Biological Network Analysis

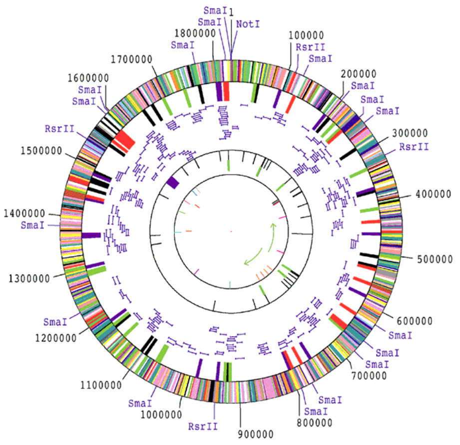
slides at
Lectures.GersteinLab.org

(See Last Slide for References & More Info.)

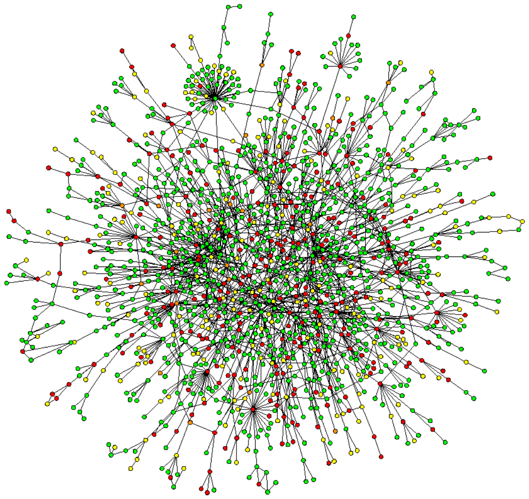
Mark
B
Gerstein

Yale

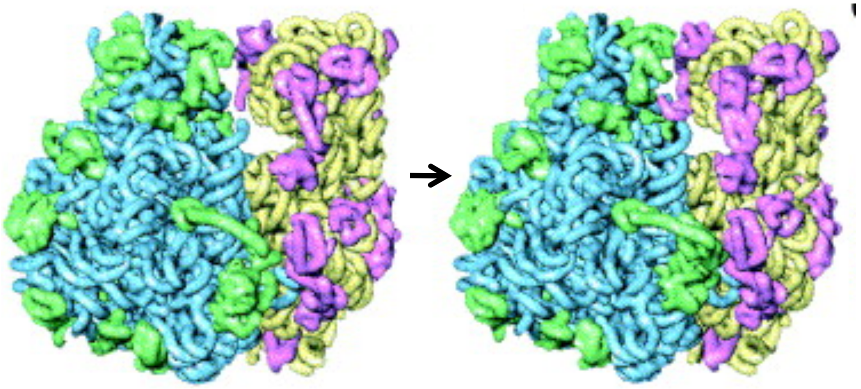
Networks occupy a midway point in terms of level of understanding



1D: Complete
Genetic Partslist



~2D: Bio-molecular
Network
Wiring Diagram



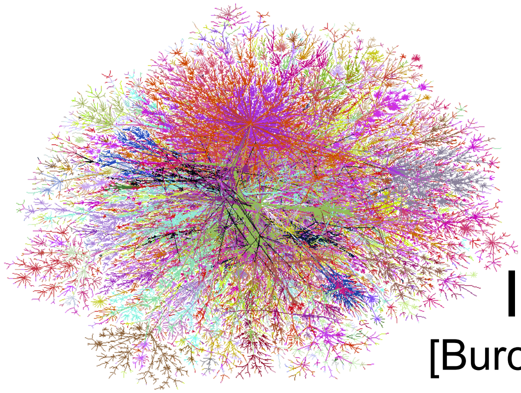
3D and 4D:
Detailed structural understanding
of cellular machinery
(e.g. ribosome in different
functional states)

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

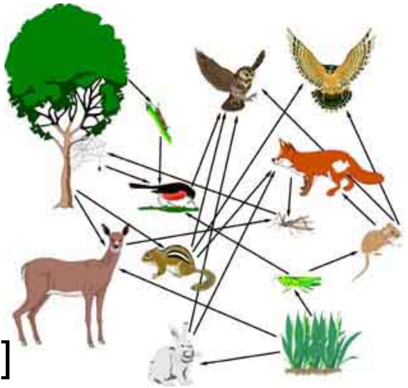
[Jeong et al. Nature, 41:411]

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

Networks as a universal language

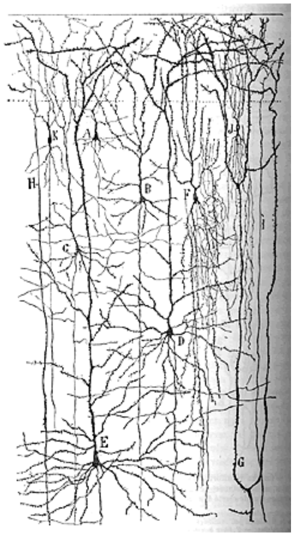


Internet
[Burch & Cheswick]

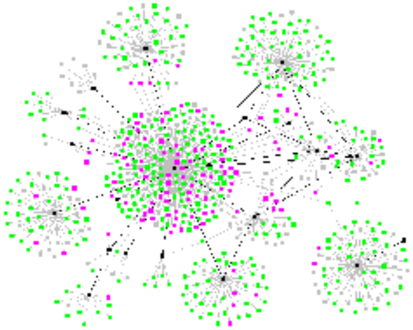


Food Web

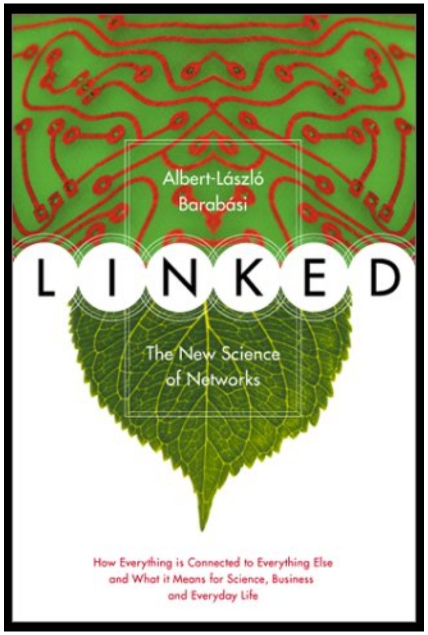
Electronic
Circuit



Neural Network
[Cajal]



Disease
Spread
[Krebs]



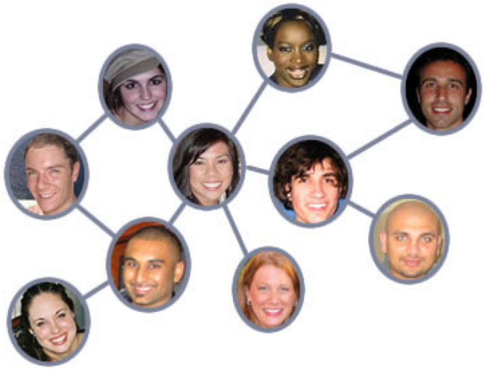
Albert-László
Barabási

LINKED

The New Science
of Networks

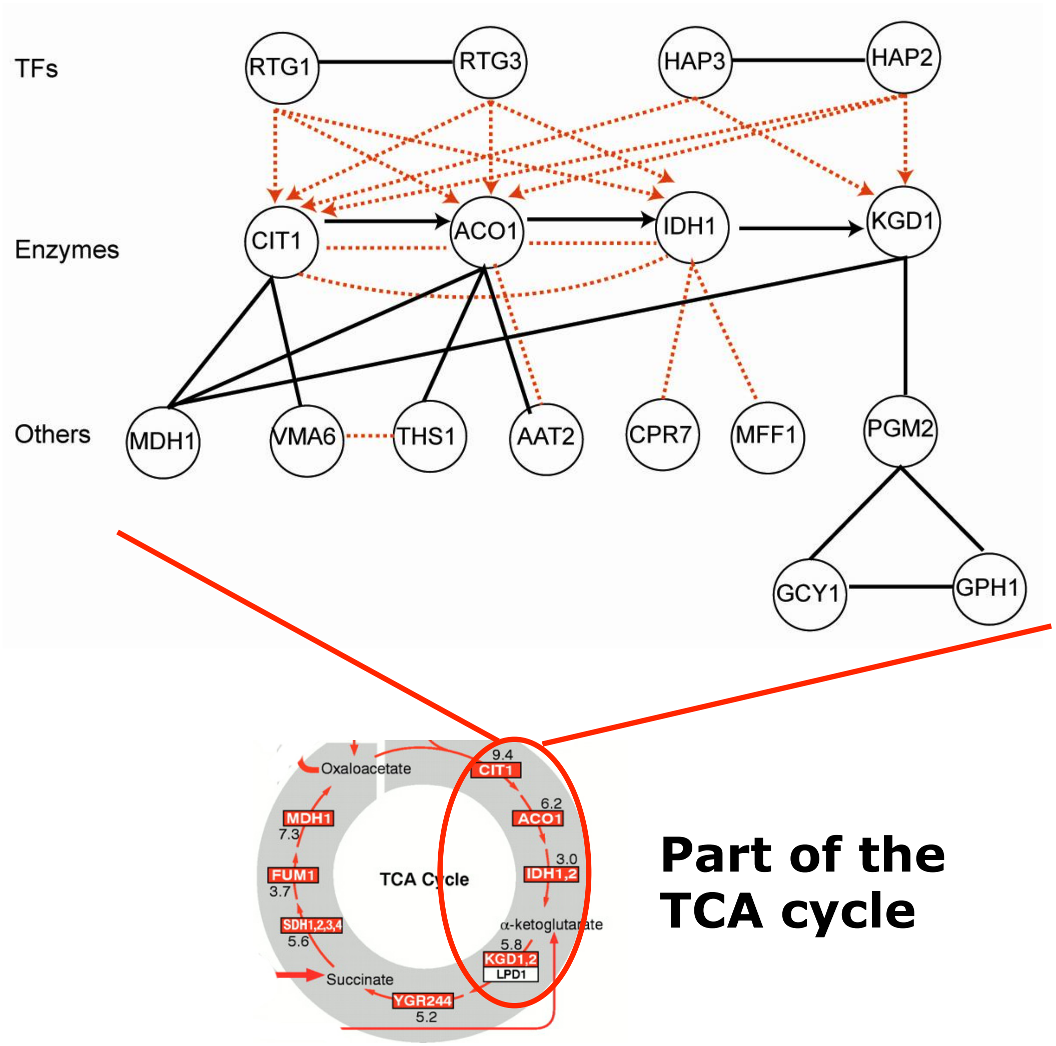
How Everything is Connected to Everything Else
and What it Means for Science, Business
and Everyday Life

Protein
Interactions
[Barabasi]



Social Network

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

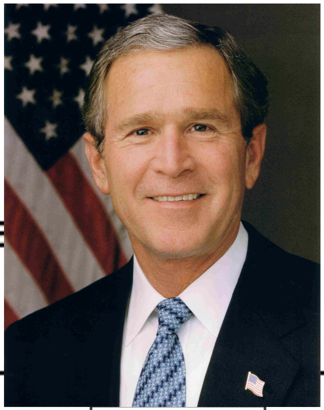
Outline: Molecular Networks

- Why Networks?
- Network Comparisons
  (reg. net. in many organisms)
  - in rel. to social hierarchy
  - Computer OS comparisons
  - Rewiring rate from comparisons amongst species
- Network Dynamics Across Environments
  (prokaryote metab. pathways)
  - Metabolic Pathways
  - Entry pts. (Mem. Proteins)

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

Social Hierarchy

THE GOVE UNITED STATES



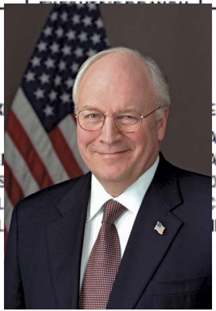
LEGISLATIVE BRANCH

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

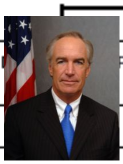
NAGEMENT AND BUDGET
ATIONAL DRUG CONTROL POLICY
LICY DEVELOPMENT
IENCE AND TECHNOLOGY POLICY
E U.S. TRADE REPRESENTATIVE

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

OF RBAN T



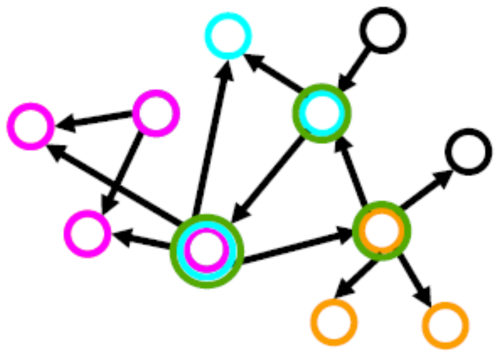
OF AIRS

INDEPENDENT ESTABLISHMENTS AND GOVERNMENT CORPORATIONS

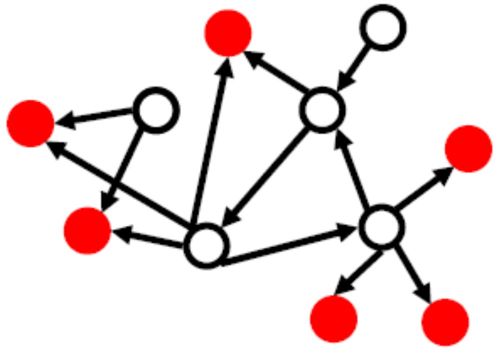
AFRICAN DEVELOPMENT FOUNDATION
CENTRAL INTELLIGENCE AGENCY
COMMODITY FUTURES TRADING COMMISSION
CONSUMER PRODUCT SAFETY COMMISSION
CORPORATION FOR NATIONAL AND COMMUNITY SERVICE
DEFENSE NUCLEAR FACILITIES SAFETY BOARD
ENVIRONMENTAL PROTECTION AGENCY
EQUAL EMPLOYMENT OPPORTUNITY COMMISSION
EXPORT-IMPORT 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

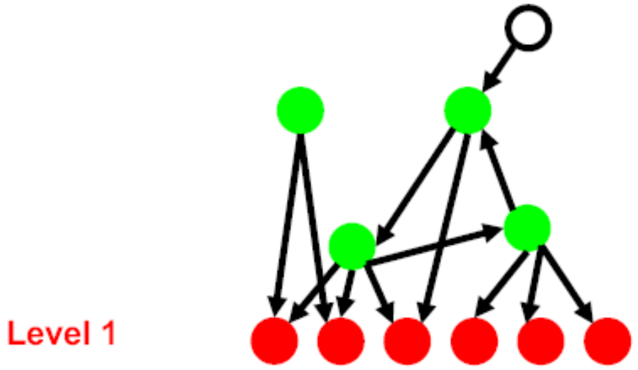
I. Example network with all 4 motifs



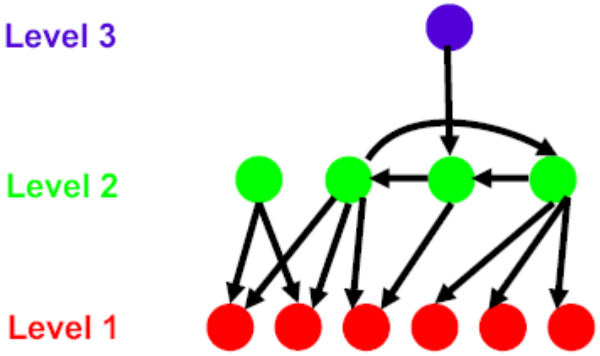
II. Finding terminal nodes (Red)



III. Finding mid-level nodes (Green)

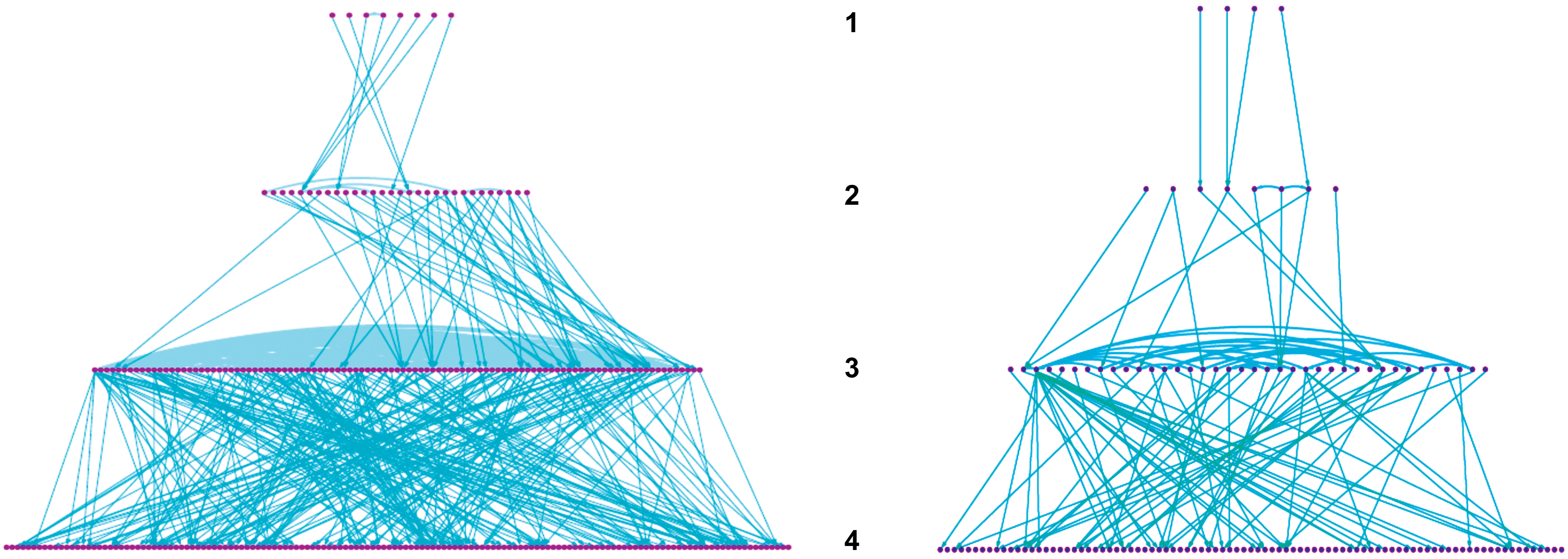


IV. Finding top-most nodes (Blue)



[Yu et al., PNAS (2006)]

