Biological Network Analysis

Mark B Gerstein

Yale

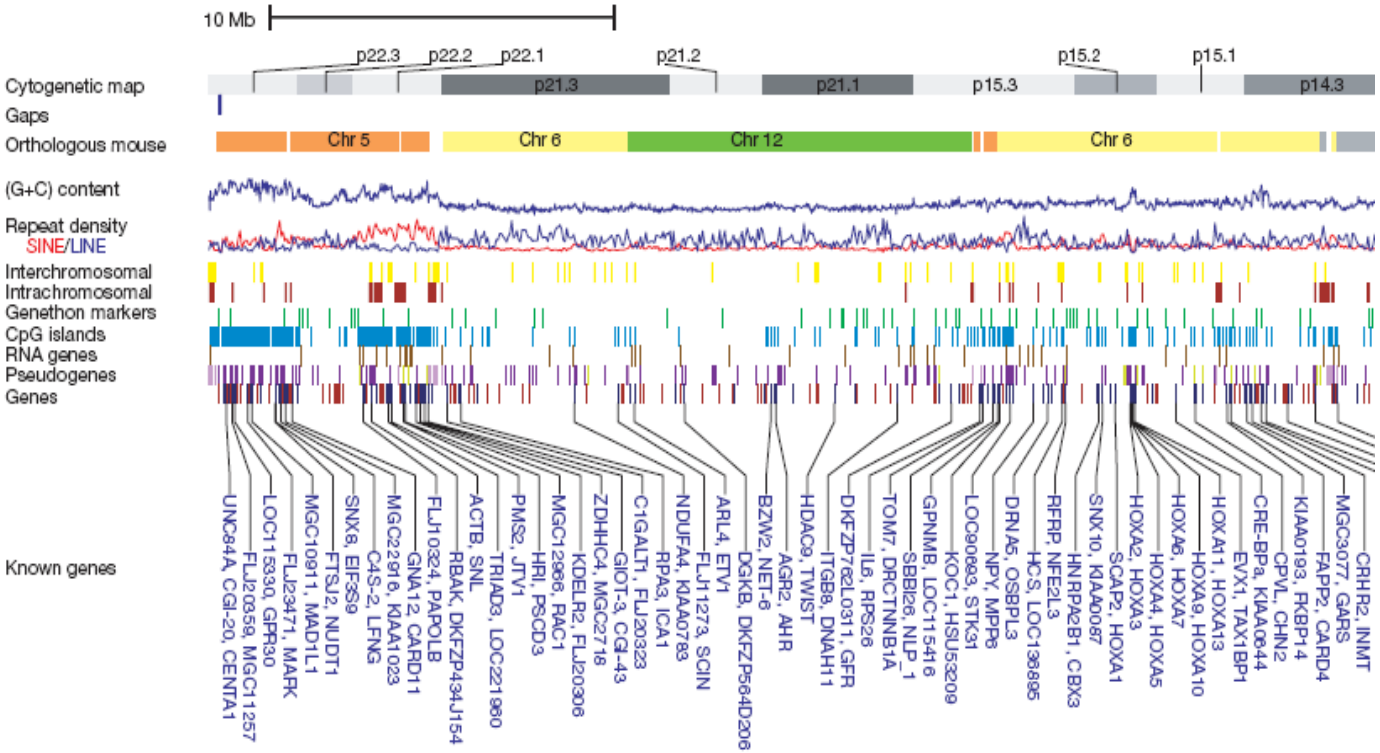
slides at

Lectures.GersteinLab.org

(See Last Slide for References & More Info.)

The problem: Grappling with Function on a Genome Scale?

sequence of human chr. 7



~1,200 protein-coding genes

(~950 pseudogenes)

[Hillier et al, Nature, 424, 157]

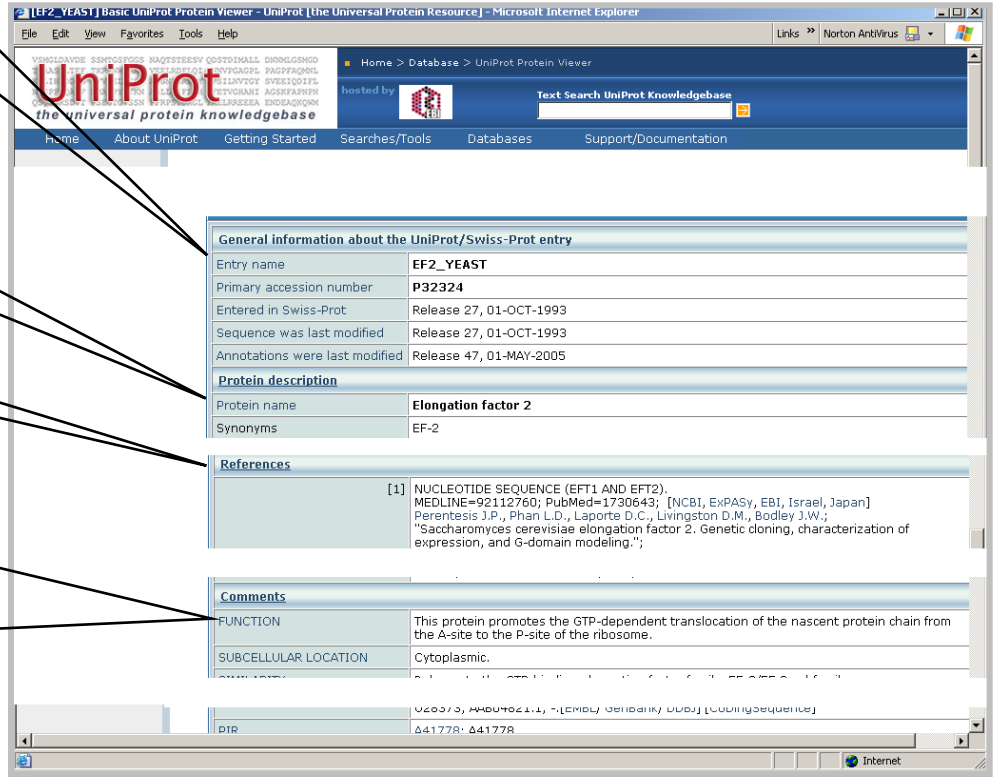
Traditional single molecule way to integrate evidence & describe function

EF2_YEAST

Descriptive Name:
Elongation Factor 2

Lots of references
to papers

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



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

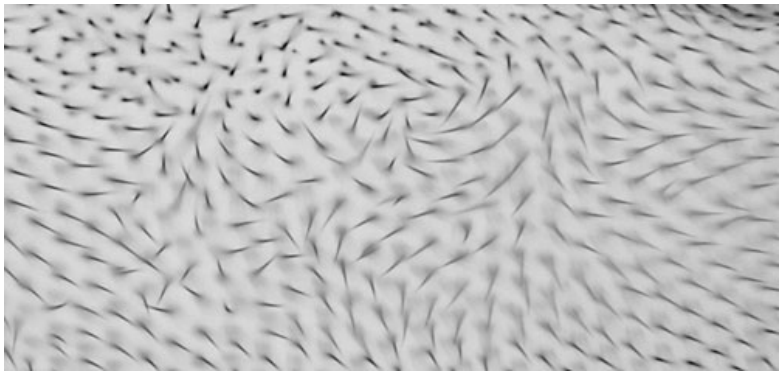
- Fundamental complexities
  - Role Conflation:
    molecular, cellular, phenotypic
  - Often >2 proteins/function
  - Also Multi-functionality:
    2 functions/protein
    - phenotypically – e.g. Pleiotropic effects such as human PKU being involved in retardation & eczema
    - cellular role – e.g.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

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

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

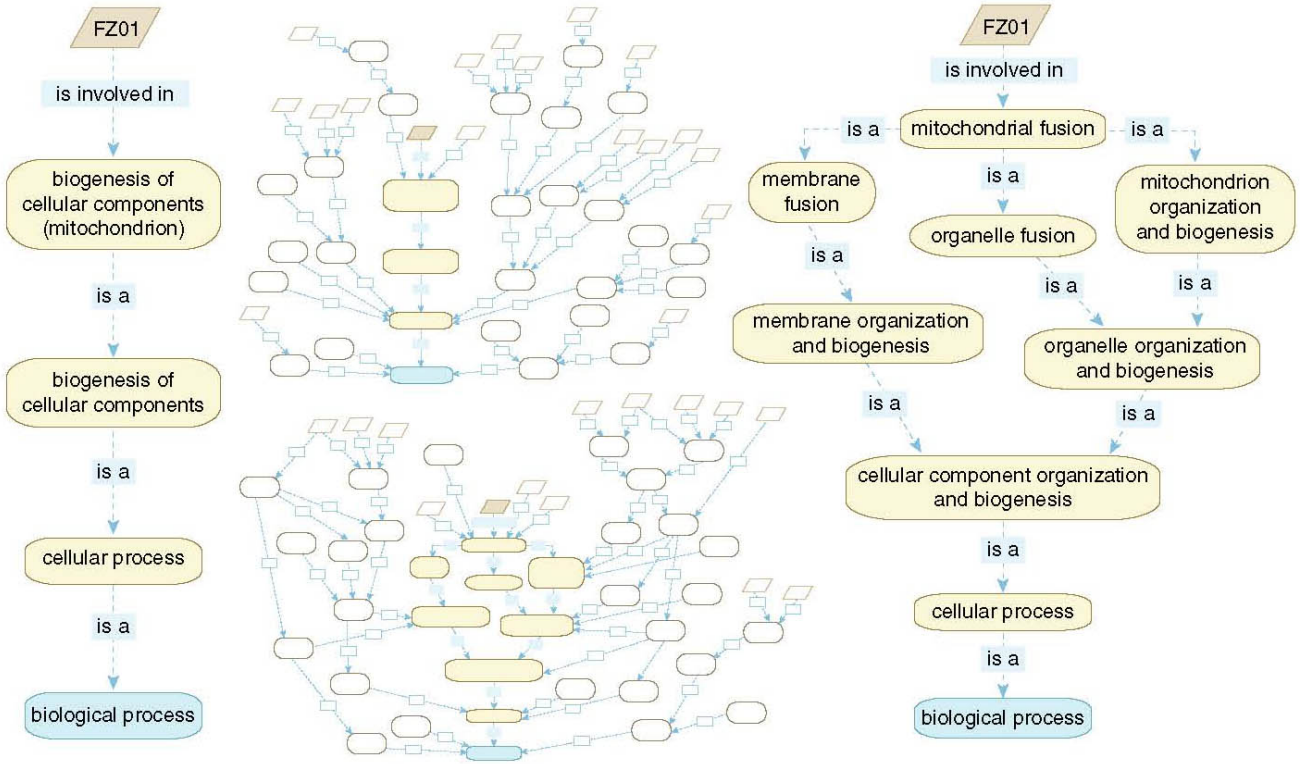
- Fundamental complexities
  - Role Conflation:
    molecular, cellular, phenotypic
  - Often >2 proteins/function
  - Also Multi-functionality:
    2 functions/protein
    - phenotypically – e.g. Pleiotropic effects such as human PKU being involved in retardation & eczema
    - cellular role – e.g.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

