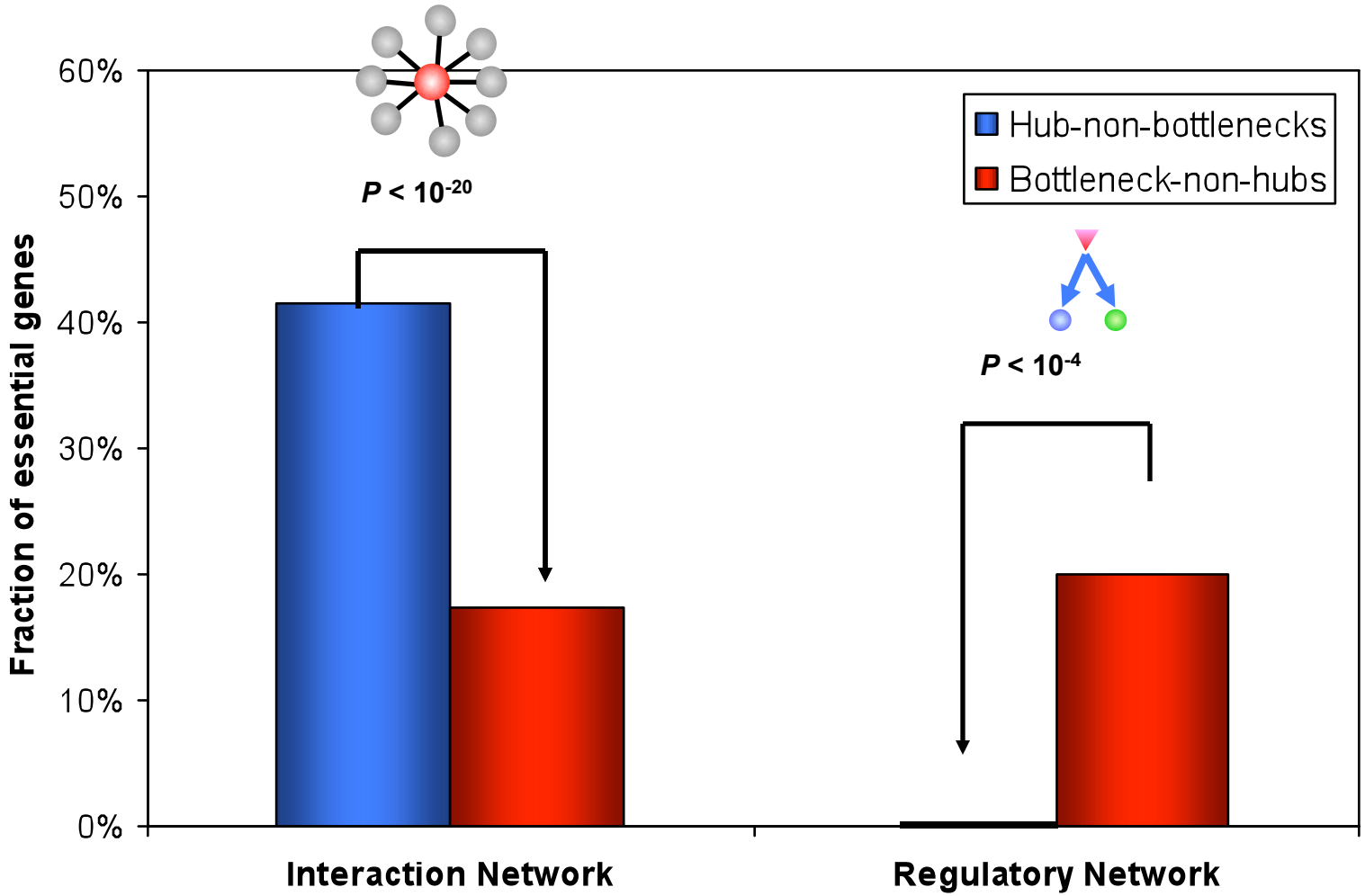
Non-hub-bottleneck **node**

Hub-non-bottleneck **node**

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

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

Bottlenecks are what matters in regulatory networks

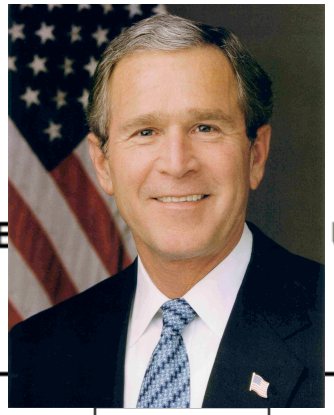


[Yu *et al.*, *PLoS Comput Biol* (2007)]

Finding Central Points in Networks #2: Tops of the Hierarchy

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

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 ESIDENT

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

ANAGEMENT AND BUDGET
ATIONAL DRUG CONTROL POLICY
OLICY DEVELOPMENT
CIENCE 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

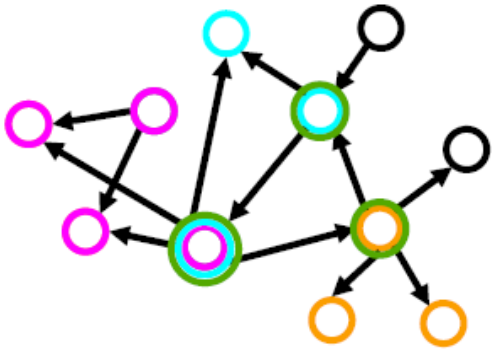
FEDERAL LABOR RELATIONS AUTHORITY
FEDERAL MARITIME COMMISSION
FEDERAL MEDIATION AND CONCILIATION SERVICE
FEDERAL MINE SAFETY AND HEALTH REVIEW COMMISSION
FEDERAL RESERVE SYSTEM
FEDERAL RETIREMENT THRIFT INVESTMENT BOARD
FEDERAL TRADE COMMISSION
GENERAL SERVICES ADMINISTRATION
INTER-AMERICAN FOUNDATION
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

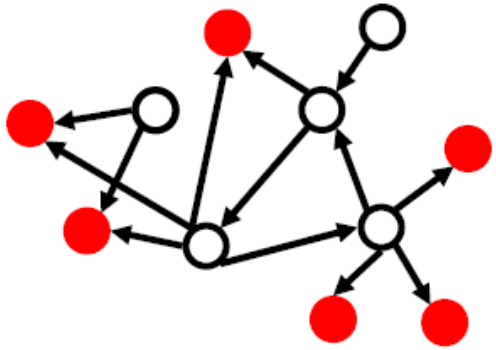
POSTAL RATE COMMISSION
RAILROAD RETIREMENT BOARD
SECURITIES AND EXCHANGE COMMISSION
SELECTIVE SERVICE SYSTEM
SMALL BUSINESS ADMINISTRATION
SOCIAL SECURITY ADMINISTRATION
TENNESSEE VALLEY AUTHORITY
TRADE AND DEVELOPMENT AGENCY
U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT
U.S. COMMISSION ON CIVIL RIGHTS
U.S.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

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

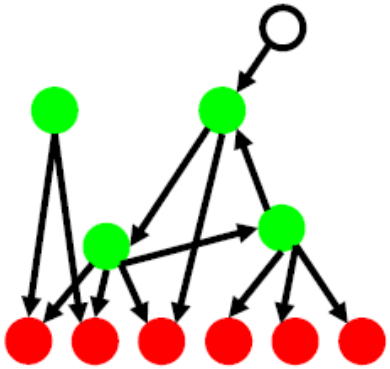
I. Example network with all 4 motifs



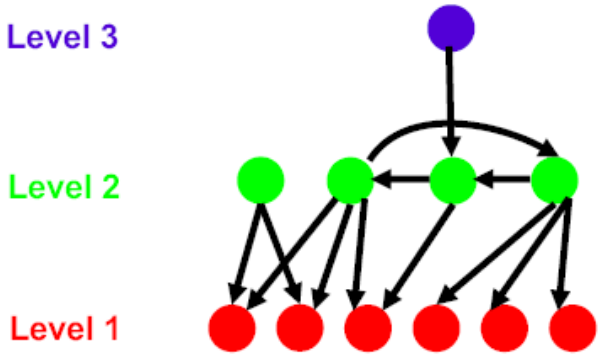
II. Finding terminal nodes (Red)



III. Finding mid-level nodes (Green)

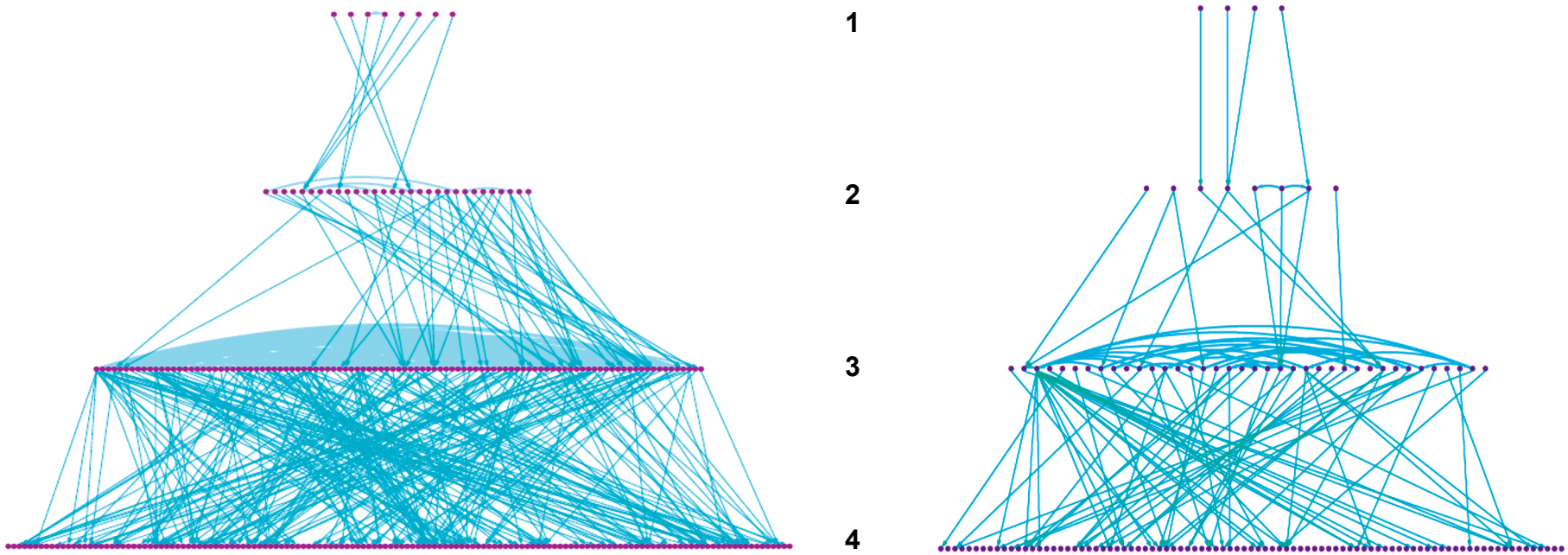


IV. Finding top-most nodes (Blue)



[Yu et al., PNAS (2006)]

Regulatory Networks have similar hierarchical structures

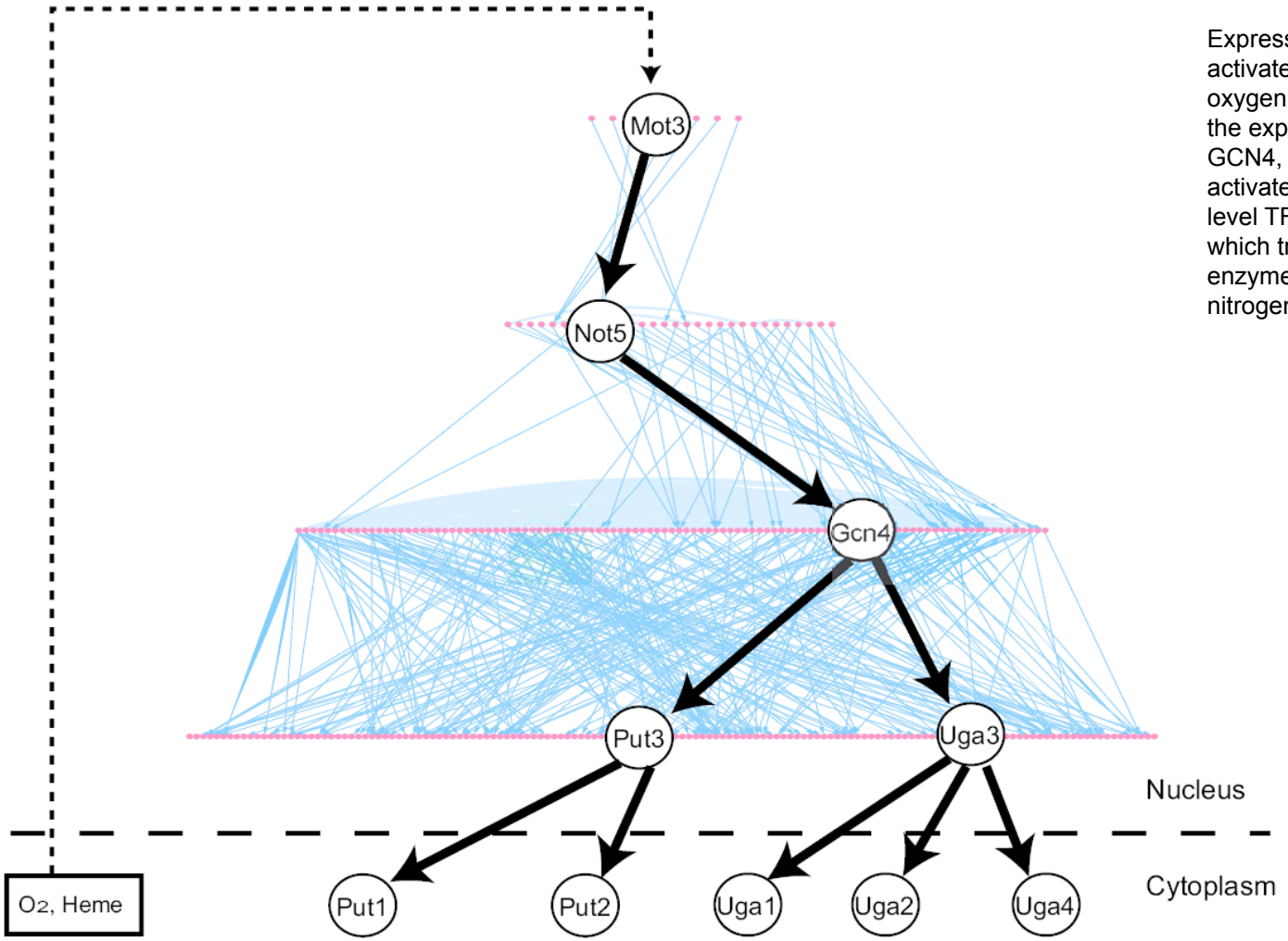


S. cerevisiae

E. coli

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

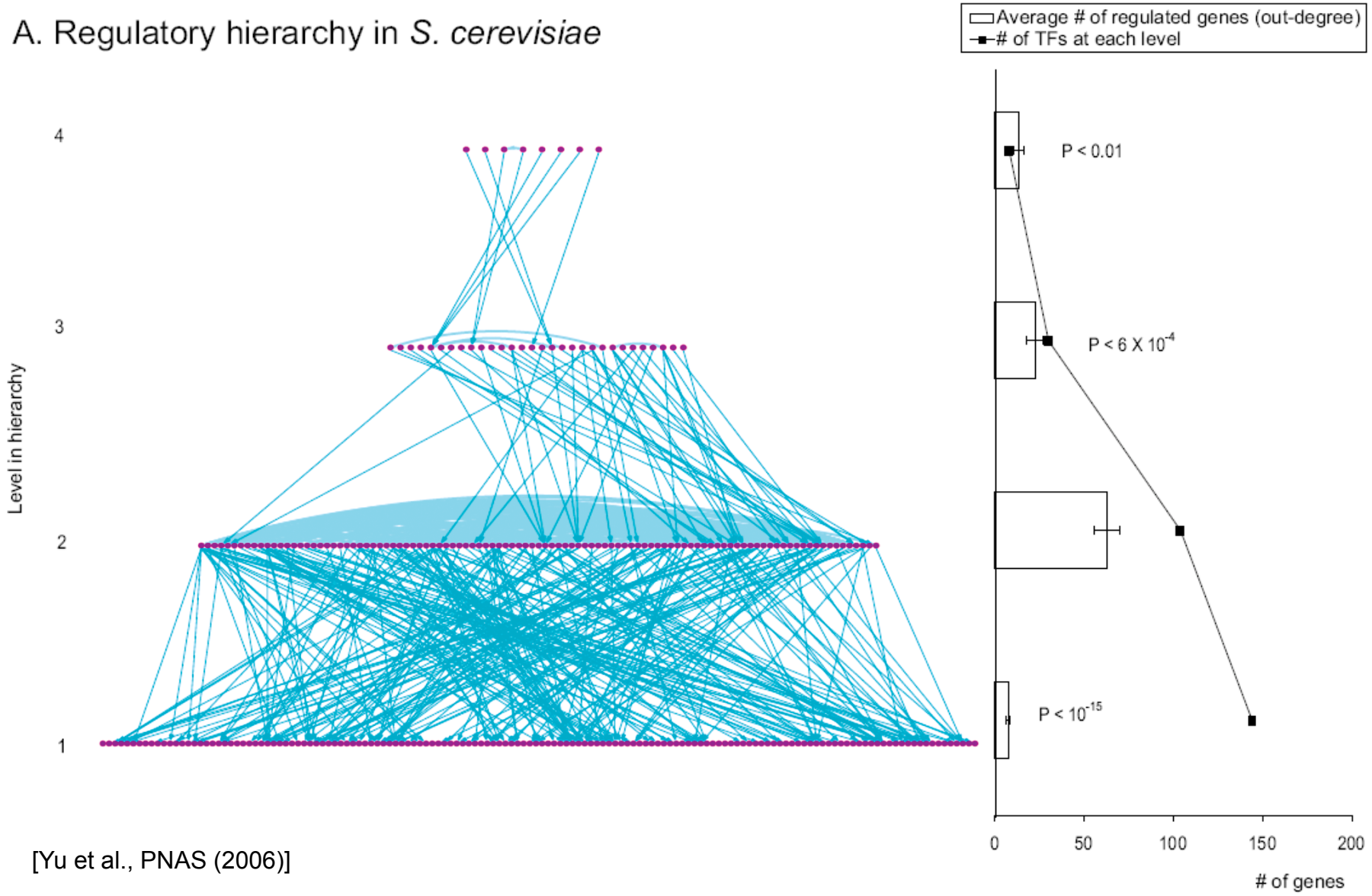
Example of Path Through Regulatory Network



Expression of MOT3 is activated by heme and oxygen. Mot3 in turn activates the expression of NOT5 and GCN4, mid-level hubs. GCN4 activates two specific bottom-level TFs, Put3 and Uga3, which trigger the expression of enzymes in proline and nitrogen utilization.

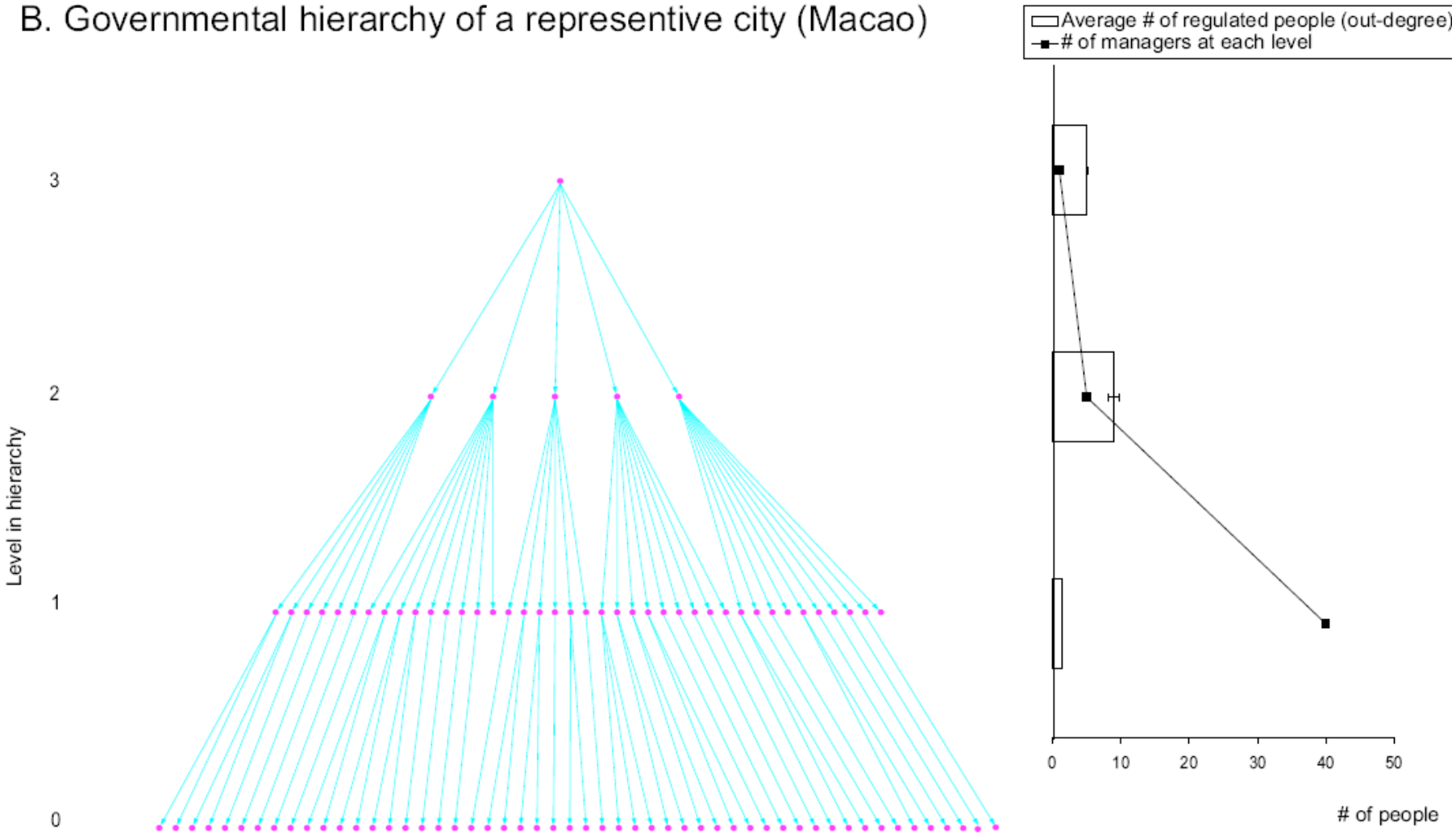
[Yu et al., PNAS (2006)]