Regulatory Networks have similar hierarchical structures

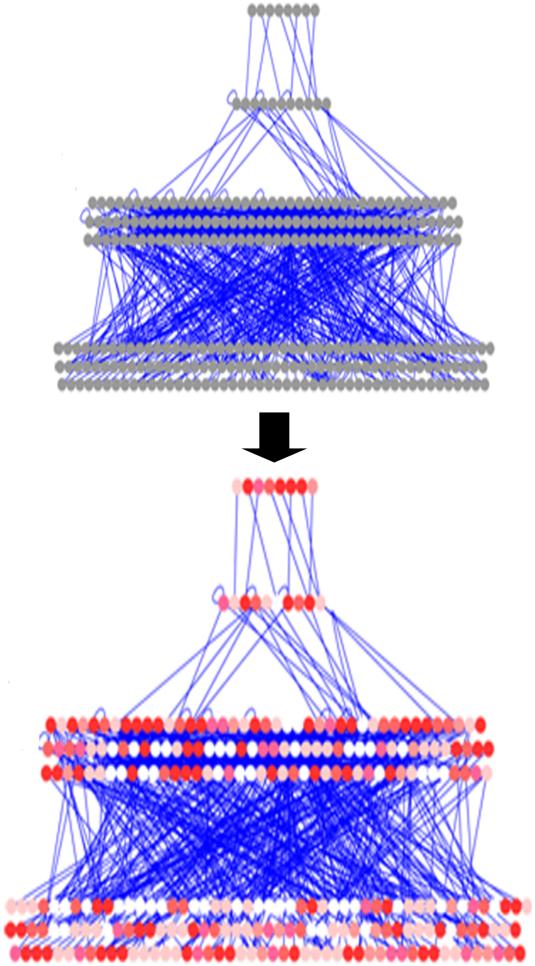


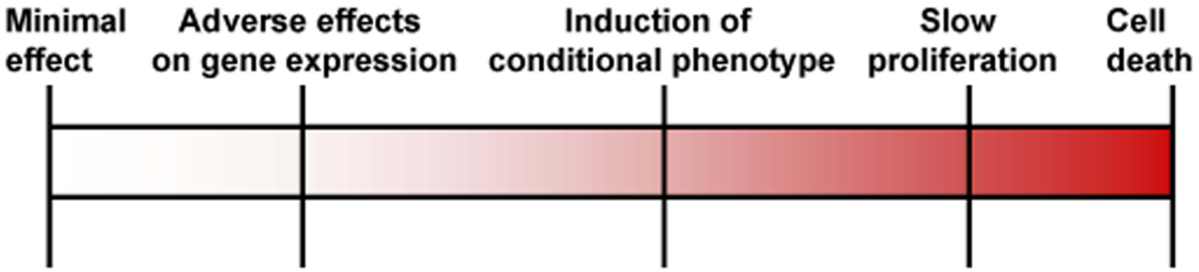
S. cerevisiae

E. coli

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

Essential vs Influential Genes





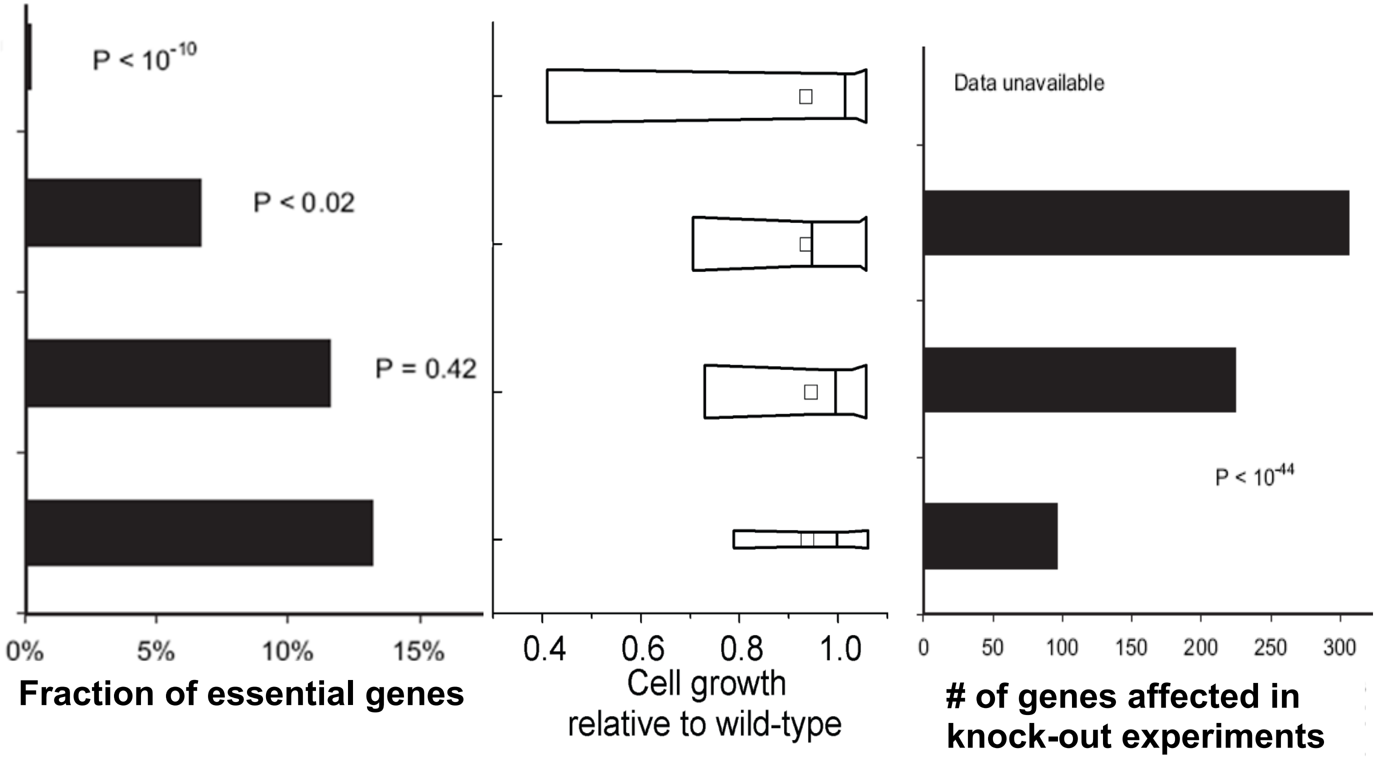
Minimal effect

Adverse effects on gene expression

Induction of conditional phenotype

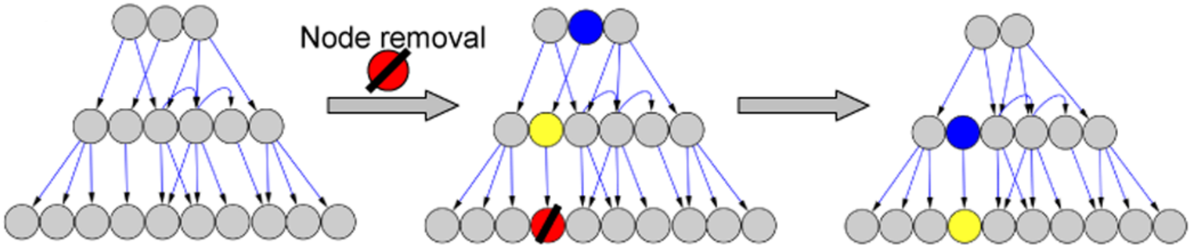
Slow proliferation

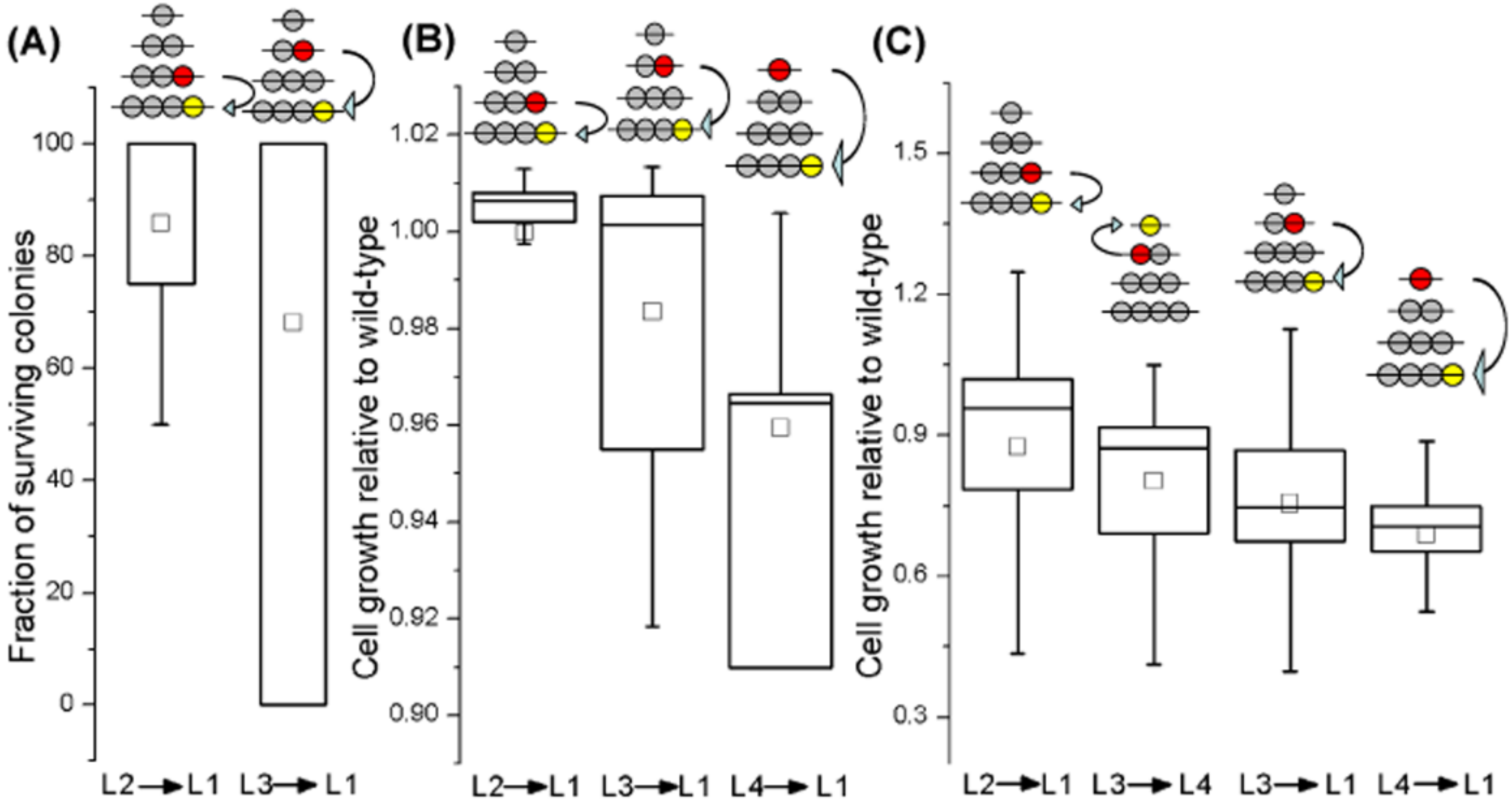
Cell death



[Yu *et al*, *PNAS*, 2006, 103, 40, 14724], [Bhardwaj et al., *Sci. Signal.*, 2010, Vol. **3**, Issue 146, p. ra79]

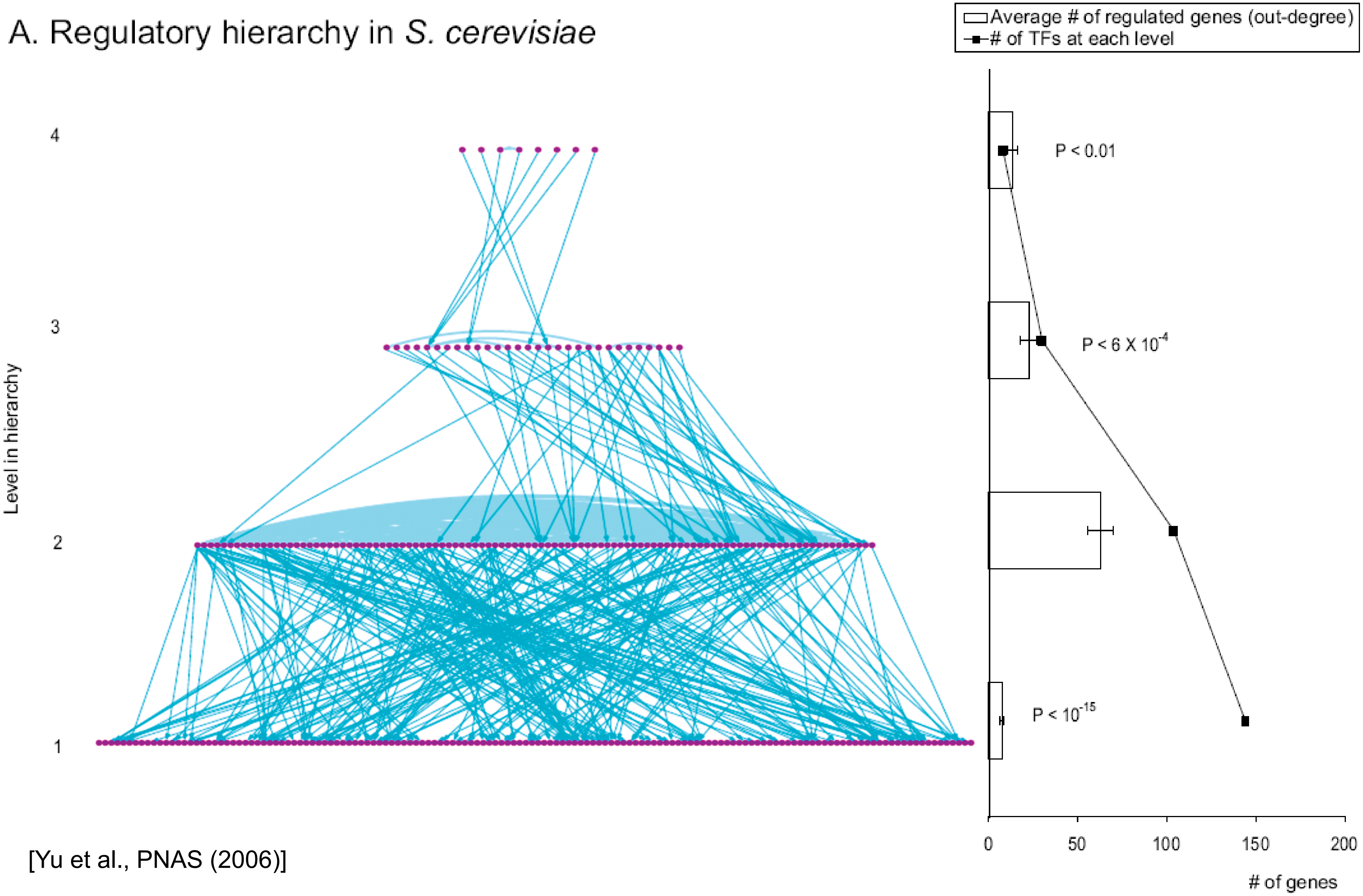
Dynamic changes in the hierarchy: Location of change is a better index





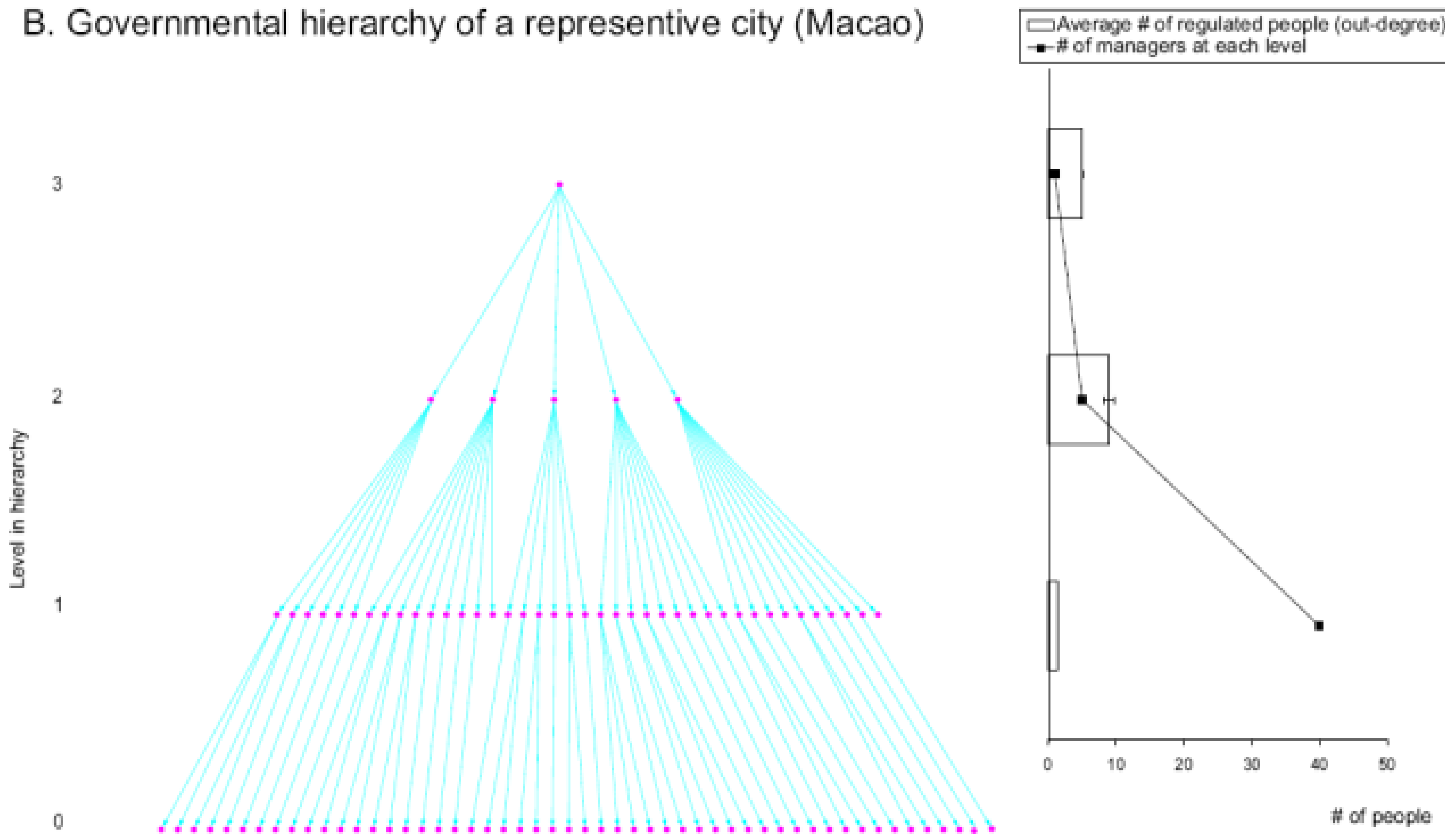
[Bhardwaj et al., *Sci. Signal.*, 2010, Vol. **3**, Issue 146, p. ra79]

Yeast Regulatory Hierarchy: the Middle-managers Rule



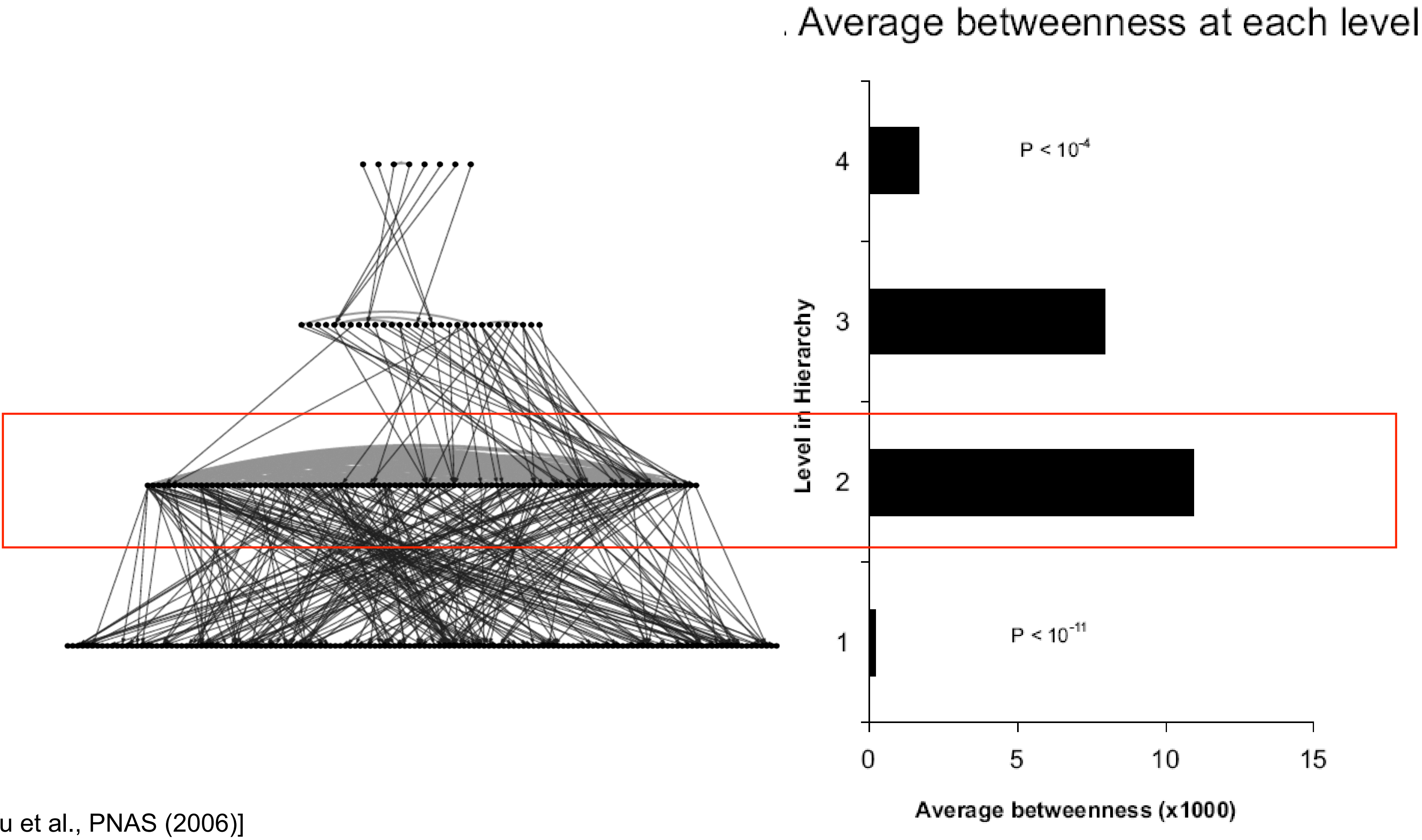
[Yu et al., PNAS (2006)]

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



B. Governmental hierarchy of a representive city (Macao)

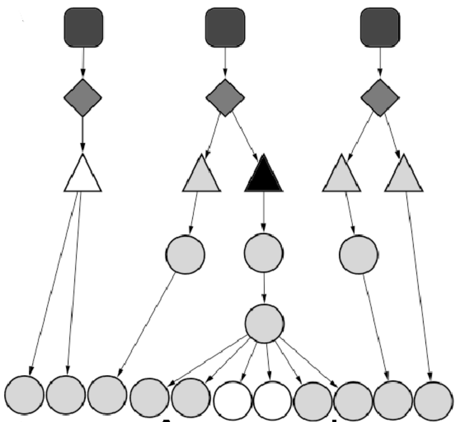
Characteristics of Regulatory Hierarchy: Middle Managers are Information Flow Bottlenecks



[Yu et al., PNAS (2006)]

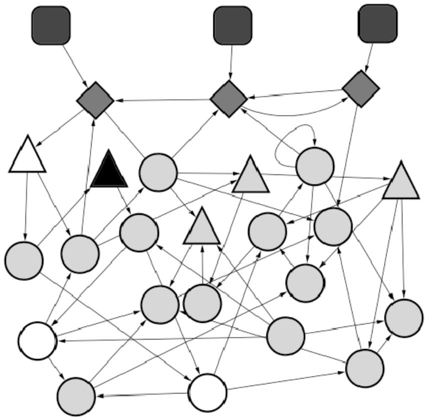
Network Comparison: Broadening the comparison to different types of hierarchies & different types of biological networks

Different kinds of Hierarchies



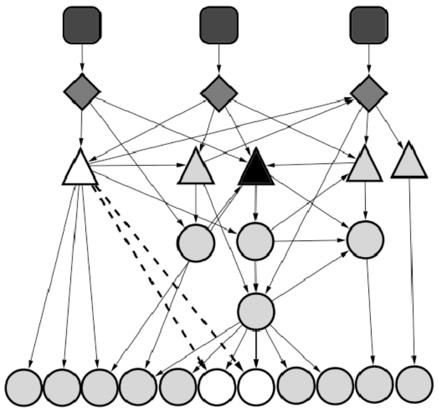
Autocratic

- 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

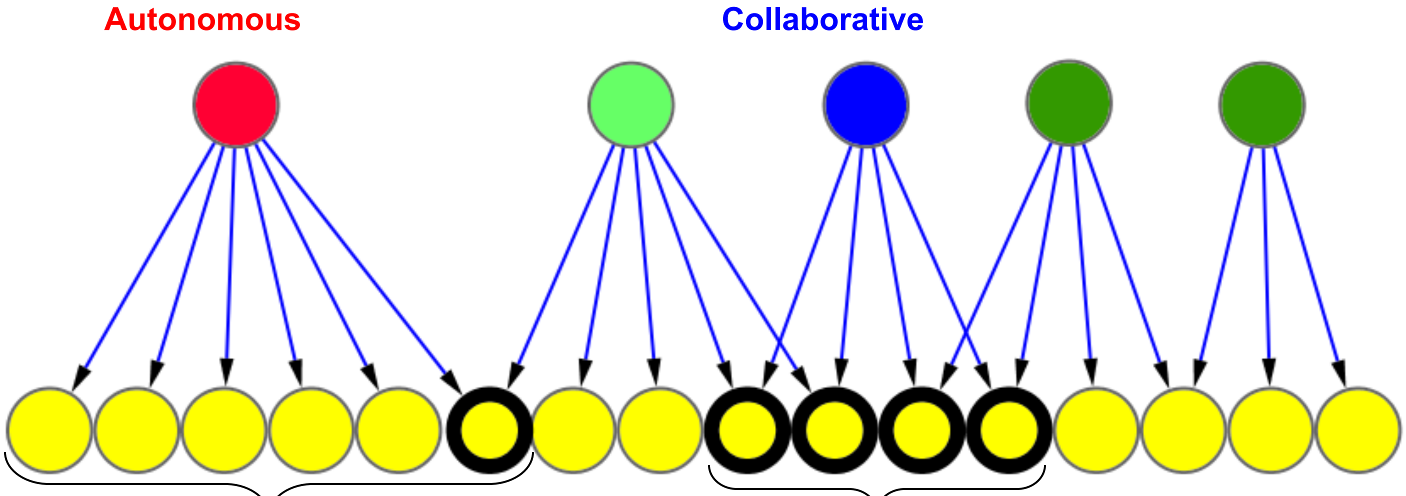
	Autocratic	Democratic	Intermediate
Betweenness (△)	1.03	3.6	3.3
Betweenness (▲)	4.1	1.08	3.4
Var. Betw. (triangles)	2.1	0.58	1.74
Var. Betw. (all)	2.9	1.4	1.9
$D_{Net\text{-}collab}$	0	0.91	0.71

⬇ **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}

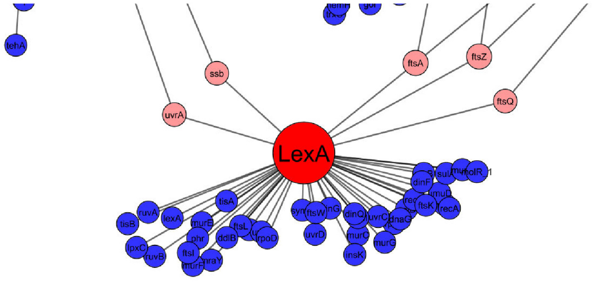
1/6=0.16

4/4=1

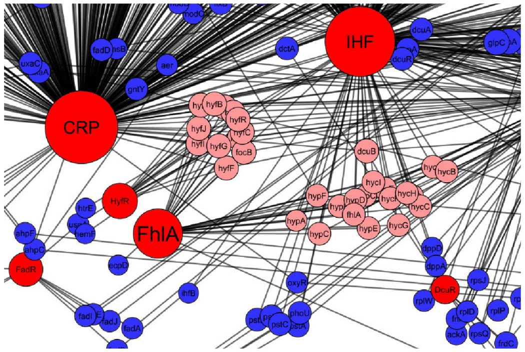
More Collaborative: Democratic

More Autonomous: Autocratic

Autonomous

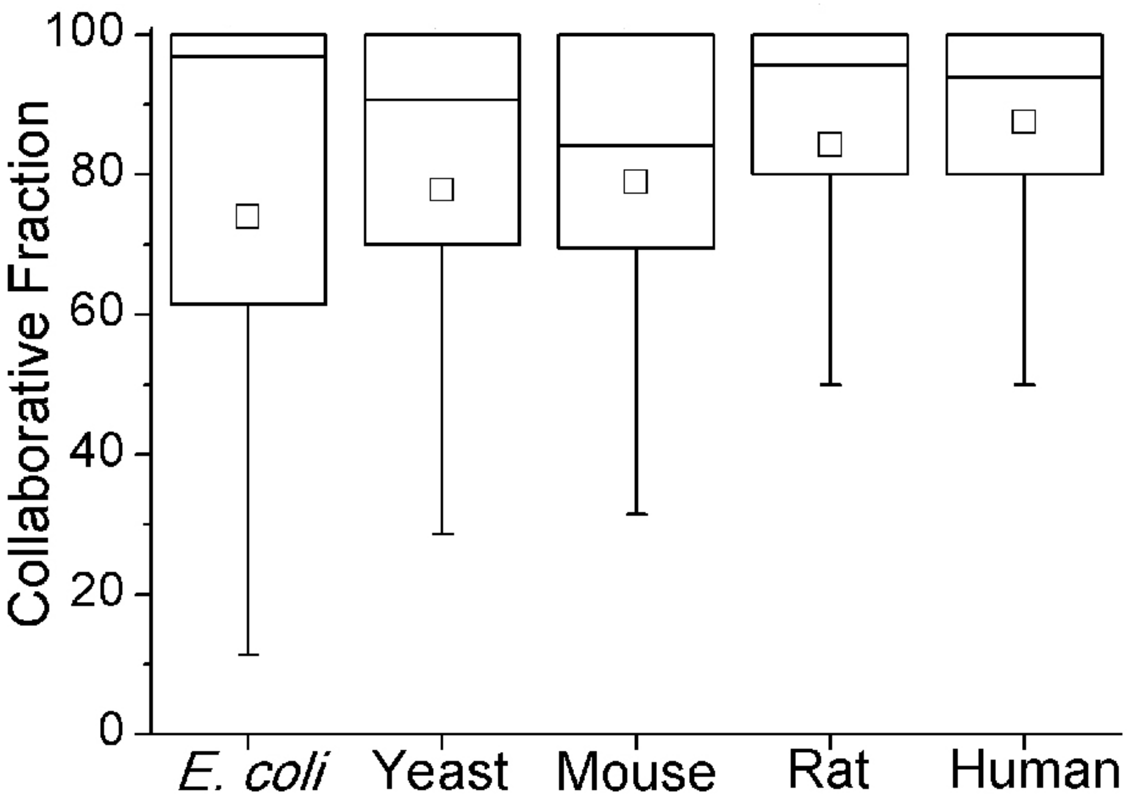


Collaborative



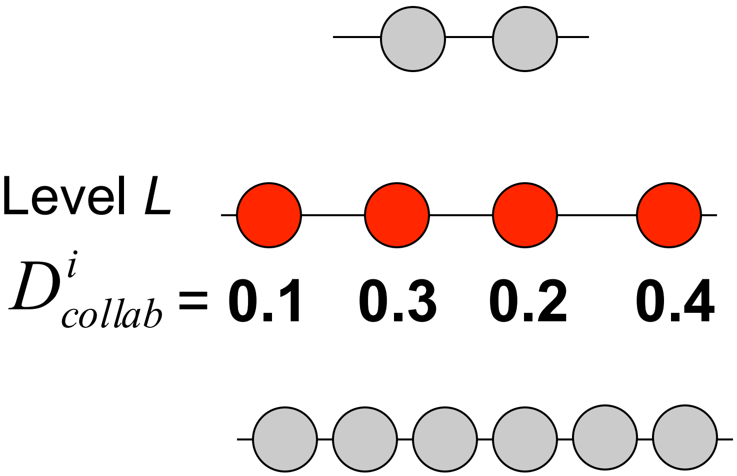
[Bhardwaj et al., PNAS (2010), in press]