- Fun terms… but do they scale?....
  - **Starry night** (P Adler, '94)



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

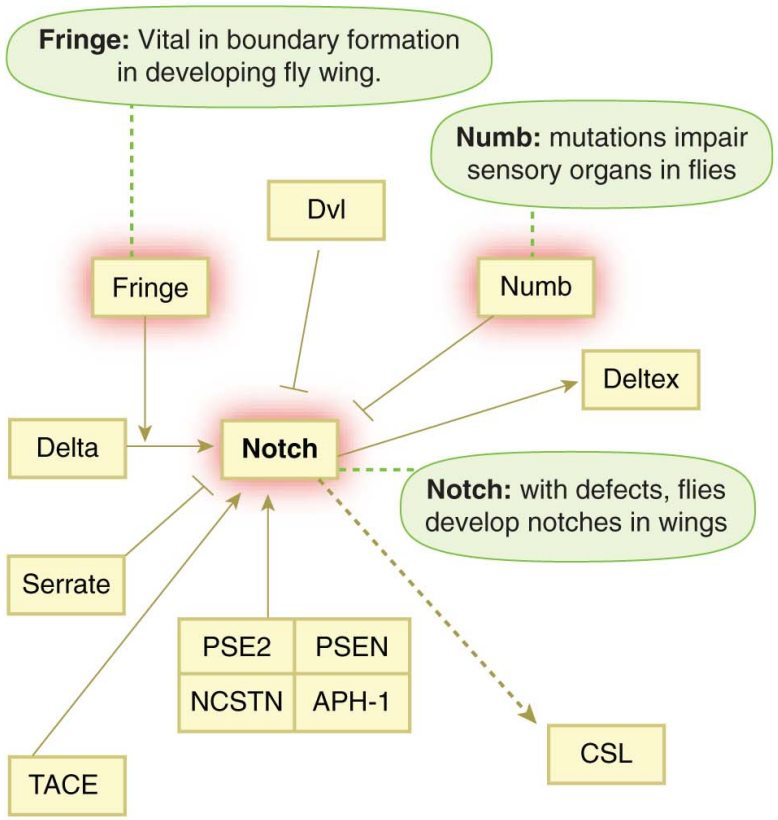
Hierarchies & DAGs of controlled-vocab terms but still have issues...



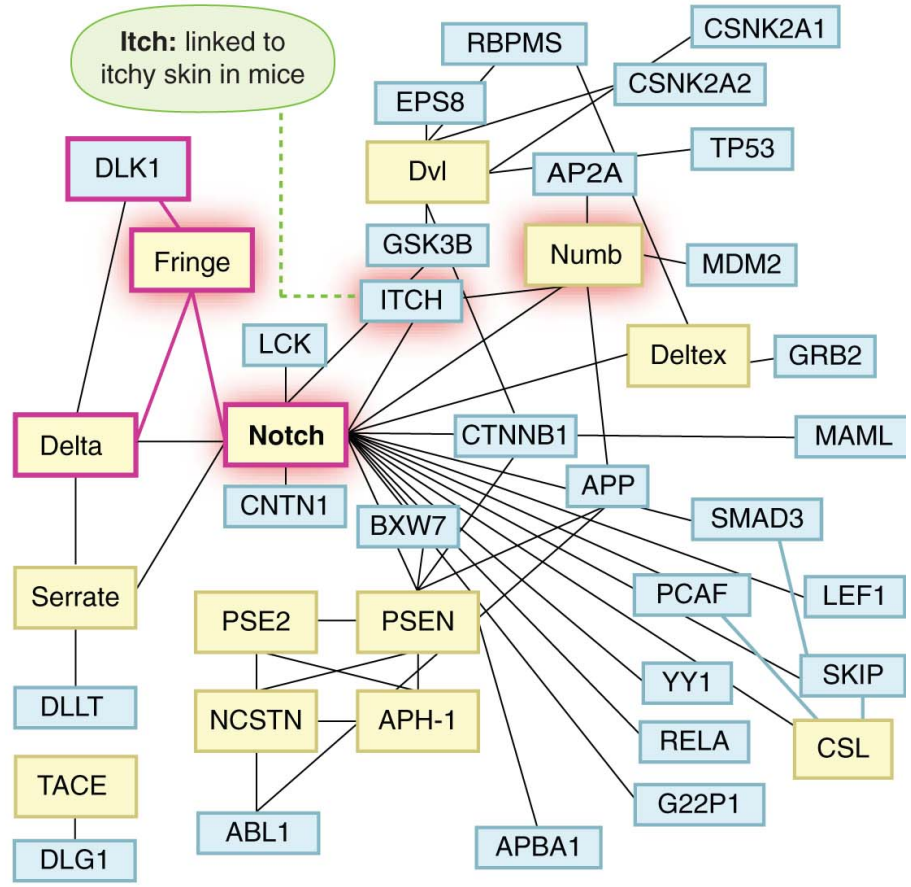
MIPS (Mewes et al.) **GO (Ashburner et al.)**

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

Networks (Old & New)



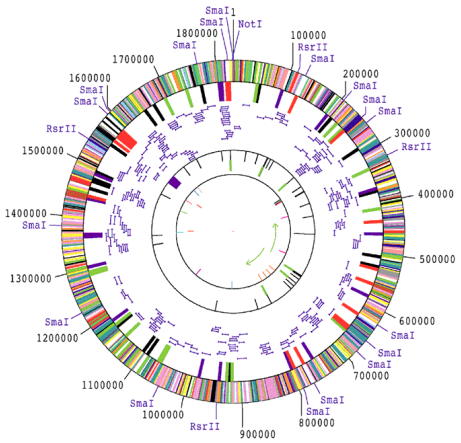
Classical KEGG pathway



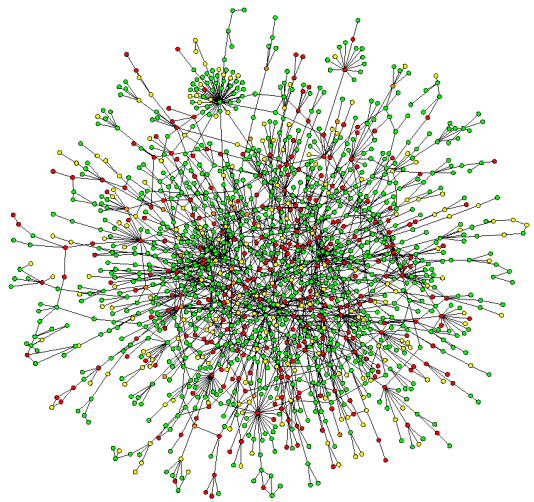
Same Genes in High-throughput Network

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

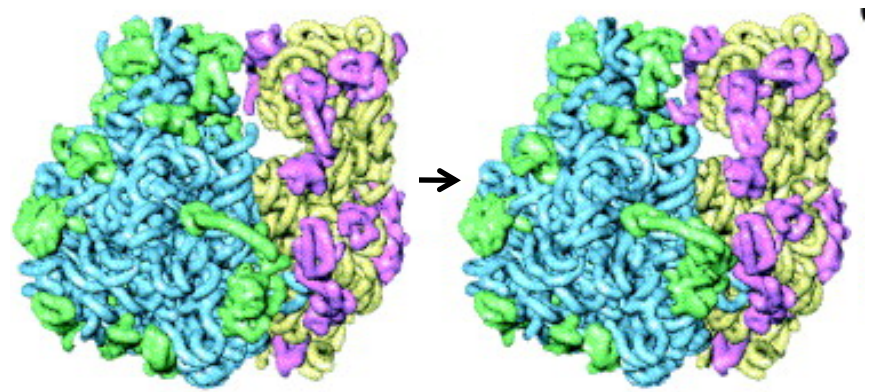
Networks occupy a midway point in terms of level of understanding