Yeast Regulatory Hierarchy: the Middle-managers Rule

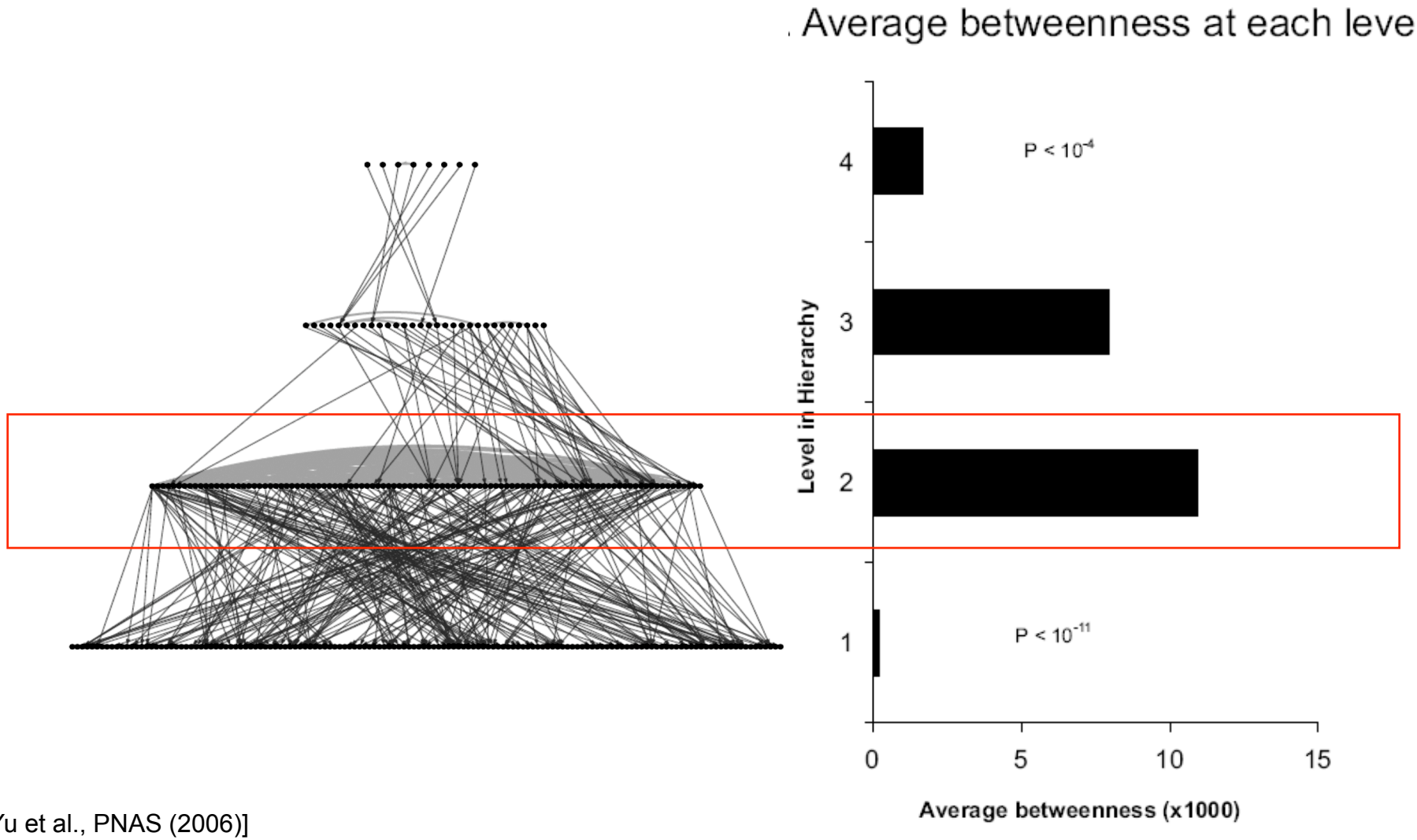


[Yu et al., PNAS (2006)]

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

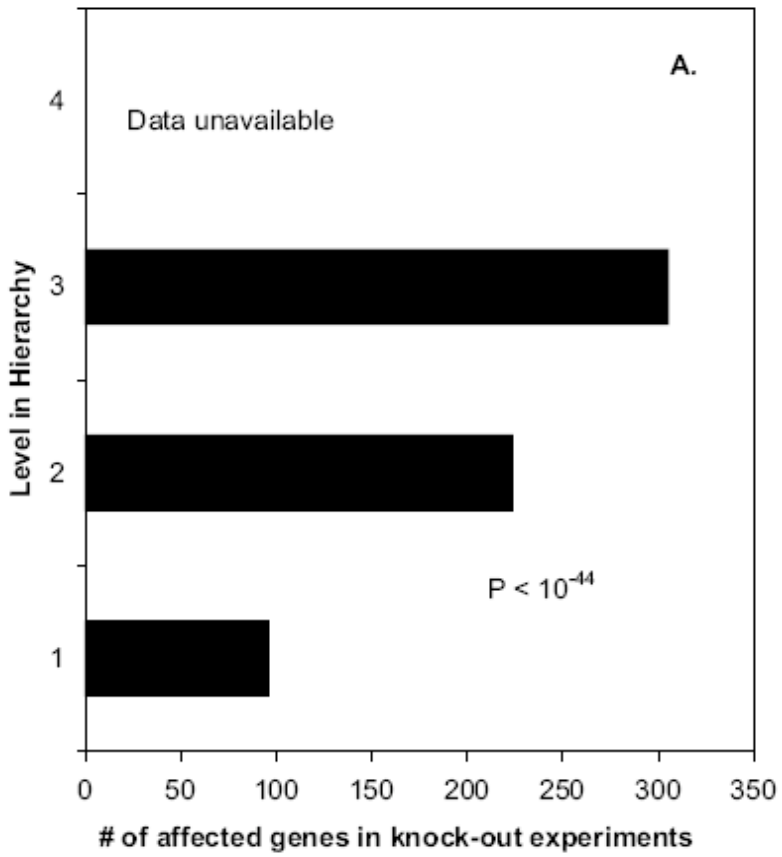


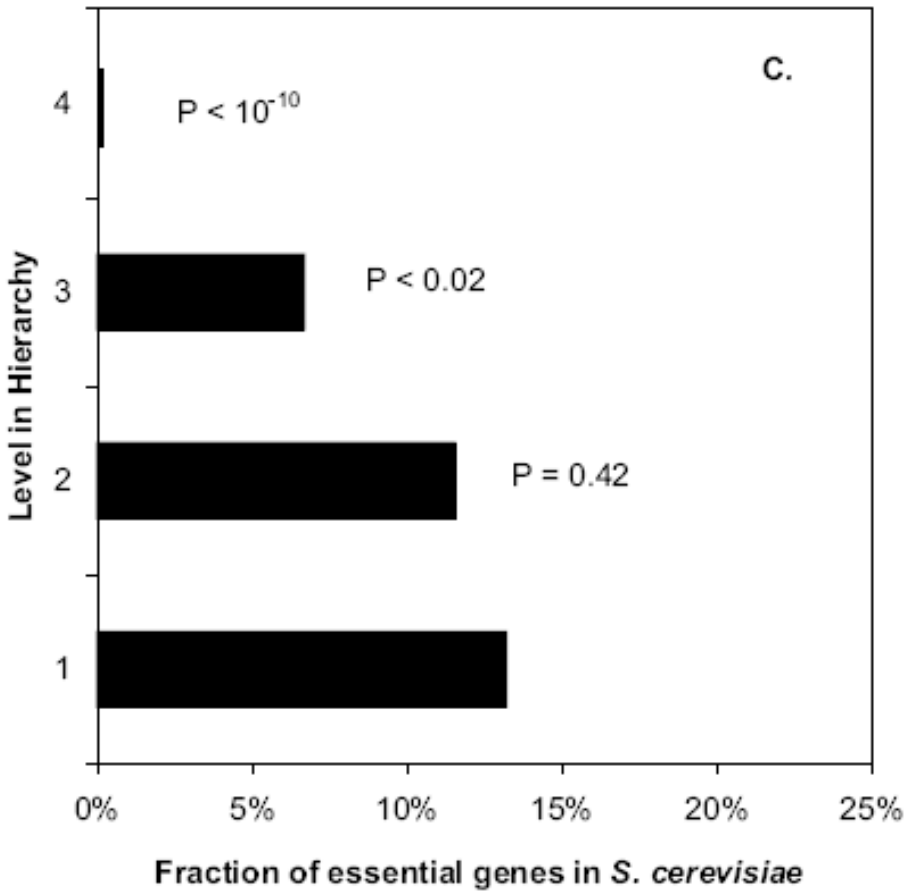
Characteristics of Regulatory Hierarchy: Middle Managers are Information Flow Bottlenecks



[Yu et al., PNAS (2006)]

Characteristics of Regulatory Hierarchy: The Paradox of Influence and Essentiality



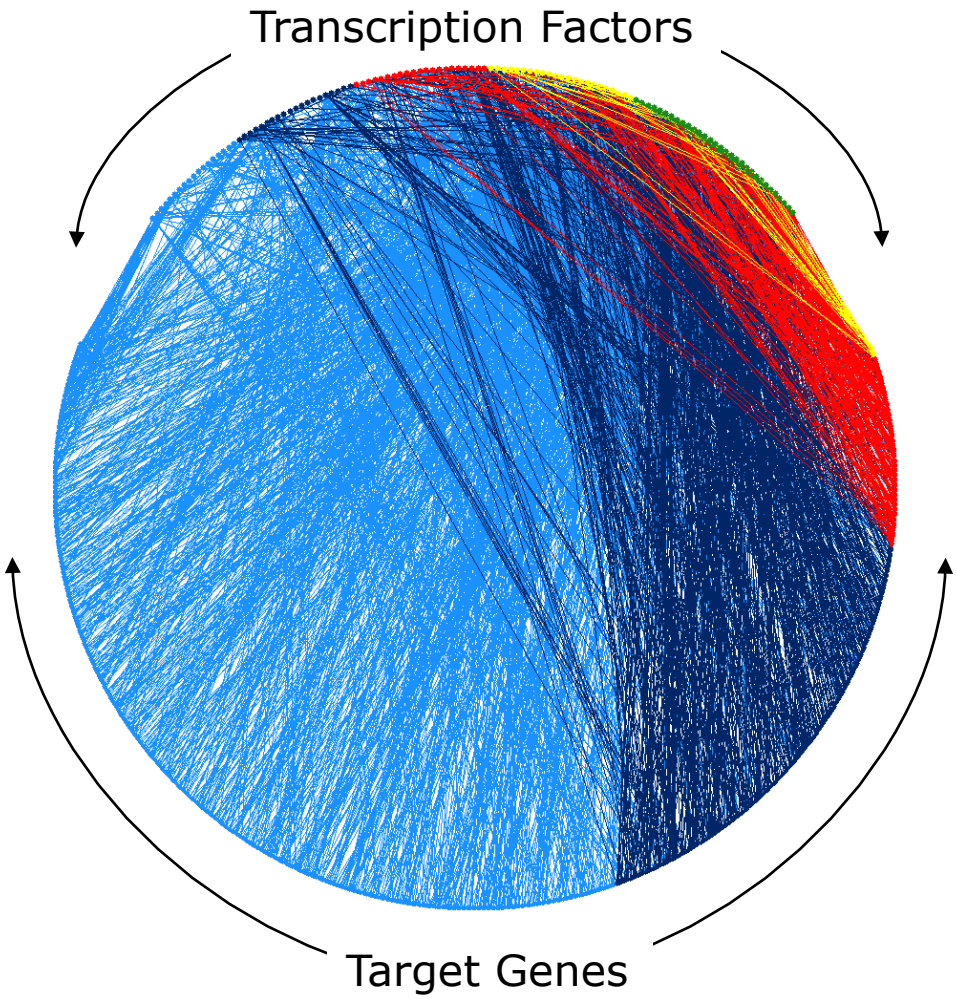


[Yu et al., PNAS (2006)]

Network Dynamics #1: Cellular States

How do networks change across different cellular states?
How can this be used to assign function to a protein?

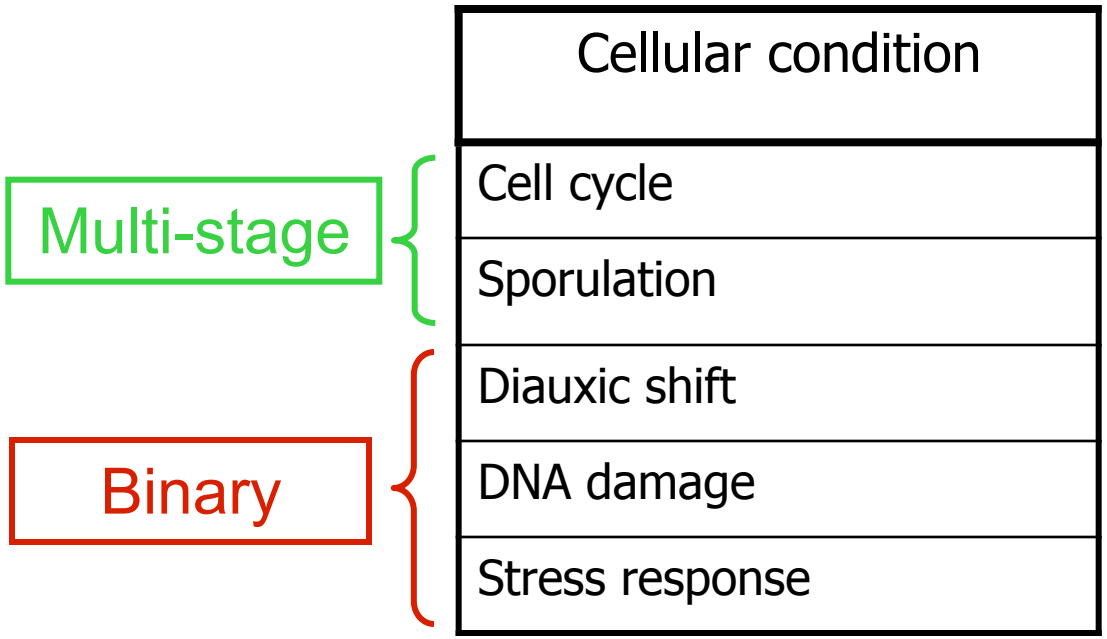
Dynamic Yeast TF network



- Analyzed network as a static entity
- But network is *dynamic*
  - ◊ Different sections of the network are active under different cellular conditions
- Integrate gene expression data

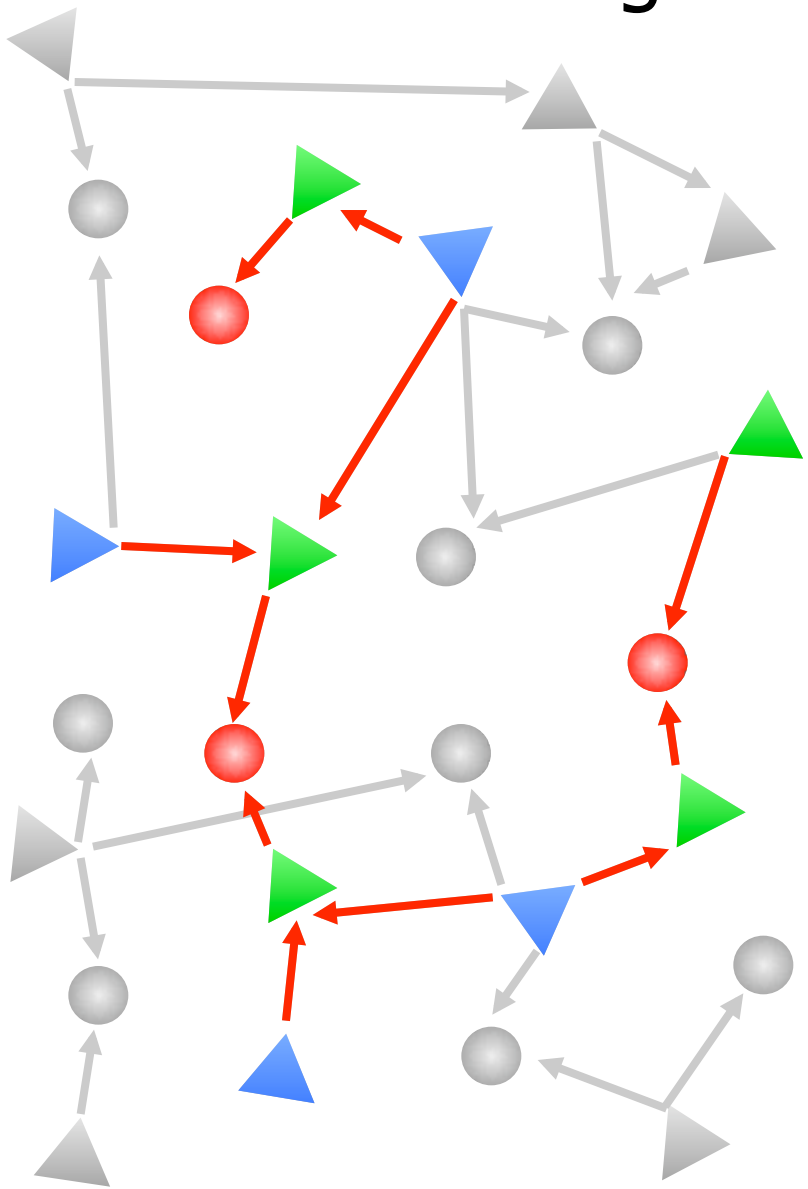
Luscombe et al. Nature 431: 308

Gene expression data for five cellular conditions in yeast



[Brown, Botstein, Davis….]

Backtracking to find active sub-network

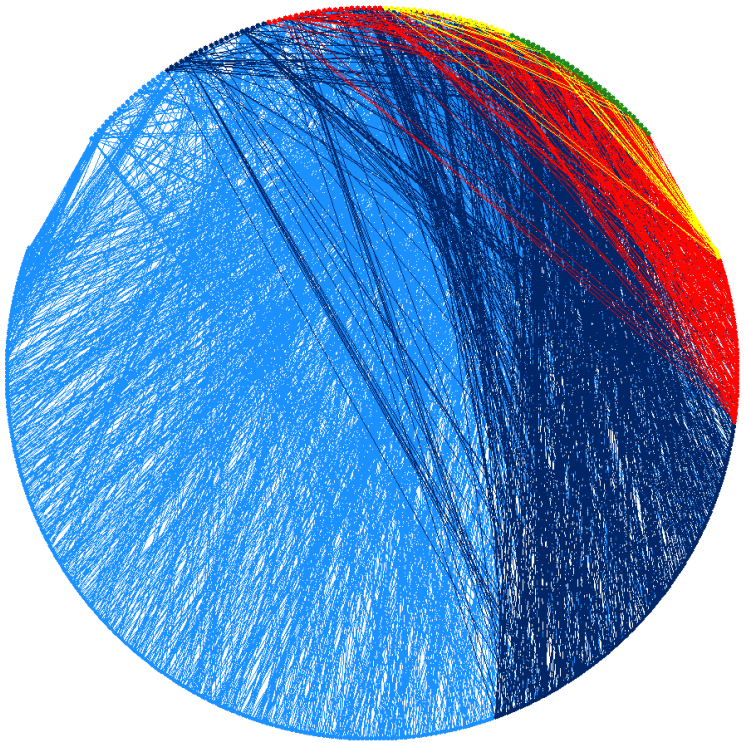


- Define differentially expressed genes
- Identify TFs that regulate these genes
- Identify further TFs that regulate these TFs

Active regulatory sub-network

Network usage under different conditions

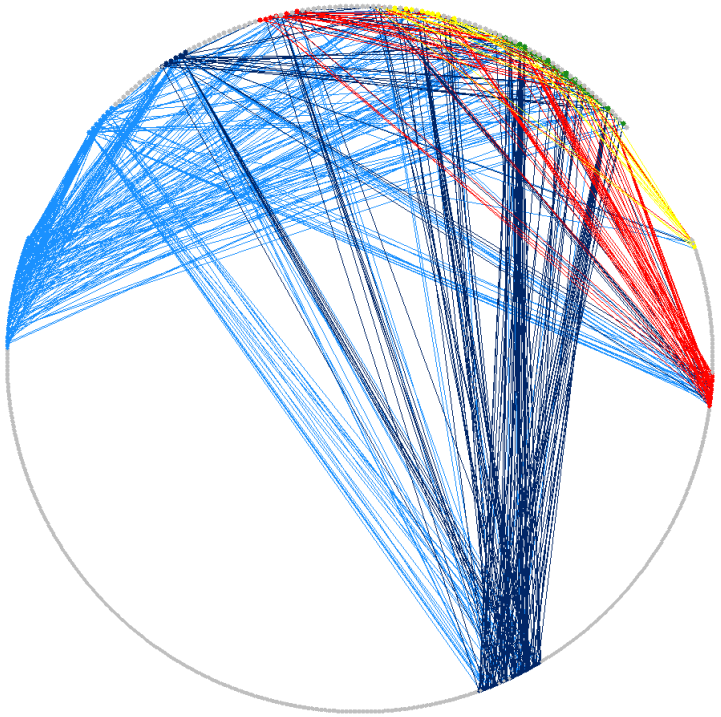
static



Luscombe et al. Nature 431: 308

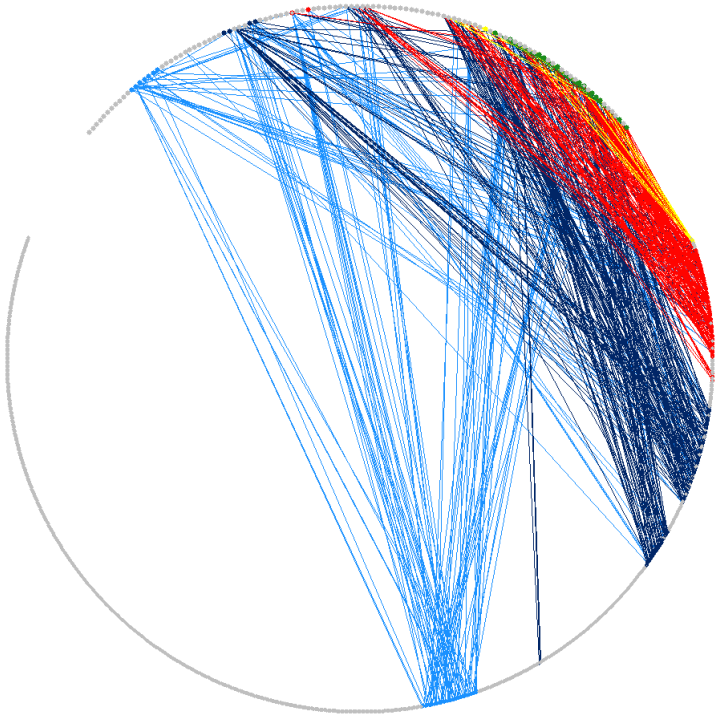
Network usage under different conditions

cell cycle



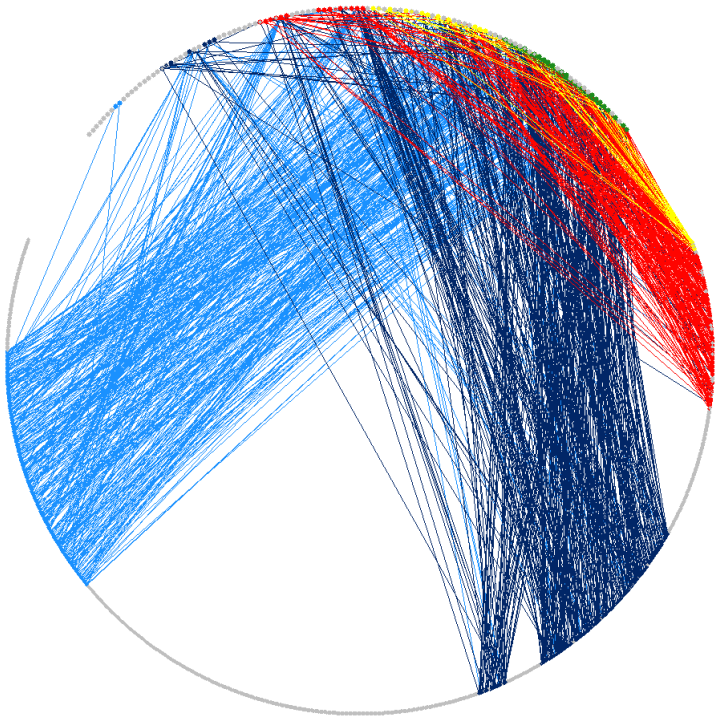
Network usage under different conditions