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



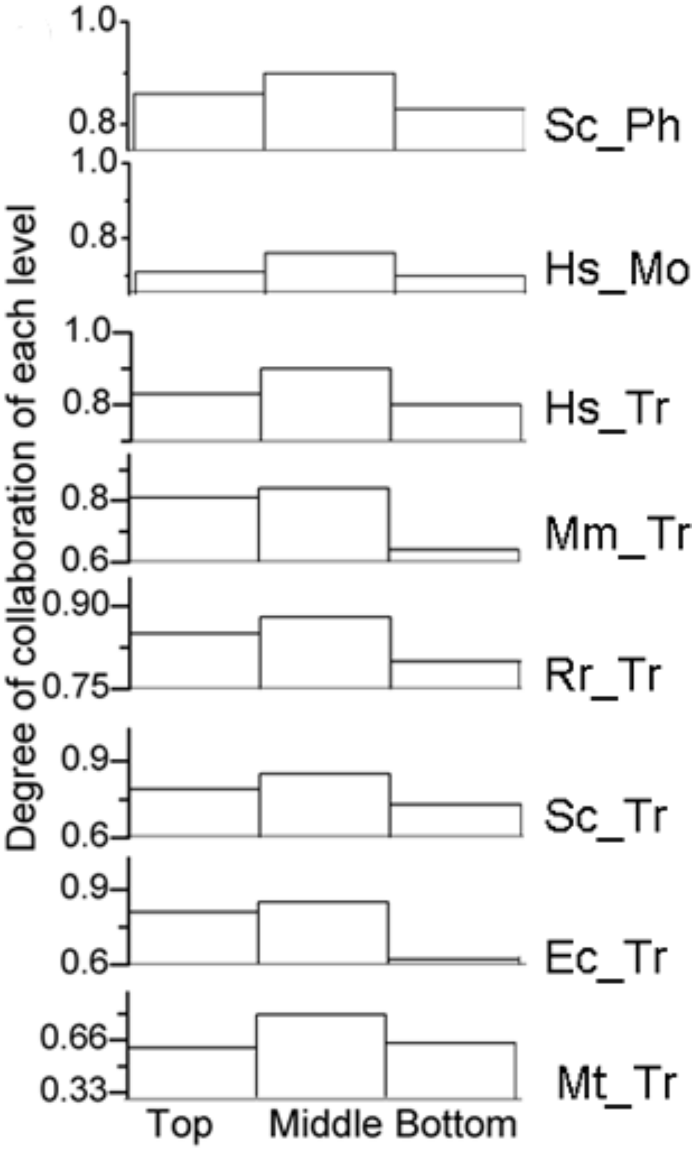
[Bhardwaj et al., PNAS (2010), in press]

Collaborative Nature of the Levels



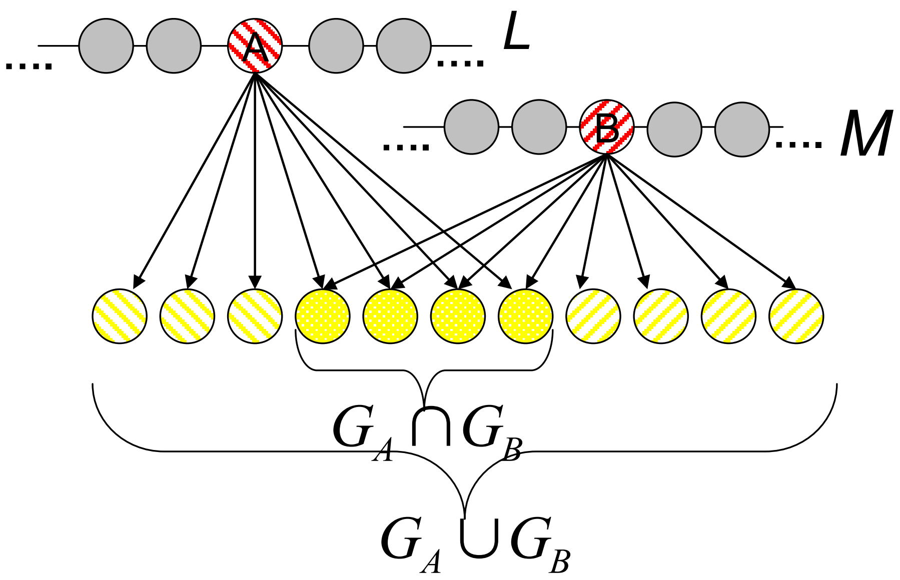
D^{i}_{collab} = **0.1 0.3 0.2 0.4**

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



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

Collaboration Between Levels

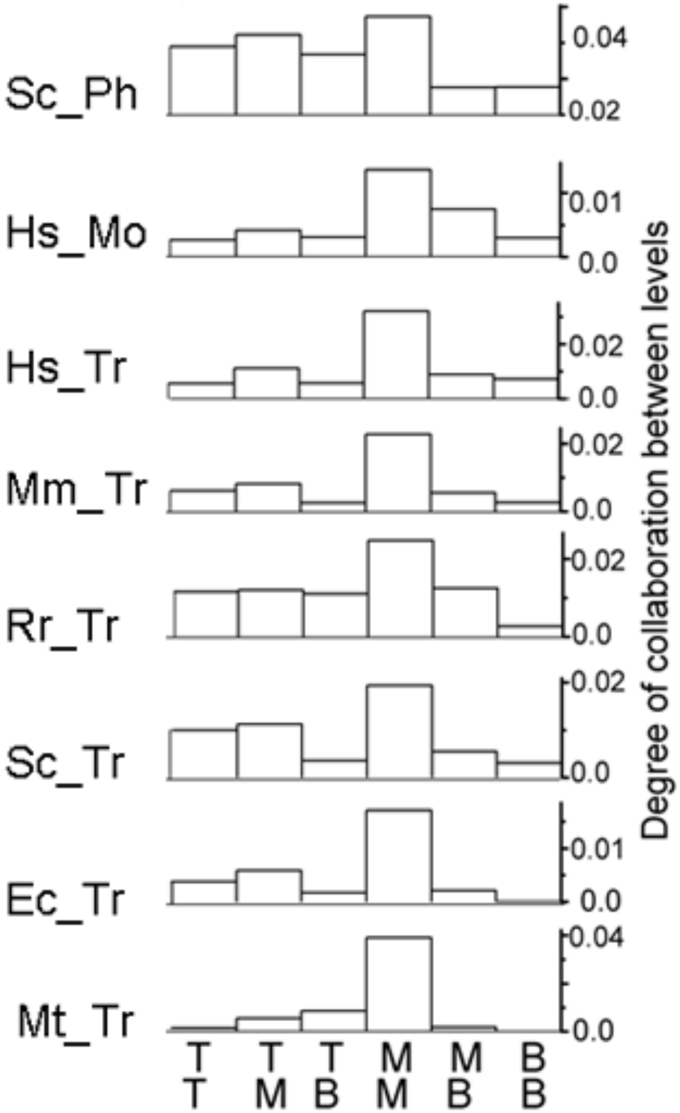


D(collab. betw. level L & M) =

avg. of $$\frac{G_A \cap G_B}{G_A \cup G_B}$$

over all pairs of TFs A in B in L & M

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



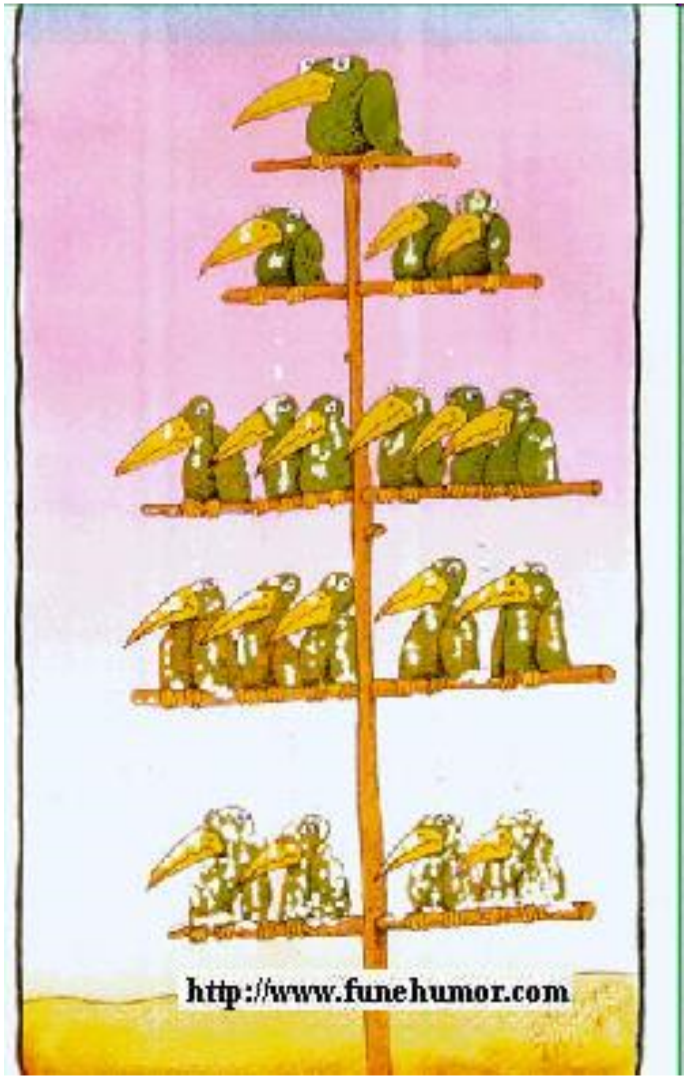
Middle Managers Interact the Most in Efficient Corporate Settings

- 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).

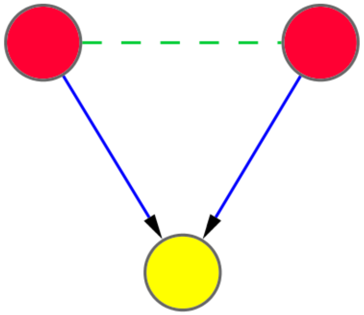
- Floyd, S. W. et al (1993)
  Dinosaurs or Dynamos? Recognizing Middle Management's Strategic Role
  The Academy of Management Executive 8, 47-57.

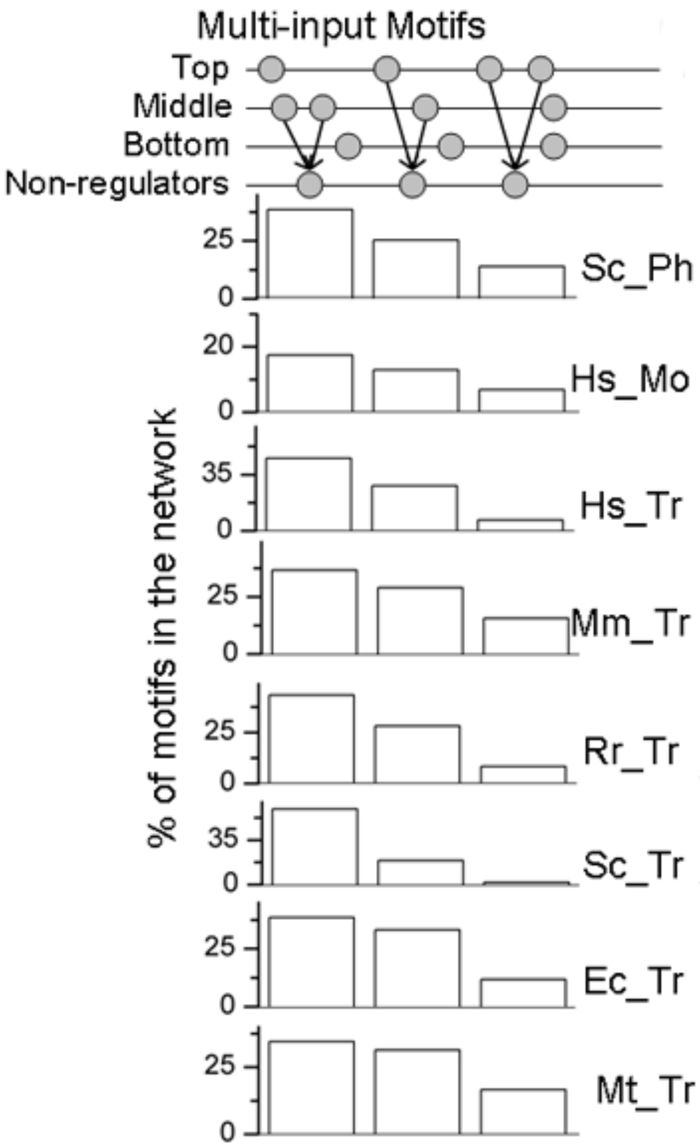
- Floyd, S. W. et al (1997)
  Middle management's strategic influence and organizational performance
  Journal of Management Studies 34, 465-485.



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

Co-regulation Instantiates a Multi-Input Motif

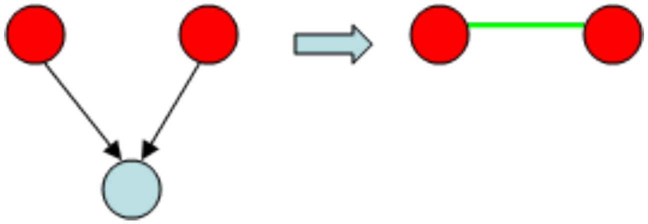




[Bhardwaj et al., PNAS (2010), in press]

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

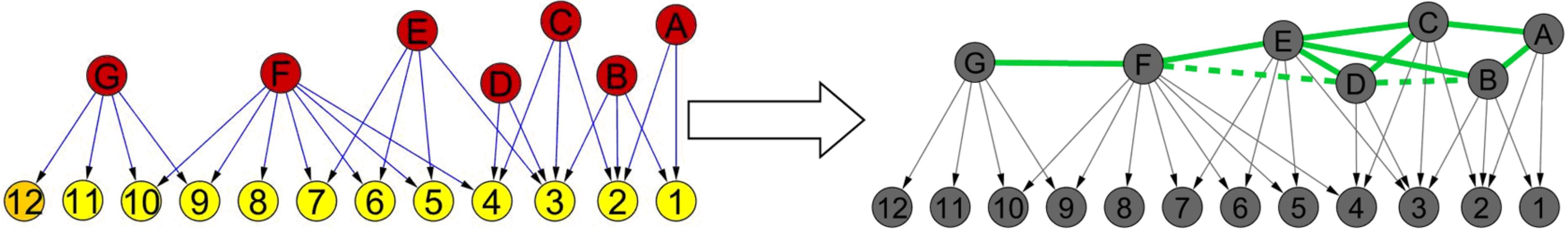
Co-regulation Partnerships



- 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]

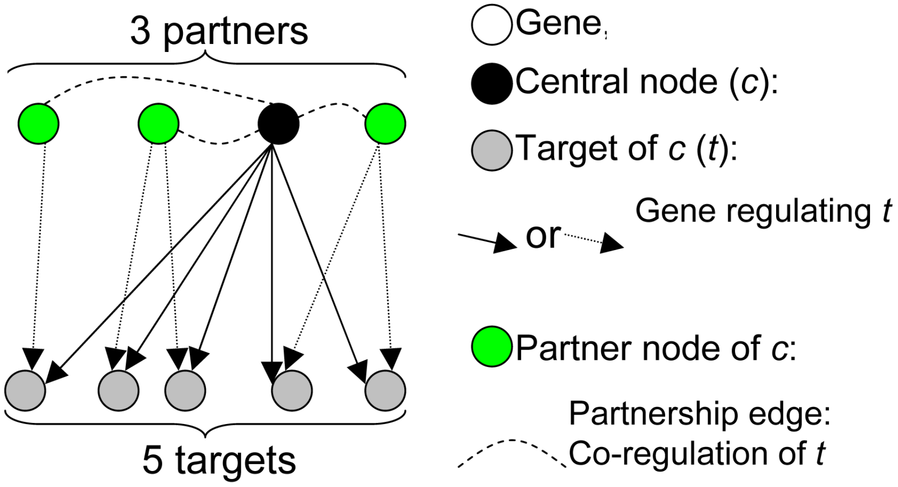
Building and Analysis of Networks

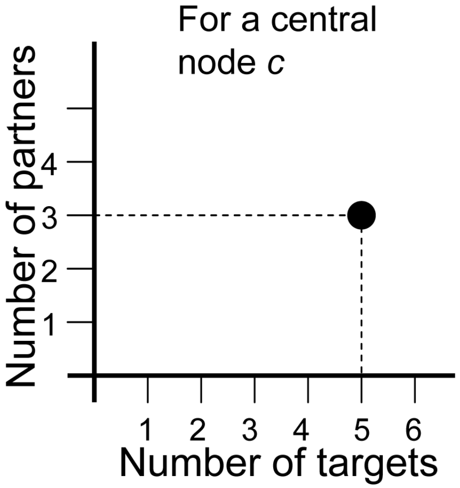


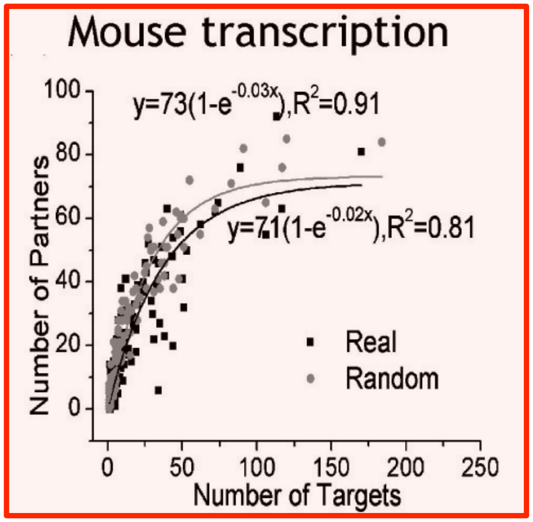
Edge placed if two regulators co-regulate

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]

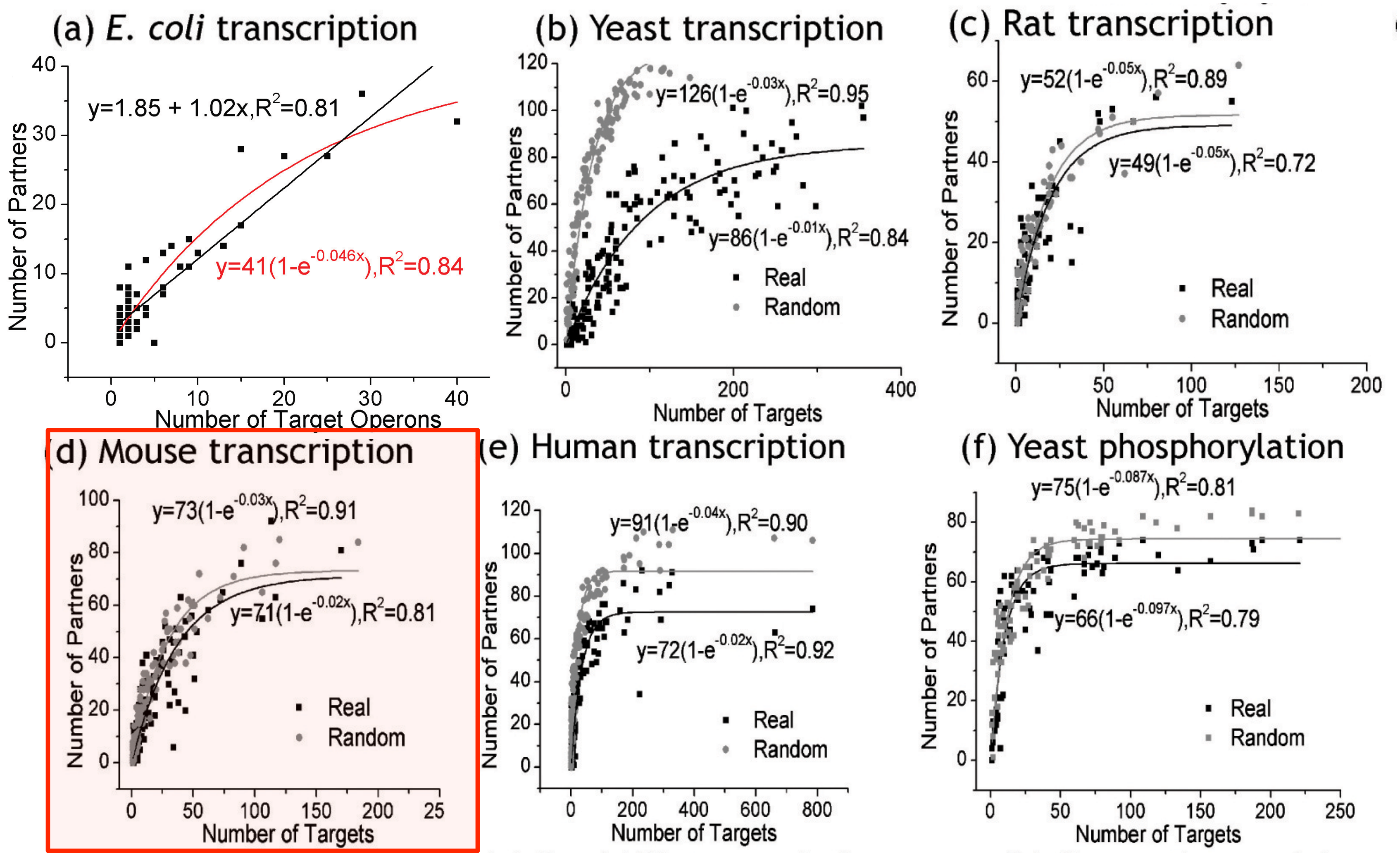






Scaling of Regulatory Partnerships with Targets

Scaling across many networks



(a) *E. coli* transcription

(b) Yeast transcription

(c) Rat transcription

(d) Mouse transcription

(e) Human transcription

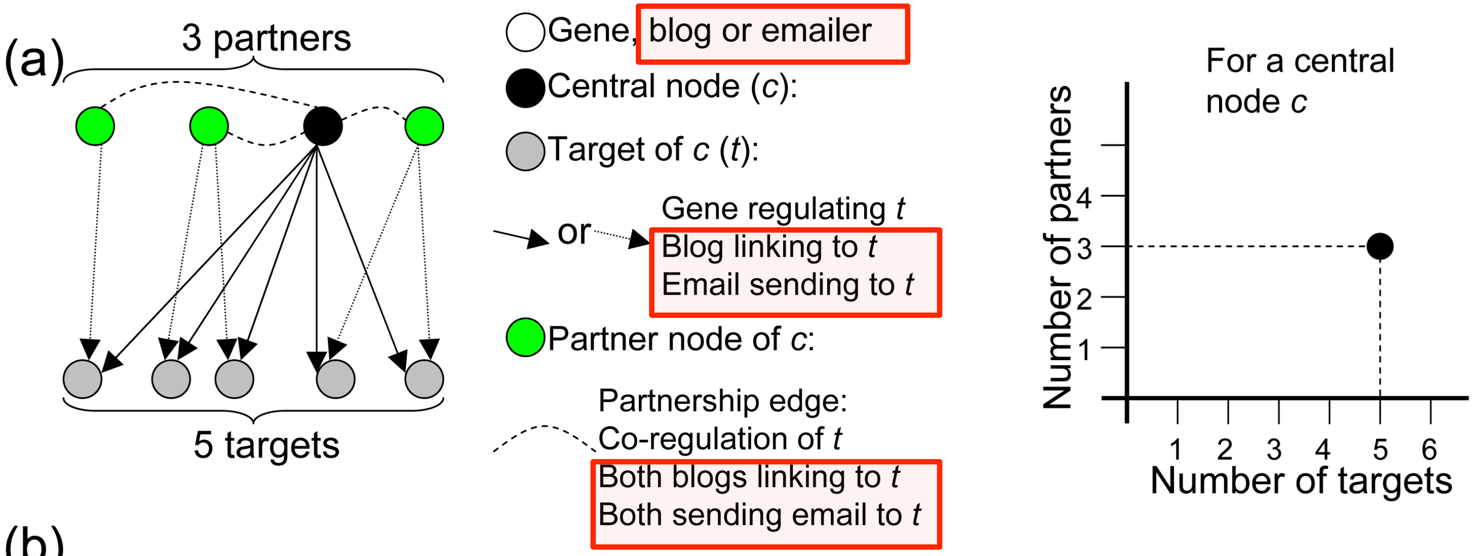
(f) Yeast phosphorylation

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

(a)



○Gene, blog or emailer

●Central node (*c*):