1D: Complete Genetic Partslist



~2D: Bio-molecular Network Wiring Diagram



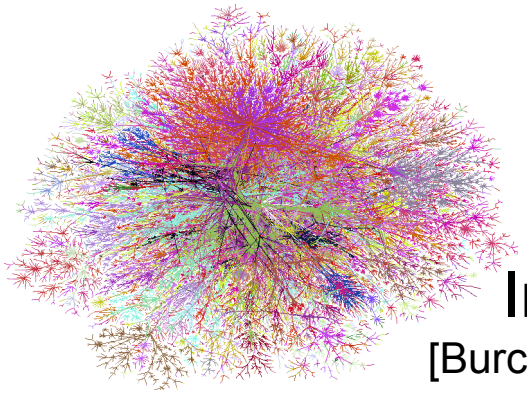
3D and 4D: Detailed structural understanding of cellular machinery (e.g. ribosome in different functional states)

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

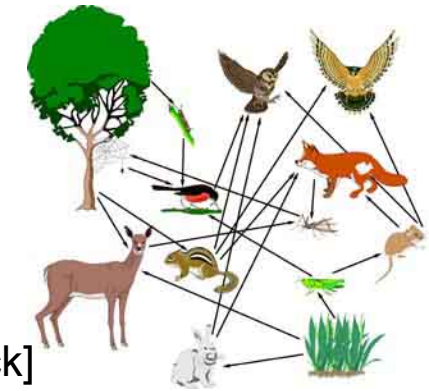
[Jeong et al. Nature, 41:411]

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

Networks as a universal language

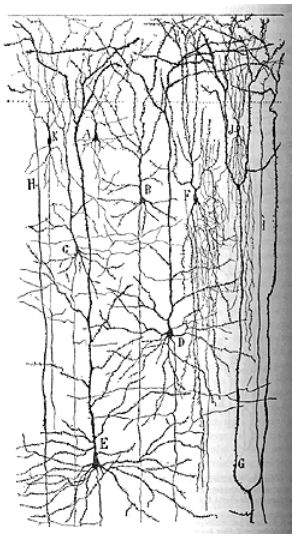


Internet
[Burch & Cheswick]

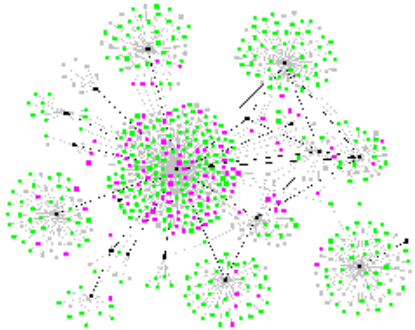


Food Web

Electronic Circuit

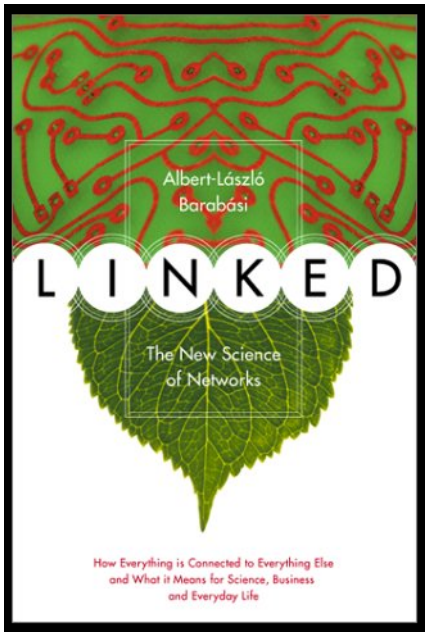


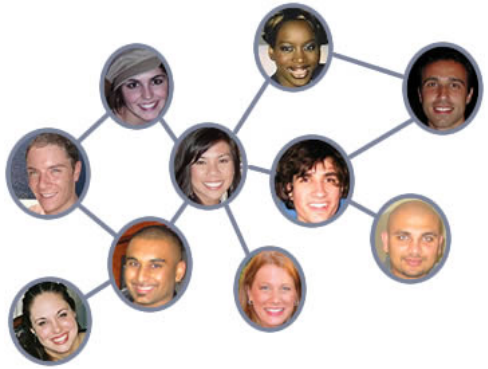
Neural Network
[Cajal]



Disease Spread
[Krebs]

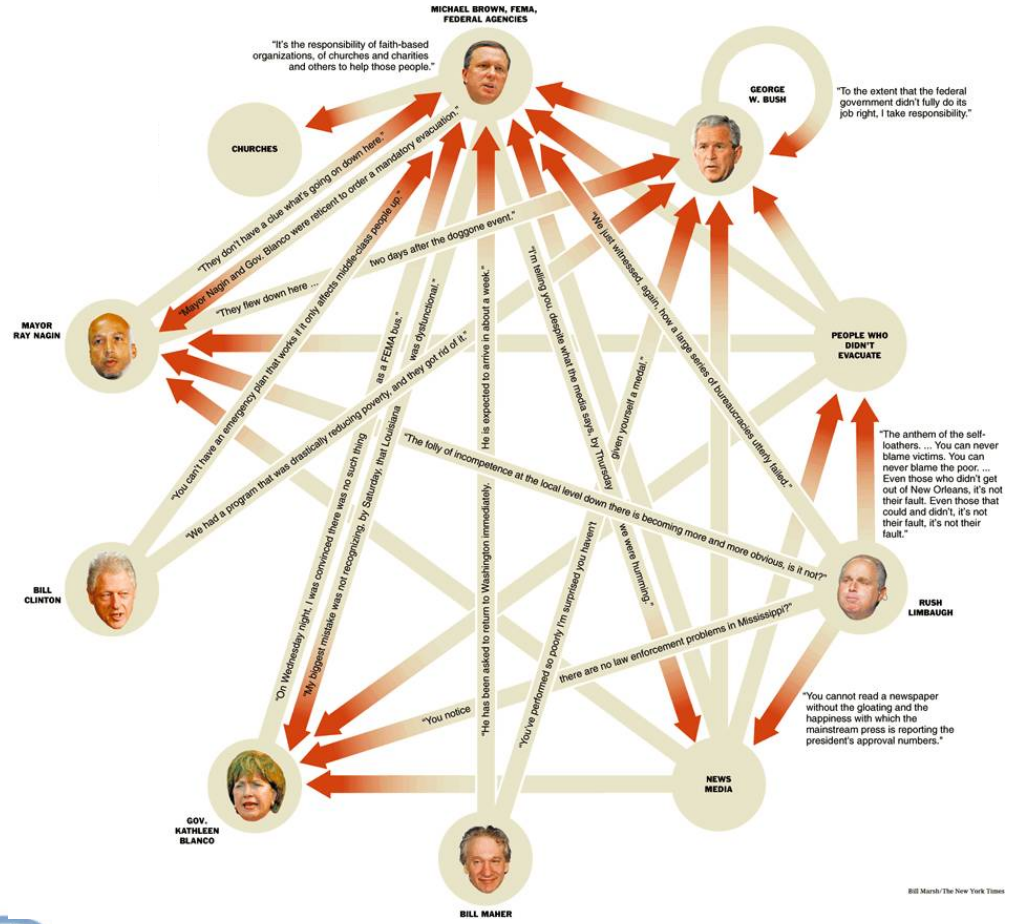
Protein Interactions
[Barabasi]