sporulation



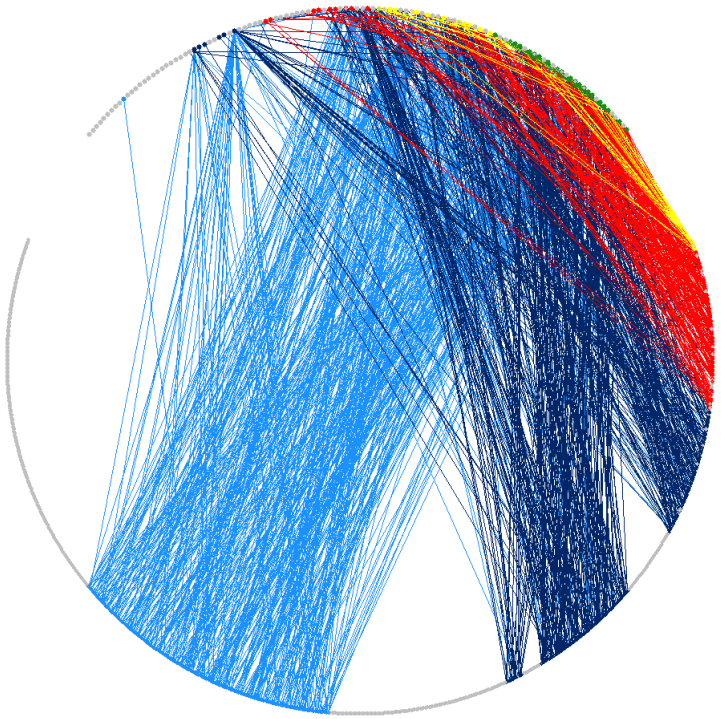
Network usage under different conditions

diauxic shift



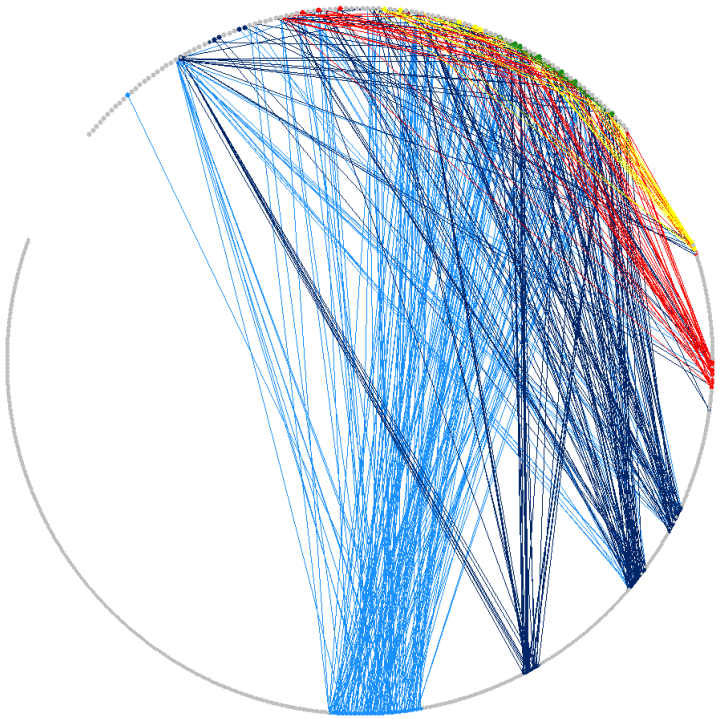
Network usage under different conditions

DNA damage



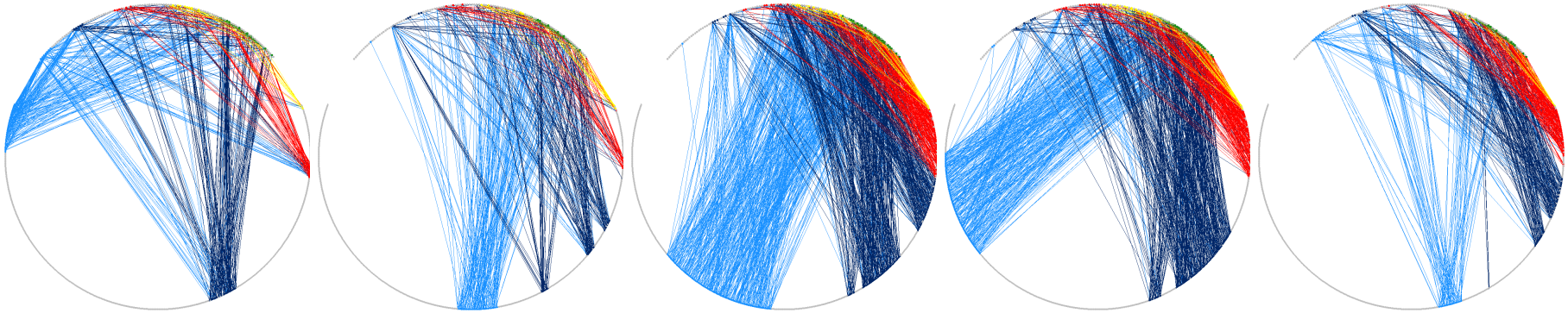
Network usage under different conditions

stress response



Network usage under different conditions

Cell cycle **Sporulation** **Diauxic shift** **DNA damage** **Stress**



SANDY:

1. Standard graph-theoretic statistics:

- Global topological measures
- Local network motifs

2. Newly derived follow-on statistics:

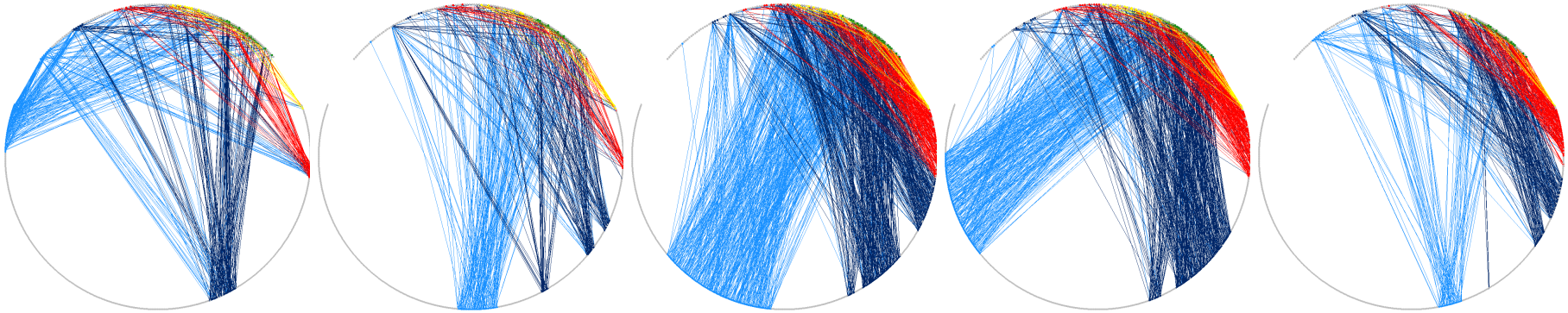
- Hub usage
- Interaction rewiring

3. Statistical validation of results

Luscombe et al. Nature 431: 308

Network usage under different conditions

Cell cycle **Sporulation** **Diauxic shift** **DNA damage** **Stress**



SANDY:

1. Standard graph-theoretic statistics:

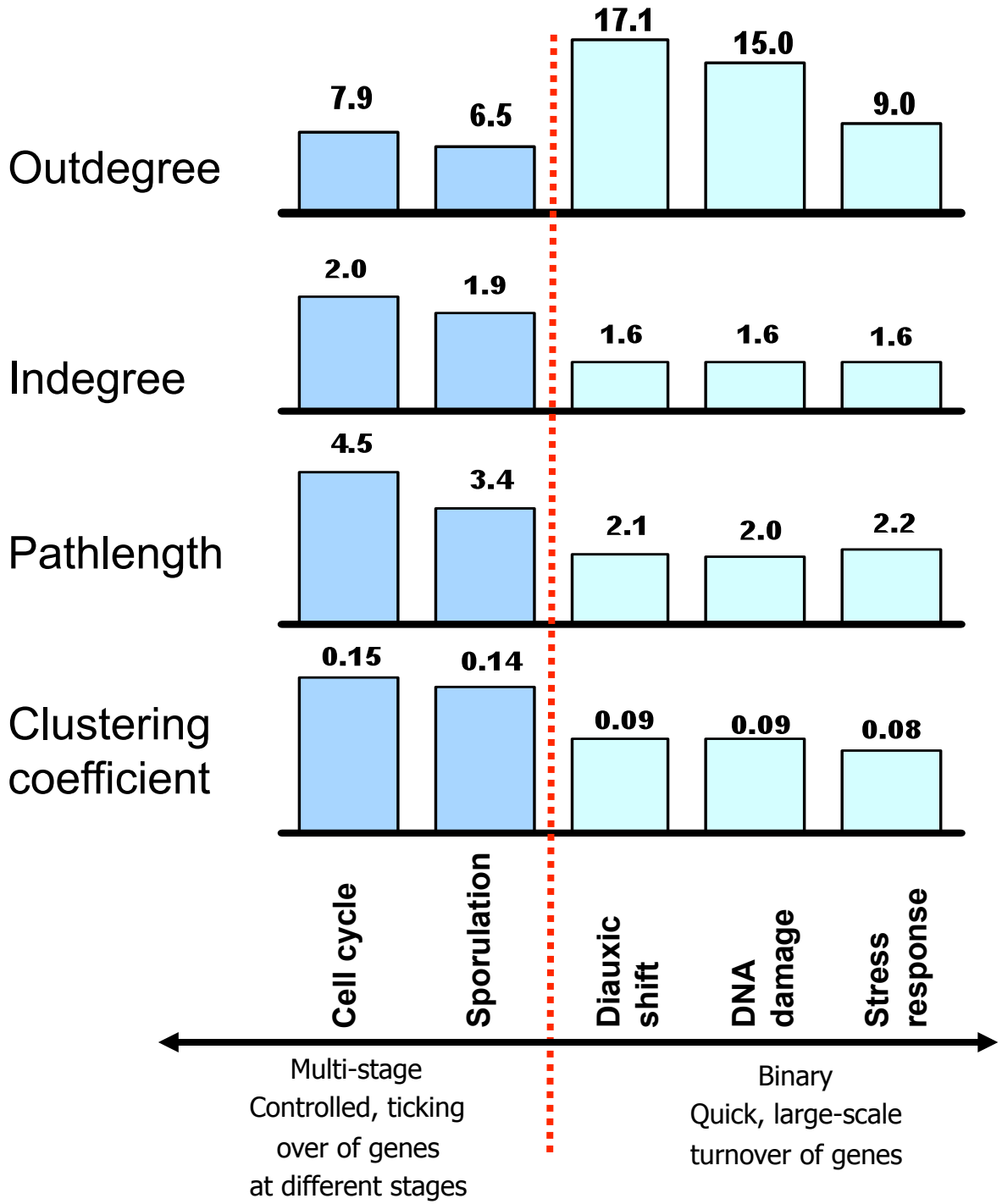
- Global topological measures
- Local network motifs

2. Newly derived follow-on statistics:

- Hub usage
- Interaction rewiring

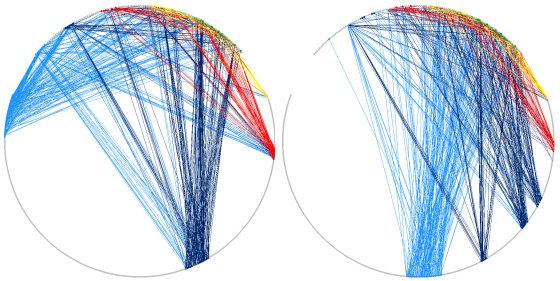
3. Statistical validation of results

Analysis of condition-specific subnetworks in terms of global topological statistics

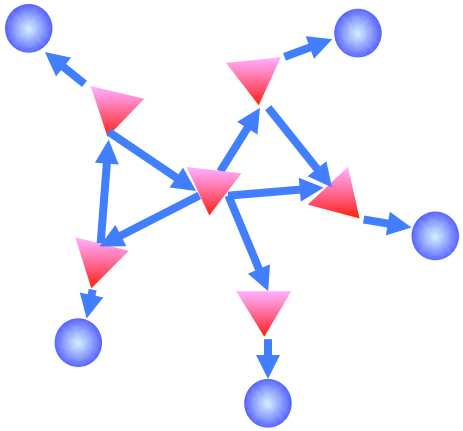


Luscombe et al. Nature 431: 308

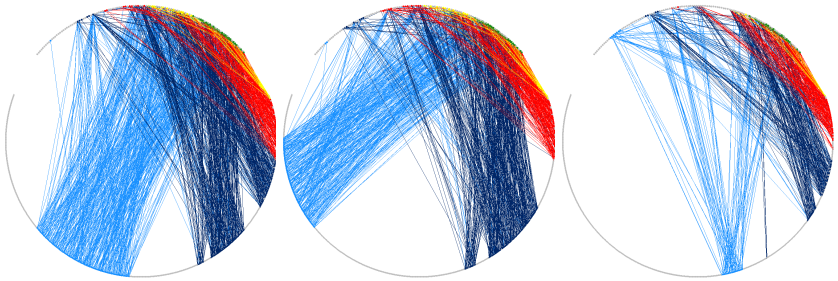
Cell cycle Sporulation



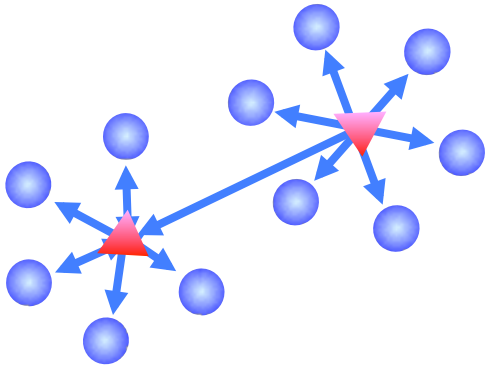
multi-stage conditions



Diauxic shift DNA damage Stress



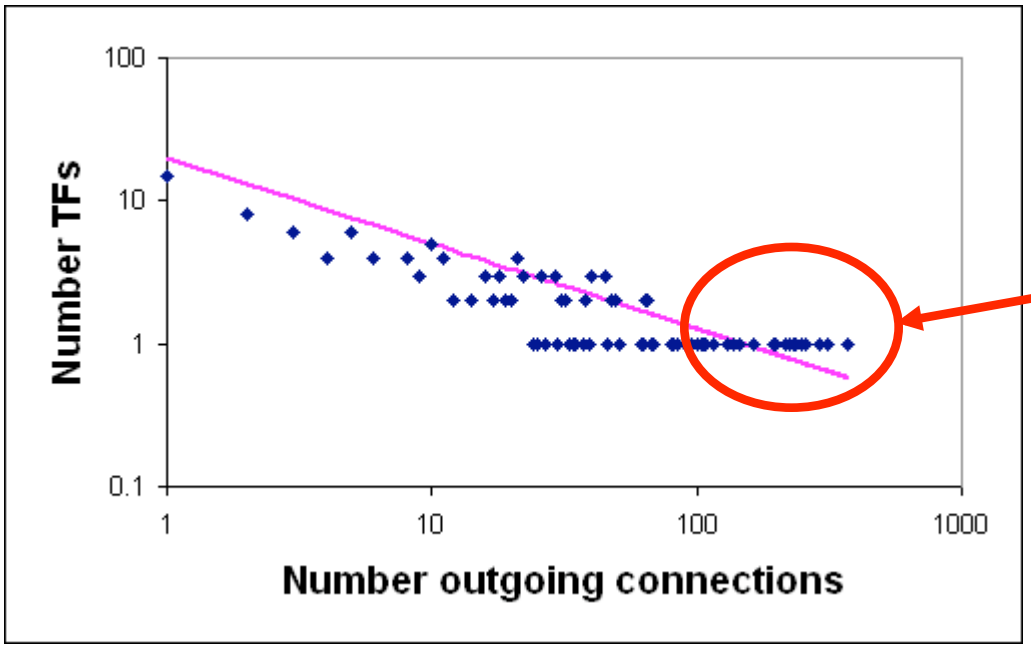
binary conditions

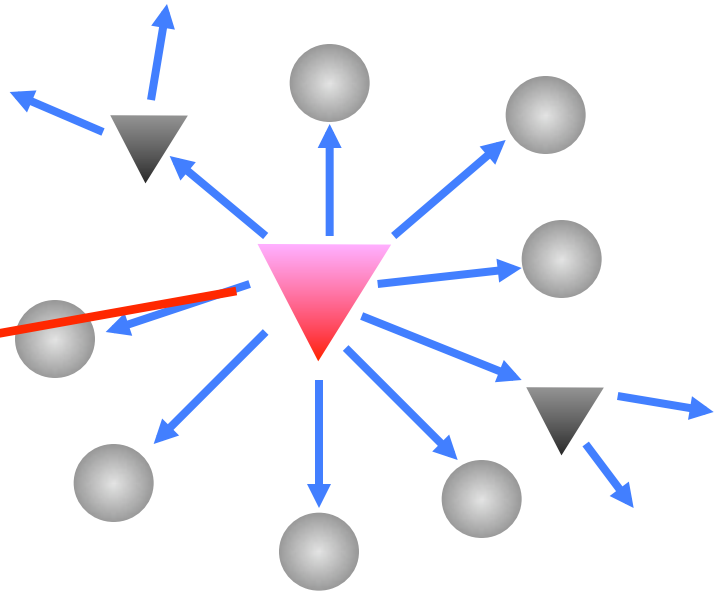


Summary

less pronounced	Hubs	more pronounced
longer	Path Lengths	shorter
more	TF inter-regulation	less
complex loops (FFLs)	Motifs	simpler (SIMs)

Transient Hubs

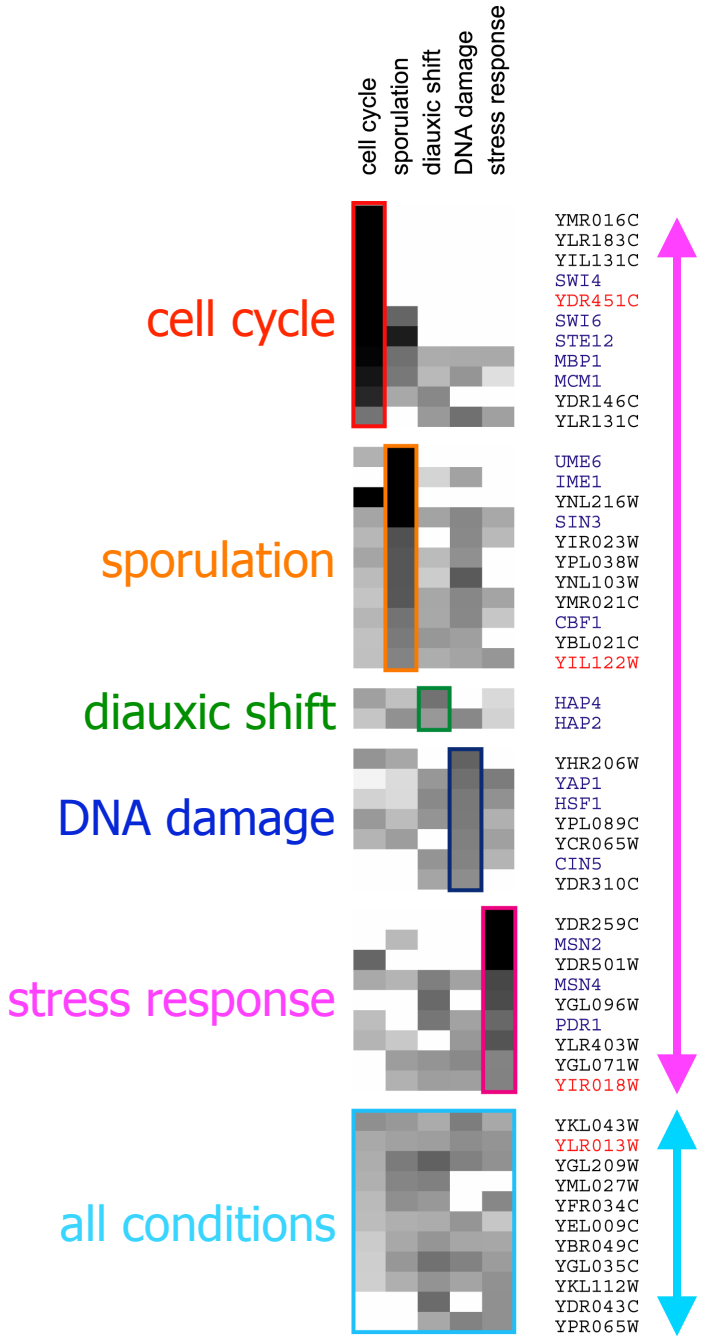




Regulatory hubs

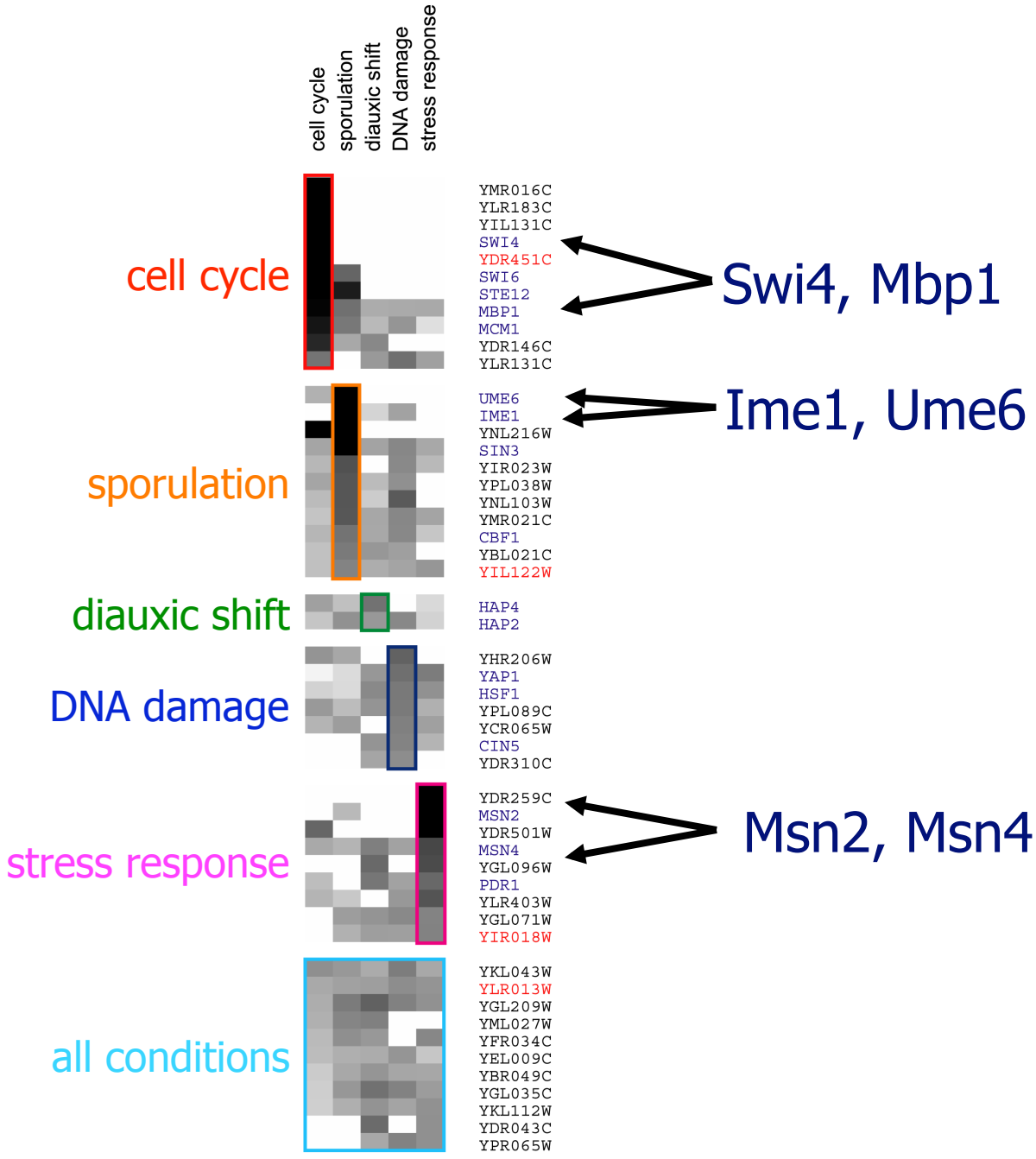
- Questions:
  - ◊ Do hubs stay the same or do they change over between conditions?
  - ◊ Do different TFs become important?
- Our Expectations
  - ◊ Literature:
    - Hubs are permanent features of the network regardless of condition
  - ◊ Random networks (sampled from complete regulatory network)
    - Random networks converge on same TFs
    - 76-97% overlap in TFs classified as hubs (*ie* hubs are permanent)