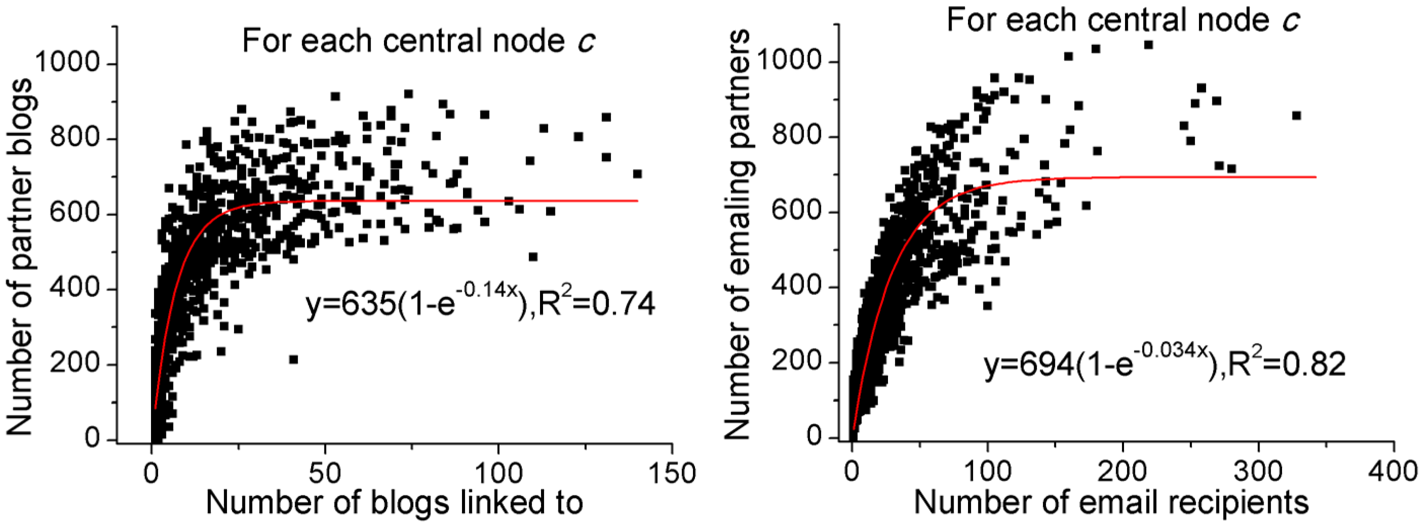
●Target of *c* (*t*):

→ or ⇢ Gene regulating *t*
Blog linking to *t*
Email sending to *t*

●Partner node of *c*:

Partnership edge:
Co-regulation of *t*
Both blogs linking to *t*
Both sending email to *t*

(b)

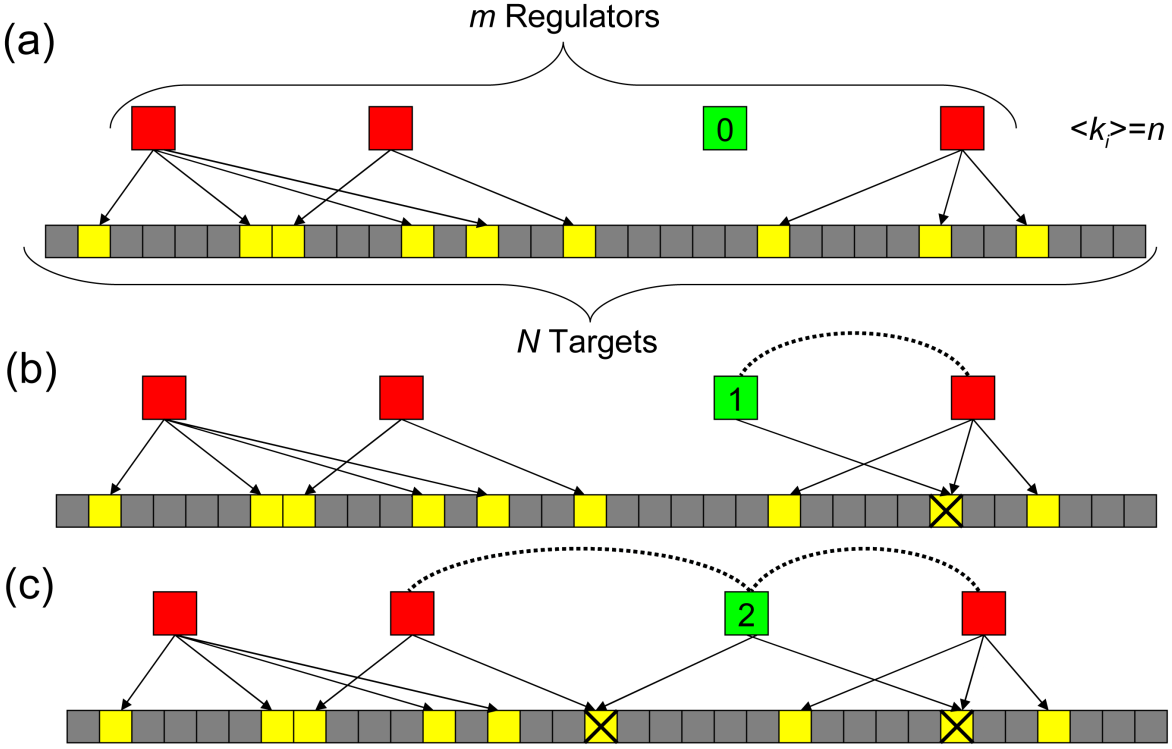


A Simple Theoretical Model

On average, each regulator has *n* targets

f_i : the number of partners

k_i :the number of targets



When f =0,

$<k_i>=n$

$$\left.\frac{\Delta f}{\Delta k}\right|_{f=0} = \frac{nm}{N}$$

When f =1,

$$\left.\frac{\Delta f}{\Delta k}\right|_{f=1} = \frac{(m-1)n}{N}$$

Generalizing,

$$\frac{\partial f}{\partial k} = \frac{(m-f)n}{N}$$

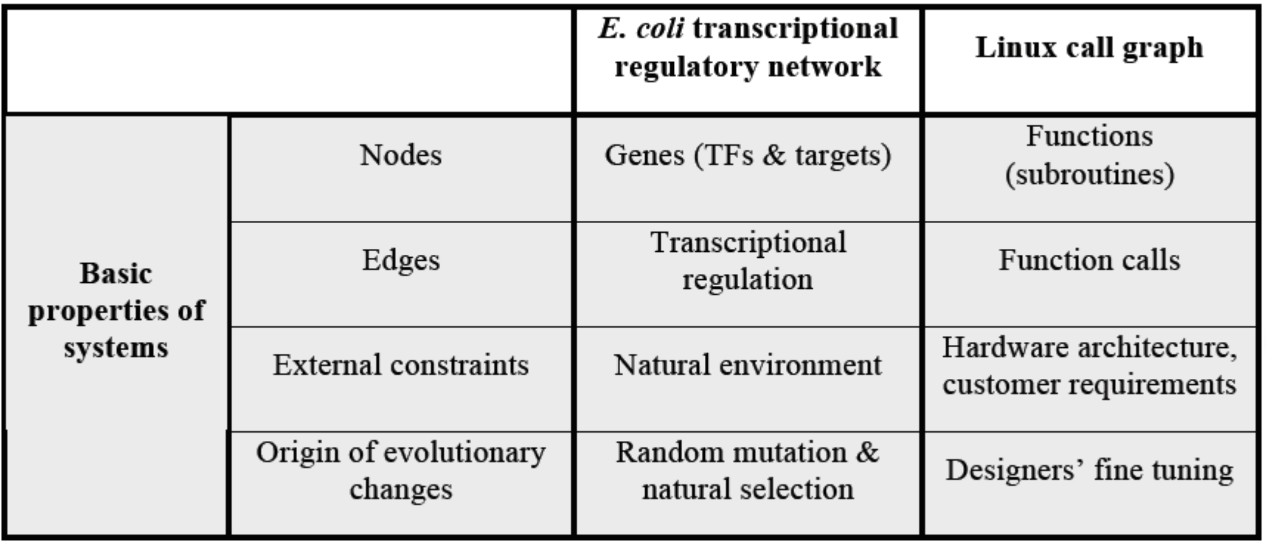
Integrating this, we get:

$$f = m(1-e^{-n/Nk}) = a(1-e^{-bk})$$

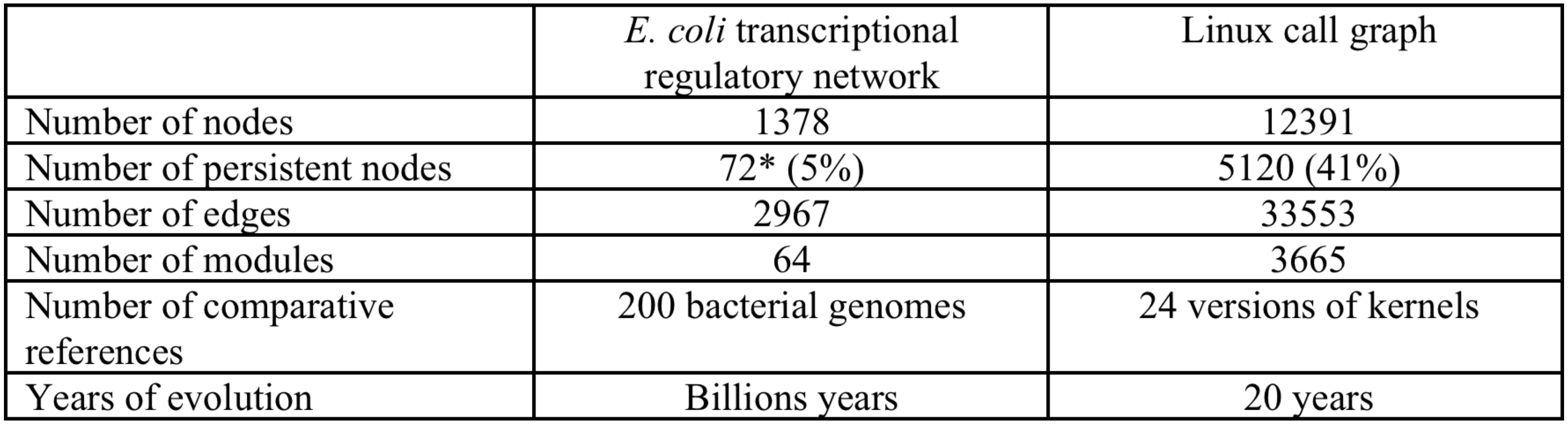
[Bhardwaj et al., PLoS Comp Biol (2010), in press]

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

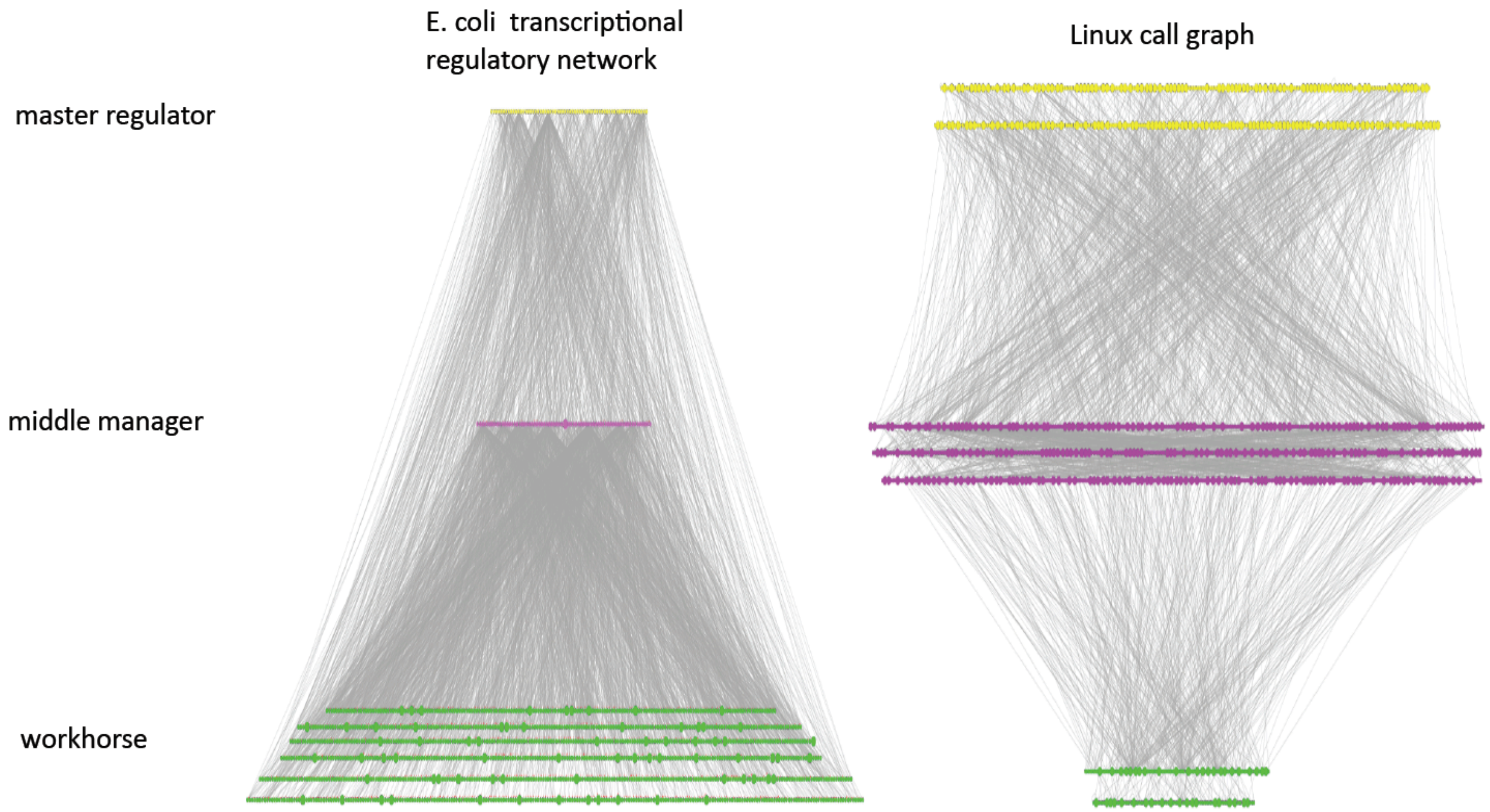


		***E. coli* transcriptional regulatory network**	**Linux call graph**
Basic properties of systems	Nodes	Genes (TFs & targets)	Functions (subroutines)
	Edges	Transcriptional regulation	Function calls
	External constraints	Natural environment	Hardware architecture, customer requirements
	Origin of evolutionary 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

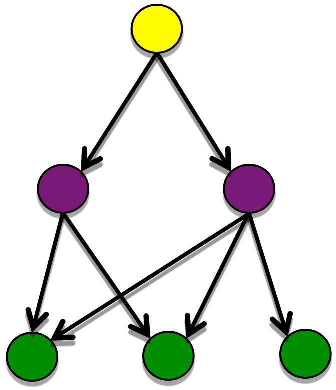
[Yan et al., PNAS (2010), in press]



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

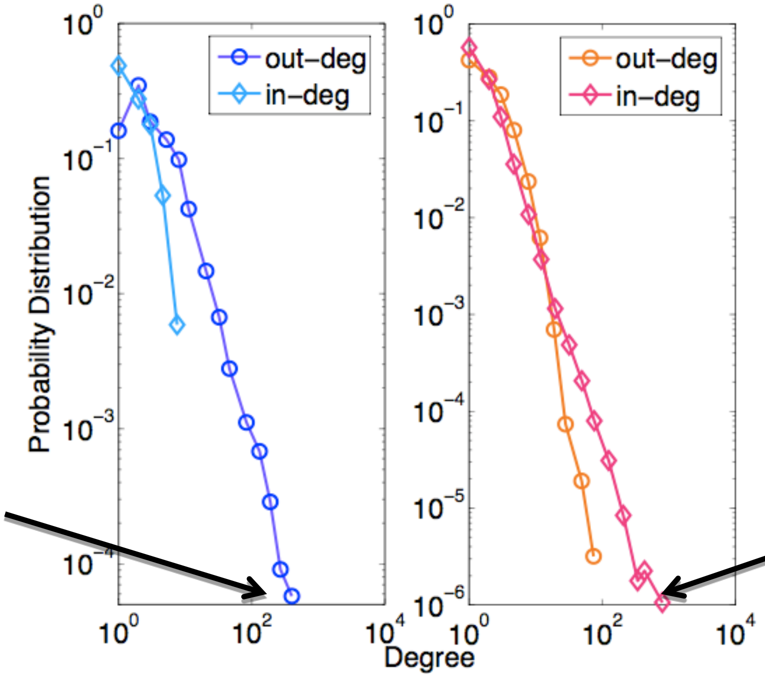
Comparison: hierarchical organization

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



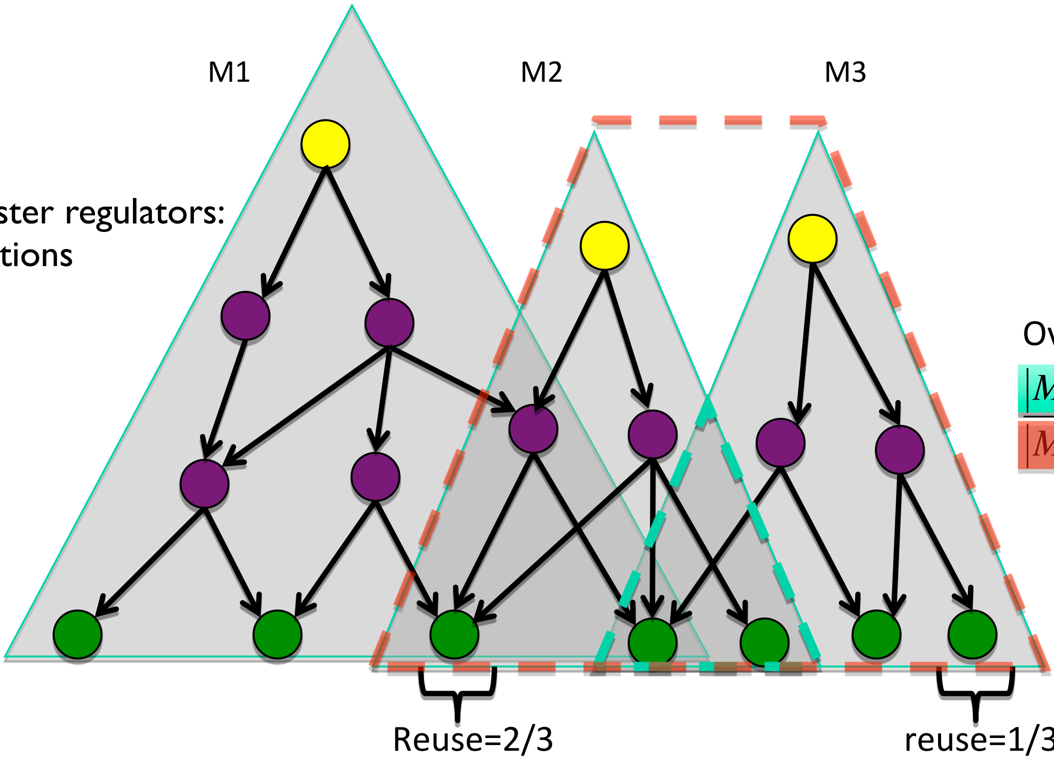
out-degree hubs
e.g. "crp"

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)=

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

TRN:
modules overlap little,
components are
less generic: “ompF”

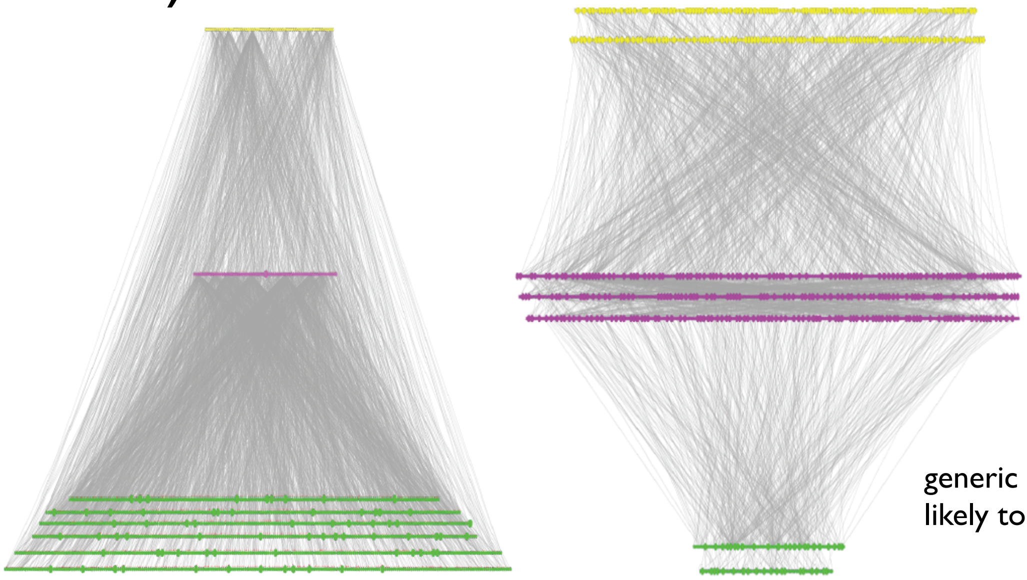
	E. Coli TRN	Linux call graph
# of modules	64	3665
Average overlap	4.3%	80.7%
Maximum node reuse	15.6%	87.5%
Average node reuse	3.5%	8.4%

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)



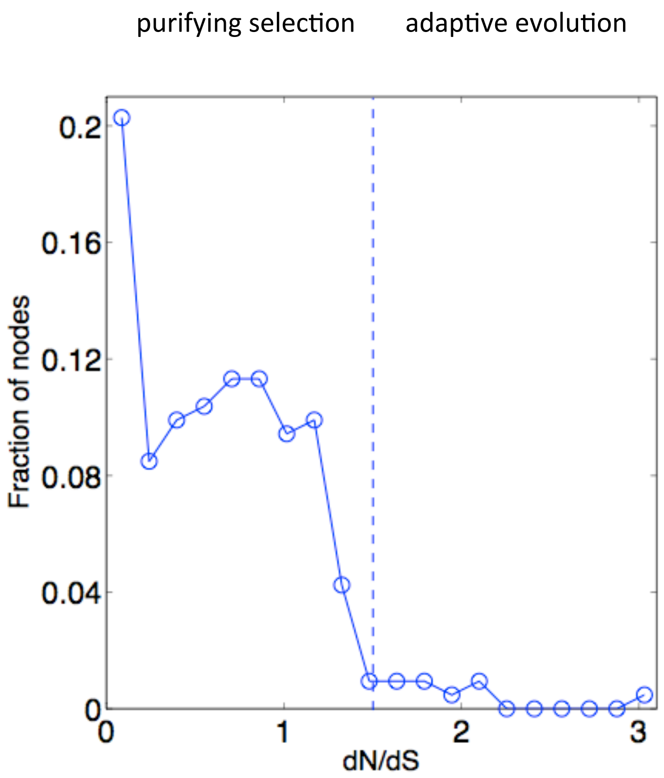
specialized proteins
are preserved
across genomes

generic guys are more
likely to be persistent

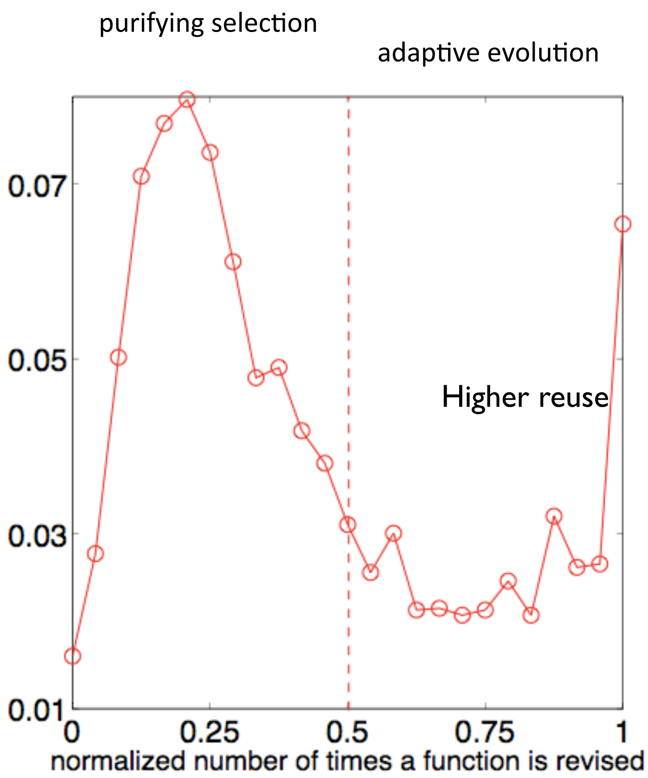
- Building of the hierarchy:
  - ◊ TRN: Bottom up. Regulatory changes are the main driving forces of evolution
  - ◊ Call graph: top down

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

Evolutionary rate of persistent functions



Persistent genes evolve slowly

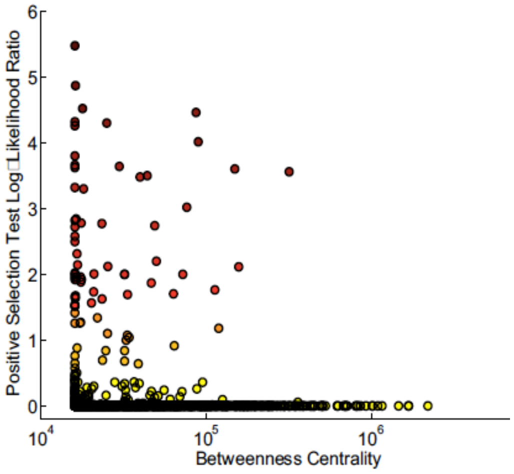


Two classes of persistent functions

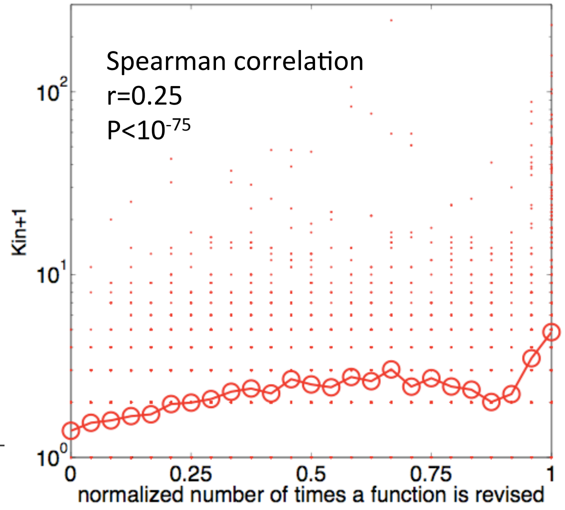
[Yan et al., PNAS (2010), in press]

Why and so what?