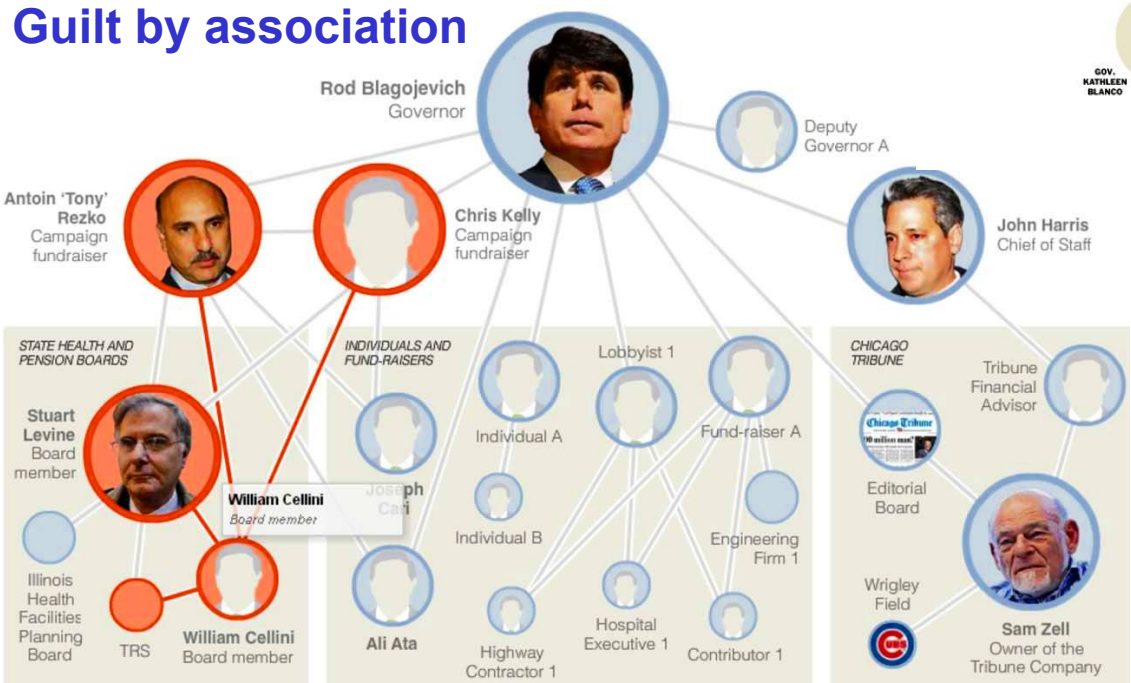


Social Network

Using the position in networks to describe function



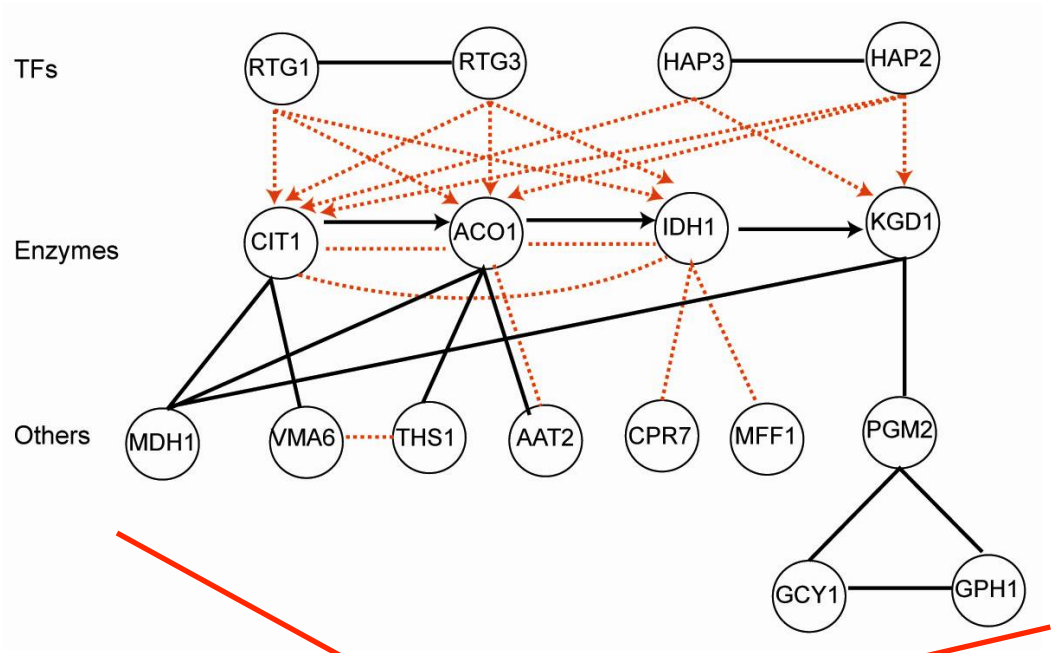
Guilt by association

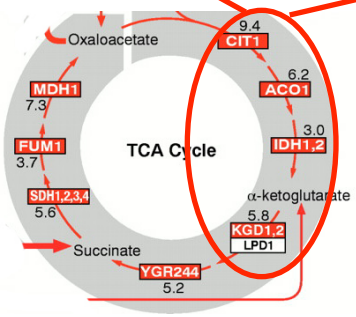


Finding the causal regulator (the "Blame Game")

[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

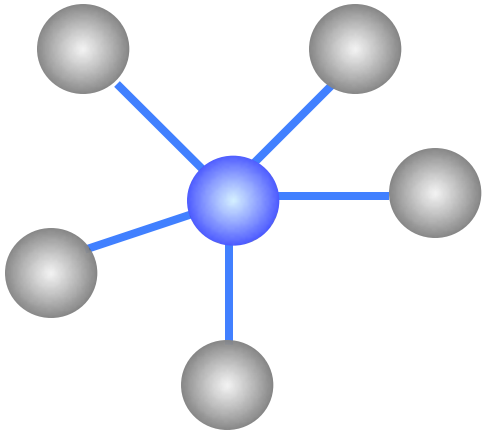
Outline: Molecular Networks

- Why Networks?
- Backround:
  Central Network Points
- Networks & Variation
  (human ppi)
- Social Network Comparisons
  (reg. net. in many organisms)
  - in rel. to social hierarchy
  - scaling in rel. to partnerships
- Computer OS Comparisons
  (E. coli reg. net)
- Network Dynamics Across Environments
  (prokaryote metab. pathways)
  - Metabolic Pathways
  - Entry pts. (Mem. Proteins)

Background: Finding Central Points in Networks

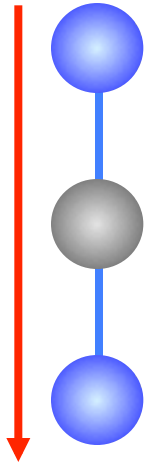
Global topological measures

Indicate the gross topological structure of the network



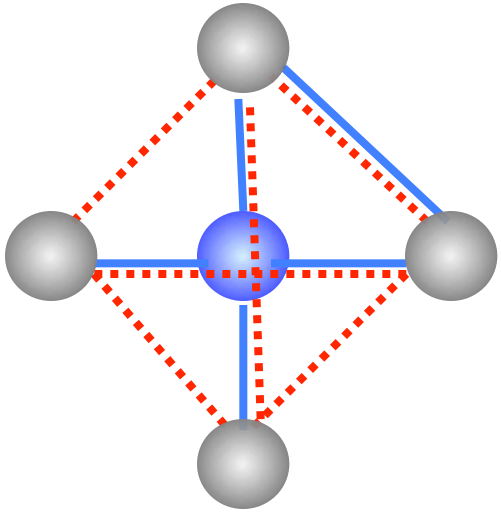
Degree (K)

5



Path length (L)

2



Clustering coefficient (C)

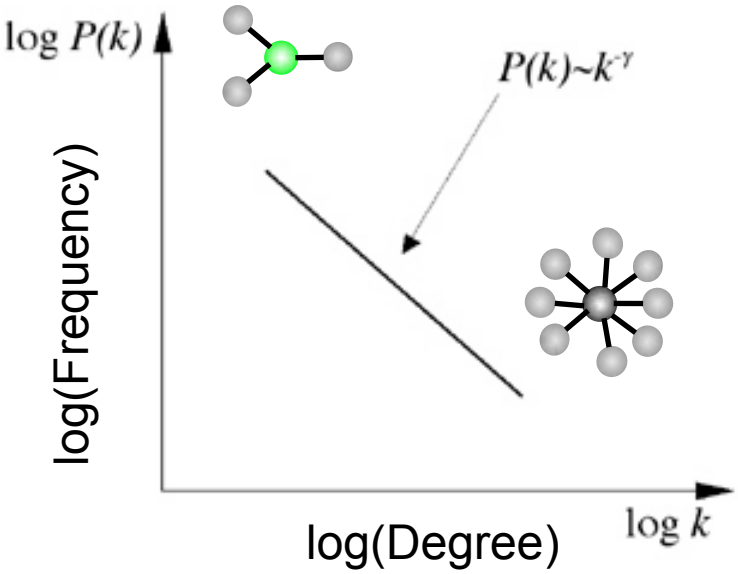
1/6

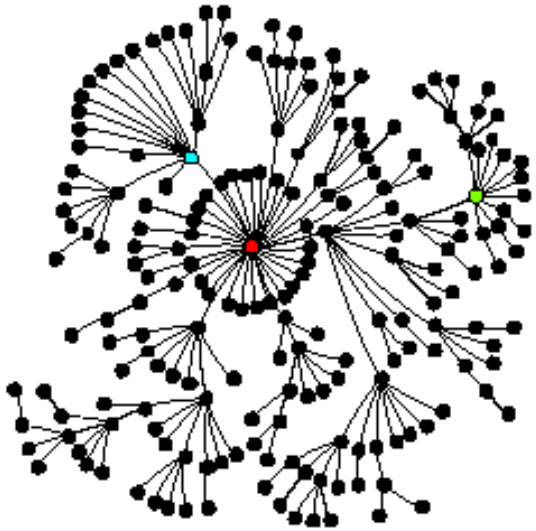
Interaction and expression networks are ***undirected***

[Barabasi]

Scale-free networks

Power-law distribution





Hubs dictate the structure of the network

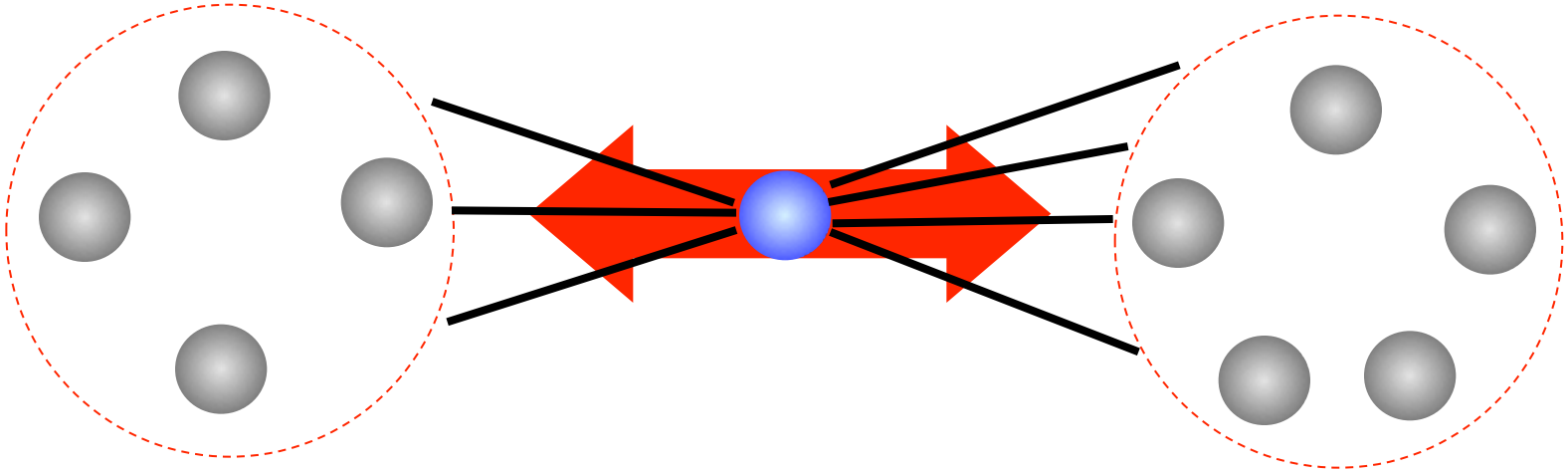
[Barabasi]

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.