Luscombe et al. Nature 431: 308

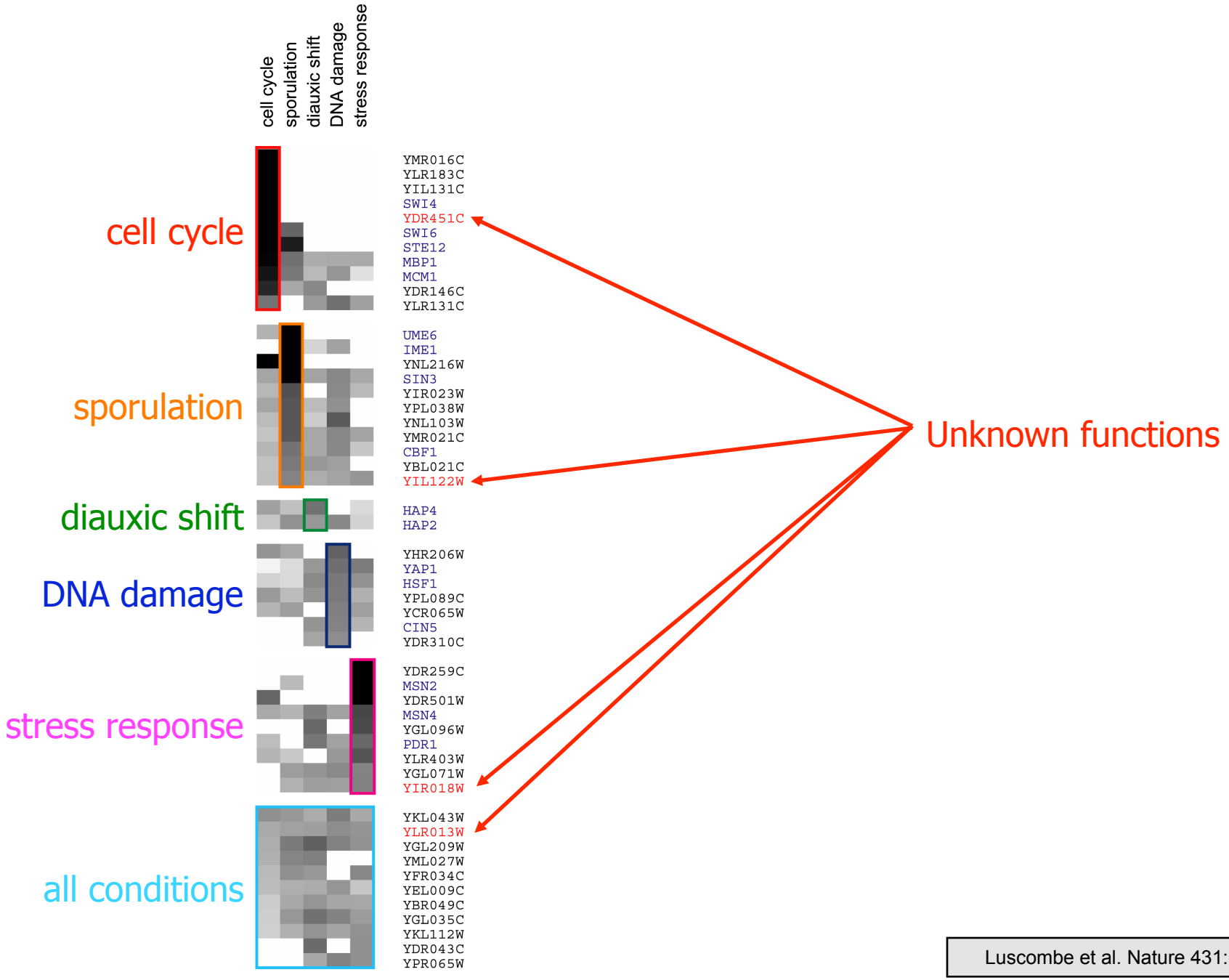


- Some permanent hubs
  - ◊ house-keeping functions
- Most are transient hubs
  - ◊ Different TFs become key regulators in the network
- Implications for condition-dependent vulnerability of network

Luscombe et al. Nature 431: 308



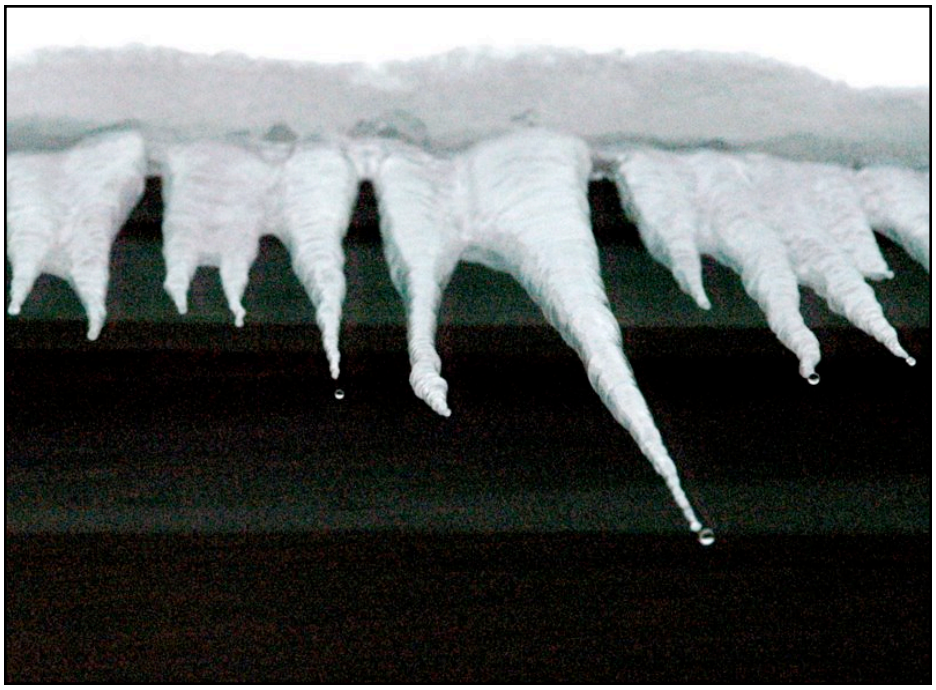
Luscombe et al. Nature 431: 308



Luscombe et al. Nature 431: 308

Network Dynamics #2: Environments

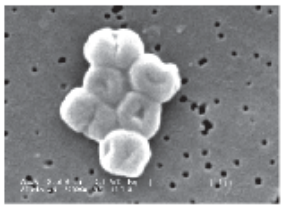
How do molecular networks change across environments?
What pathways are used more ?
Used as a biosensor ?



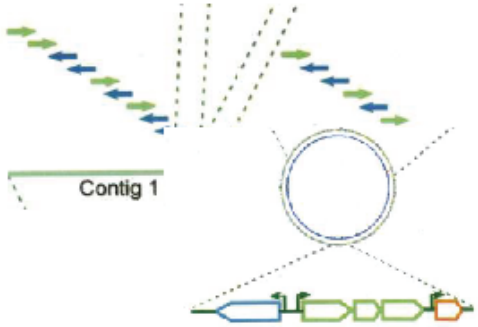
What is metagenomics?

Genomics Approach

Culture Microbes → Extract DNA → Sequence → Assemble and Annotate



ATCGTATA
CGCGAAG
ACGTCTGA
AGTGCTGCT

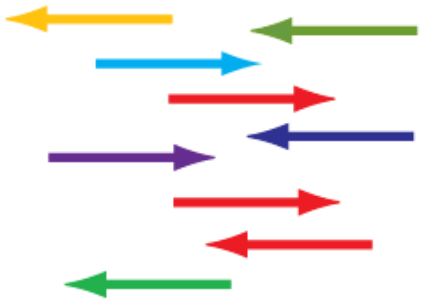


PROBLEM: Estimated that less than 1% can be cultured in the lab

Metagenomics Approach

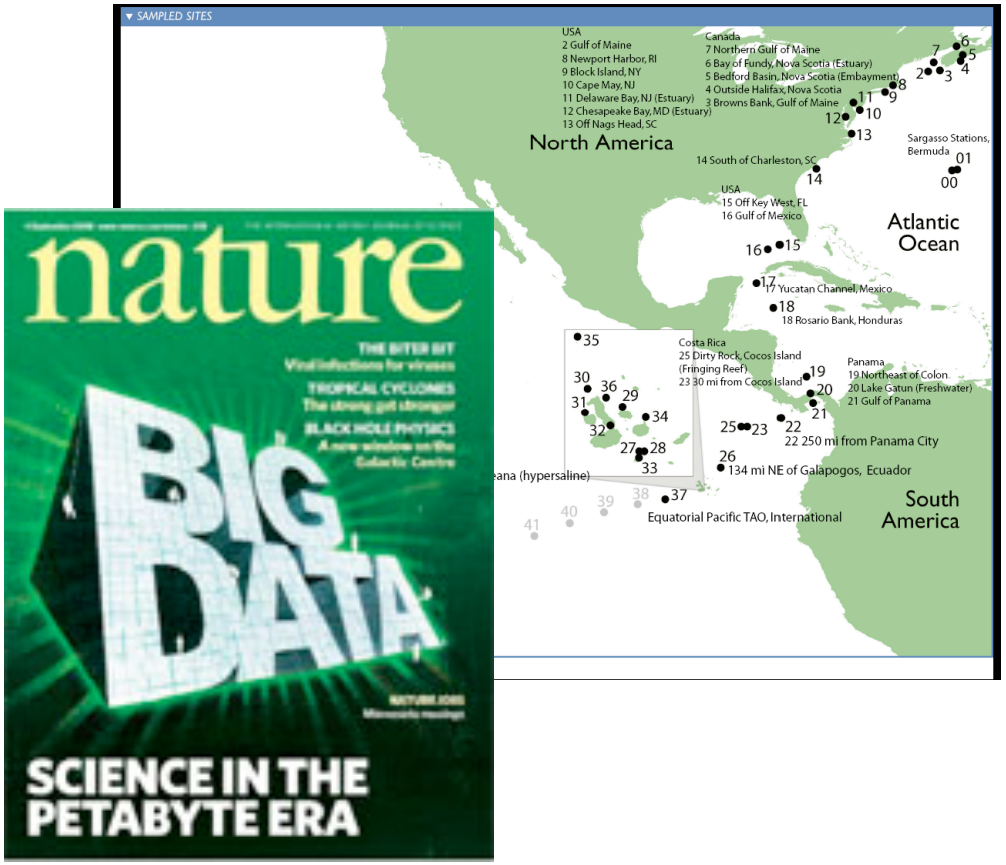
Collect Sample → Extract DNA → Sequence → Partially Assemble and Annotate

ATCGTGATAGATGATAGTAGA
ATGCTGCATGCATCTAGCACT
ACAGTAGCTAGCTACGTACTA
CAGCTGACTAGCTAGCTAGCT
ACGTAGCATGCTAGCTAGCAG
ACGTACGTAGCTAGCTAGCTAG
ACGTACGTACGTAGCTAGCATC
AGTCGACTGAGCCAGTGATGAT
ACGATGCATGAGCAGATGCTAC
AGATCGTAGCATGCTAGCATGCT
ACGTACGTAGCTAGCTAGCTAAG
AGCTAGCATGCTAGTAGCATGAG
ACGATGCTAGCTAGCTAGCTGATA
TCGATCAGCATGCTACGATGCAAG
ACGATCGATGCTAGCTAGCTAGCAT
AGCTAGCTAGTCAGCTAGCTAGATG



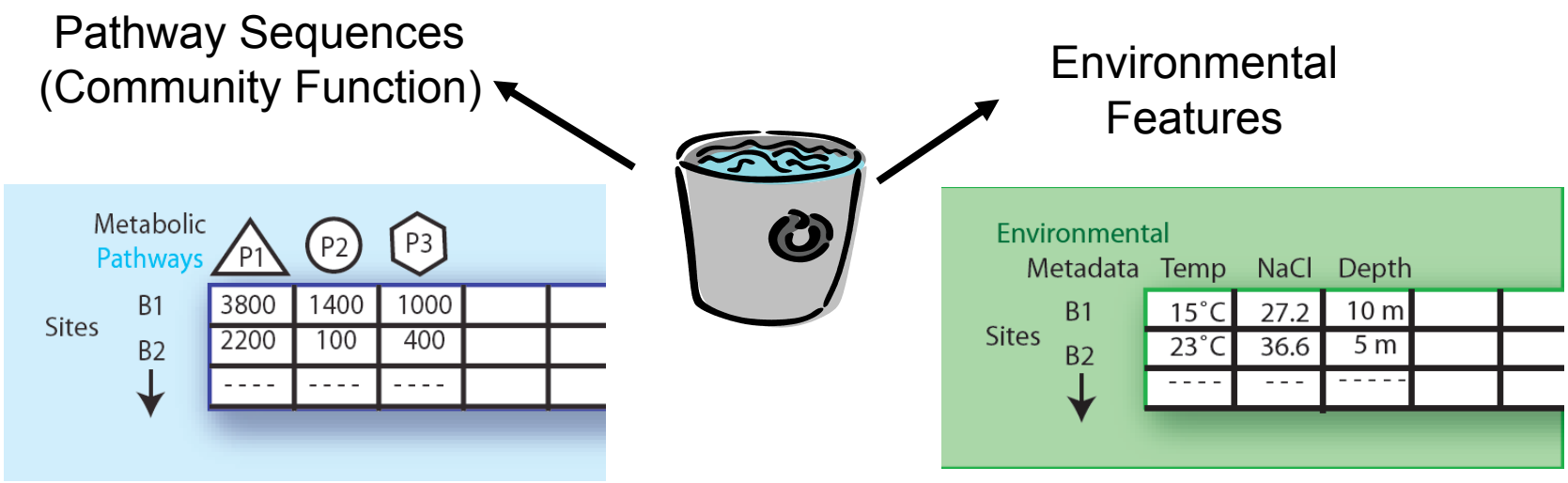
PROBLEM: Lose information about which gene belongs to which microbe.

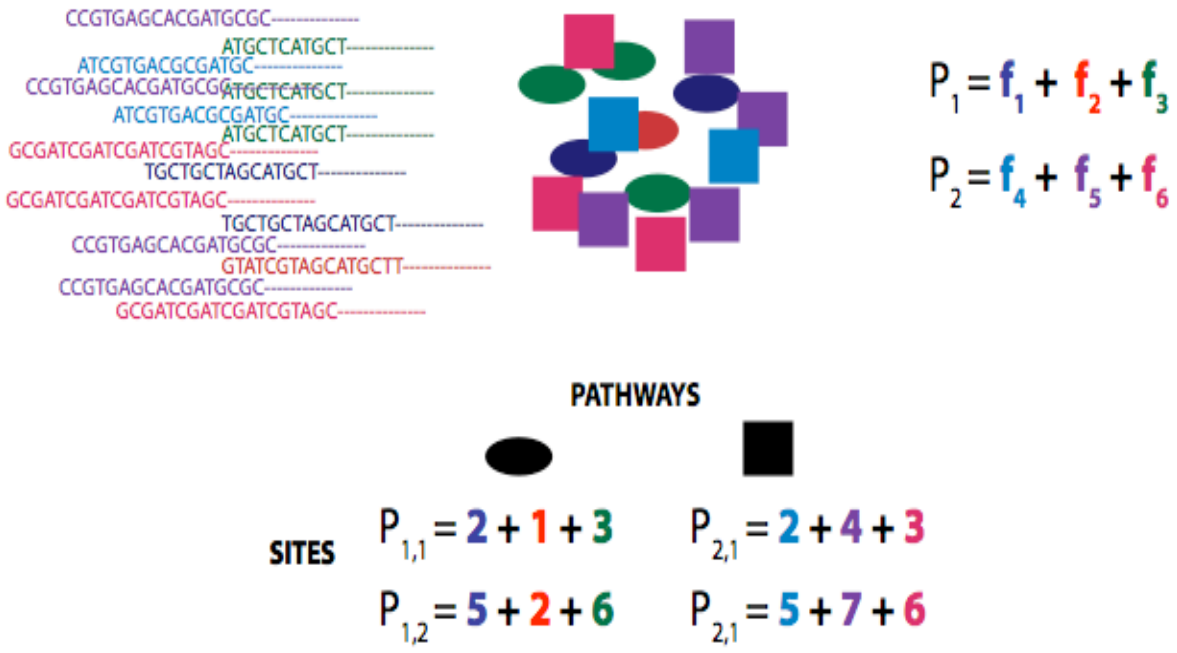
Global Ocean Survey Statistics (GOS)



6.25 GB of data
7.7M Reads
1 million CPU hours
to process

Rusch, et al., PLOS Biology 2007



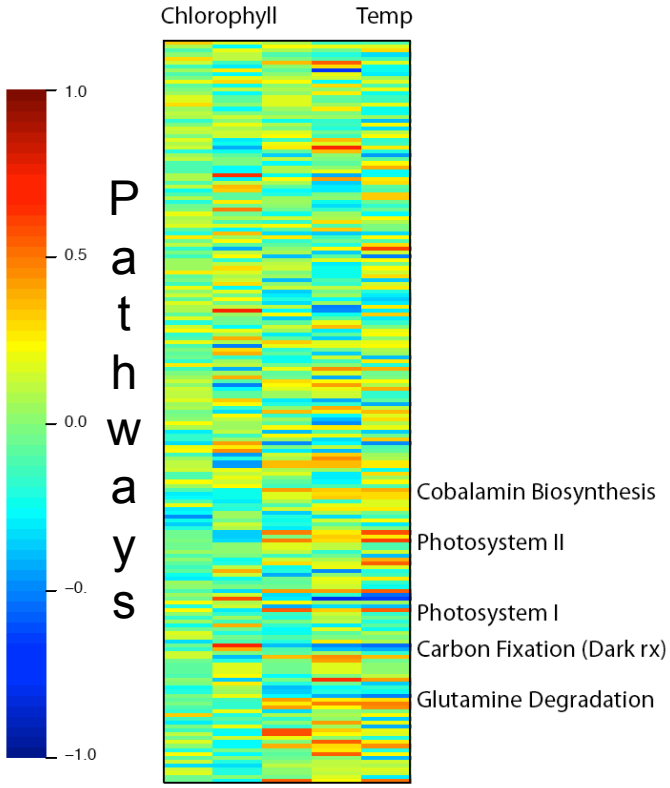


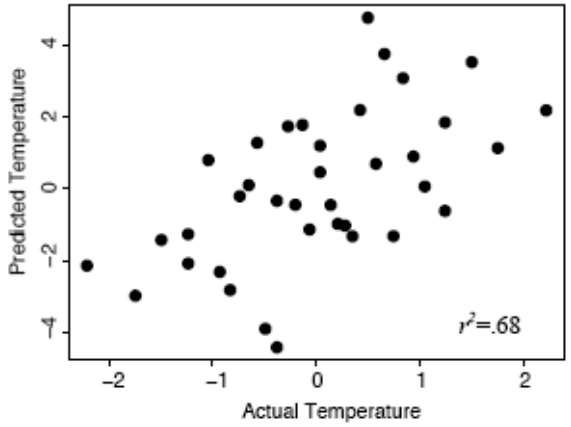
Expressing data as matrices indexed by site, env. var., and pathway usage

[Rusch et. al., (2007) PLOS Biology; Gianoulis et al., PNAS (in press, 2009]

Simple Relationships: Pairwise Correlations

Environmental Features





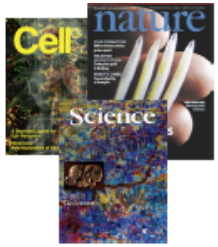
[Gianoulis et al., PNAS (in press, 2009)]

Canonical Correlation Analysis: Simultaneous weighting

Score	# of papers published
GRE	

Undergraduate Performance Index (UPI)	Graduate School Performance Index (GPI)
GRE GPA 	

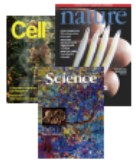
UPI = a GRE + b  + c GPA

GPI = a'  + b'  + c' 

[Gianoulis et al., PNAS (in press, 2009)]

Canonical Correlation Analysis: Simultaneous weighting

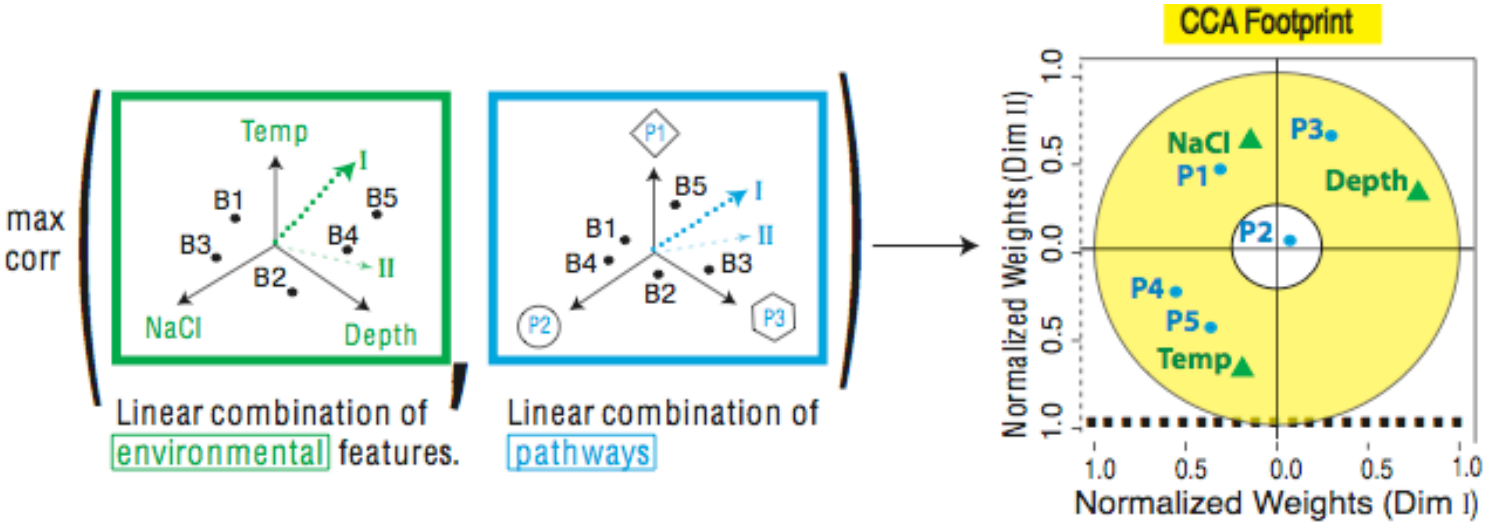
Score	# of papers published
GRE	

Undergraduate Performance Index (UPI)	Graduate School Performance Index (GPI)
GRE GPA 	  

Environmental Features	Metabolic Pathways
Temp etc	Photosynthesis etc
Chlorophyll	Lipid Metabolism

[Gianoulis et al., PNAS (in press, 2009)]

Environmental-Metabolic Space



The goal of this technique is to interpret cross-variance matrices
We do this by defining a change of basis.