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



Kim et.al. PNAS 2007

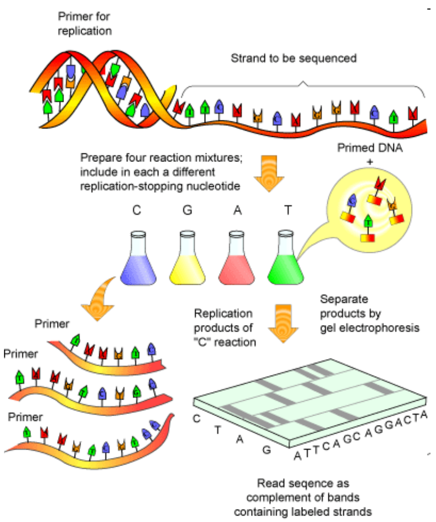


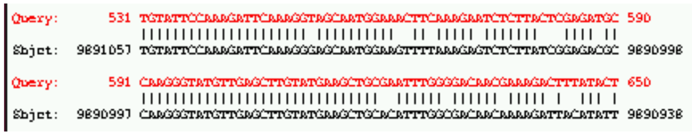
- Independent modules:
  - robust
  - costly: the system needs a variety of tools for different tasks
- Overlap modules (reuse):
  - Less robust:
    - Breakdown of a generic component is harmful to the whole system
    - Fragile in the sense any change in a module may require compensating changes in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

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

Network Comparisons: Direct Comparisons between Species to Estimate Rewiring Rates of Different Types of Networks

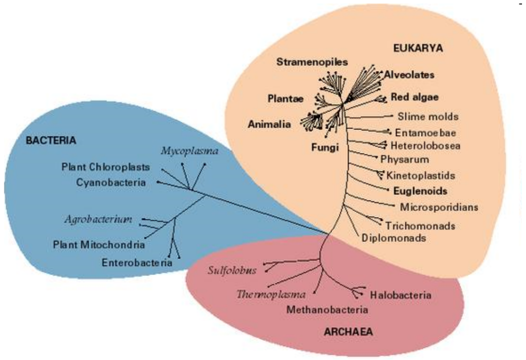
Stages of sequence research





```
Query:      531 TGTATTCCAAAGATTCAAAGGTAGCAATGGAAACTTCAAAGAATCTCTTACTCGGAGATGC 590
                |||||||||||||||||||| ||||||||||  || ||||| |||||||    |||| ||
Sbjct:  9891057 TGTATTCCAAAGATTCAAAGGGAGCAATGGAAGTTTTAAAGAGTCTCTTATCGGAGACGC 9890998

Query:      591 CAAGGGTATGTTGAGCTTGTATGAAGCTGCGAATTTGGGGACAACGAAAGACTTTATACT 650
                ||||||||||||||||||||||||||||||  |||||| |||||| ||||| |  ||| |
Sbjct:  9890997 CAAGGGTATGTTGAGCTTGTATGAAGCTGCACATTTGGCGACAACAAAAGATTACATATT 9890938
```



Sequencing

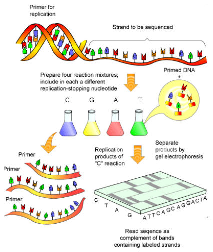
Alignment

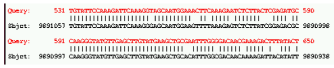
Evolution

How about networks?

http://www.scq.ubc.ca/genome-projects-uncovering-the-blueprints-of-biology/

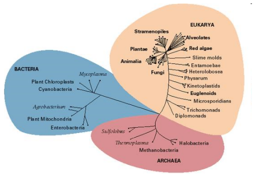
Network follows sequence

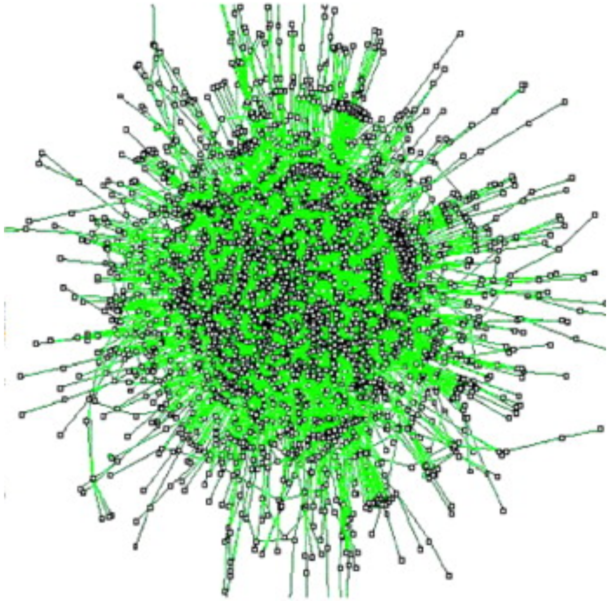


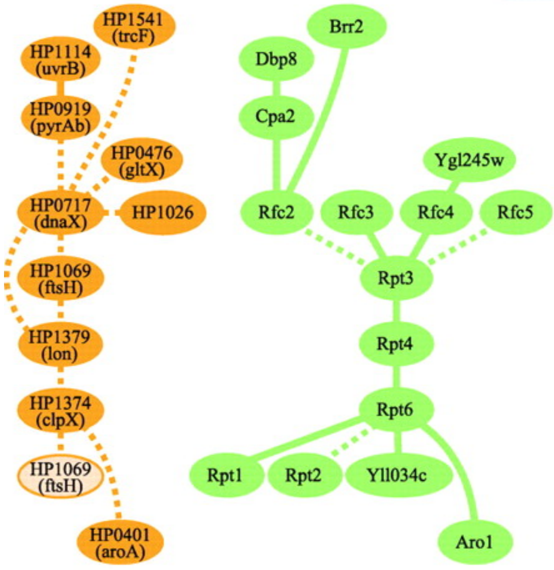


```
Query:      531 TGTATTCCAAAGATTCAAAGGTAGCAATGGAAACTTCAAAGAATCTCTTACTCGAGATGC 590
                ||||||||||||||||||||| ||||||||||  || ||||| |||||||   |||| ||
Sbjct:  9891057 TGTATTCCAAAGATTCAAAGGGAGCAATGGAAGTTTTAAAGAGTCTCTTATCGGAGACGC 9890998

Query:      591 CAAGGGTATGTTGAGCTTGTATGAAGCTGCGAATTTGGGGACAACGAAAGACTTTATACT 650
                ||||||||||||||||||||||||||||||  |||||| |||||| ||||| |  ||| |
Sbjct:  9890997 CAAGGGTATGTTGAGCTTGTATGAAGCTGCACATTTGGCGACAACAAAAGATTACATATT 9890938
```







***

Kelley BP et al., PNAS, 2003

Social network rewiring

Understand how fast the networks (relationships between individual entities) change

The Ronald Reagan family poses in the Red Room for a family portrait on his inaugural, January 20, 1981.

Fast

Moderate

Slow

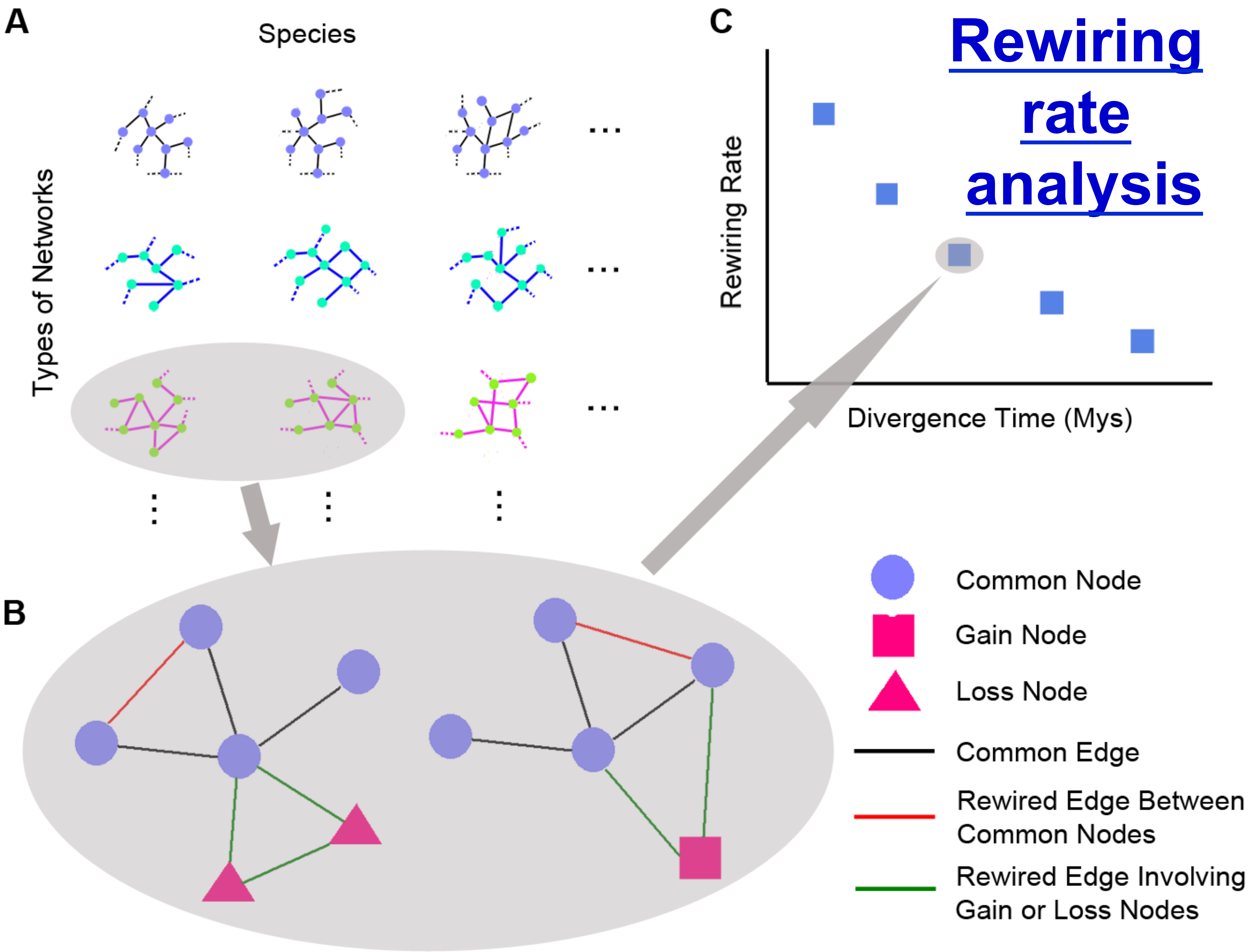
Systematic study of biological network rewiring

- Idea
  - ◊ Quantify network rewiring in a unified way for all types of biological networks
  - ◊ Compare rewiring rates w.r.t. divergence and network types
  - ◊ Compare network rewiring to sequence change
- Goal
  - ◊ Understand evolutionary characteristics of biological networks
  - ◊ Implications on species evolution

Current network data

Species	scer	calb	spom	other fungi	worm	fly	mouse	human
PPI	yes	no	yes	no	yes	yes	no	yes
Genetic	yes	no	yes	no	yes	yes	no	yes
TF	yes	yes	no	yes	yes	yes	yes	yes
miRNA	Not Exist				yes	yes	yes	yes
Phospho	yes	yes	yes	no	no	no	no	yes
Metabolic	yes	yes	yes	yes	yes	yes	yes	yes

Rewiring rate analysis



Define rewiring rate

- Given two networks with certain time divergence

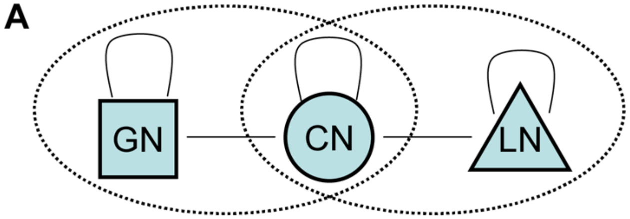
$$\text{Rewiring rate} = \frac{\text{R}}{\text{C} \times \text{Time divergence}}$$

R: total number of rewired edges

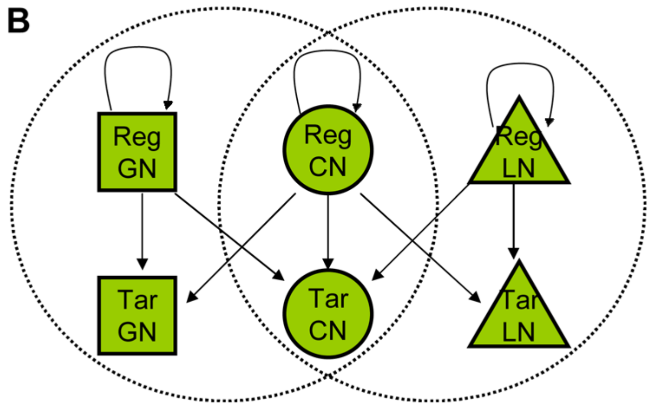
C: total number of possible rewiring edges (number of edges in a complete full network)

- Pros
  - ◊ Normalized by the size of two comparing networks
  - ◊ Considering overlap of nodes
  - ◊ Unified measure applicable to all types of biological networks

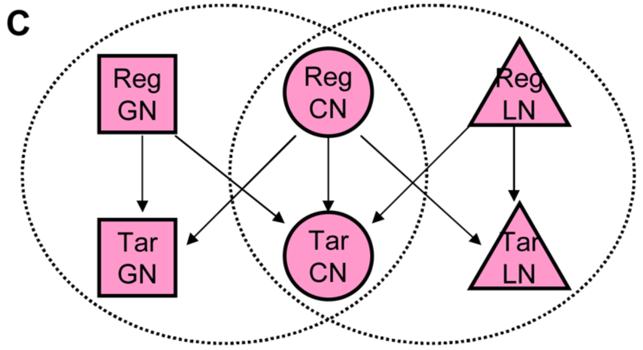
Number of total possible changes



$$\frac{\text{CNs}\times(\text{CNs}-1)+\text{GNs}\times(\text{GNs}-1)+\text{LNs}\times(\text{LNs}-1)}{2}+\text{CNs}\times(\text{GNs}+\text{LNs})$$

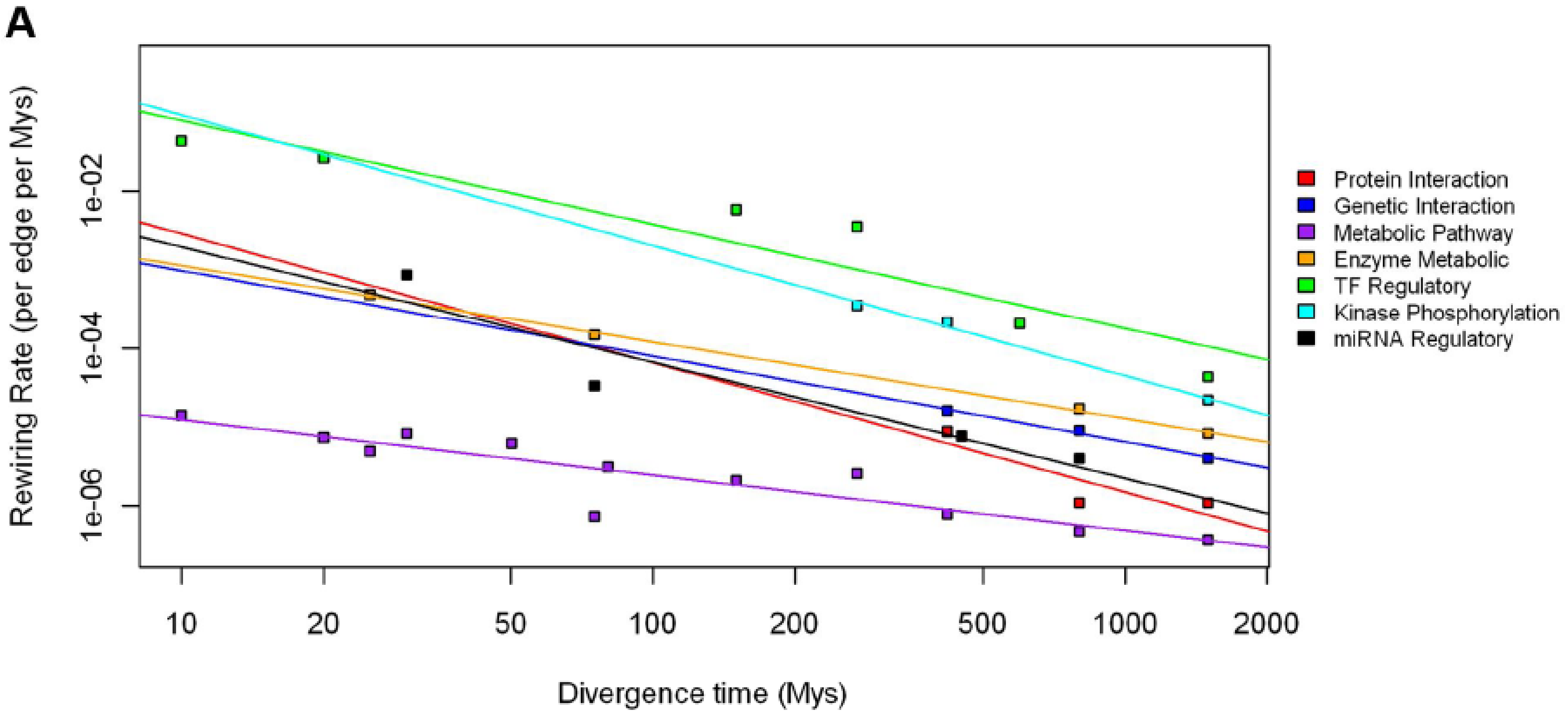


$$\frac{\text{Reg CNs}\times(\text{Reg CNs - 1})+\text{Reg GNs}\times(\text{Reg GNs - 1})+\text{Reg LNs}\times(\text{Reg LNs - 1})}{2}+$$
$$\text{Reg CNs}\times\text{Tar CNs}+\text{Reg GNs}\times\text{Tar GNs}+\text{Reg LNs}\times\text{Tar LNs}+$$
$$\text{Reg CNs}\times(\text{Tar GNs}+\text{Tar LNs})+\text{Tar CNs}\times(\text{Reg GNs}+\text{Reg LNs})$$



$$\text{Reg CNs}\times\text{Tar CNs}+\text{Reg GNs}\times\text{Tar GNs}+\text{Reg LNs}\times\text{Tar LNs}+$$
$$\text{Reg CNs}\times(\text{Tar GNs}+\text{Tar LNs})+\text{Tar CNs}\times(\text{Reg GNs}+\text{Reg LNs})$$

Rewiring rate bands



Linear in Log-Log scale

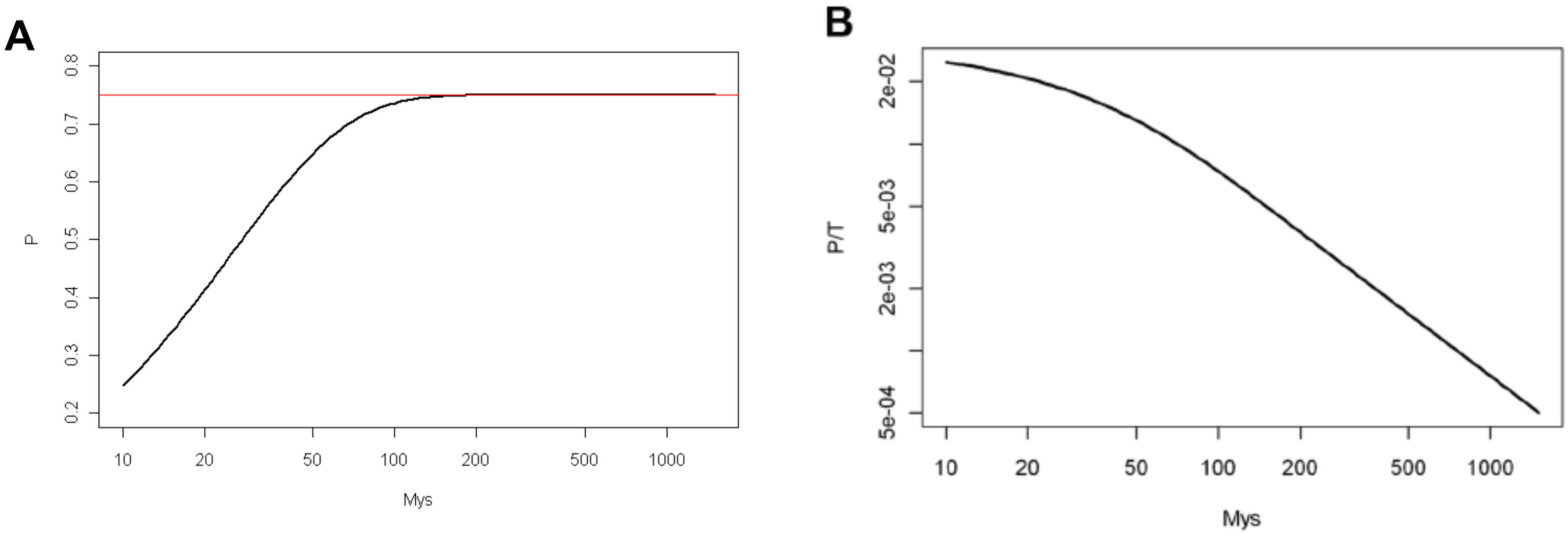
Biological Network	Linear Regression Model	Correlation Coefficient
Transcription factor regulatory network	$\log(\mathrm{r}) = (-1.32 \pm 0.63) \times \log(t) + (0.22 \pm 1.43)$	-0.95
Kinase phosphorylation network	$\log(\mathrm{r}) = (-1.66 \pm 1.97) \times \log(t) + (0.62 \pm 5.45)$	-0.96
miRNA regulatory network	$\log(\mathrm{r}) = (-1.47 \pm 1.61) \times \log(t) + (-1.24 \pm 3.68)$	-0.59
Protein interaction network	$\log(\mathrm{r}) = (-1.64 \pm 11.83) \times \log(t) + (-0.90 \pm 34.43)$	-0.77
Genetic interaction network	$\log(\mathrm{r}) = (-1.09 \pm 1.59) \times \log(t) + (-1.92 \pm 4.61)$	-0.97
Metabolic enzyme network	$\log(\mathrm{r}) = (-0.97 \pm 0.09) \times \log(t) + (-1.98 \pm 0.22)$	-0.54
Metabolic pathway network	$\log(\mathrm{r}) = (-0.70 \pm 0.24) \times \log(t) + (-4.21 \pm 0.5)$	-0.75

95% confidence intervals for the fitted parameters are computed for linear models.