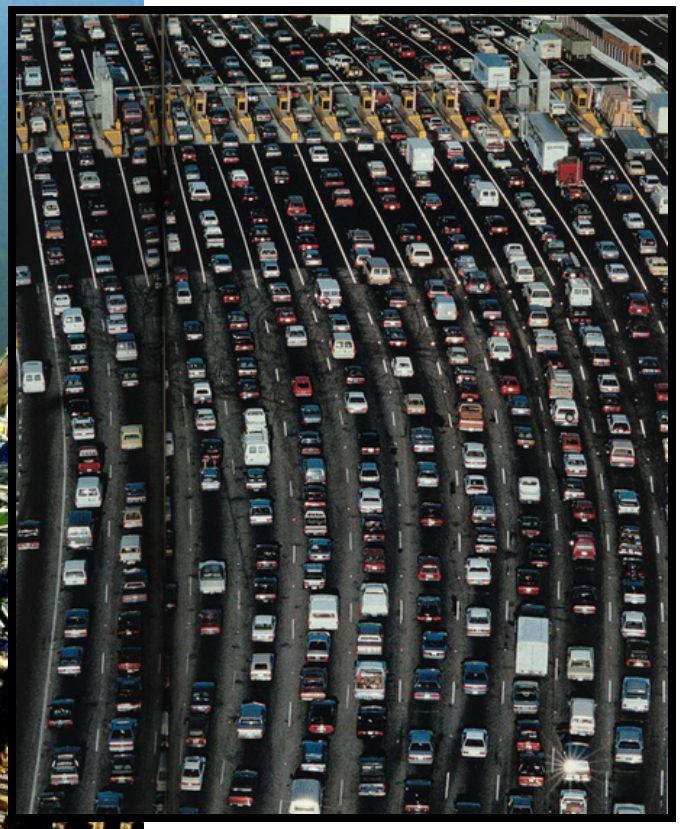
Freeman LC (1977) Set of measures of centrality based on betweenness. Sociometry 40: 35–41.

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



Networks & Variation

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

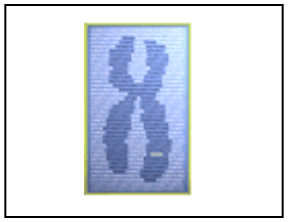
METHODOLOGY: MAP SNP AND CNV DATA ONTO ENSEMBL GENES, AND THEN MAP ENSEMBL GENES TO THE KNOWN INTERACTOME

ILLUSTRATIVE

Hapmap/Perlegen

SNPs

Database of Genomic Variants



CNVs + SDs

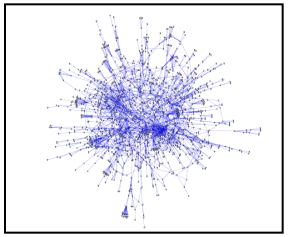
Map to ENSEMBL genes

Ensembl Genes

ENSG000XXXX:
rsSNP00XXX
CNV_XXX
DN/DS XXXX
Recombination rate

Map to proteins in the interaction network

Interactome



~30000 interactions from HPRD and Y2H screens

Result

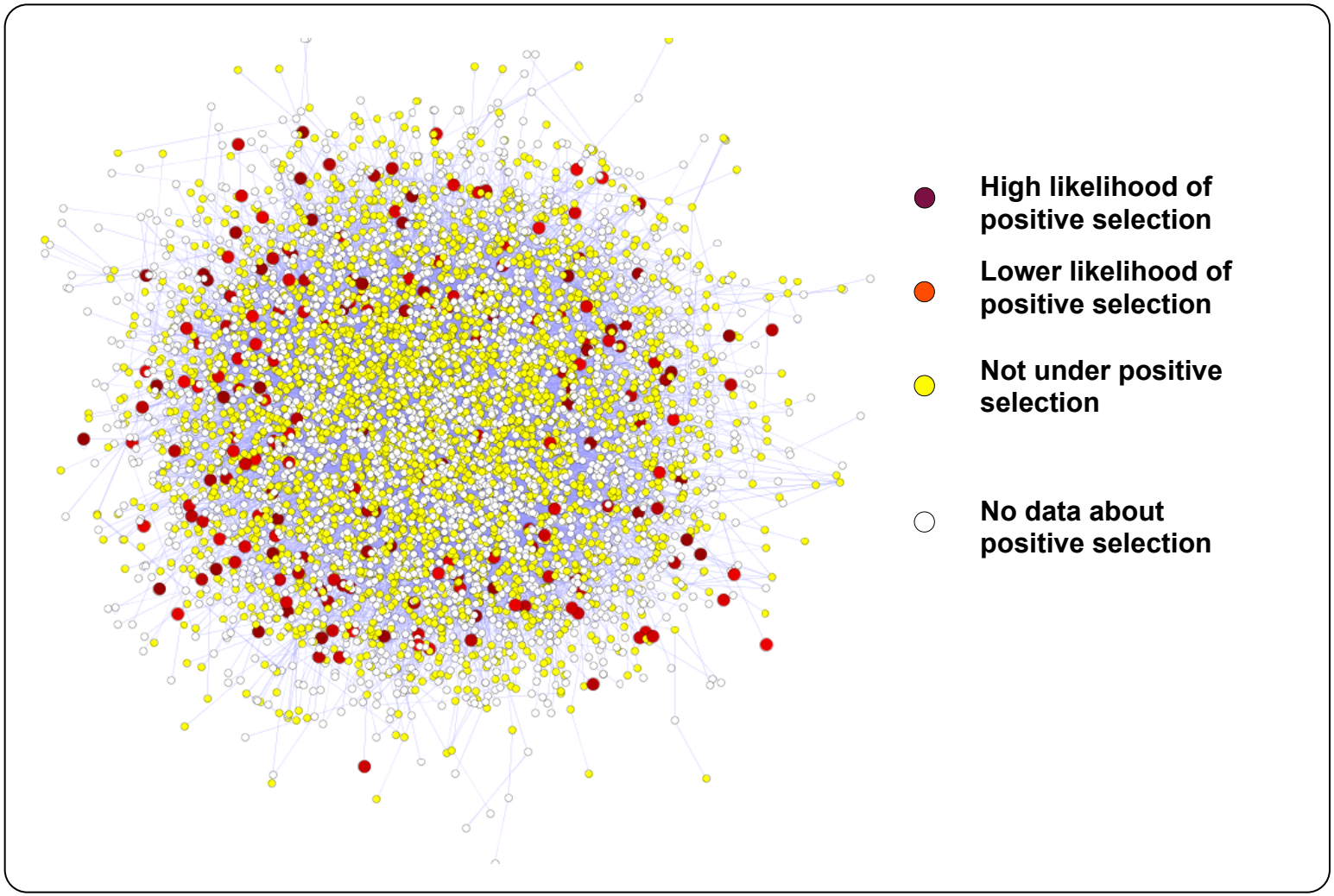
- Dataset of network position / parameters (e.g. degree centrality or betweenness centrality) in relationship to SNPs, CNV's, recombination rates and positive selection tests

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

Source: PMK

POSITIVE SELECTION LARGELY TAKES PLACE AT THE NETWORK PERIPHERY

Positive selection in the human interactome

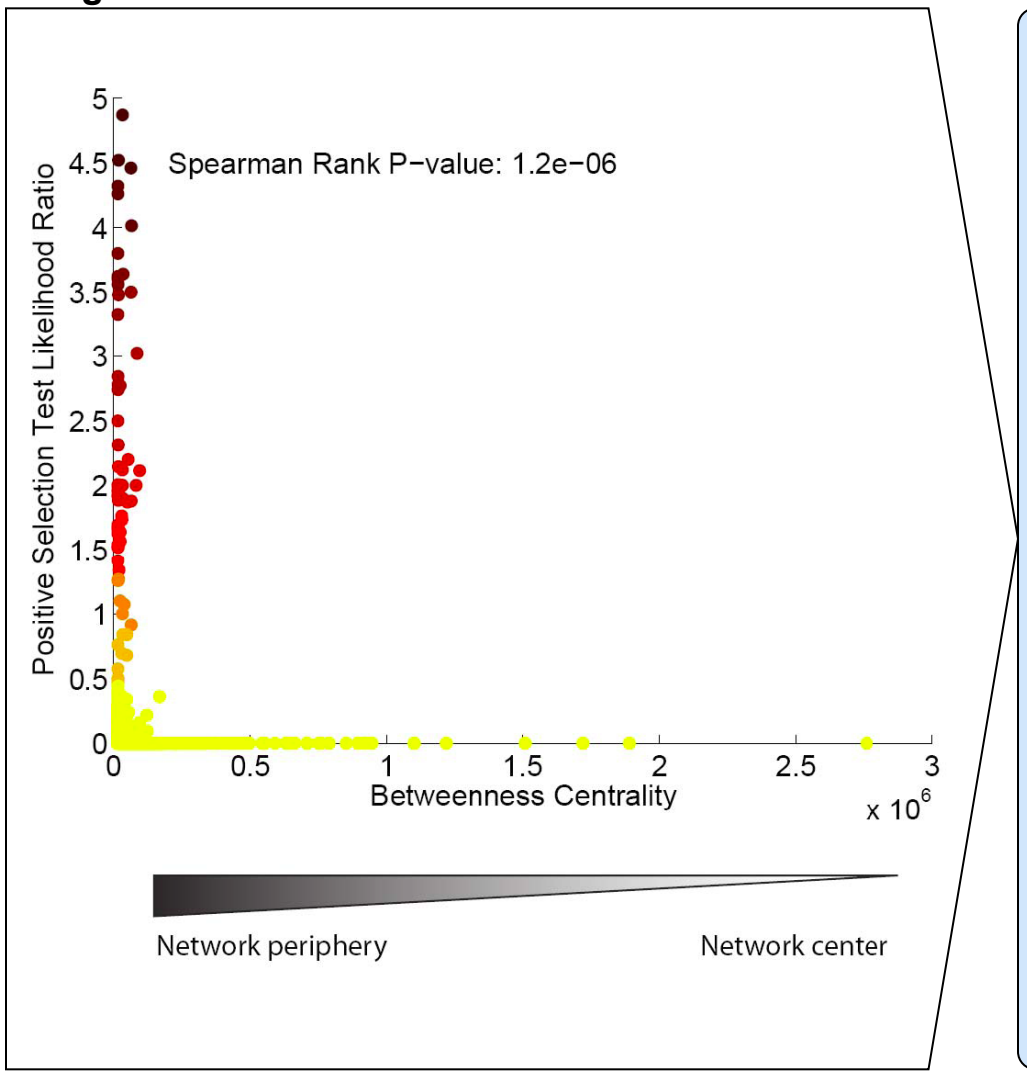


Source: Nielsen et al. *PLoS Biol.* (2005), HPRD, and Kim et al. PNAS (2007)