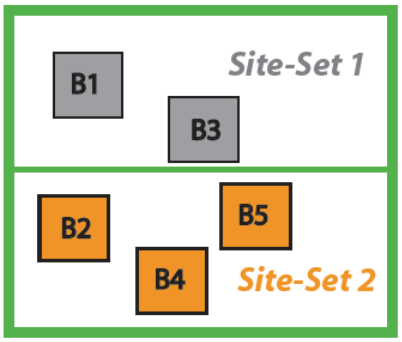
$$\text{Given } X=\{x_1,x_2,\ldots,x_n\} \text{ and } Y=\{y_1,y_2,\ldots,y_m\}$$

$$C=\begin{matrix}\Sigma_X & \Sigma_{X,Y}\\ \Sigma_Y & \Sigma_{Y,X}\end{matrix} \qquad \max_{a,b} Corr(U,V)=\frac{a'\Sigma_{12}b}{\sqrt{a'\Sigma_{11}a}\sqrt{b'\Sigma_{22}b}}$$

[Gianoulis et al., PNAS (in press, 2009)]

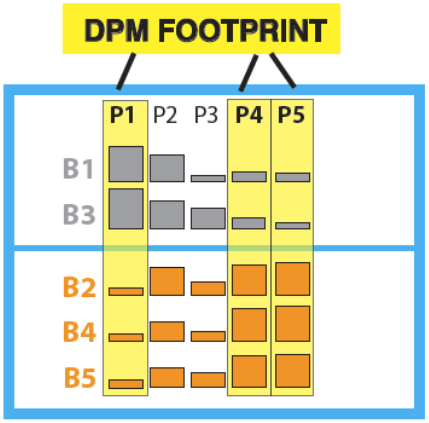
DPM: Discriminative Partition Matching

Environment



Cluster (Partition)

Metabolism

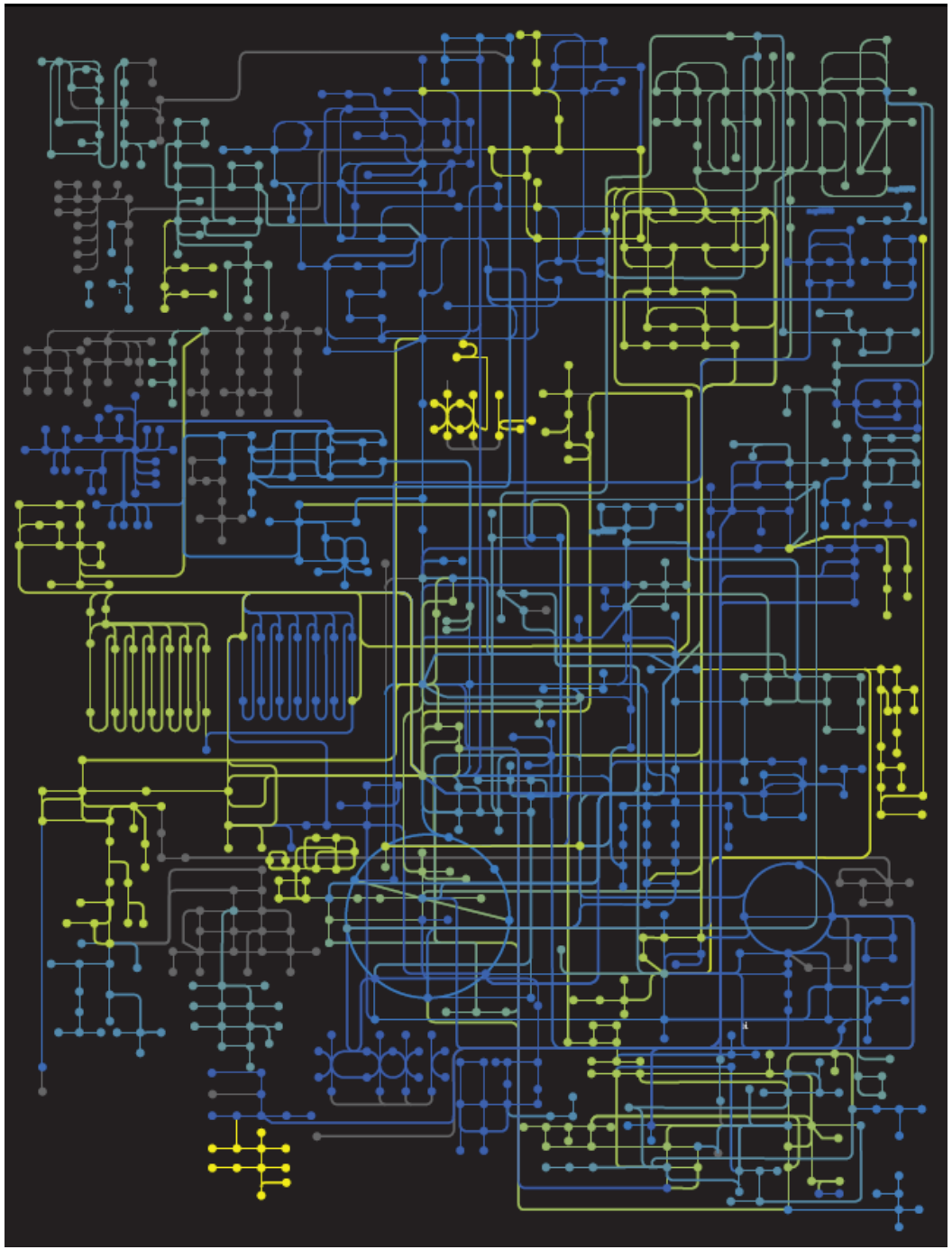


Test

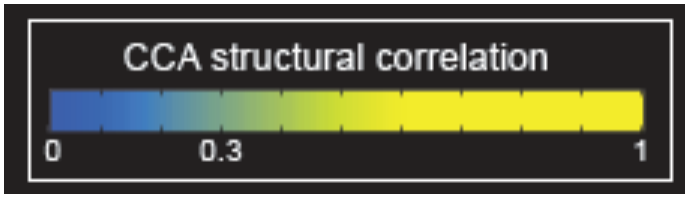
Taurine biosynthesis
Heme biosynthesis
Asparagine degradation
Nitrogen fixation
Acylglycerol degradation
Asparagine biosynthesis
Cysteine Metabolism

Functional class	pval
InfoStorage & Processing	.07
Cellular Process	.08
Metabolism	4x10-14

[Gianoulis et al., PNAS (in press, 2009)]

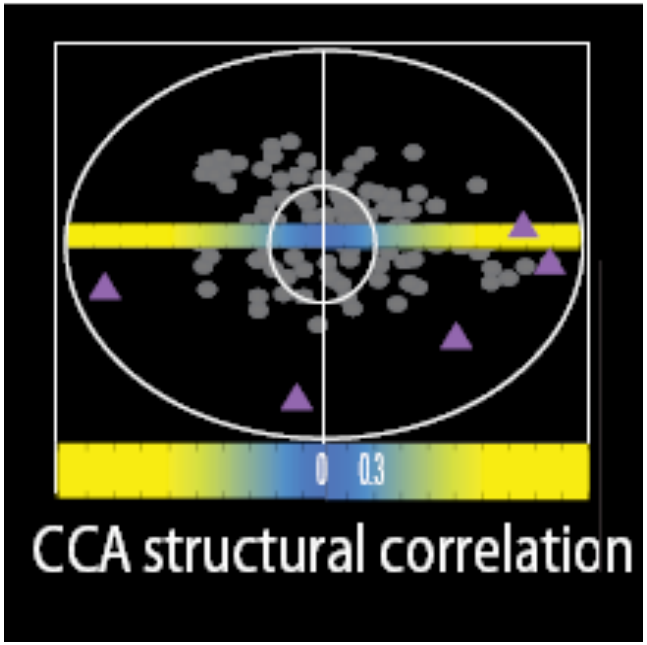


Strength of Pathway co-variation with environment



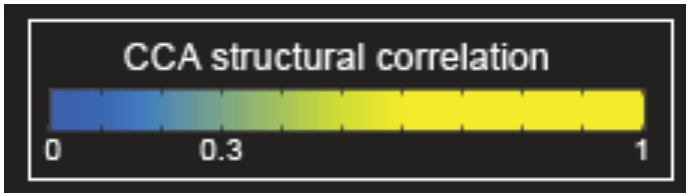
Environmentally invariant

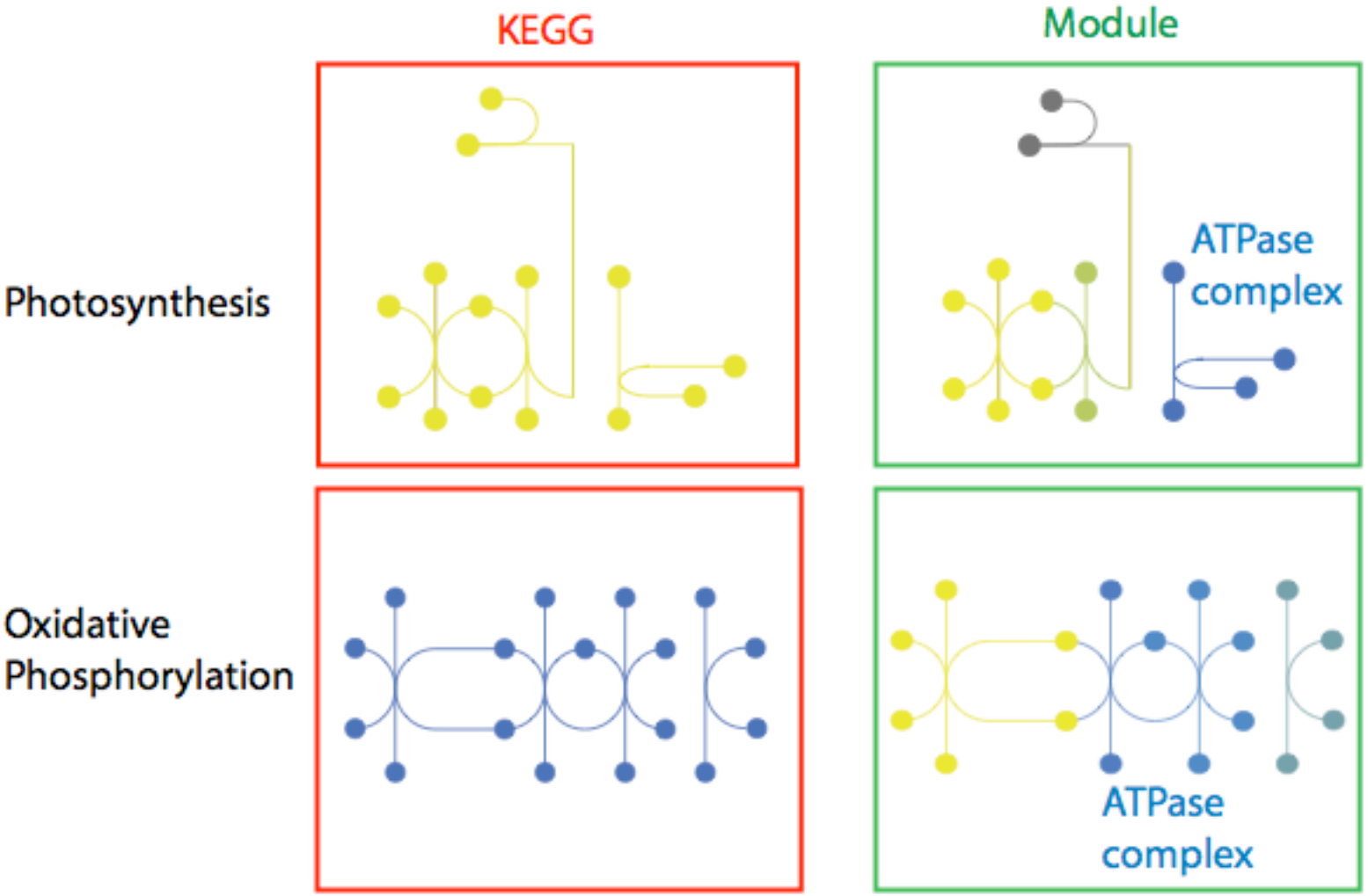
Environmentally variant



[Gianoulis et al., PNAS (in press, 2009)]

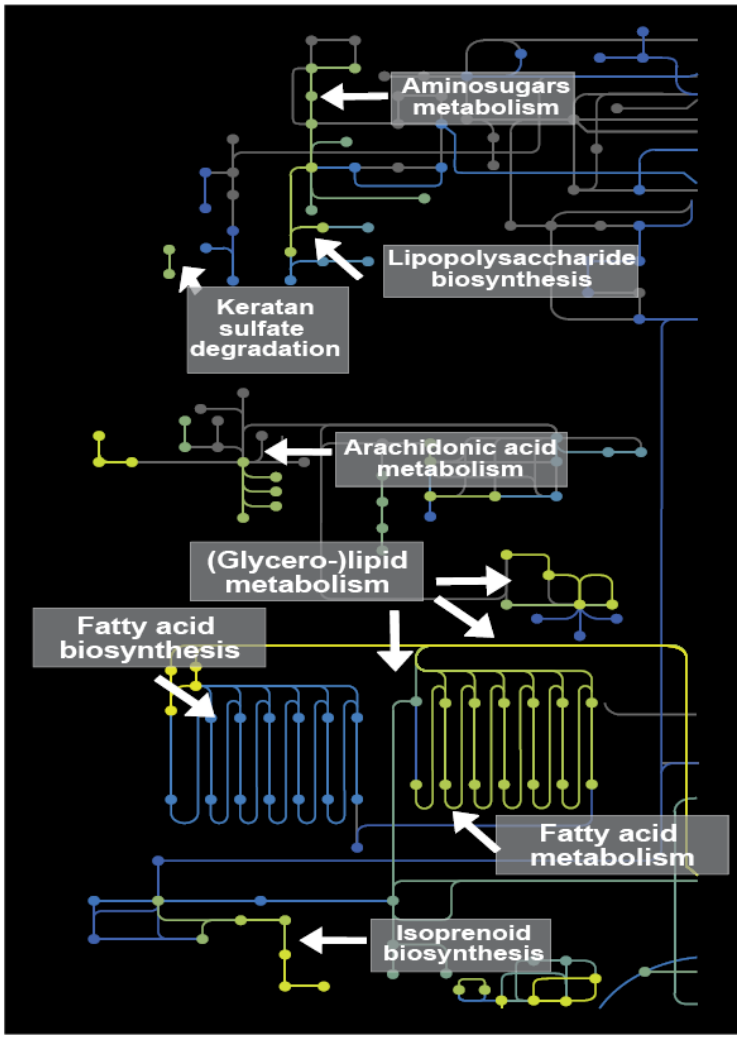
Conclusion #1: energy conversion strategy, temp and depth

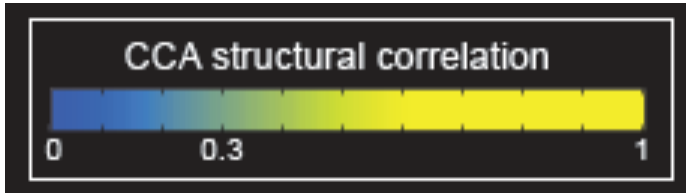




[Gianoulis et al., PNAS (in press, 2009)]

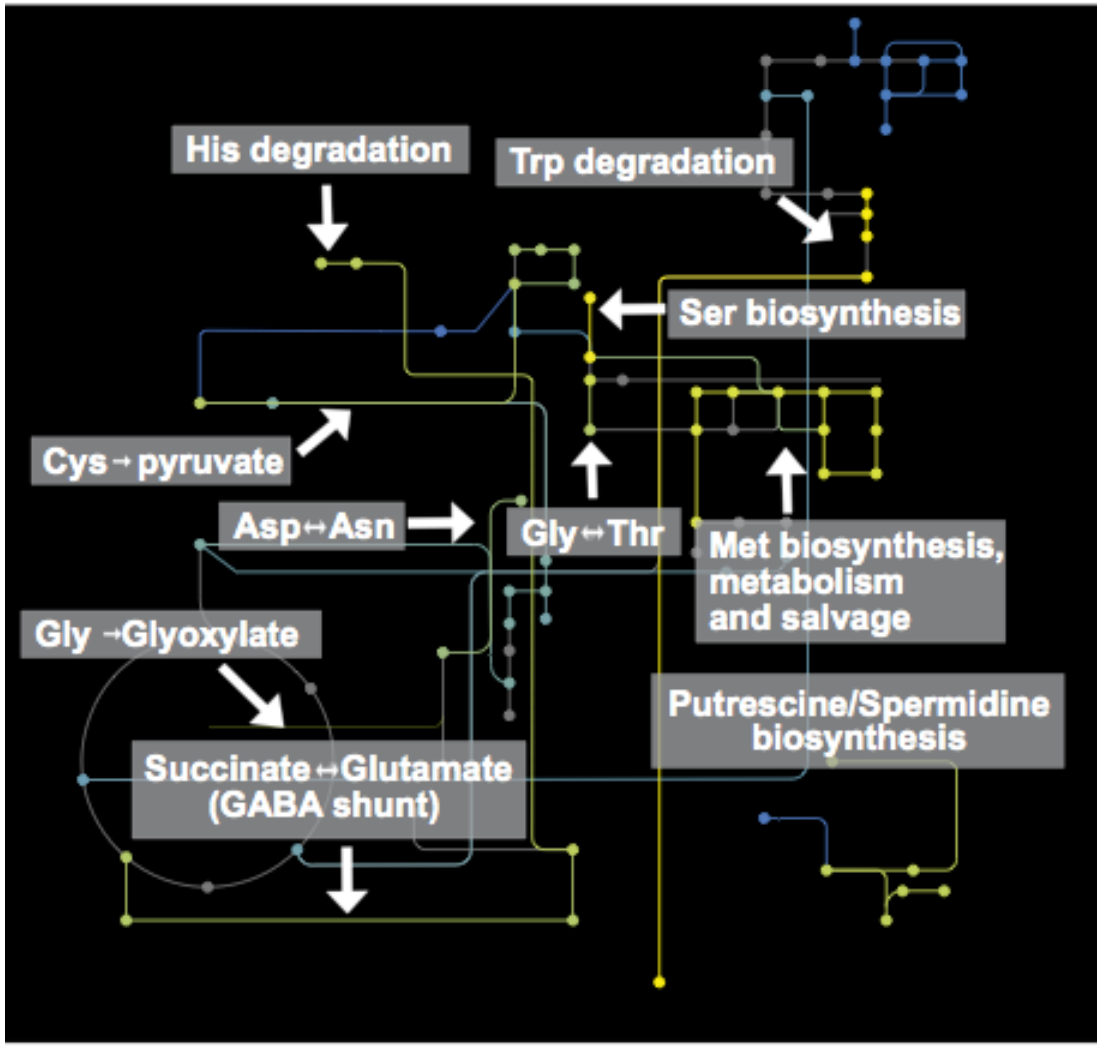
Conclusion #2: Outer Membrane components vary the environment





[Gianoulis et al., PNAS (in press, 2009)]

Conclusion #3: Covariation of AA biosynthesis and Import



Why is their fluctuation
in amino acid metabolism?
Is there a feature(s) that
underlies those that are
environmentally-variant
as opposed to those which are not?