Sequence evolution

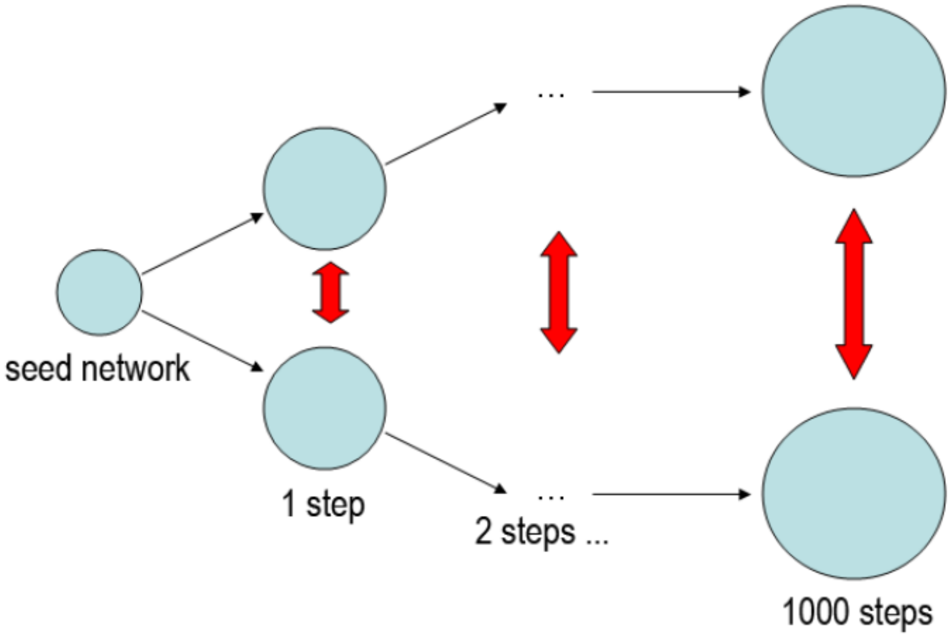
- Jukes-Cantor model

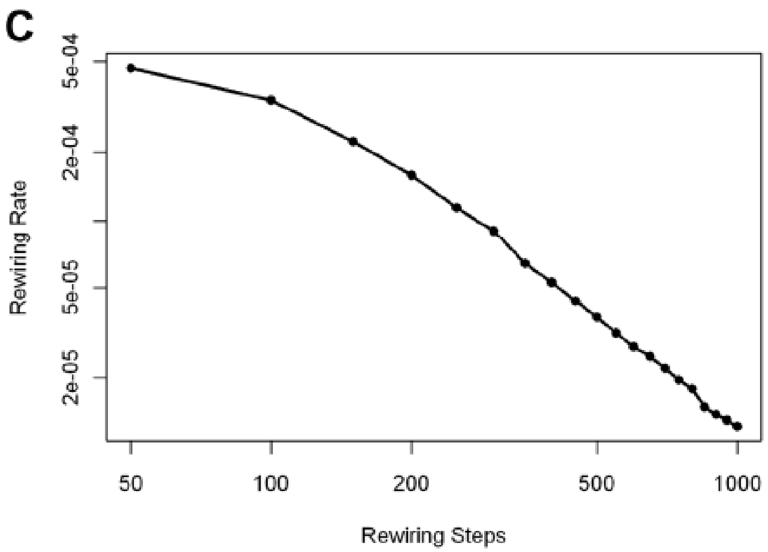
$$P = \frac{3}{4} - \frac{3}{4}e^{-8\alpha T}$$



Saturation of nucleotide changes

Simulated network rewiring





Hypothetical networks are rewired by random simulation, add/remove nodes and edges with preferential attachment.

Rewiring rates of networks

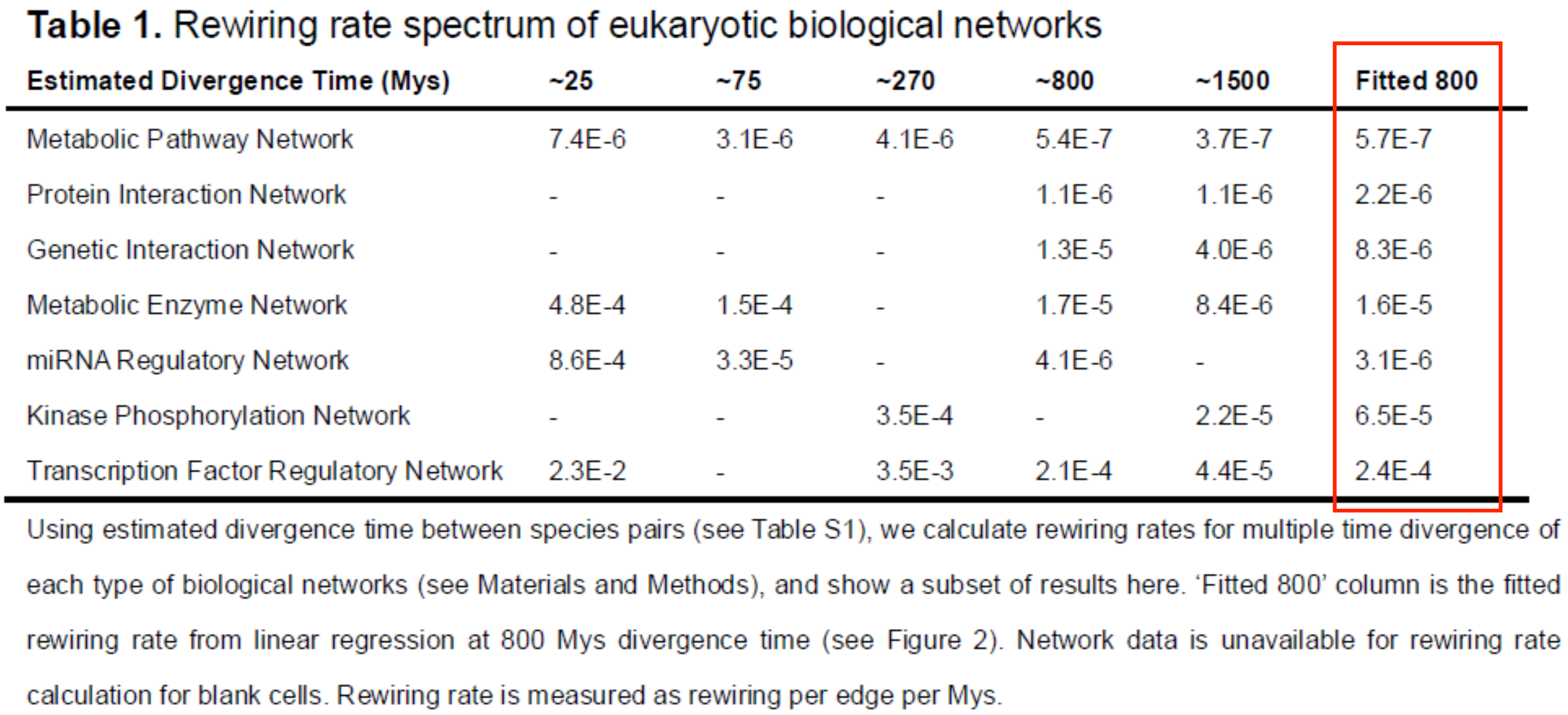


Table 1. Rewiring rate spectrum of eukaryotic biological networks

Estimated Divergence Time (Mys)	**~25**	**~75**	**~270**	**~800**	**~1500**	**Fitted 800**
Metabolic Pathway Network	7.4E-6	3.1E-6	4.1E-6	5.4E-7	3.7E-7	5.7E-7
Protein Interaction Network	-	-	-	1.1E-6	1.1E-6	2.2E-6
Genetic Interaction Network	-	-	-	1.3E-5	4.0E-6	8.3E-6
Metabolic Enzyme Network	4.8E-4	1.5E-4	-	1.7E-5	8.4E-6	1.6E-5
miRNA Regulatory Network	8.6E-4	3.3E-5	-	4.1E-6	-	3.1E-6
Kinase Phosphorylation Network	-	-	3.5E-4	-	2.2E-5	6.5E-5
Transcription Factor Regulatory Network	2.3E-2	-	3.5E-3	2.1E-4	4.4E-5	2.4E-4

Using estimated divergence time between species pairs (see Table S1), we calculate rewiring rates for multiple time divergence of each type of biological networks (see Materials and Methods), and show a subset of results here. 'Fitted 800' column is the fitted rewiring rate from linear regression at 800 Mys divergence time (see Figure 2). Network data is unavailable for rewiring rate calculation for blank cells. Rewiring rate is measured as rewiring per edge per Mys.

Instantaneous vs. averaged rates

Sequences (Jukes-Cantor model)

$$P = \frac{3}{4} - \frac{3}{4} e^{-8\alpha t} \Rightarrow \frac{dP}{dt}\,(\text{nucleotide change rate}) = \frac{6\alpha}{\exp(8\alpha t)} \Rightarrow \frac{dP}{dt}_{\,t \rightarrow 0} = 6\alpha$$

Networks

$$\text{Rewiring rate} = \frac{\Delta \text{R}}{\text{C} \times \Delta \text{Time divergence}}$$

At very small time divergence, ~10 Mys, very few networks are available for close species.

Our rewiring rate quantification is an averaged rate measure for the purpose of comparison between network types.

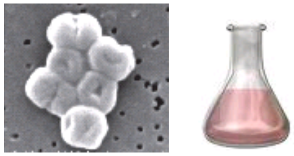
Network Dynamics Across Environments: Metabolic Pathways

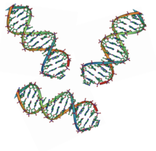
How do molecular networks change across environments?
What pathways are used more ?
Used as a biosensor ?

What is Metagenomics?

Traditional Genomics

Select organism and culture

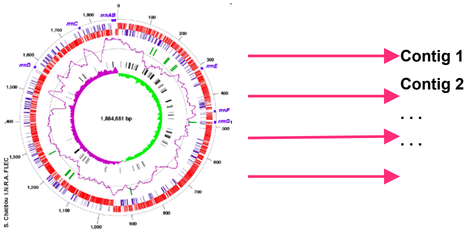




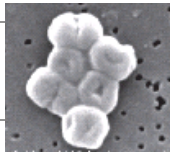
Extract DNA and sequence

atgctcgatctcg
atcgatctcgctg
atgccgatctaa

Assemble and annotate

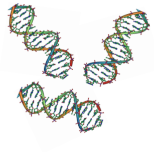


Estimated that less than 1% of microbes can be cultured



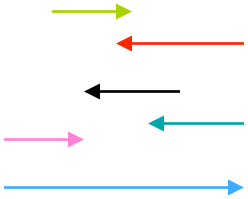
Metagenomics

Collect sample from environment



atgctcgatctcg
atcgatctcgctg
atgccgatctaa

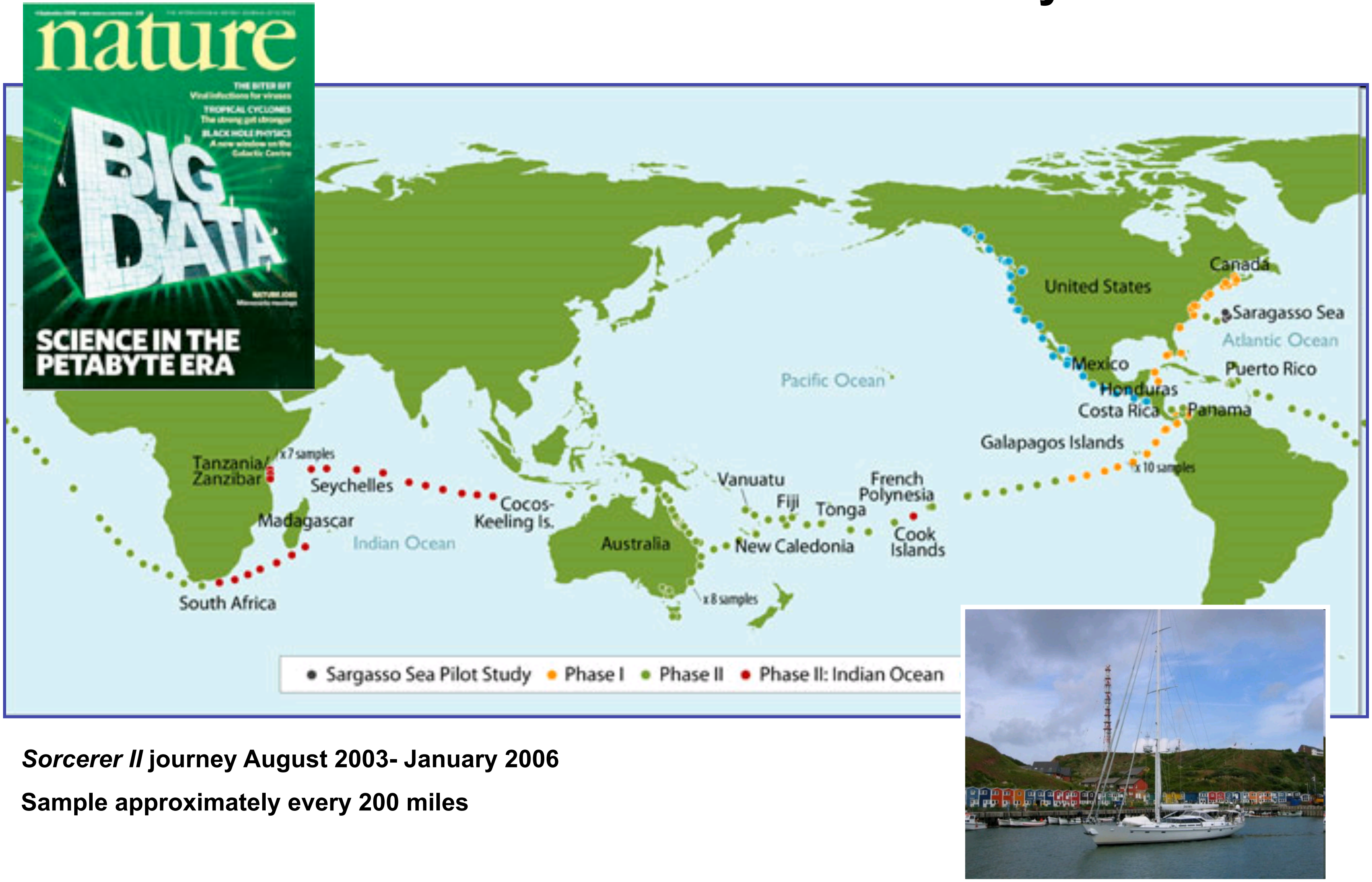
Extract DNA and sequence



Assemble and annotate

Lose information about which gene belongs to which microbe

Sorcerer II Global Ocean Survey

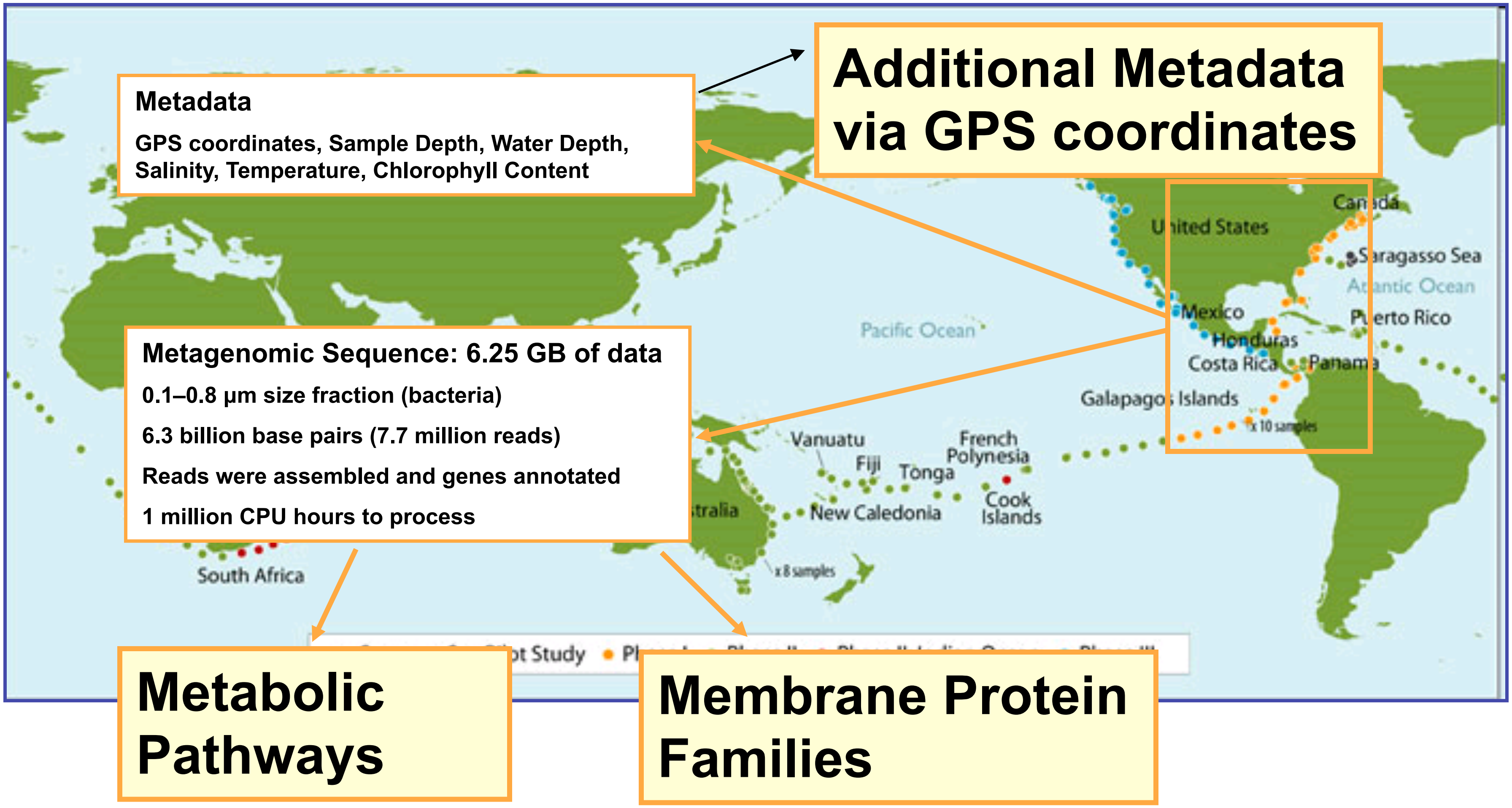


***Sorcerer II* journey August 2003- January 2006**

Sample approximately every 200 miles

Rusch, et al., *PLOS Biology* 2007

Sorcerer II Global Ocean Survey



Metadata

GPS coordinates, Sample Depth, Water Depth, Salinity, Temperature, Chlorophyll Content

Additional Metadata via GPS coordinates

Metagenomic Sequence: 6.25 GB of data

0.1–0.8 µm size fraction (bacteria)

6.3 billion base pairs (7.7 million reads)

Reads were assembled and genes annotated

1 million CPU hours to process

Metabolic Pathways

Membrane Protein Families

Rusch, et al., *PLOS Biology* 2007

Extracting Environmental Data from Other Sources

Sample Depth: 1 meter

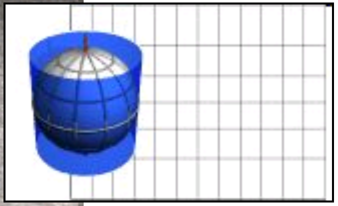
Water Depth: 32 meters

Chlorophyll: 4.0 ug/kg

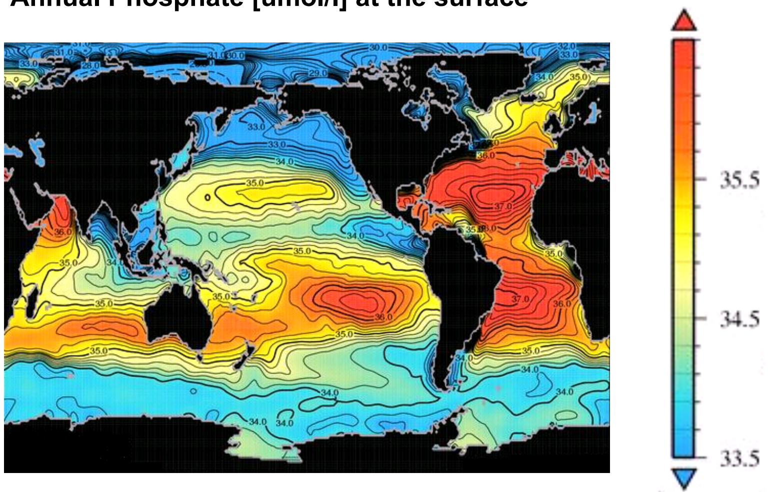
Salinity: 31 psu

Temperature: 11 C

Location: 41°5'28"N, 71°36'8"W



Annual Phosphate [umol/l] at the surface



World Ocean Atlas 2005
NOAA/NODC

Nutrient Features Extracted:

Phosphate

Silicate

Nitrate

Apparent Oxygen Utilization

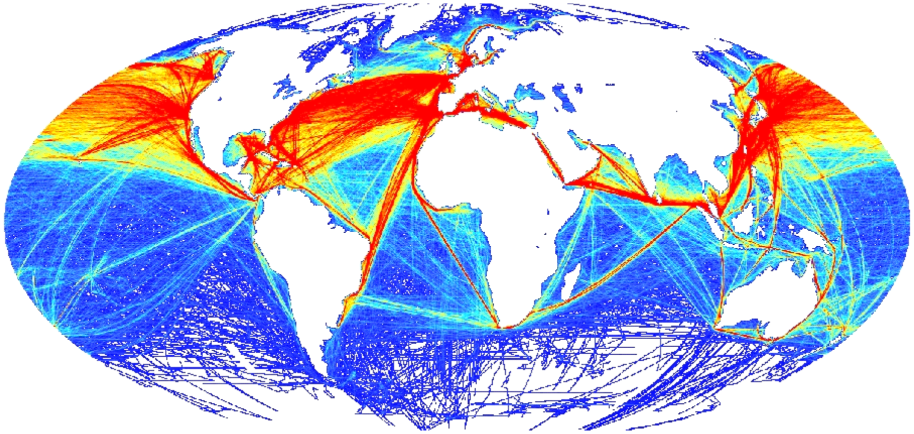
Dissolved Oxygen

40% of Oceans are Impacted by Humans

*** Resolution is 1 km square**

*** Value of a activity at a particular location is determined by the type of ecosystem present:**

Impact = ∑ Features * Ecosystem * impact weight



Shipping

Anthropogenic Features Extracted:

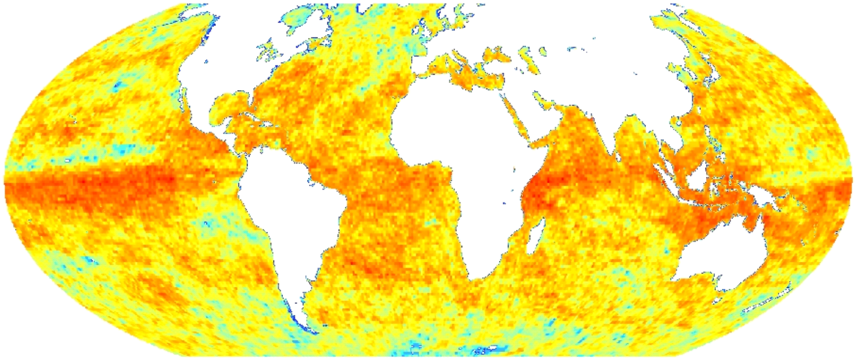
Ultraviolet radiation

Shipping

Pollution

Climate Change

Ocean Acidification



Climate Change

Halperin et al. *Science* 2008

Mapping Raw Metagenomic Reads to a Matrix of Familes or Pathways for each Site

READS → PROTEIN FAMILIES → PATHWAYS

$P_1 = f_1 + f_2 + f_3$

$P_2 = f_4 + f_5 + f_6$

PATHWAYS

● ■

SITES

$P_{1,1} = 2 + 1 + 3$ $P_{2,1} = 2 + 4 + 3$

$P_{1,2} = 5 + 2 + 6$ $P_{2,1} = 5 + 7 + 6$

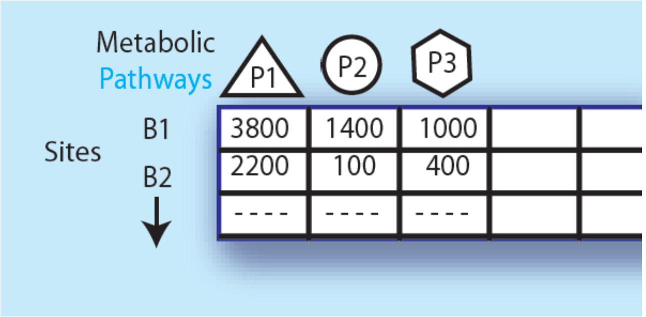
Families Matrix

	Fam 1	Fam 2	…	…	Fam 151
Site 1	.01	.02			
Site 2	0	.01			
…					
Site 29					

$\frac{\#\text{ counts Fam 2}}{\#\text{total protein counts at site 2}}$

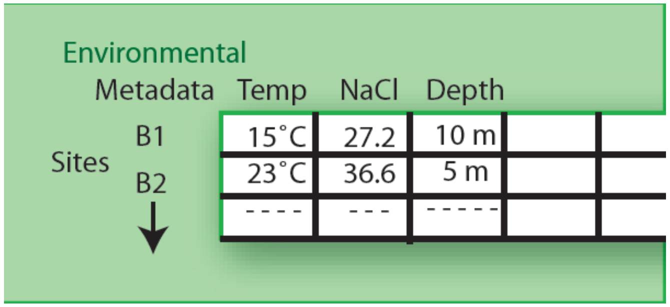
Patel et. al., *Genome Research* 2010

Pathway Sequences
(Community Function)



Sites	Metabolic Pathways	P1	P2	P3
	B1	3800	1400	1000
	B2	2200	100	400
	↓	----	----	----