CENTRAL PROTEINS ARE LESS LIKELY TO BE UNDER POSITIVE SELECTION

Hubs

Degree vs. Positive Selection



Reasoning

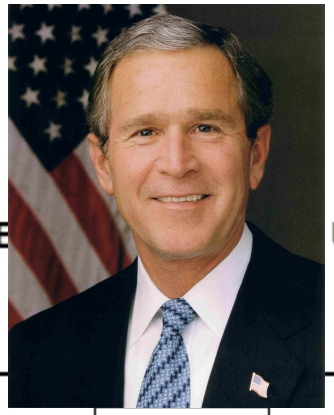
- Peripheral genes are likely to under positive selection, whereas hubs aren't
- This is likely due to the following reasons:
  - Hubs have stronger structural constraints, the network periphery doesn't
  - Most recently evolved functions (e.g. "environmental interaction genes" such as sensory perception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

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

Social Hierarchy



THE GOVE UNITED STATES

LEGISLATIVE BRANCH

THE CONGRESS

SENATE HOUSE

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

EX ESIDENT

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
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











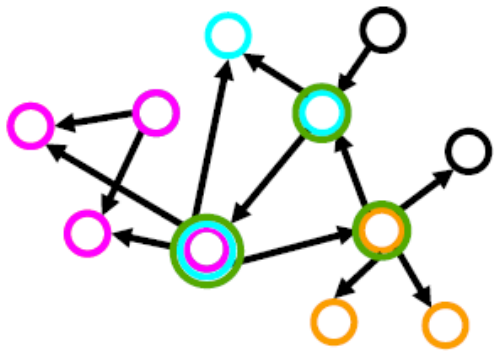



INDEPENDENT ESTABLISHMENTS AND GOVERNMENT CORPORATIONS

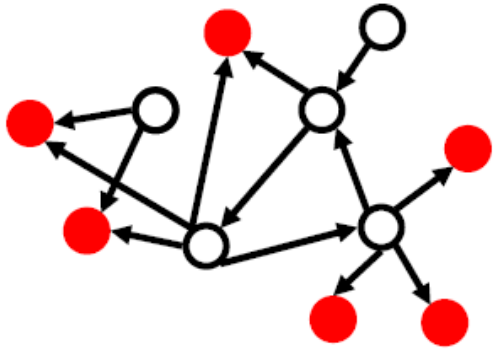
AFRICAN DEVELOPMENT FOUNDATION
CENTRAL INTELLIGENCE AGENCY
COMMODITY FUTURES TRADING COMMISSION
CONSUMER PRODUCT SAFETY COMMISSION
CORPORATION FOR NATIONAL AND COMMUNITY SERVICE
DEFENSE NUCLEAR FACILITIES SAFETY BOARD
ENVIRONMENTAL PROTECTION AGENCY
EQUAL EMPLOYMENT OPPORTUNITY COMMISSION
EXPORT-IMPORT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

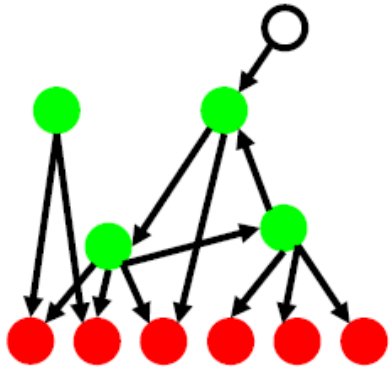
I. Example network with all 4 motifs



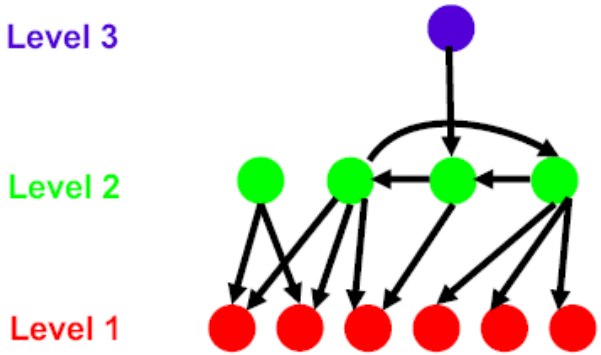
II. Finding terminal nodes (Red)



III. Finding mid-level nodes (Green)

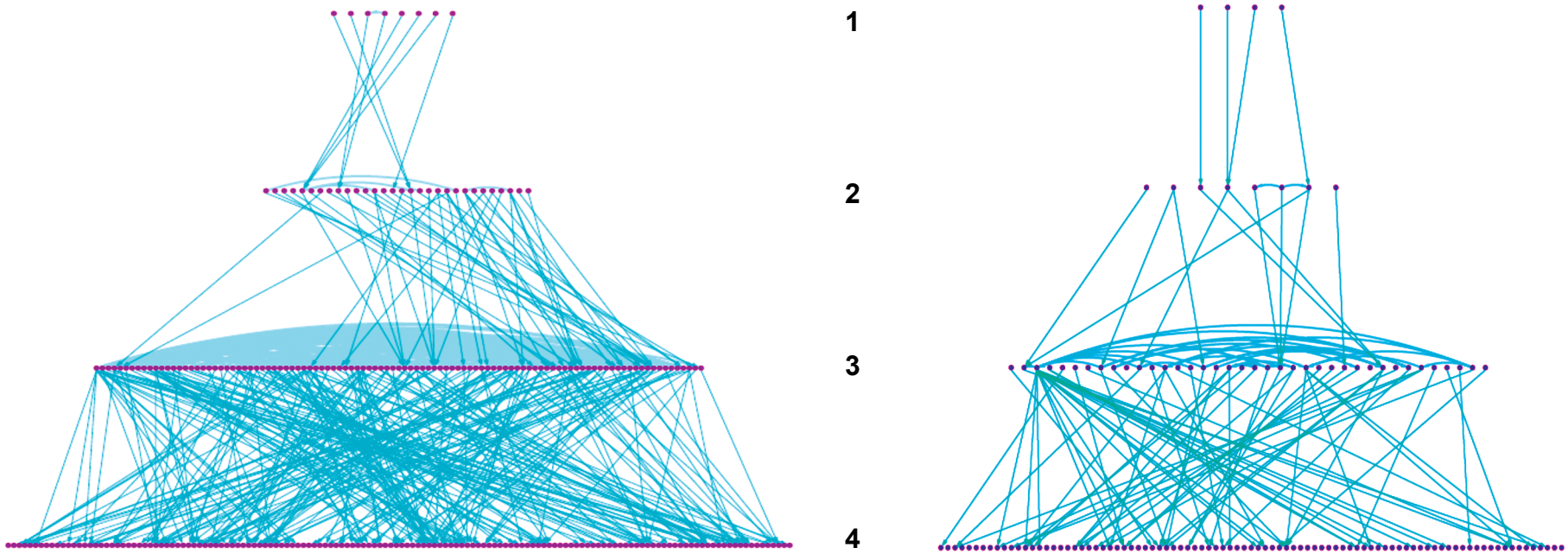


IV. Finding top-most nodes (Blue)



[Yu et al., PNAS (2006)]