[Gianoulis et al., PNAS (in press, 2009)]

Conclusion #4: Cofactor (Metal) Optimization

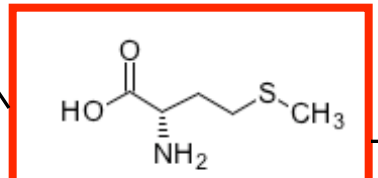
IS DEPENDENT-ON

Methionine synthesis

Cobalamin biosynthesis

Cobalt transporters

Methionine salvage, synthesis, and uptake, transport



C00073

Methionine

IS NEEDED FOR

Methionine degradation

S-adenosyl Methionine Biosynthesis (synthesize SAM one of the most important methyl donors)

Polyamine biosynthesis

RELIES ON

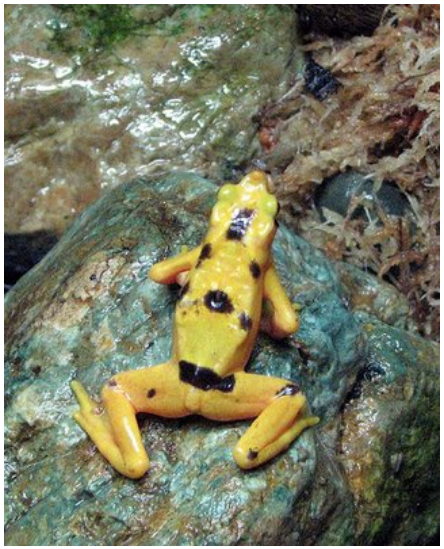
Methionine Salvage

Spermidine/Putrescine transporters

Arg/His/Ornithine transporters

[Gianoulis et al., PNAS (in press, 2009)]

Biosensors:
Beyond Canaries in a Coal Mine



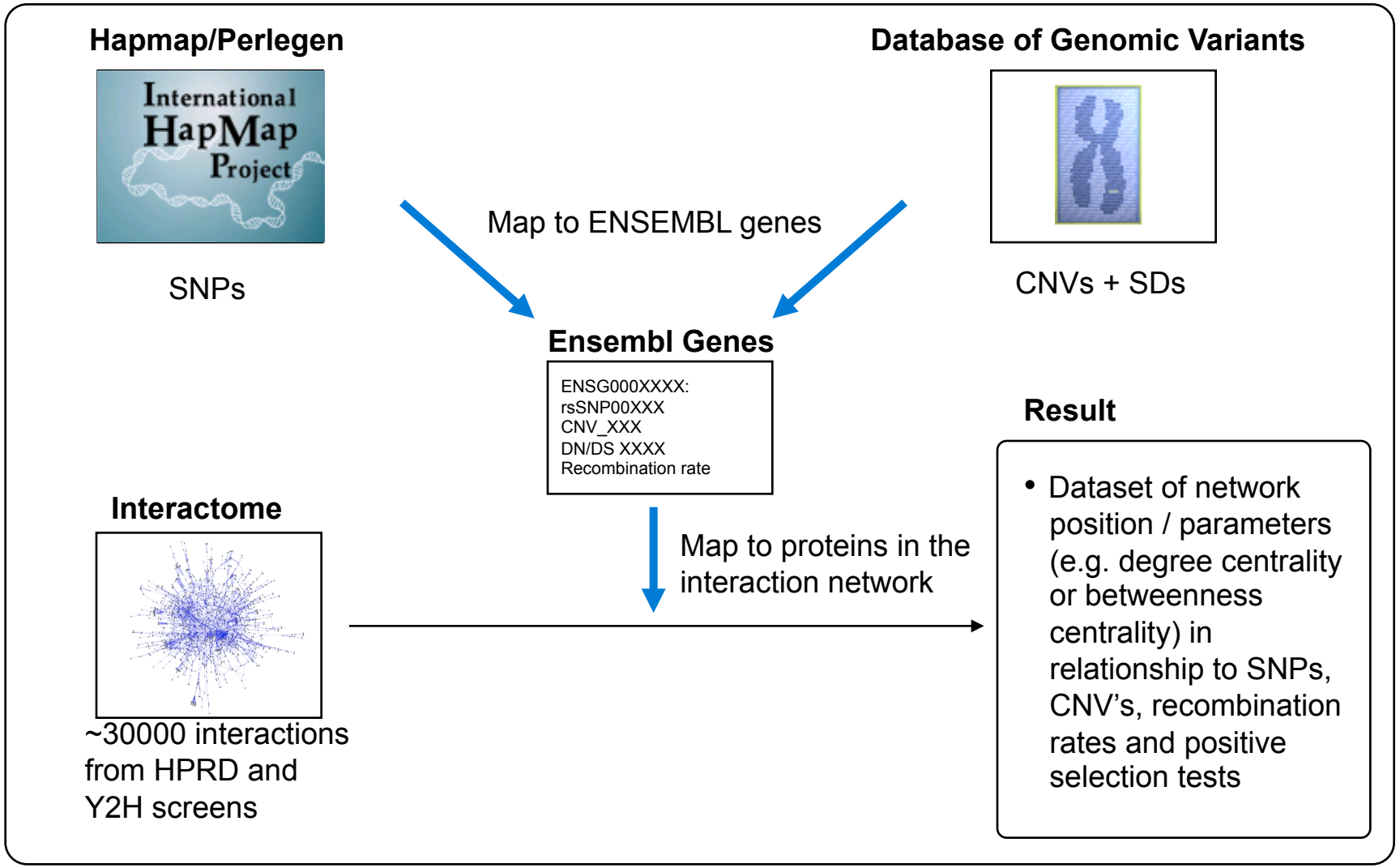
[Gianoulis et al., PNAS (in press, 2009)]

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

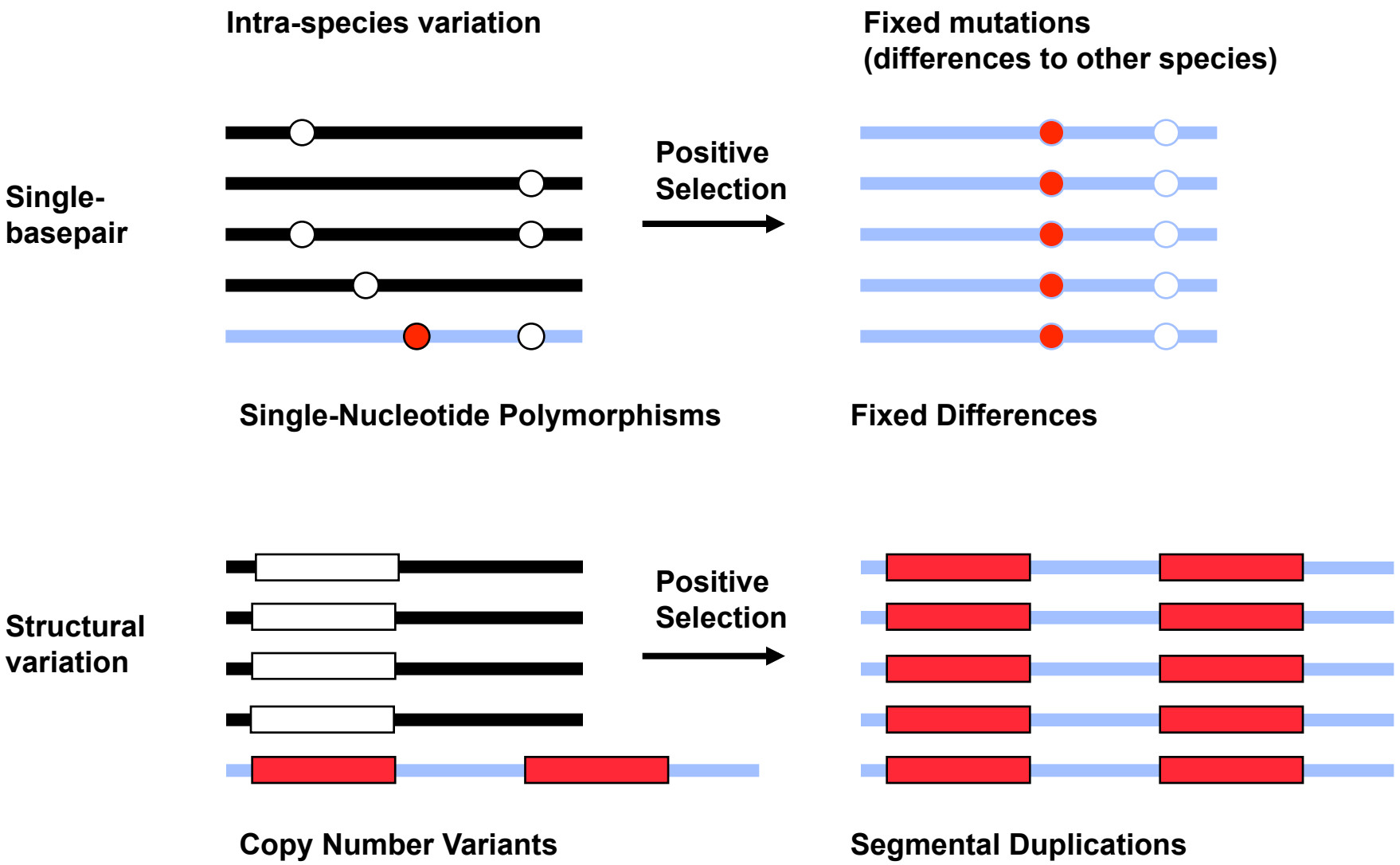
ILLUSTRATIVE



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

Source: PMK

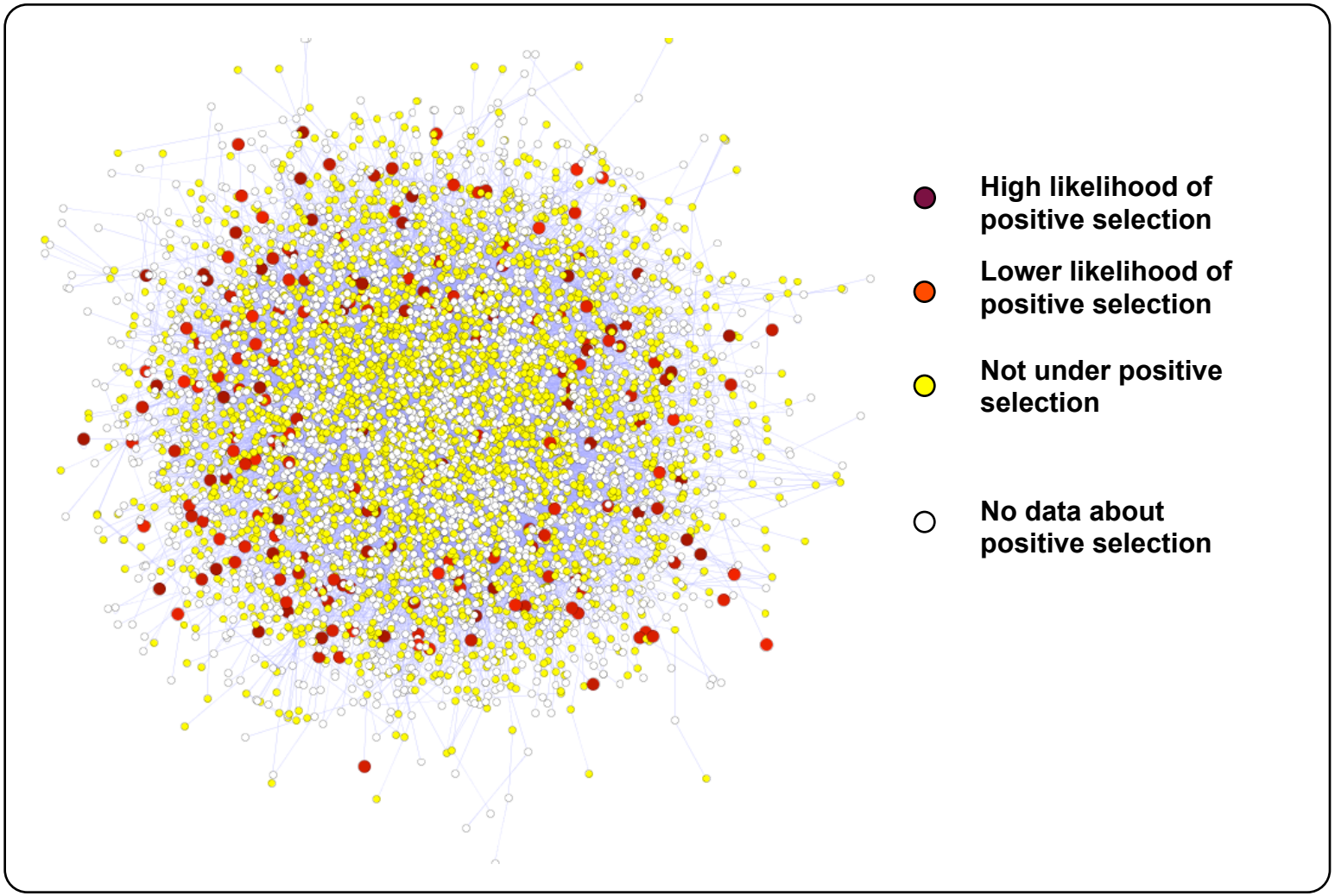
ADAPTIVE EVOLUTION CAN BE SEEN ON TWO DIFFERENT LEVELS



Source: PMK

POSITIVE SELECTION LARGELY TAKES PLACE AT THE NETWORK PERIPHERY

Positive selection in the human interactome

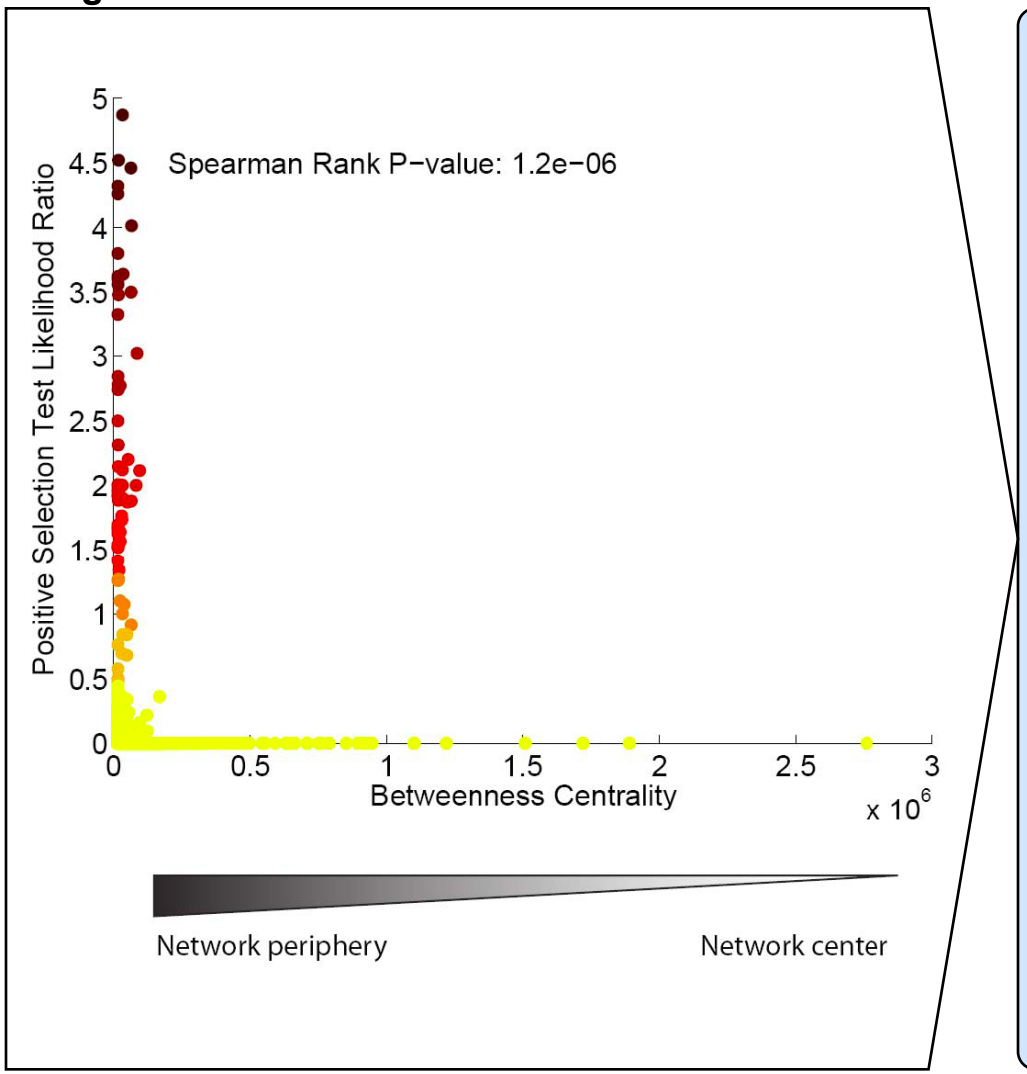


Source: Nielsen et al. *PLoS Biol.* (2005), HPRD, and Kim et al. PNAS (2007)

CENTRAL PROTEINS ARE LESS LIKELY TO BE UNDER POSITIVE SELECTION

Hubs

Degree vs. Positive Selection



Reasoning

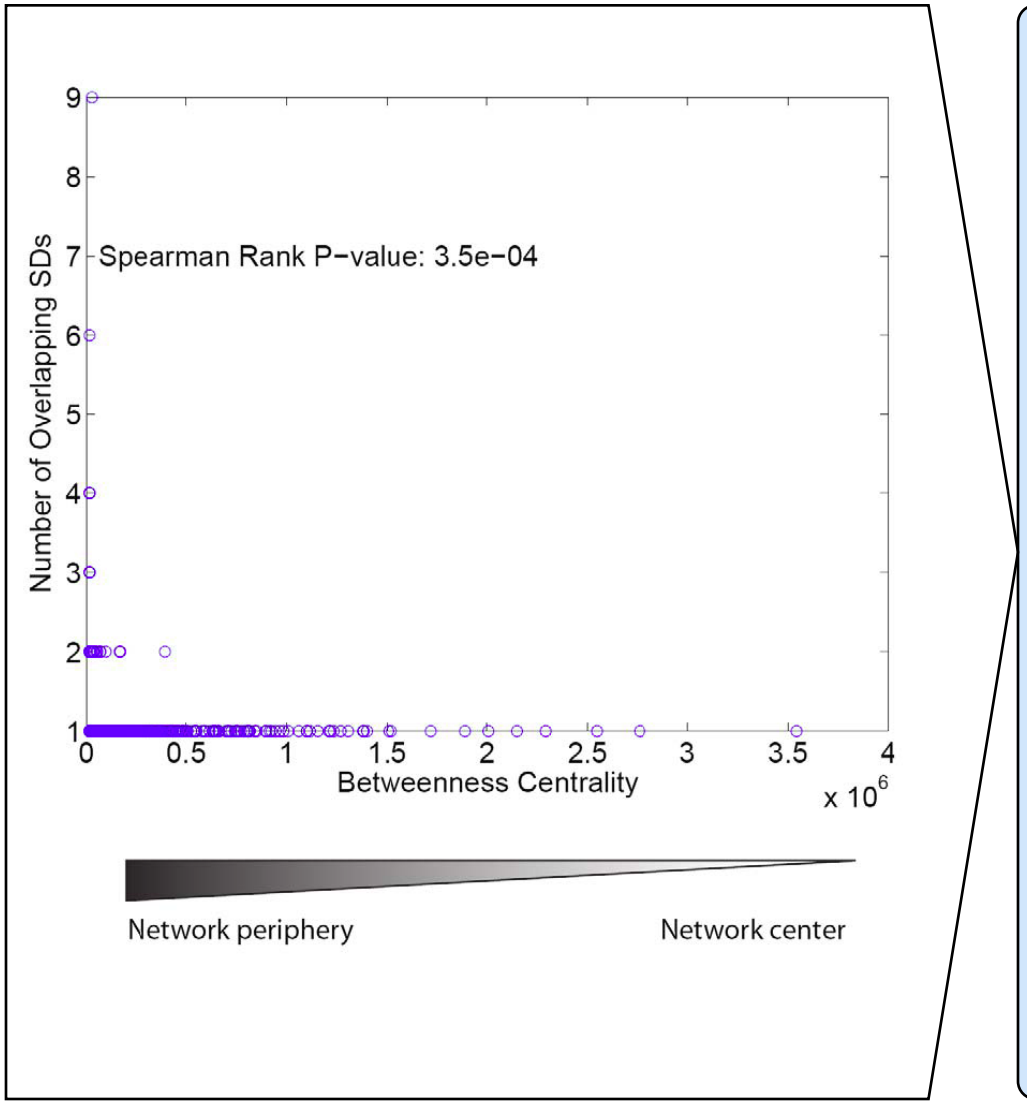
- Peripheral genes are likely to under positive selection, whereas hubs aren't
- This is likely due to the following reasons:
  - Hubs have stronger structural constraints, the network periphery doesn't
  - Most recently evolved functions (e.g. "environmental interaction genes" such as sensory perception 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

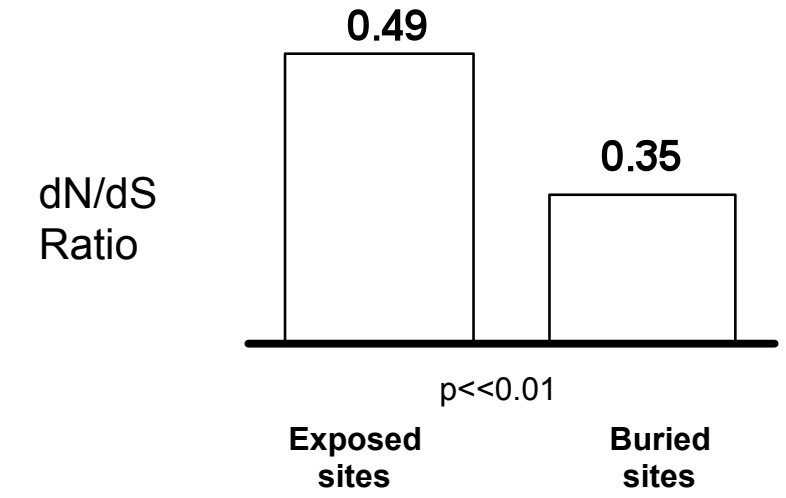
- This result also confirms our initial hypothesis – peripheral nodes tend to lie in regions rich in SDs.
- Since segmental duplications are a different mechanism of ongoing evolution, the less constrained peripheral proteins are enriched in them.
- Note that despite the small size of our dataset for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

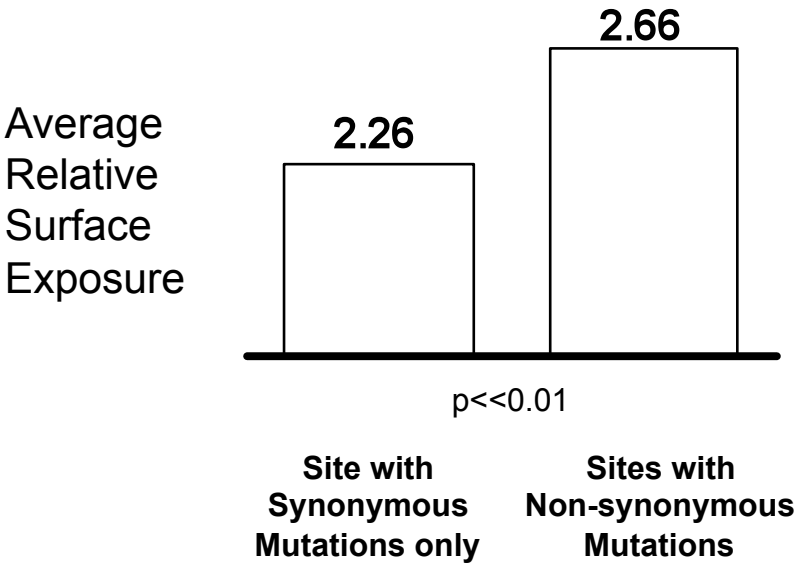
*Specifically, a number of the SDs are likely not fixed, but rather common CNVs in the reference genome

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

Why do we observer this? Perhaps central hub proteins are involved in more interactions & have more surface buried.