Environmental
Features

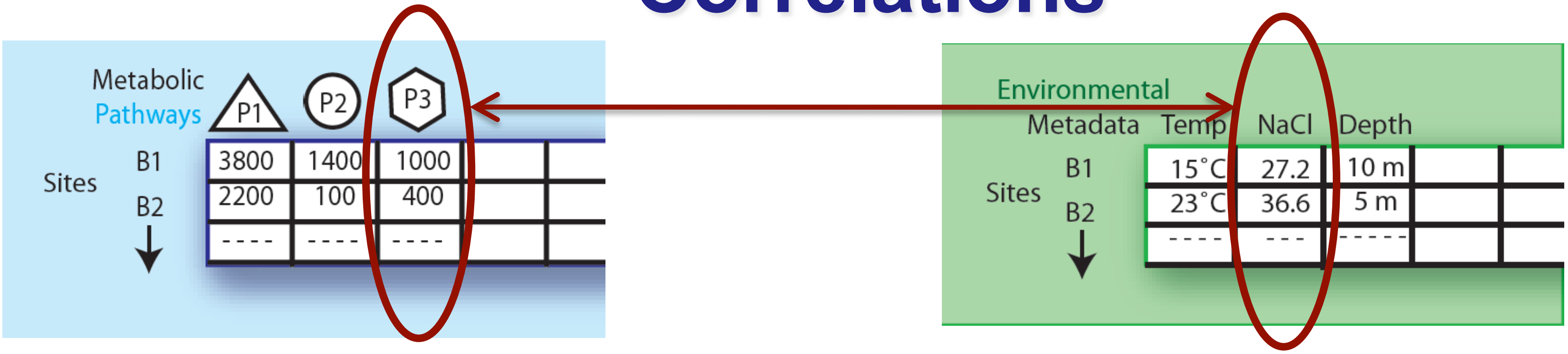


Sites	Environmental Metadata	Temp	NaCl	Depth
	B1	15°C	27.2	10 m
	B2	23°C	36.6	5 m
	↓	----	---	-----

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

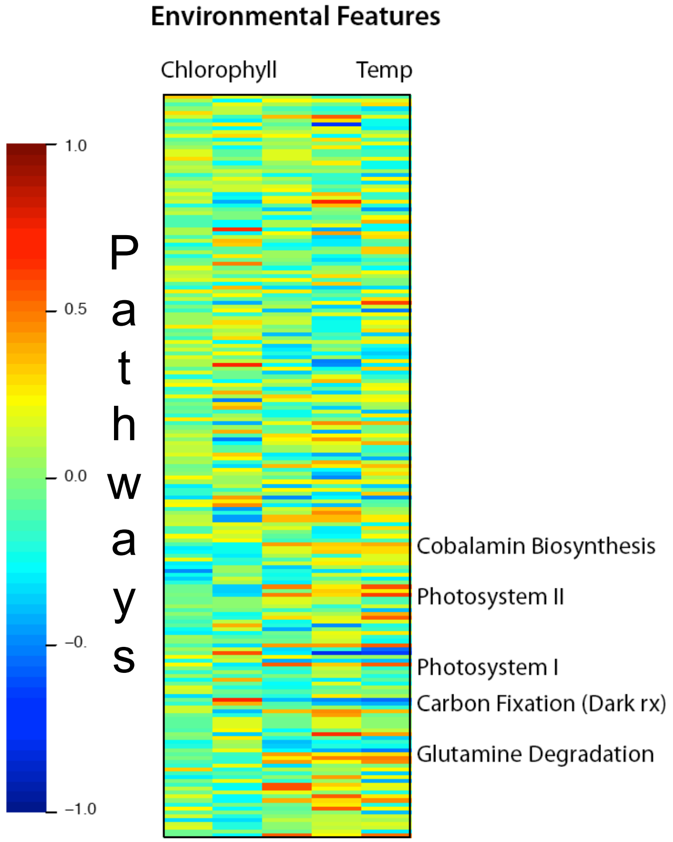
[Rusch et. al., (2007) PLOS Biology;
Gianoulis et al., PNAS (in press, 2009)]

Simple Relationships: Pairwise Correlations



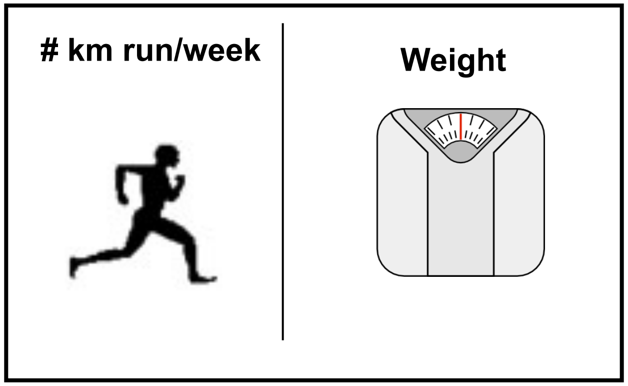
Metabolic Pathways		P1	P2	P3	
Sites	B1	3800	1400	1000	
	B2	2200	100	400	
	↓	----	----	----	

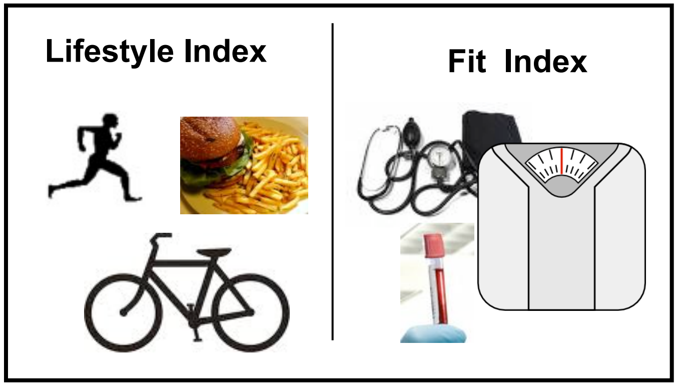
Environmental Metadata		Temp	NaCl	Depth	
Sites	B1	15˚C	27.2	10 m	
	B2	23˚C	36.6	5 m	
	↓	----	---	-----	



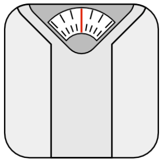
[Gianoulis et al., PNAS (in press, 2009)]

Canonical Correlation Analysis: Simultaneous weighting

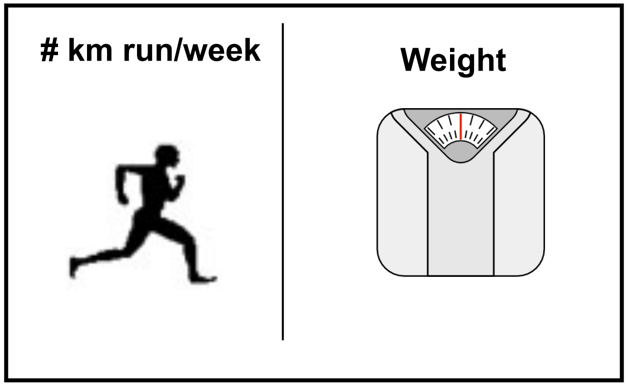


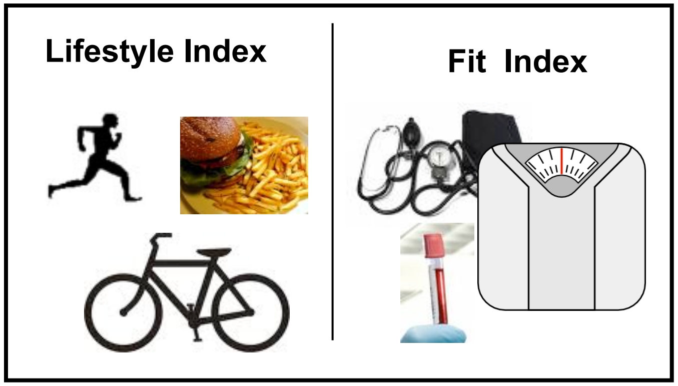


Lifestyle Index = a  +b  + c 

Fit Index = a  + b  +c 

Canonical Correlation Analysis: Simultaneous weighting





Lif

Fi

Environmental Features

Temp

etc

Chlorophyll

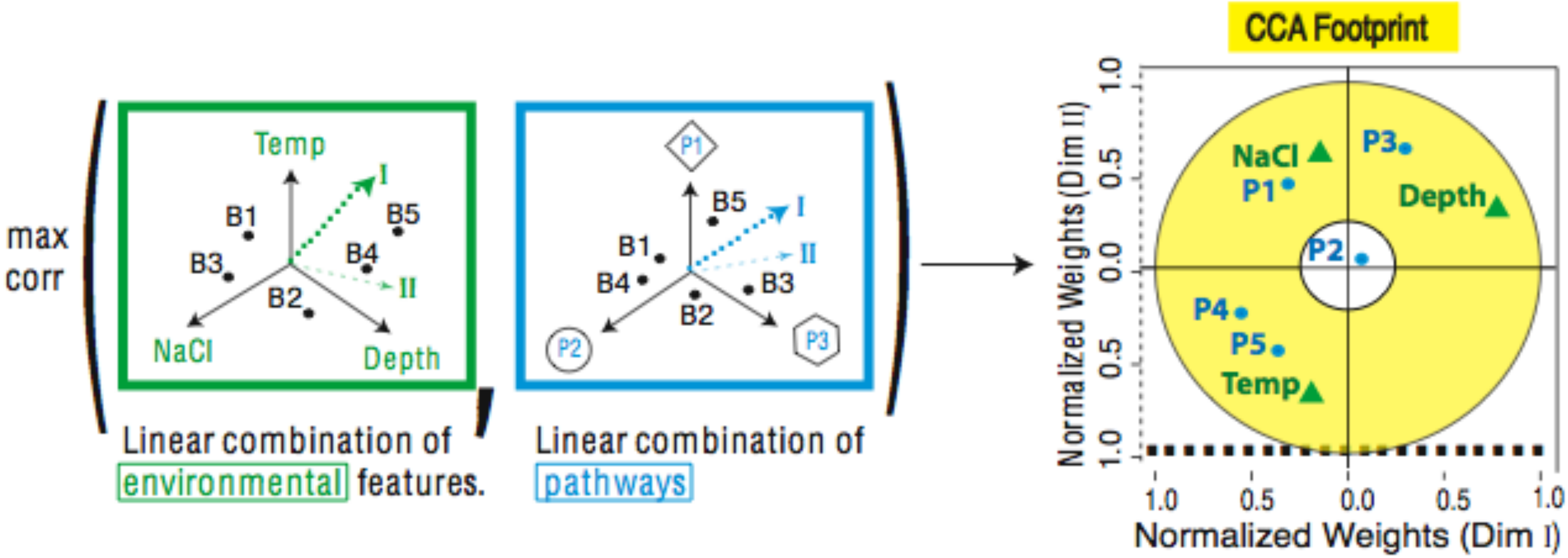
Metabolic Pathways/ Protein Families

Photosynthesis

etc

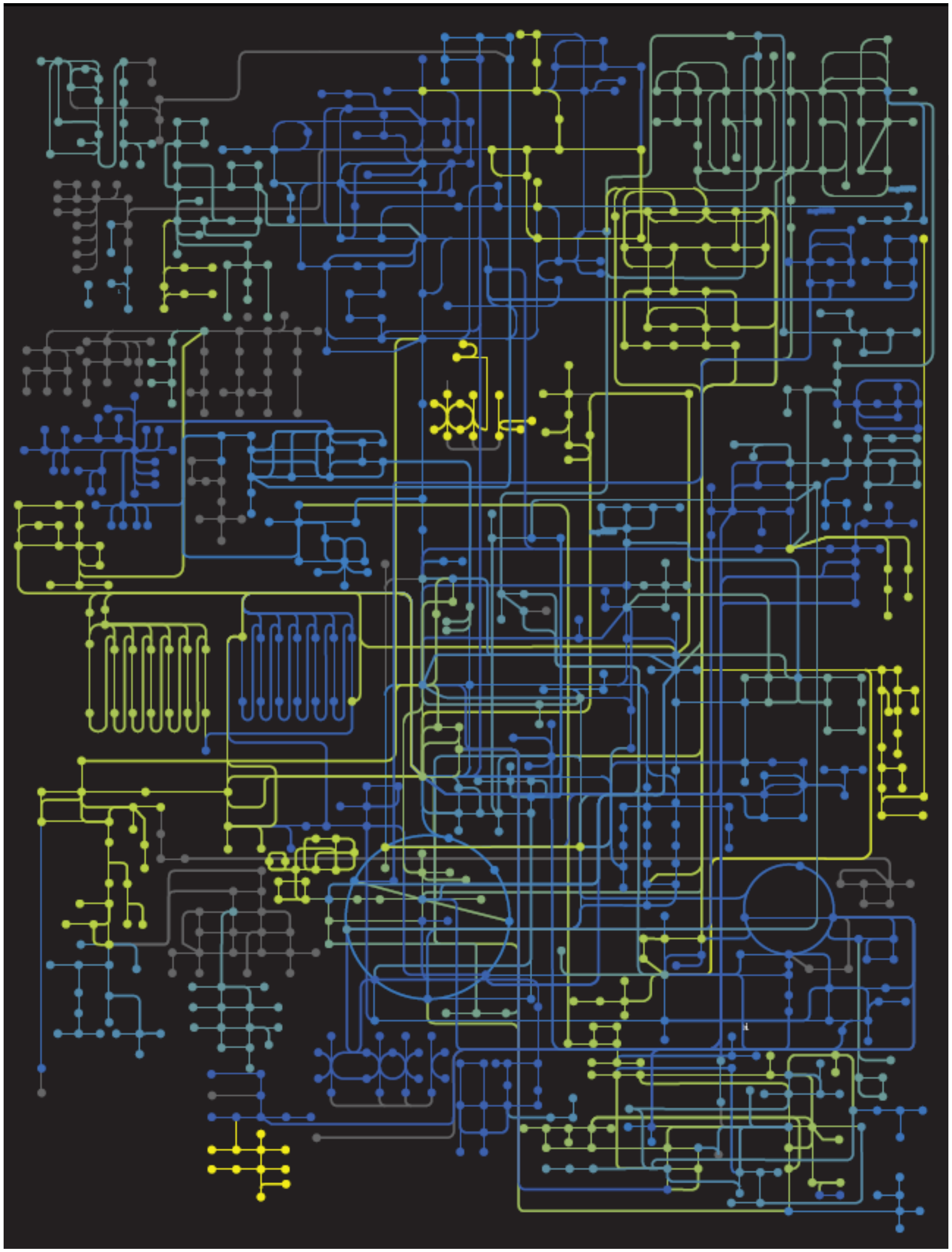
Lipid Metabolism

CCA: Finding Variables with Large Projections in "Correlation Circle"

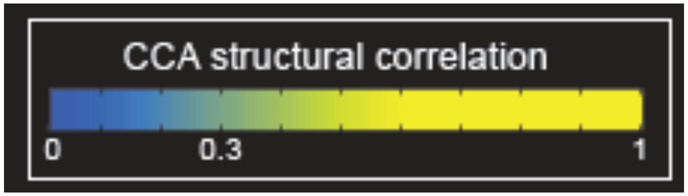


The goal of this technique is to interpret cross-variance matrices
We do this by defining a change of basis.

Gianoulis et al., PNAS 2009

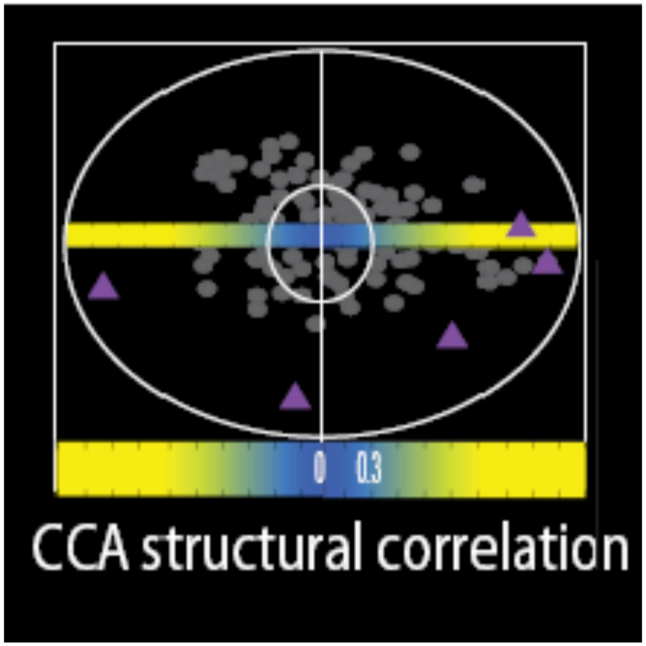


Strength of Pathway co-variation with environment



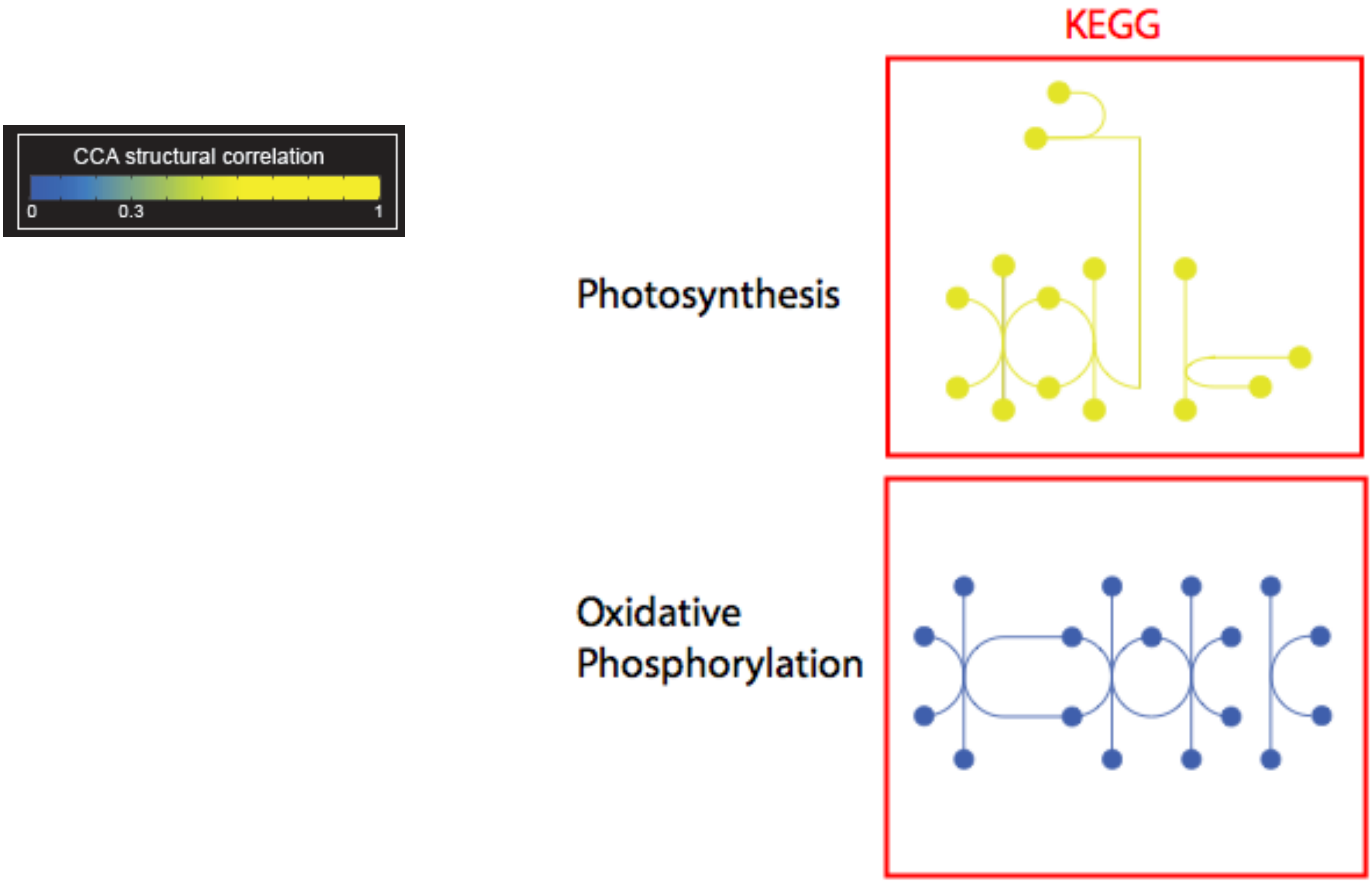
Environmentally invariant

Environmentally variant



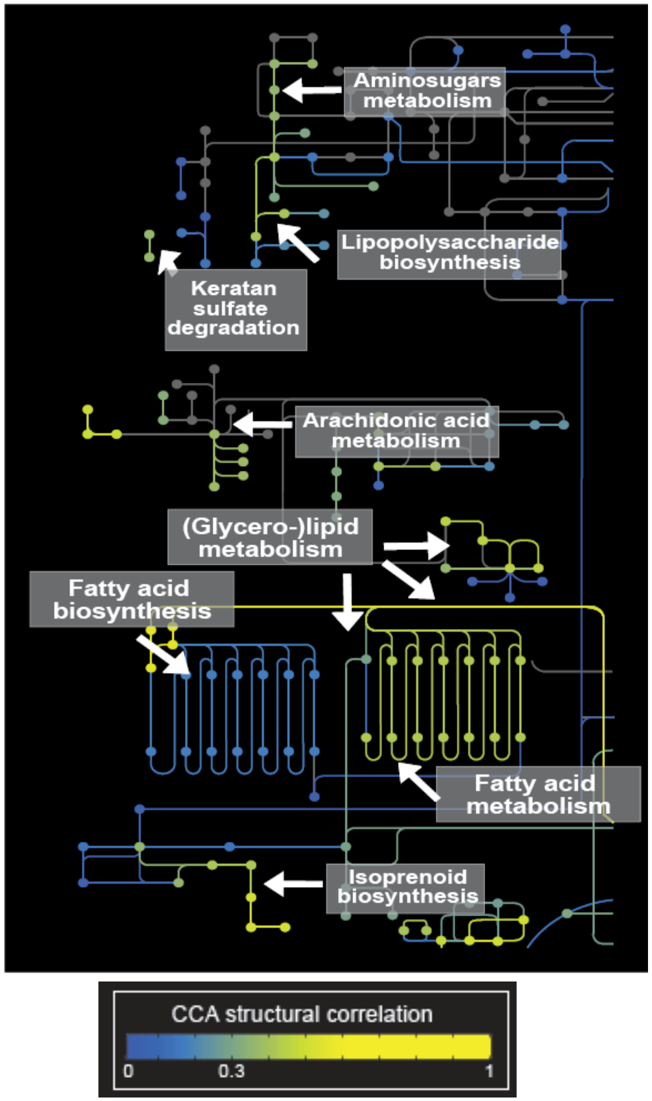
Gianoulis et al., *PNAS* 2009

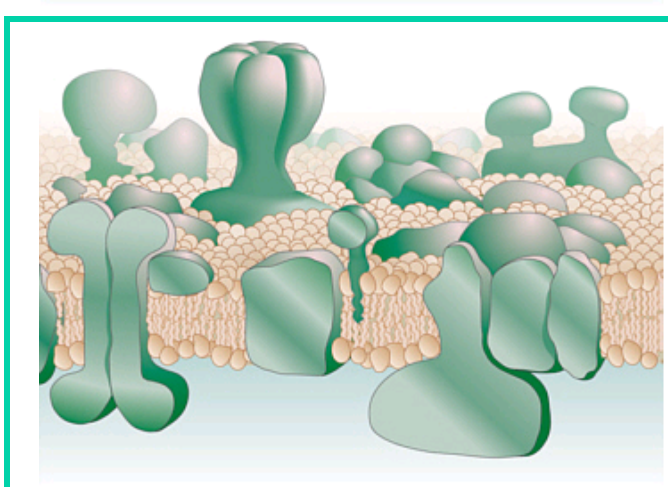
Conclusion #1: energy conversion strategy, temp and depth



Gianoulis et al., *PNAS* 2009

Conclusion #2: Outer Membrane components vary with the environment



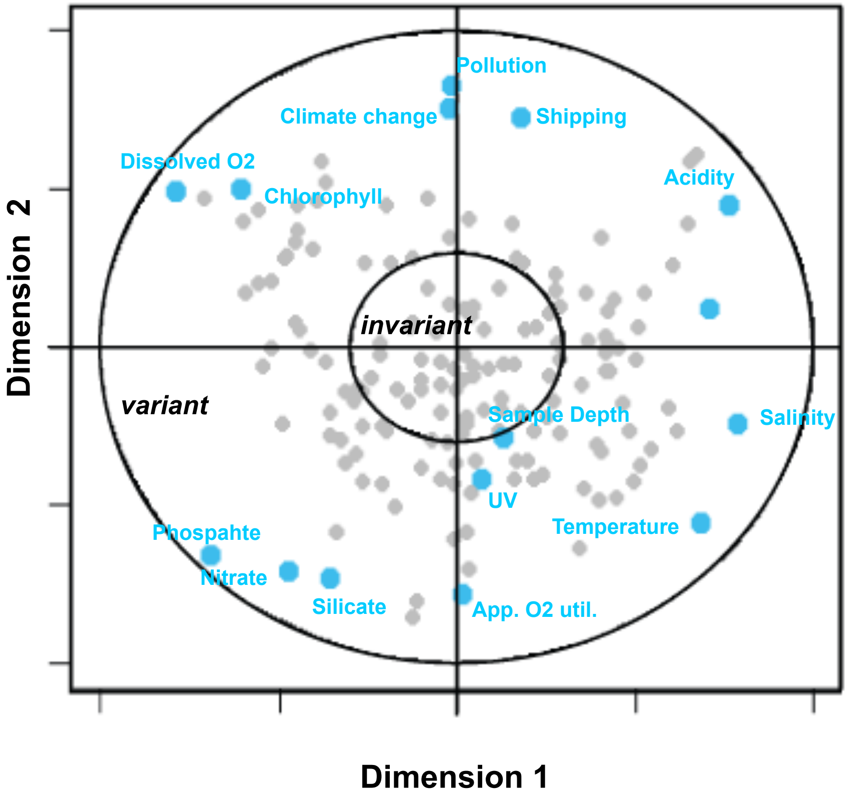


Membrane proteins interact with the environment, transporting available nutrients, sensing environmental signals, and responding to changes