Regulatory Networks have similar hierarchical structures



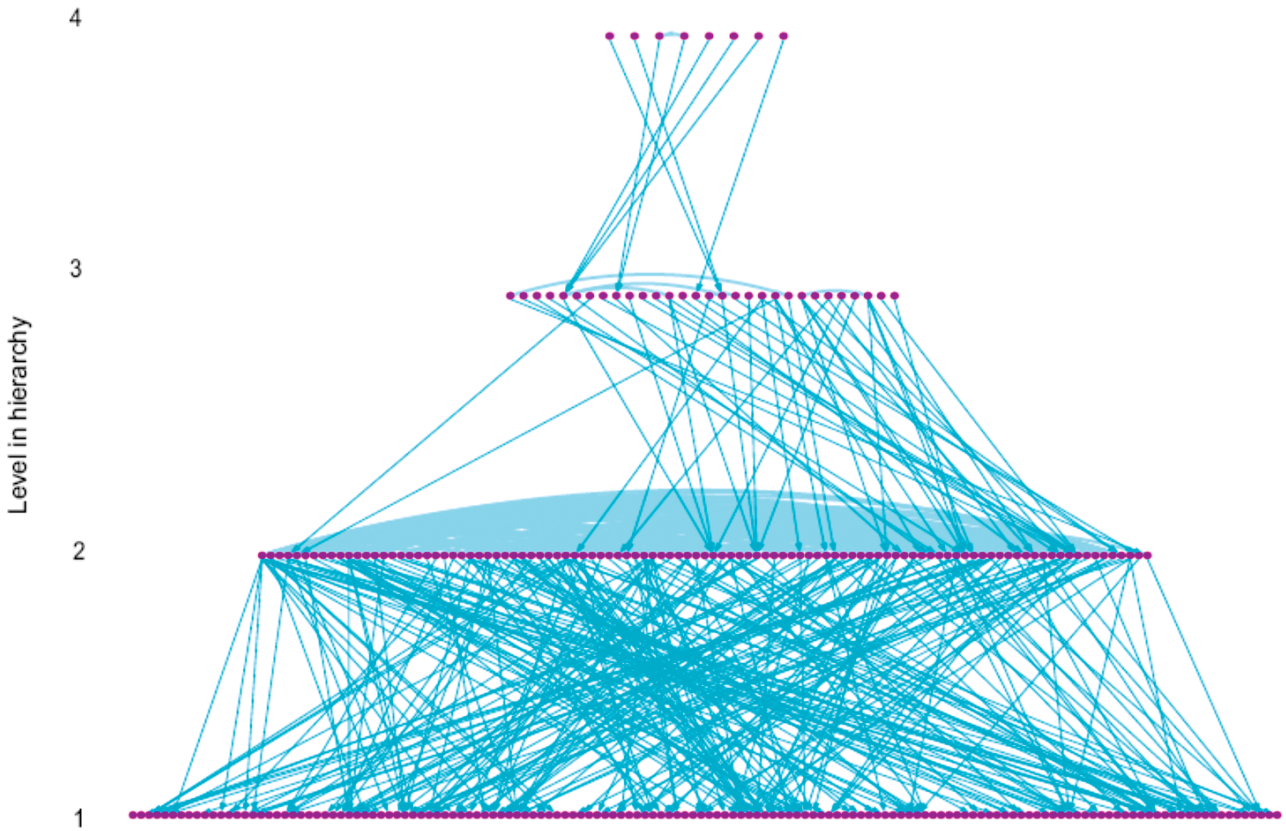
S. cerevisiae

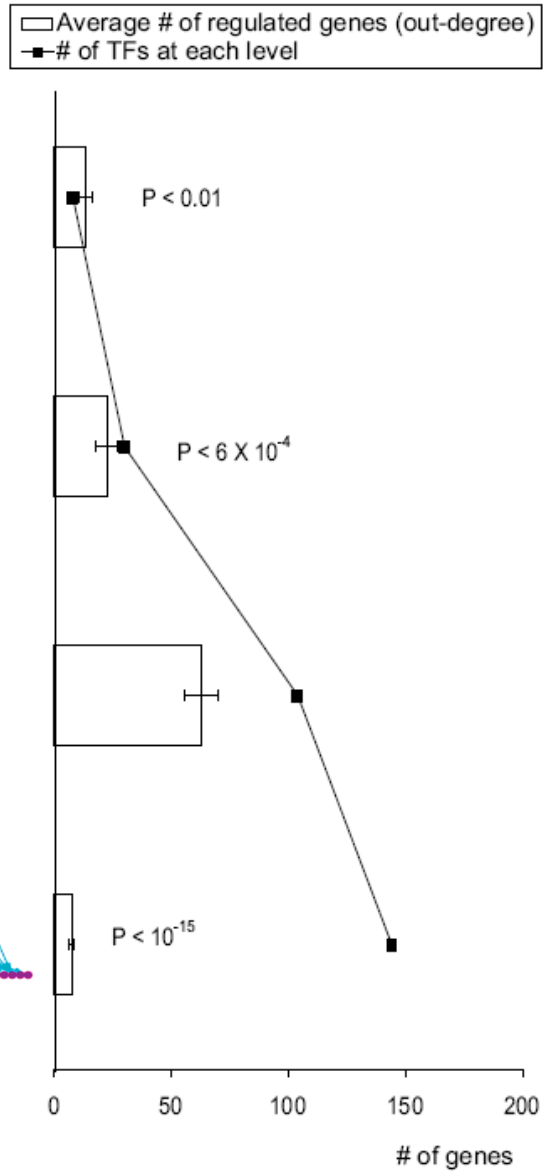
E. coli

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

Yeast Regulatory Hierarchy: the Middle-managers Rule

A. Regulatory hierarchy in *S. cerevisiae*

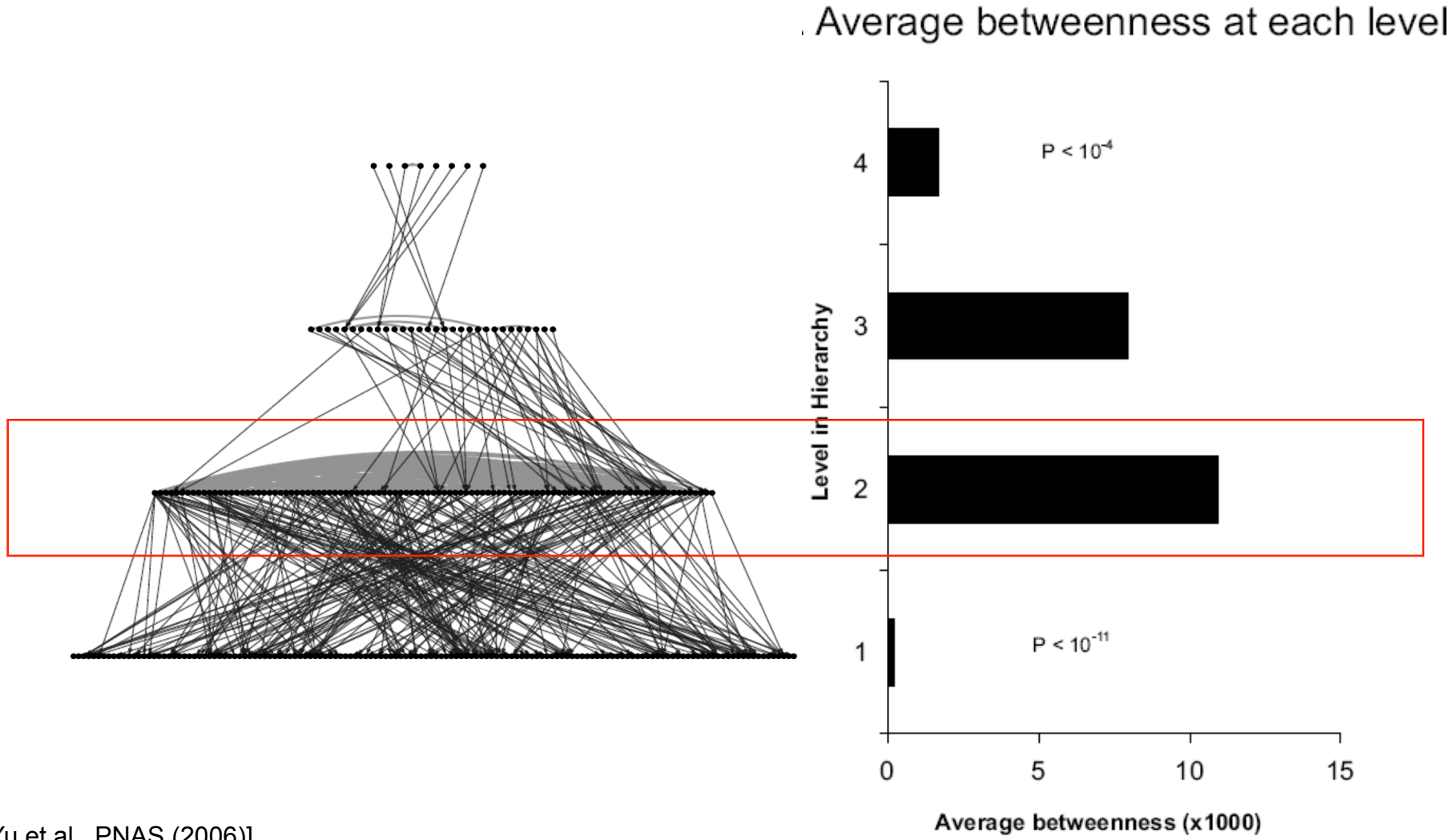




[Yu et al., PNAS (2006)]

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

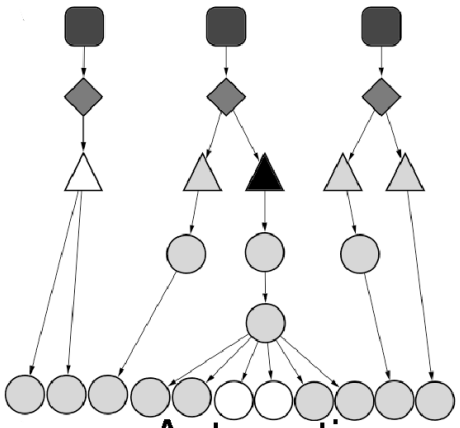
Characteristics of Regulatory Hierarchy: Middle Managers are Information Flow Bottlenecks



[Yu et al., PNAS (2006)]

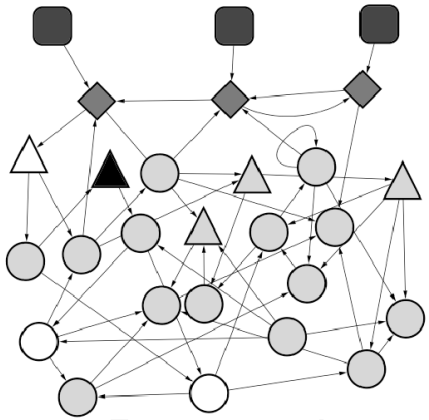
Social Network Comparison #2
Broadening the comparison to different types of hierarchies & different types of biological networks