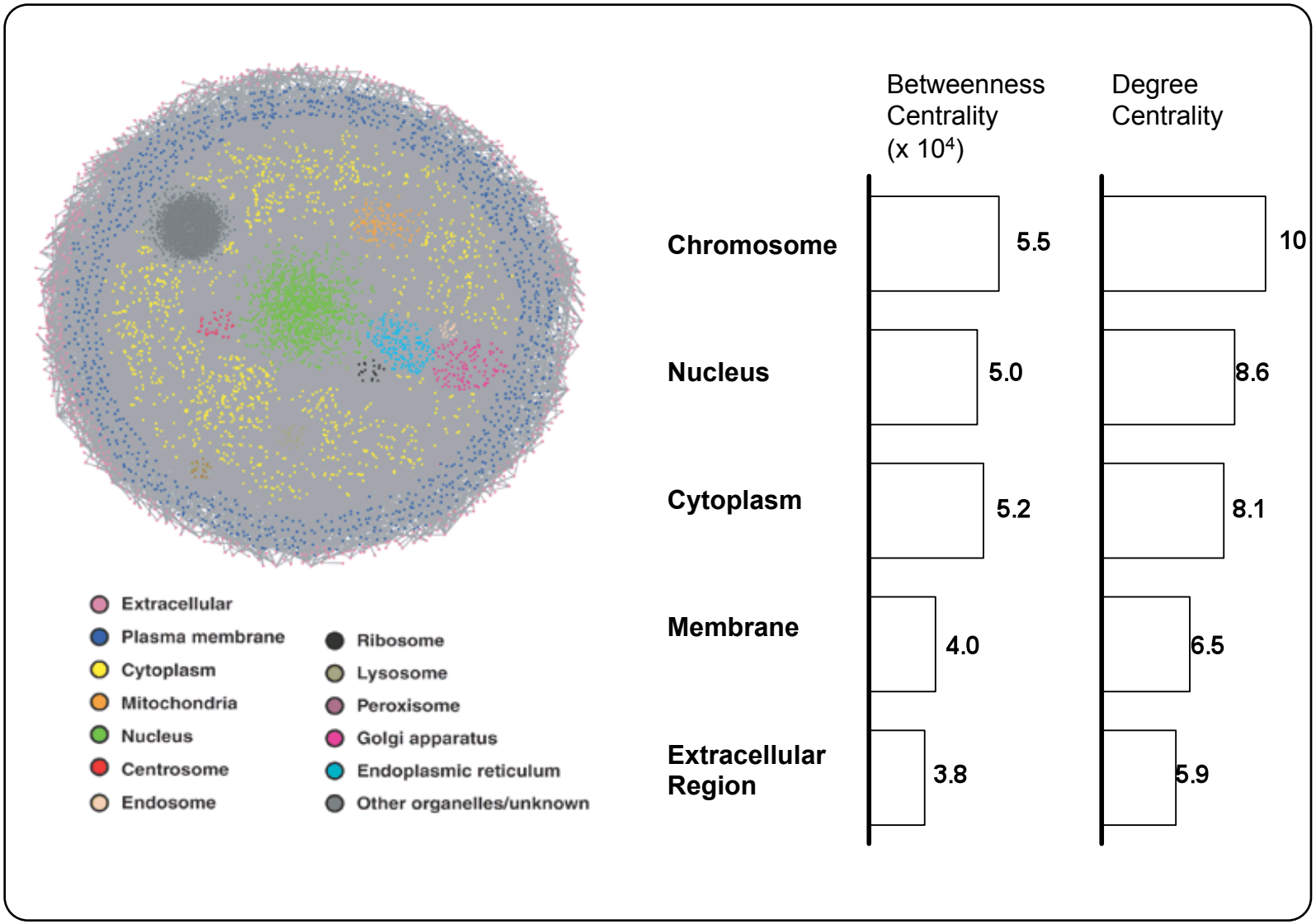
BURIED SITES ARE CONSERVED AND MUCH LESS LIKELY TO HARBOR NON-SYNONYMOUS MUTATIONS





Source: Kim et al. PNAS (2007)

Another explanation: THE NETWORK PERIPHERY CORRESPONDS TO THE CELLULAR PERIPHERY



Source: Gandhi et al. (*Nature Genetics* 2006), Kim et al. PNAS (2007)

IS RELAXED CONSTRAINT OR ADAPTIVE EVOLUTION THE REASON FOR THE PREVALENCE OF BOTH SELECTED GENES AND SDs AT THE NETWORK PERIPHERY?

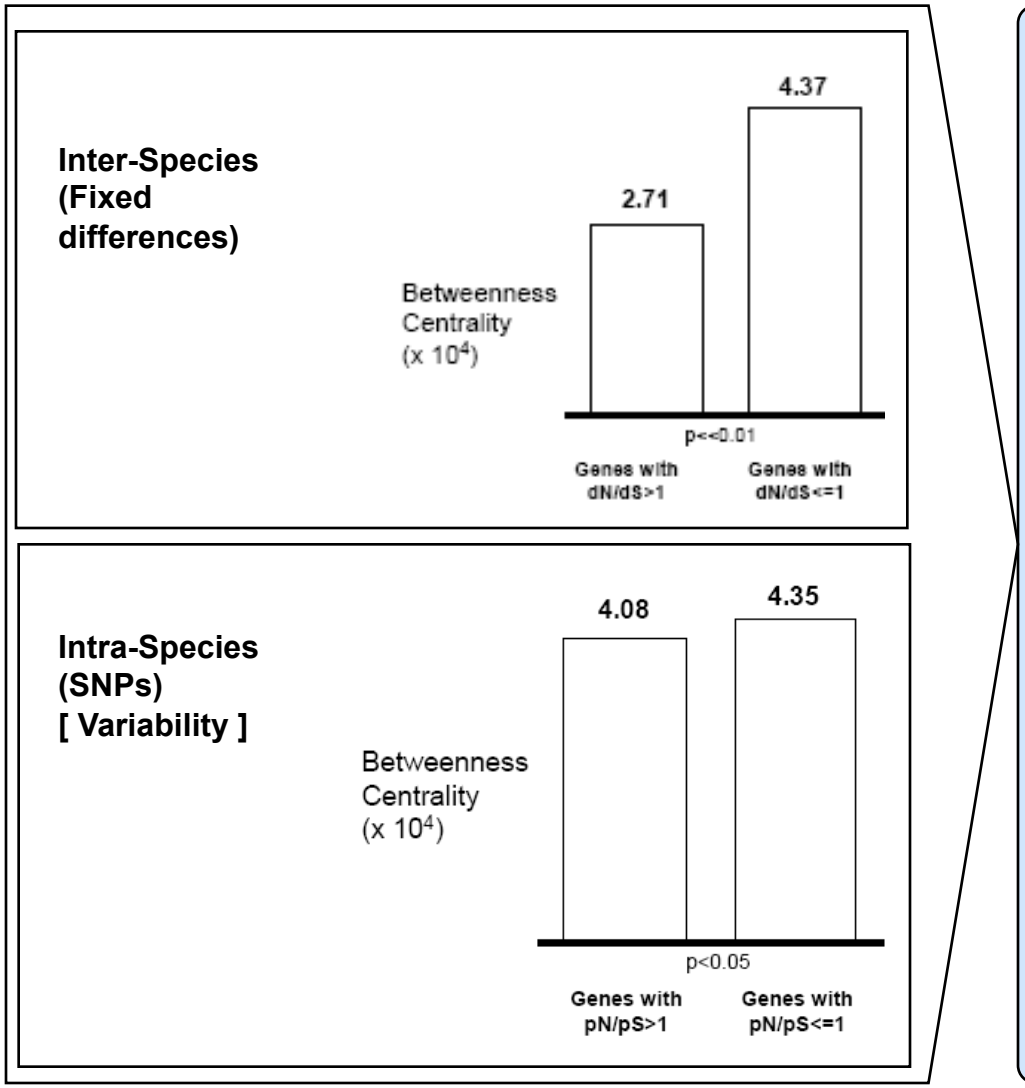
ILLUSTRATIVE

	Relaxed Constraint	Adaptive Evolution
Inter-Species Variation (Fixed differences)	• Increases inter-species variation – more variable loci are under less negative selection • Can be seen in higher Ka/Ks ratio or SD occurrence	• Increases inter-species variation – more variable loci are under less negative selection • Can be seen in higher Ka/Ks ratio or SD occurrence
Intra-Species Variation (Polymorphisms)	• Increases intra-species variation – for the very same reason • Can be seen in both SNPs or CNVs	• Should not have effects on intra-species variation

Source: Kim et al. PNAS (2007)

SOME, BUT NOT ALL OF THE SINGLE-BASEPAIR SELECTION AT THE PERIPHERY IS DUE TO RELAXED CONSTRAINT

Inter vs. Intra-Species Variation in Networks



Reasoning

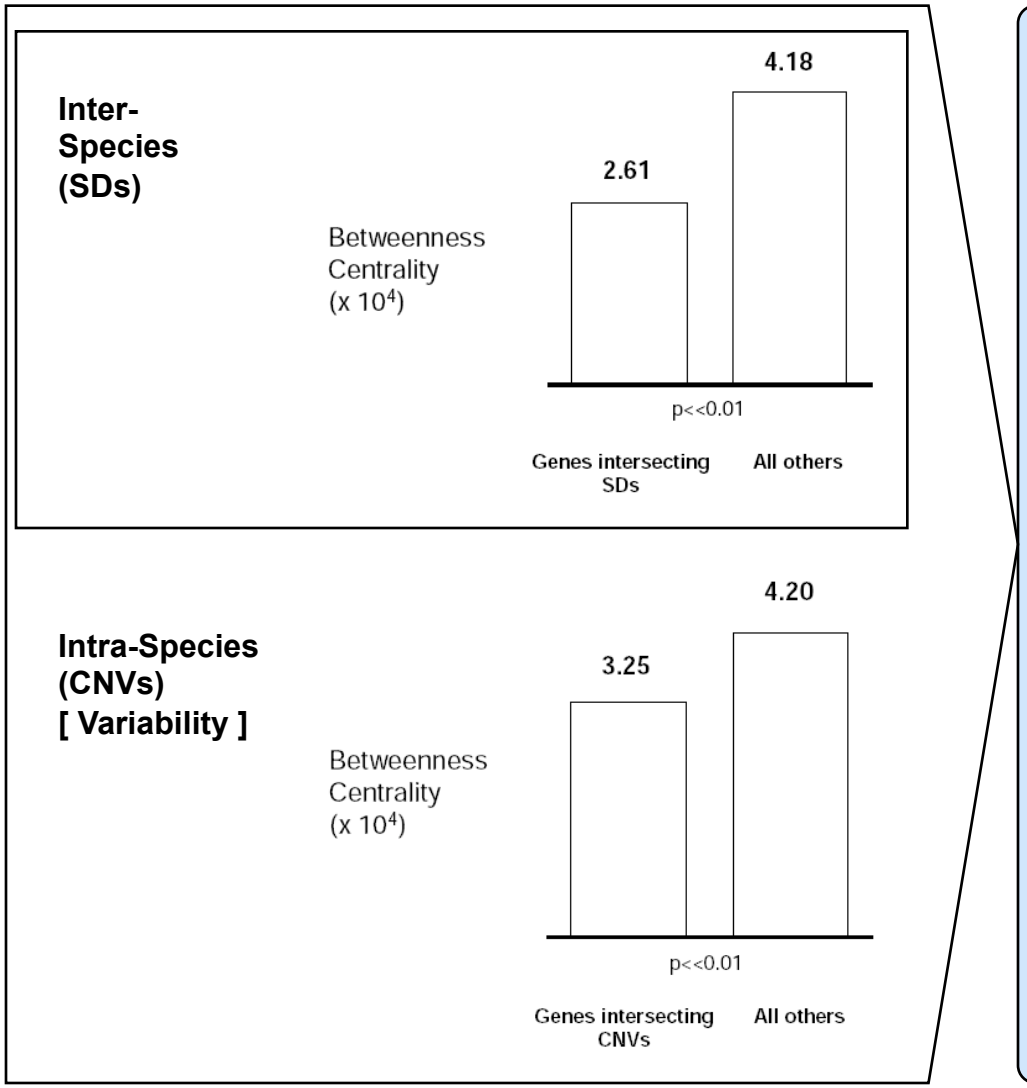
- There is a difference in **variability** (in terms of SNPs) between the network periphery and the center
- However, this difference is much smaller than the difference in **selection**
- This most likely means, that part of the effect we're seeing is due to relaxed constraint (and higher variability)
- But, not the entire effect*

*But it's hard to quantify

Source: Kim et al. (2007) PNAS

Similar Results for Large-scale Genomic Changes (CNVs and SDs)

Inter vs. Intra-Species Variation in Networks



Reasoning

- There a small difference in **variability** (in terms of CNVs) between the network periphery and the center
- But, there is a (as shown before) marked difference in fixed (and hence, presumably, **selected**) SDs at the network periphery and center

Source: Kim et al. (2007) PNAS

Outline: Molecular Networks

- Why Networks?
- Predicting Networks (yeast)
  - ◊ Propagating known information
- Network Structure: Key Positions (yeast)
  - ◊ Hubs & Bottlenecks
  - ◊ Tops of Hierarchy
- Dynamics & Variation of Networks
  - ◊ Across cellular states (yeast)
  - ◊ Across environments (in prokaryotes)
- Protein Networks & Human Variation

Conclusions on Networks: Predictions

- Predicting Networks
  - ◊ Extrapolating from the Training Set
  - ◊ Principled ways of using known information in the fullest possible fashion
    - Prediction Propagation
    - Multi-level learning

Conclusions:
Analysis of Network Structure

- Centrality Measures in Protein Network
  - ◊ Hubs & Bottlenecks
  - ◊ Importance of later in regulatory networks
- Regulatory Network Hierarchies
  - ◊ Middle managers dominate, sitting at info. flow bottlenecks
  - ◊ Paradox of influence and essentiality
  - ◊ Topmost proteins sit at center of interaction network

Conclusions: Network Dynamics across Cellular States

- Merge expression data with Networks
- Active network markedly different in different conditions
- Identify transient hubs associated with particular conditions
- Use these to annotate genes of unknown function

Conclusions: Networks Dynamics across Environments

- Developed and adapted techniques to connect quantitative features of environment to metabolism.
- Applied to available aquatic datasets, we identified footprints that were predictive of their environment (potentially could be used as biosensor).
- Strong correlation exists between a community's energy conversion strategies and its environmental parameters (e.g. temperature and chlorophyll).
- Suggest that limiting amounts of cofactor can (partially) explain increased import of amino acids in nutrient-limited conditions.

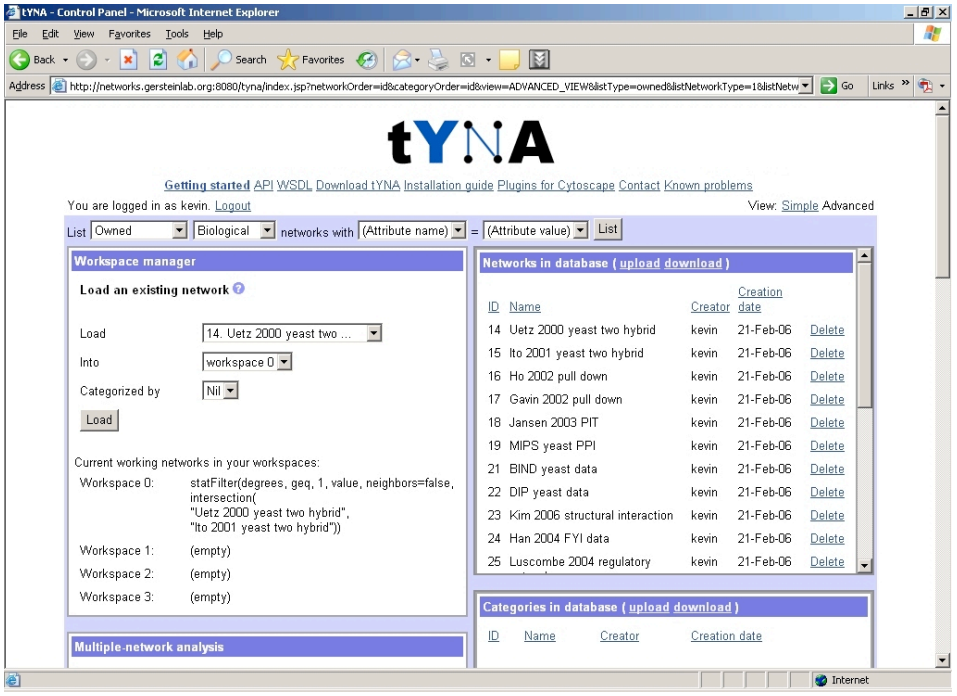
Conclusions: Connecting Networks & Human Variation

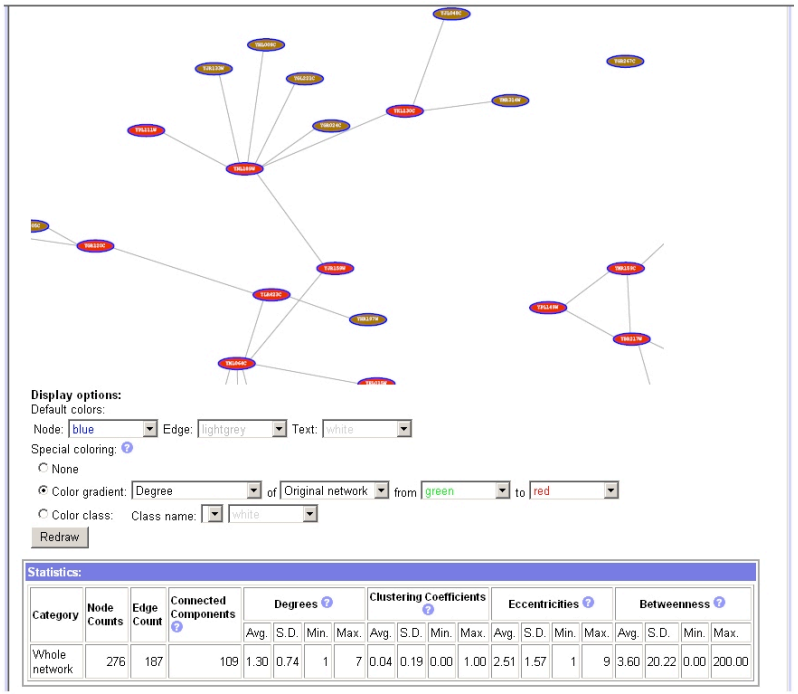
- We find ongoing evolution (positive selection) at the network periphery.
  - ◊ This trend is present on two levels:
    - On a sequence level, it can be seen as positive selection of peripheral nodes
    - On a structural level, it can be seen as the pattern of SDs that display significantly higher allele frequencies in non-central genes
  - ◊ 2 possible mechanisms for this : adaptive evolution at cellular periphery & relaxation of structural constraints at the network periphery
    - We show that the latter can only explain part of the increased variability,,,

- an automated web tool

tYNA

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



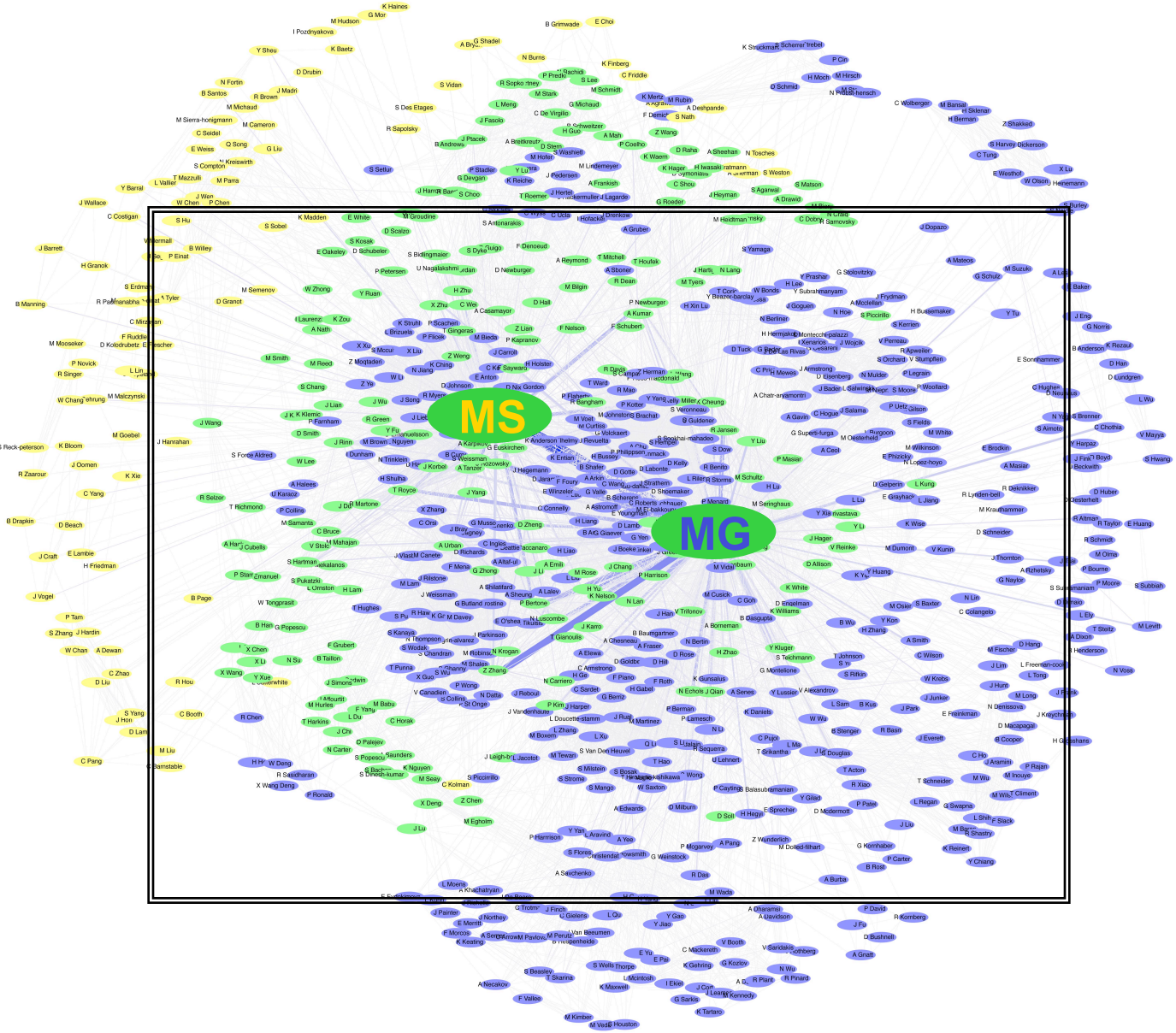


Normal website + Downloaded code (JAVA)
+ Web service (SOAP) with Cytoscape plugin