Gianoulis et al., *PNAS* 2009
Patel et al. *Genome Research* 2010

Network Dynamics Across Environments: Membrane Proteins (Pathway Entry Points)

Membrane Proteins: Sensing and Responding the Environment



- 2.3 million predicted membrane proteins
- 1.2 million unique
- 850,000 mapped to 151 membrane protein COGs

107 variant membrane protein families

44 invariant membrane protein families

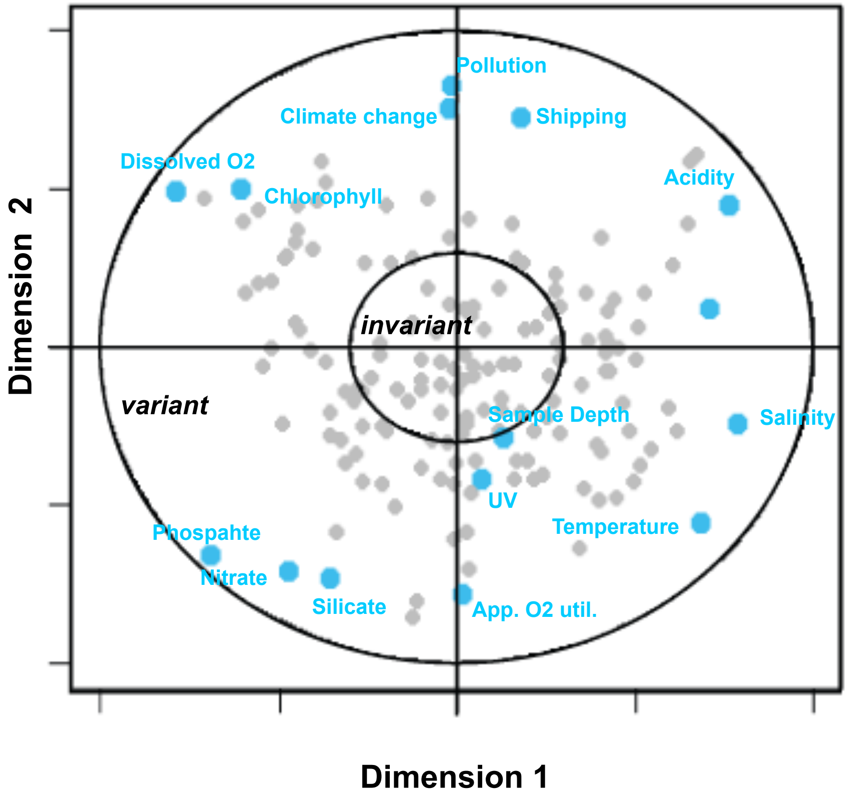
20% have NO KNOWN FUNCTION

Patel and Gianoulis et al., (in press) *Genome Research*

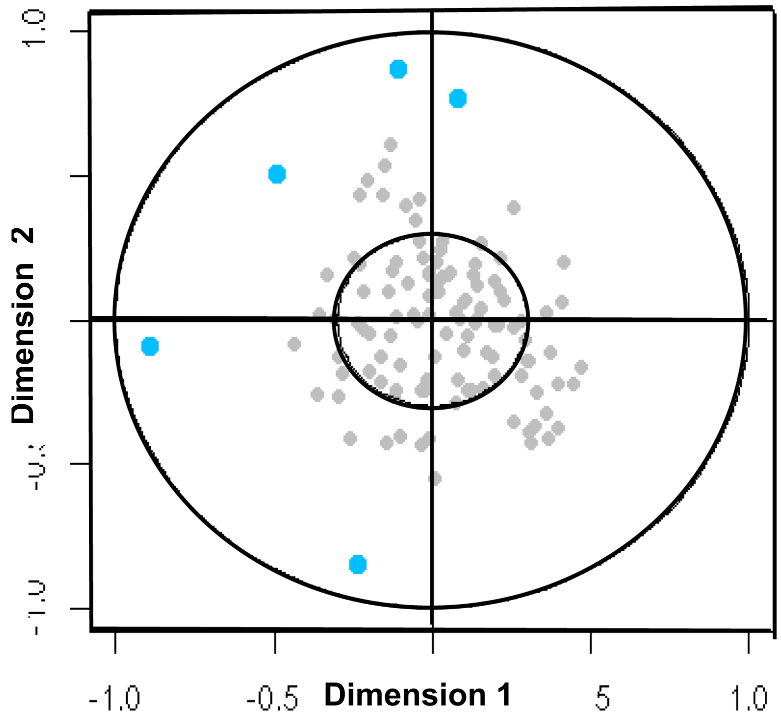
Membrane Proteins co-vary more than Metabolic Pathways

Median absolute structural Correlation Coefficient

Membrane Proteins = 0.3

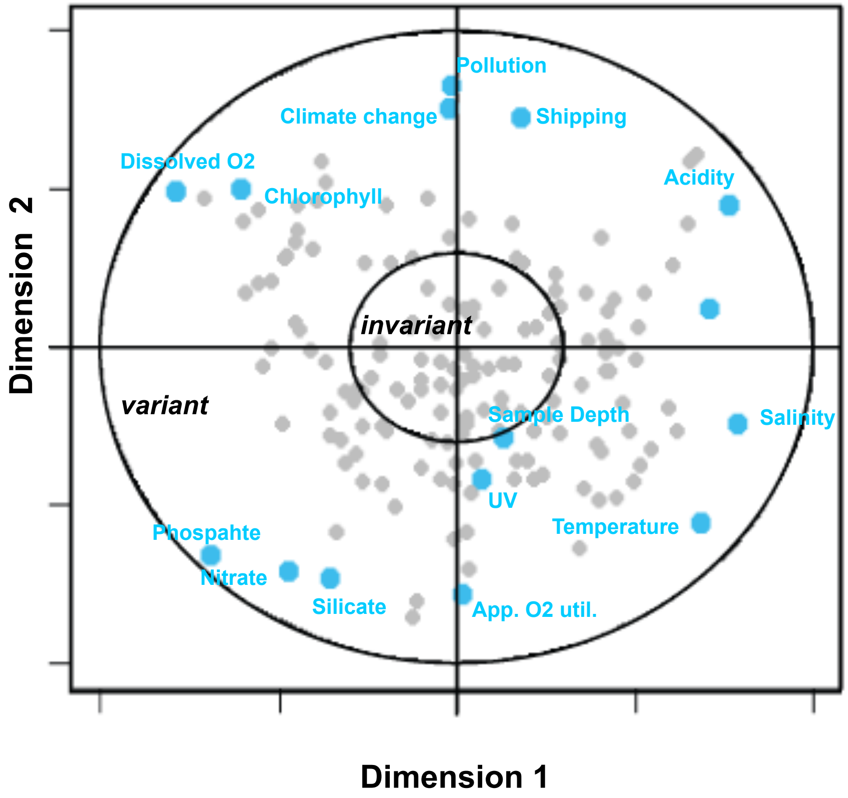


Metabolic Pathways = 0.17



Patel and Gianoulis et al., (in press) *Genome Research*

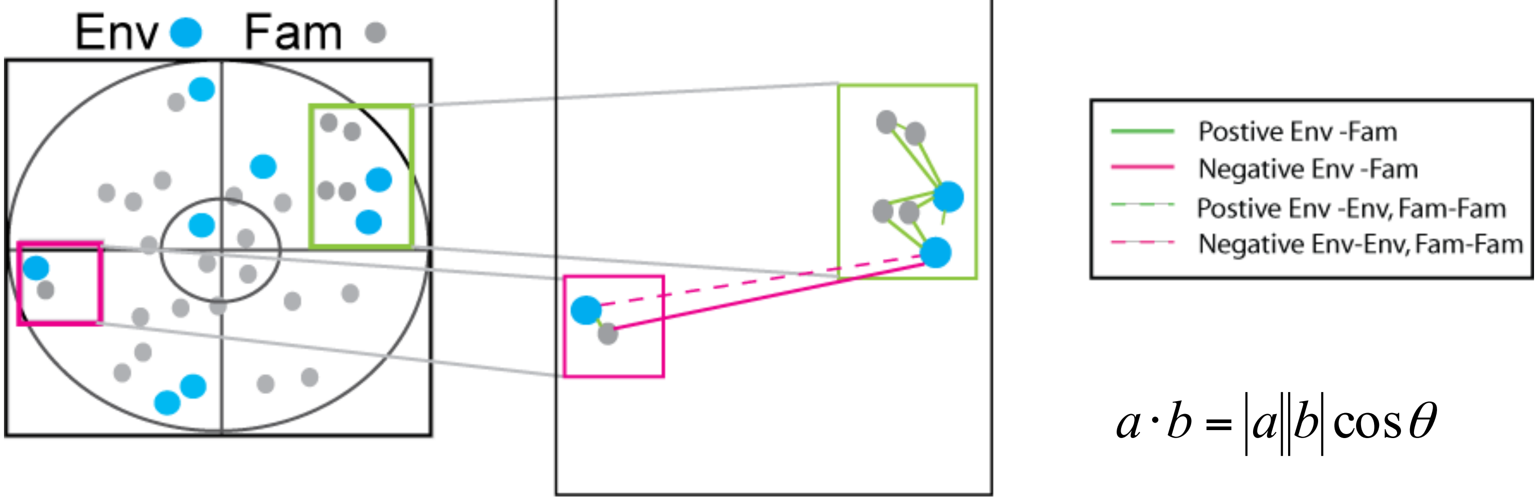
CCA Limitations



Both the strength and the directionality of relationships between nodes is difficult to decipher in this format.

Patel and Gianoulis et al., (in press) *Genome Research*

Protein Families and Environmental Features Network (PEN)

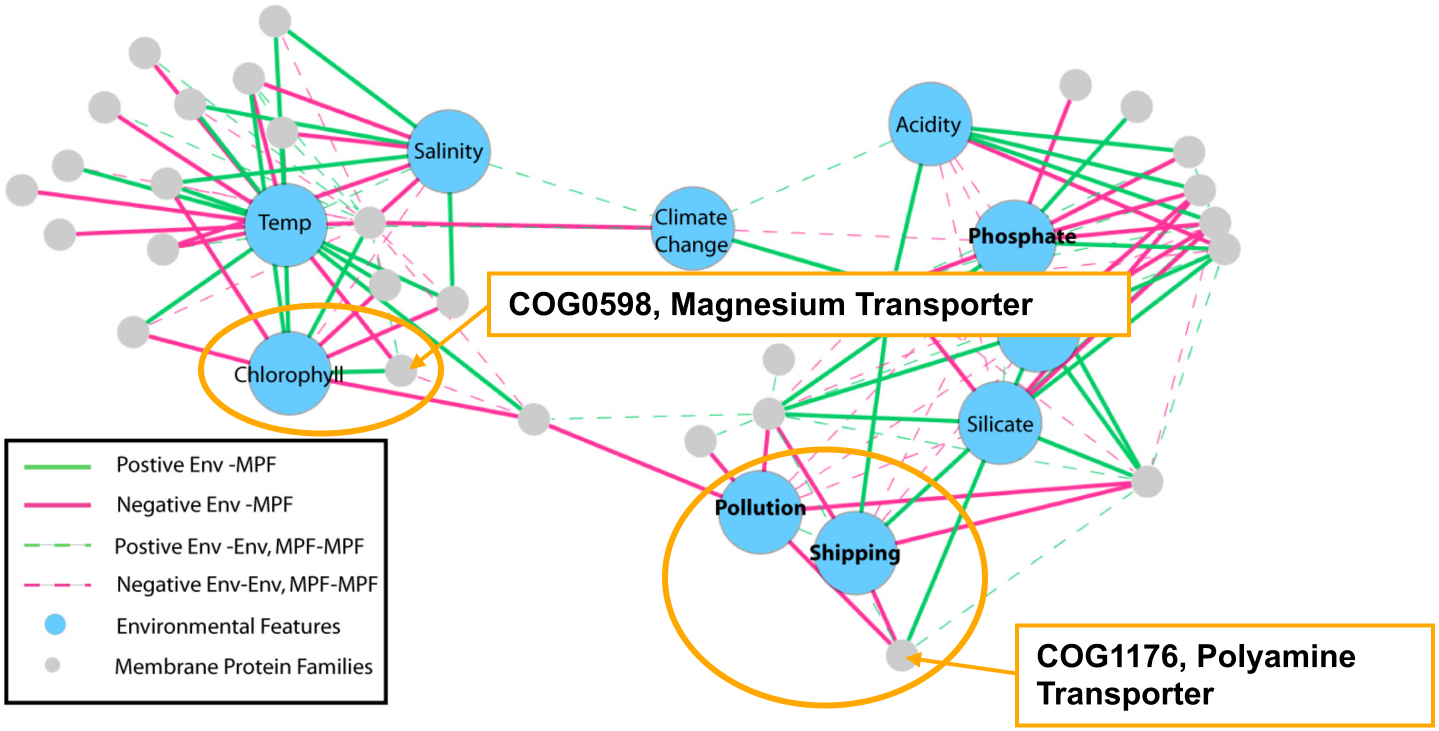


$$a \cdot b = |a||b|\cos\theta$$

Distance: Dot product between 1st and 2nd Dimension of CCA

Patel et. al., *Genome Research* 2010

Protein Families and Environmental Features Network (PEN)



COG0598, Magnesium Transporter

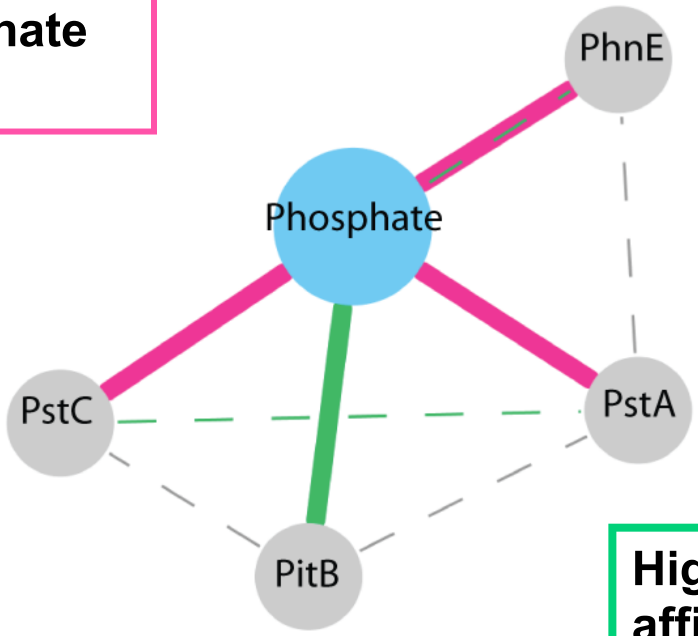
COG1176, Polyamine Transporter

"Bi-modules": groups of environmental features and membrane proteins families that are associated

UV, dissolved oxygen, apparent oxygen utilization, sample depth, and water depth are not in the network

Bi-module 1: Phosphate/Phosphate Transporters

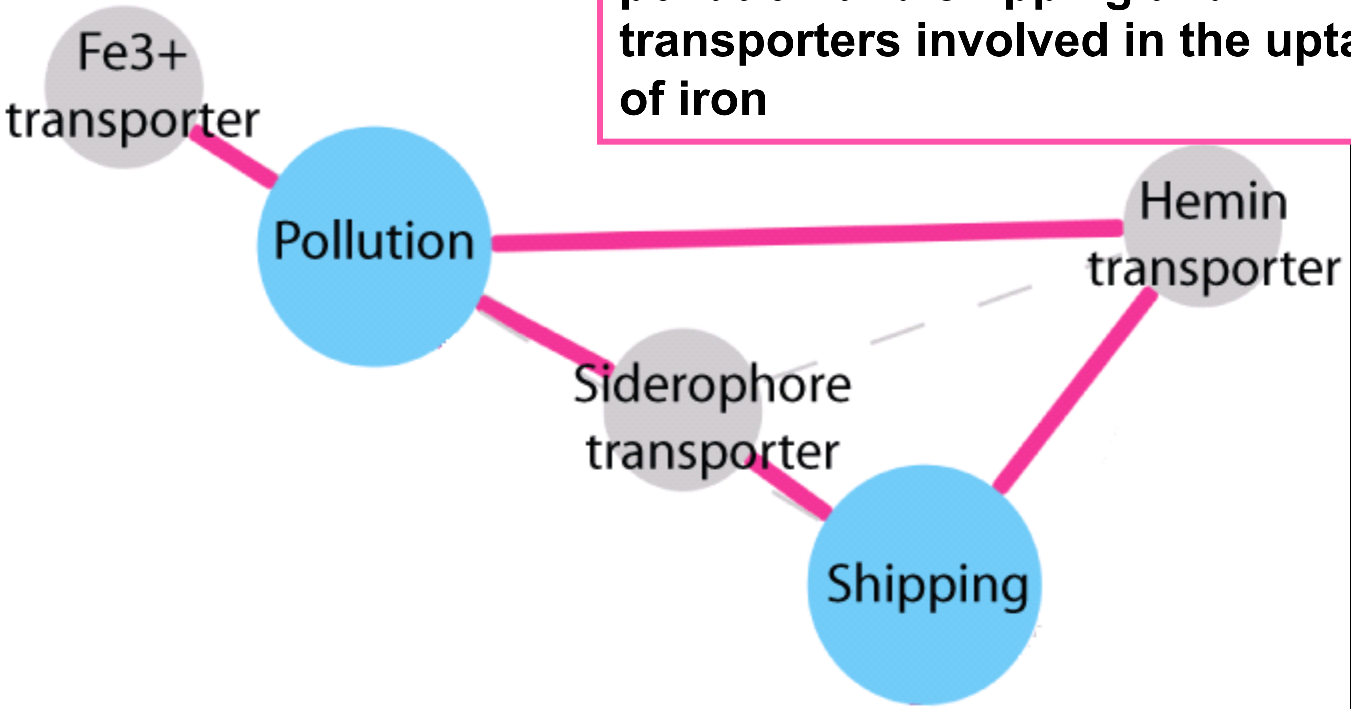
Low Phosphate, high affinity phosphate transporters which are induced during phosphate limitation



High Phosphate, low affinity inorganic phosphate ion transporter which are constitutively expressed

Patel et. al., *Genome Research* 2010

Bi-module 2: Iron Transporters/Pollution/Shipping

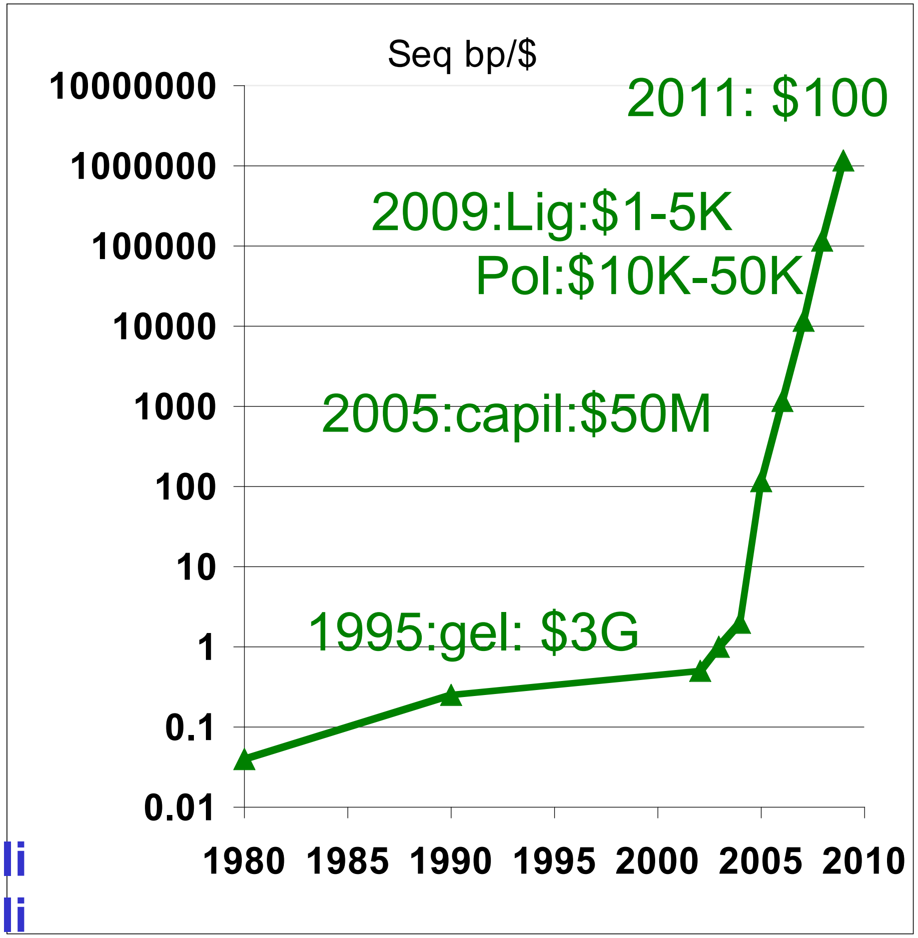


Negative relationship between areas of high ocean-based pollution and shipping and transporters involved in the uptake of iron

Pollution and Shipping may be a proxy for iron concentrations →

Patel et. al., *Genome Research* 2010

Biosensors: 4 logs in 4 years
Beyond Canaries in a Coal Mine



$1000 Human ~ $1 E. coli
$100 Human ~$.10 E. coli

Microbial Community

jgi.doe.gov

Carr and Church, Nat Biotech 2009

Outline: Molecular Networks

- Why Networks?
- Network Comparisons
  (reg. net. in many organisms)
  - in rel. to social hierarchy
  - Computer OS comparisons
  - Rewiring rate from comparisons amongst species
- Network Dynamics Across Environments
  (prokaryote metab. pathways)
  - Metabolic Pathways
  - Entry pts. (Mem. Proteins)

Conclusions: Comparison of Social and Regulatory Hierarchies

- Middle managers dominate, sitting at info. flow bottlenecks
- Democratic v Autocratic
- Collaborative (locally democratic) fraction of networks increases with organism complexity
- Middle managers most collaborative
- Most interaction occur between 2 middle managers (as seen in efficient corporate hierarchies)

- Various types of networks can be compared amongst different organisms. In general, TR rewrites faster than PPI which rewires faster than metabolic

		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)

Conclusions: Network Dynamics Across Environments

- Developed approach to connect quantitative features of environment to usage of pathways & families
  - CCA + PEN
- Applied to available aquatic datasets, identified footprints predictive of environment (potentially useful as biosensor)
- Integration of geospatial data can highlight unexpected trends as anthropogenic factors seem to be reflected in microbial function

- Specific Conclusions
  - Strong correlation exists between a community's energy conversion strategies & env. parameters (e.g. temperature & chlorophyll)
  - Relation between Fe and P transporters & amt. of chemical in environment
    - For Fe illustrates impact of pollution & shipping

N Bhardwaj
K-K Yan
P Patel
T Gianoulis
H Yu
C Shou

A Paccanaro
P Kim
K Yip
R Bjornson
G Fang
Y Xia
J Korbel
J Raes
P Bork
D Engelman
M Snyder

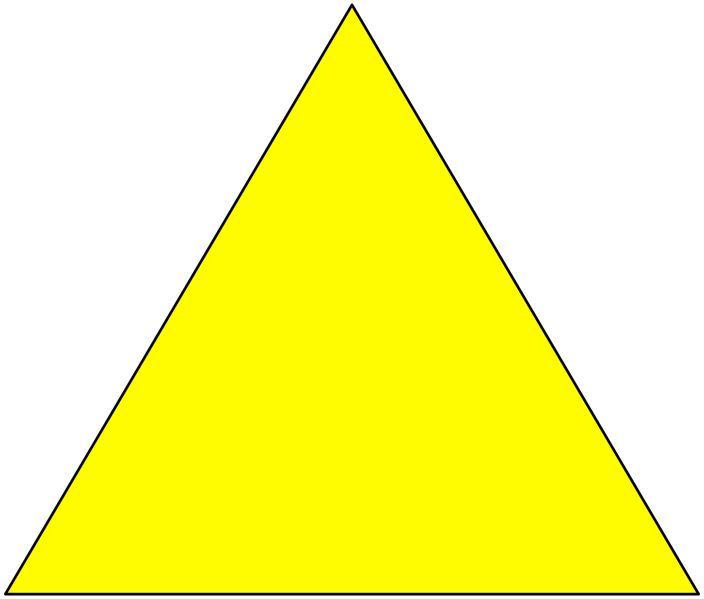
Acknowledgements

Job opportunities currently for postdocs & students

Networks.GersteinLab.org

Default Theme

- Default Outline Level 1
  - Level 2



More Information on this Talk

SUBJECT: Networks

DESCRIPTION:

NOTES:
This PPT should work on mac & PC. Paper references in the talk were mostly from Papers.GersteinLab.org.

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 .