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

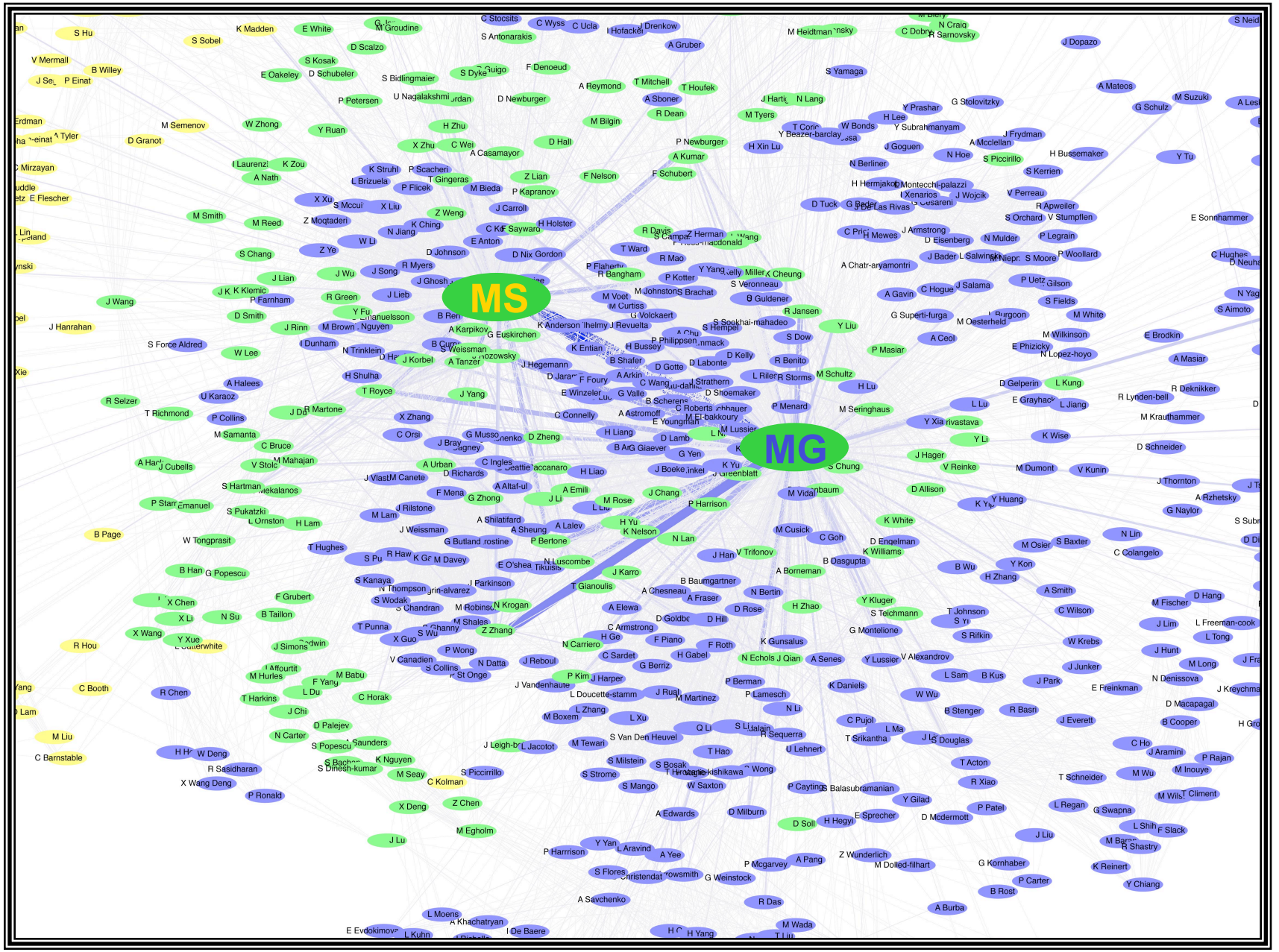
Acknowledgements

Networks.GersteinLab.org



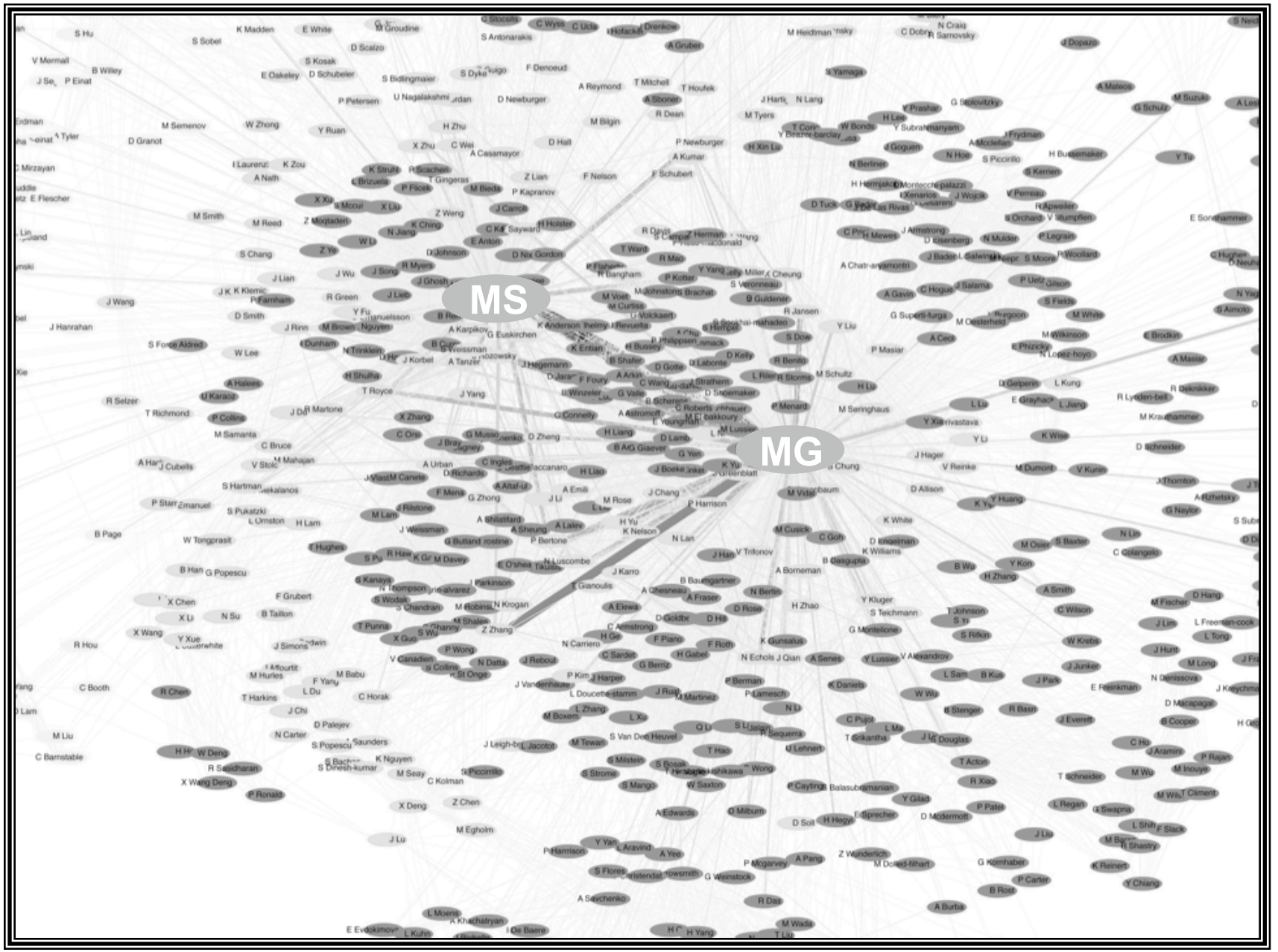
Acknowledgements

Networks.GersteinLab.org



Acknowledgements

Networks.GersteinLab.org

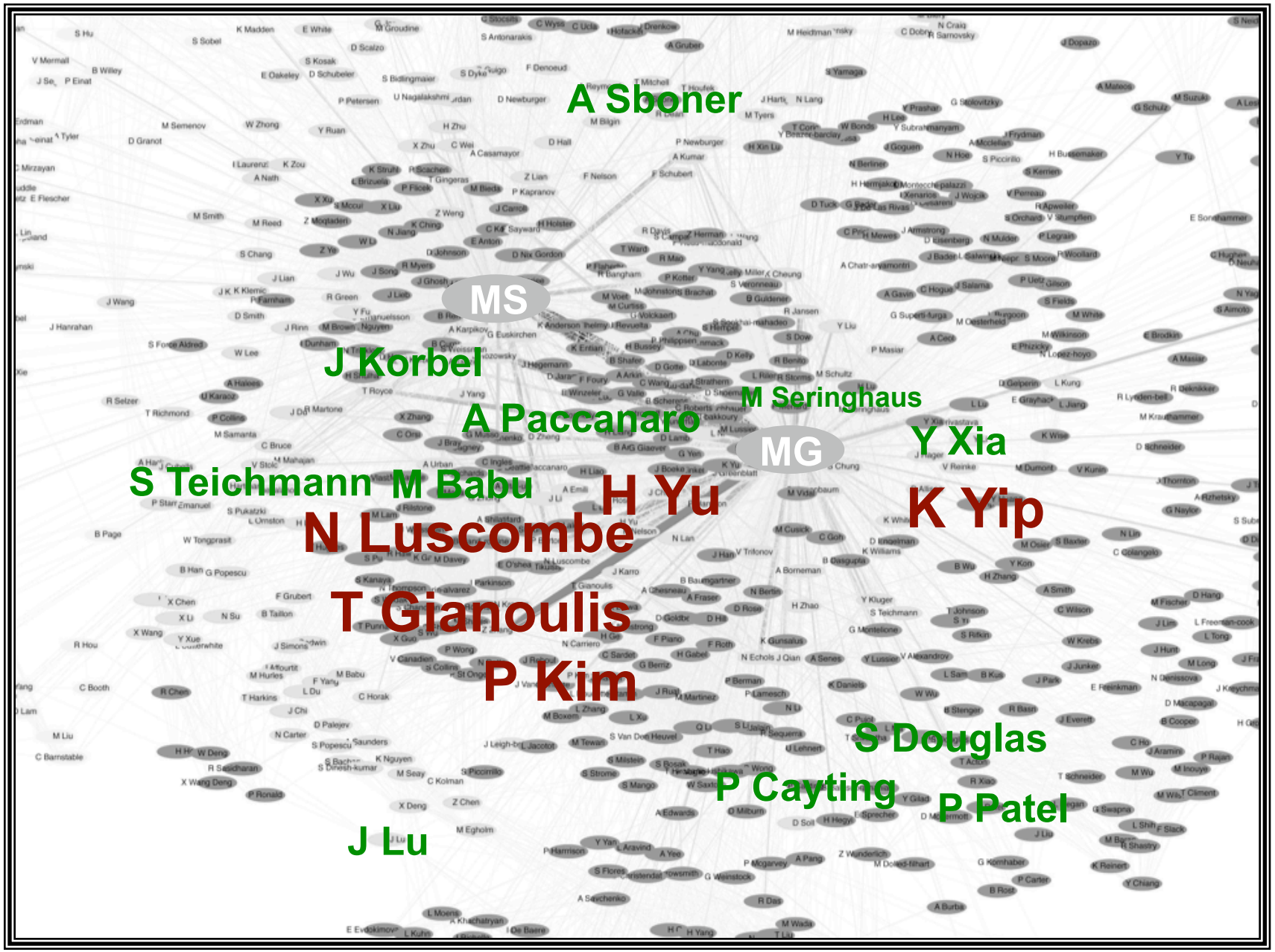


P Bork, J Raes

Acknowledgements

Networks.GersteinLab.org

Job opportunities currently for postdocs & students

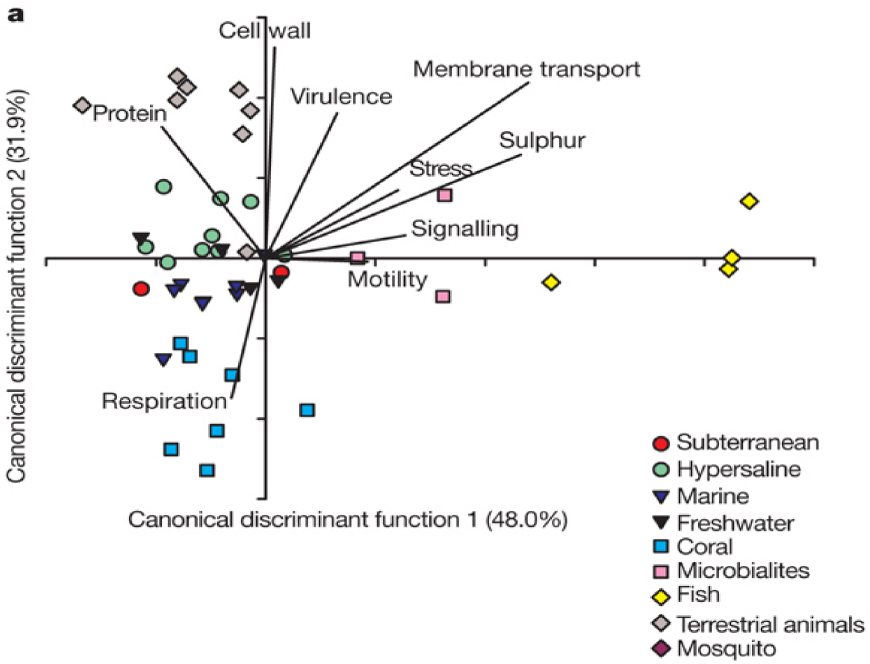


Extra

Comparative Metagenomics

Water

Soil



Do the proportions of pathways represented in these two samples differ?

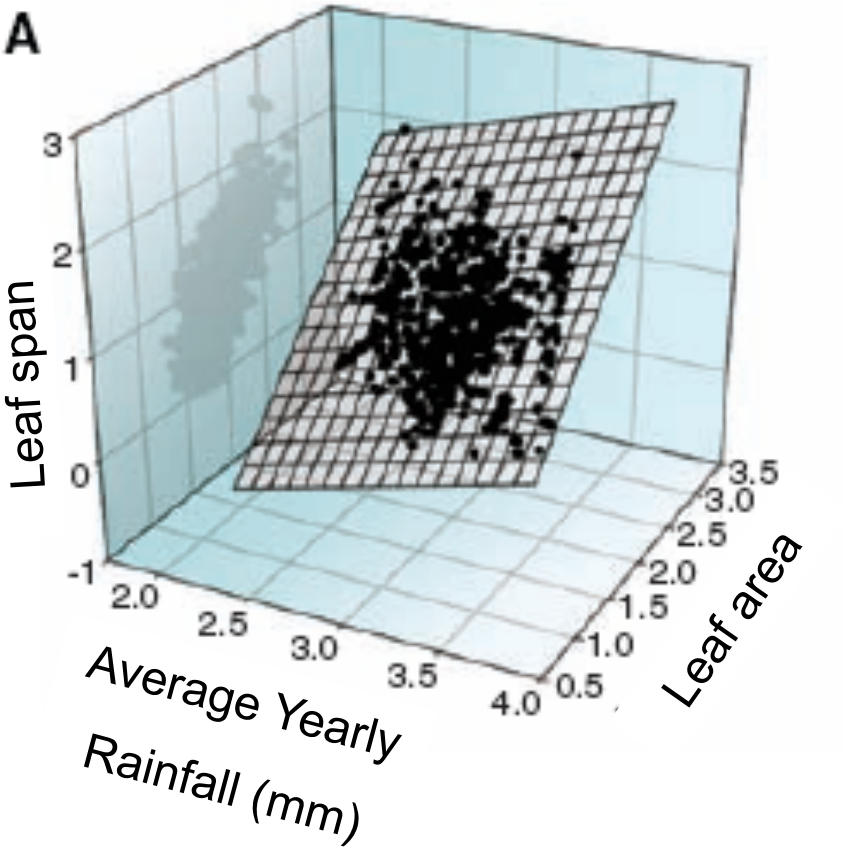
Dinsdale et. al., Nature 2008

Trait-based Biogeography

Charles River, MA

Long Island Sound, CT

Do the proportions of pathways represented in these two samples CHANGE as a function of their environments?



Green et. al., Science 2008

Interactome

GENOME

protein-DNA interactions

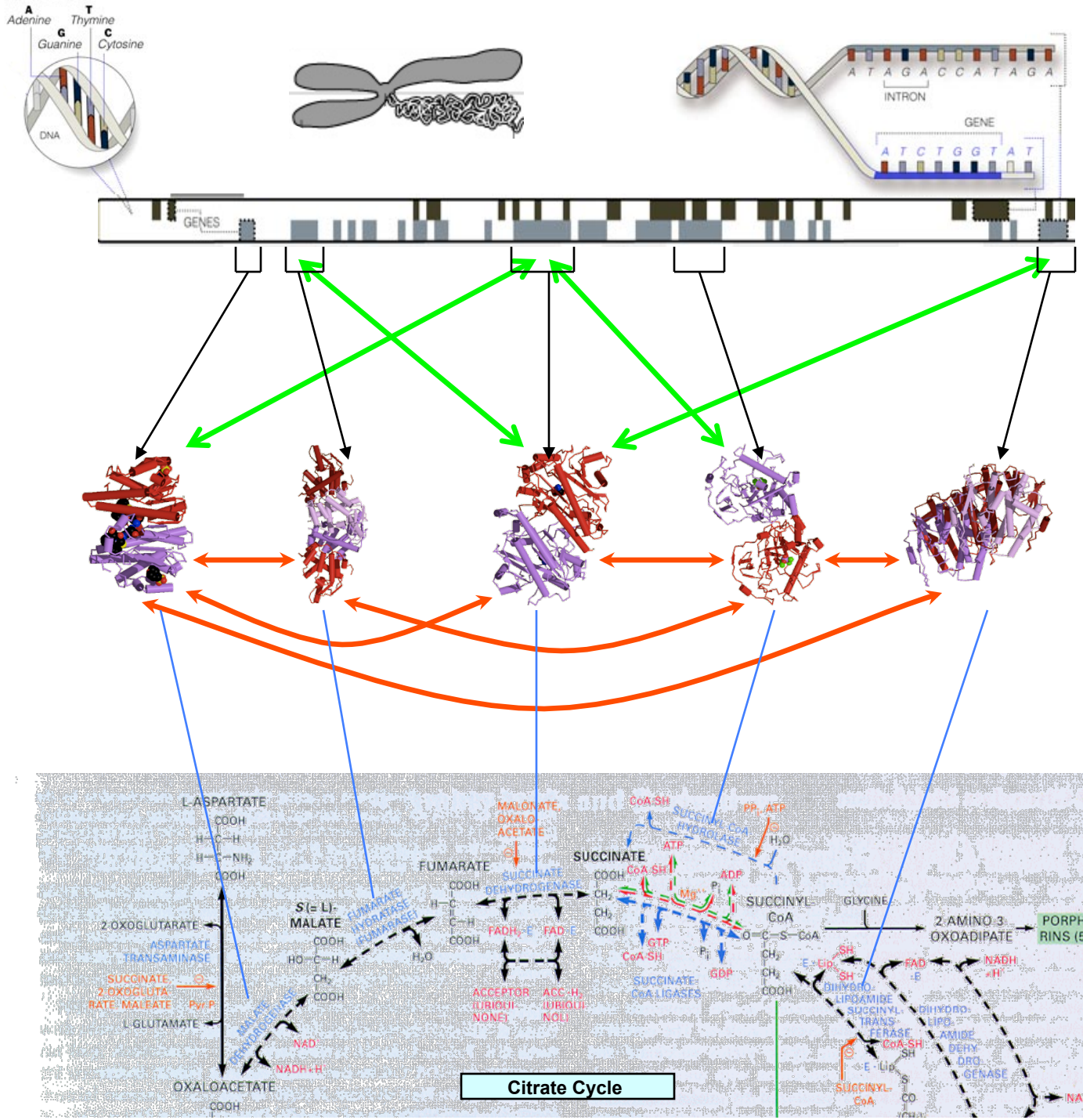
PROTEOME

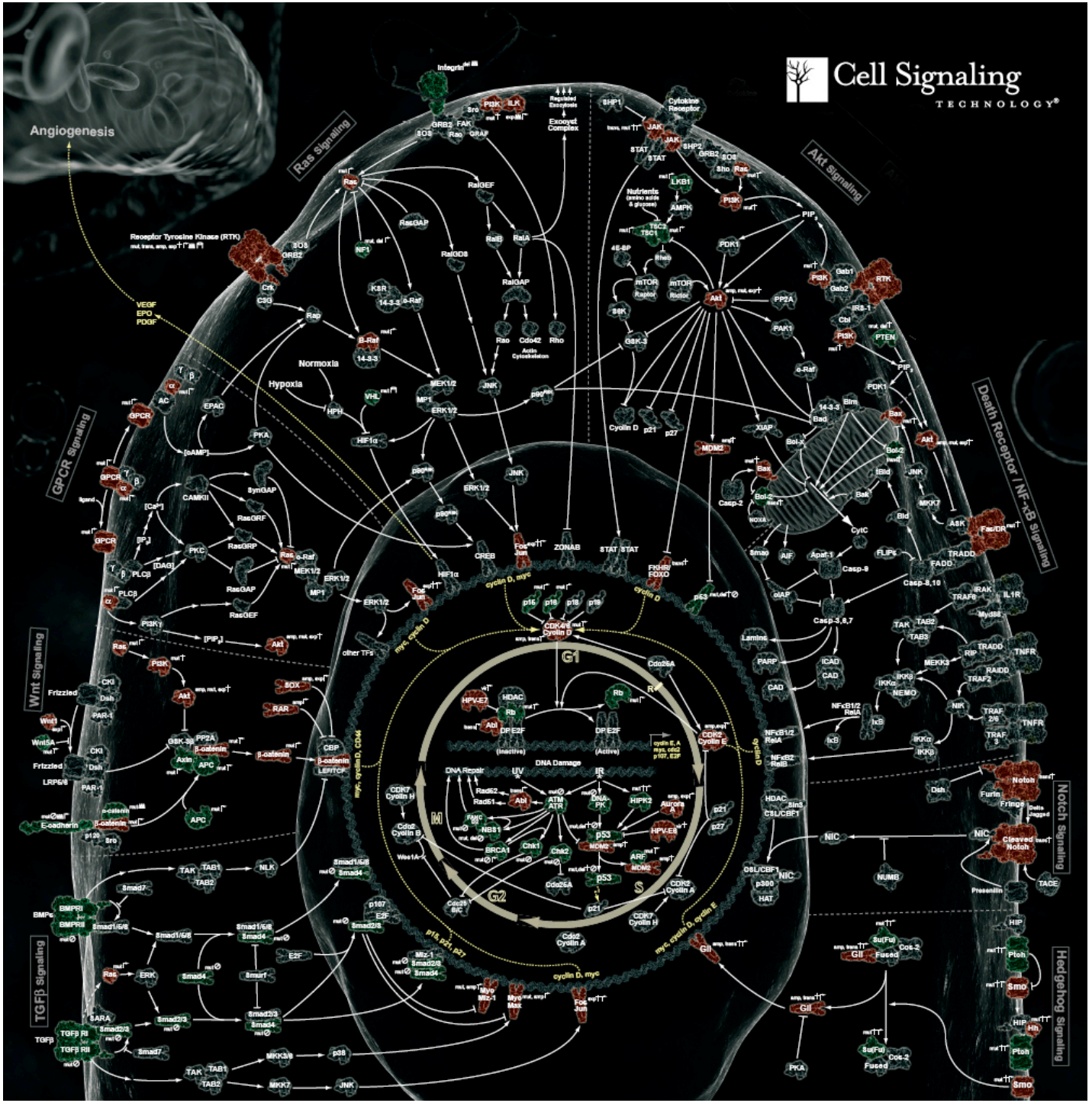
protein-protein interactions

METABOLISM

Protein-small molecule interactions

[From H Yu]





Networks help us understand biological processes

[From H Yu]

More Information on this Talk

TITLE: Understanding Protein Function on a Genome-scale through the Analysis of Molecular Networks

SUBJECT: Networks

DESCRIPTION:
Talk at Stanford Research Institute (SRI) 2009.05.21, 15:30-16:30; [I:**SRI**] (Full networks talk, I:**RECOMB09** with **nethierarchy*** and more **metagenomics***. Fits with a rush into 60' w. 10' questions. (PPT works on mac & PC.)

(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** .