Different kinds of Hierarchies



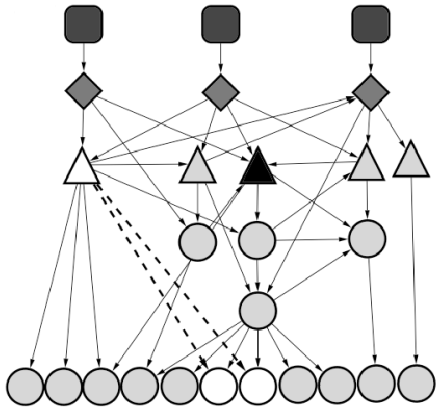
Autocratic

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



Democratic

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



Intermediate

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

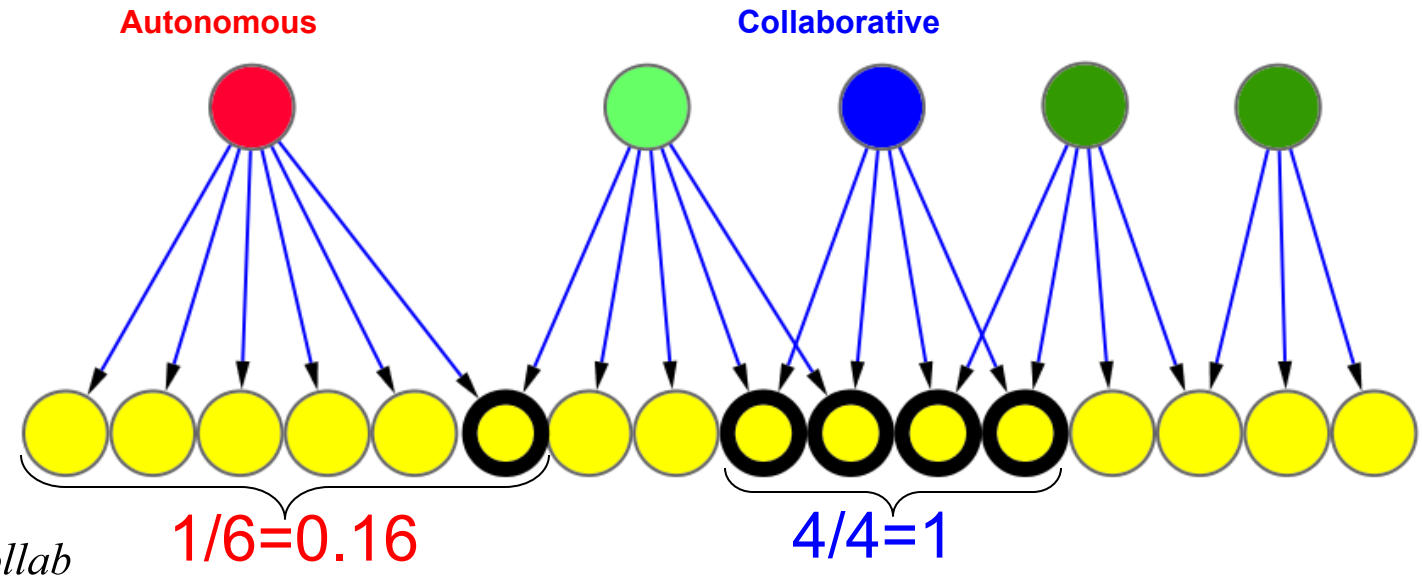
	Autocratic	Democratic	Intermediate
Betweenness (△)	1.03	3.6	3.3
Betweenness (▲)	4.1	1.08	3.4
Var. Betw. (triangles)	2.1	0.58	1.74
Var. Betw. (all)	2.9	1.4	1.9
$D_{Net\text{-}collab}$	0	0.91	0.71

Decreases σ^2 of 'stress'

Increases 'collaboration'

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

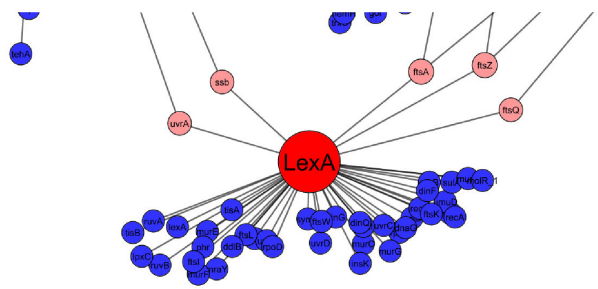
Collaborative Nature of the Nodes



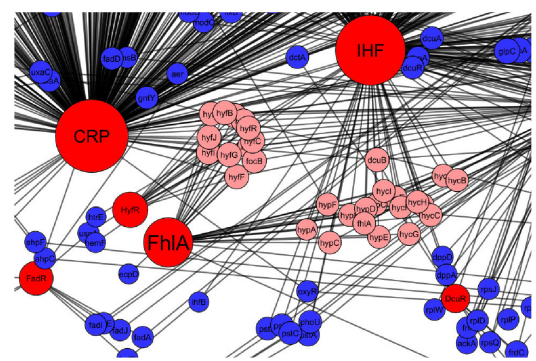
Degree of Collaboration D^{i}_{collab}

More Collaborative: Democratic
More Autonomous: Autocratic

Autonomous

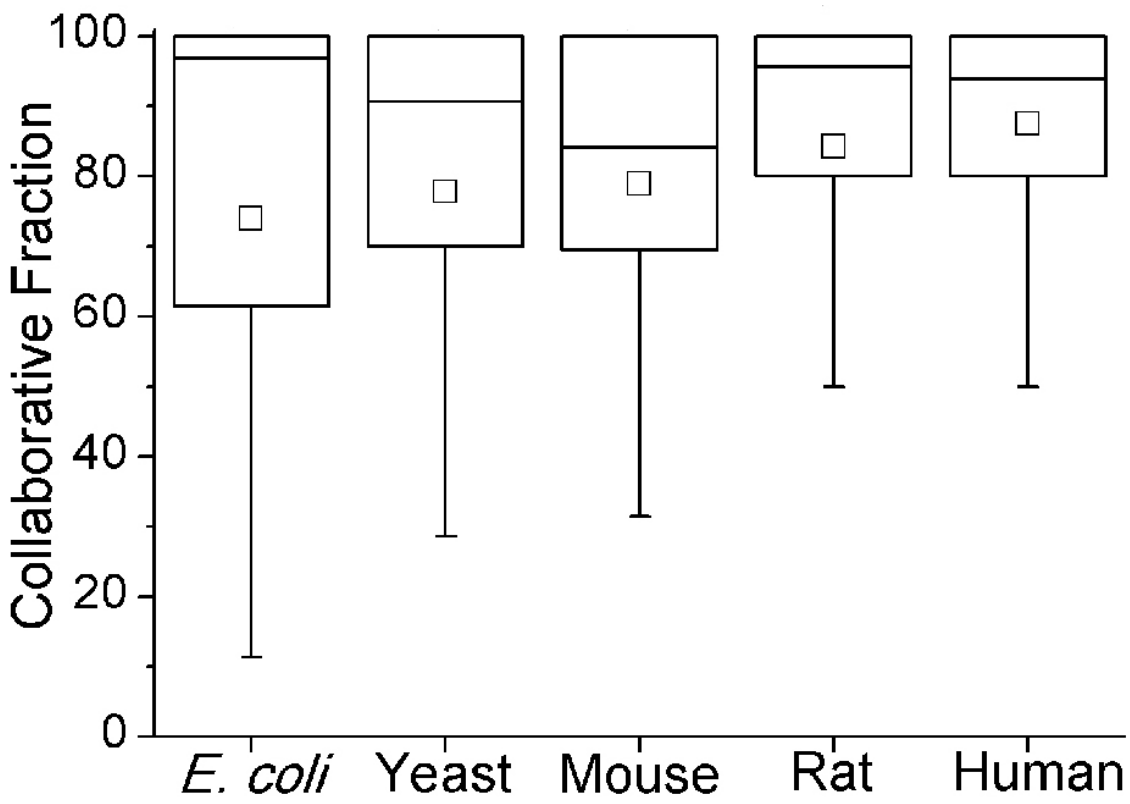


Collaborative



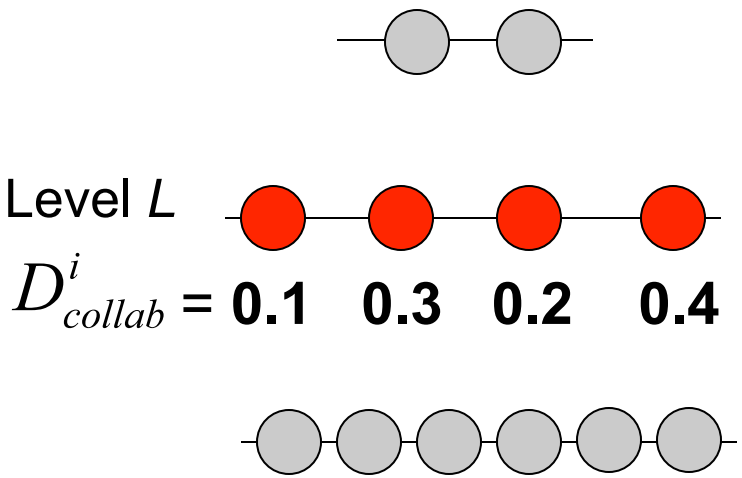
[Bhardwaj et al., PNAS (2010), in press]

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

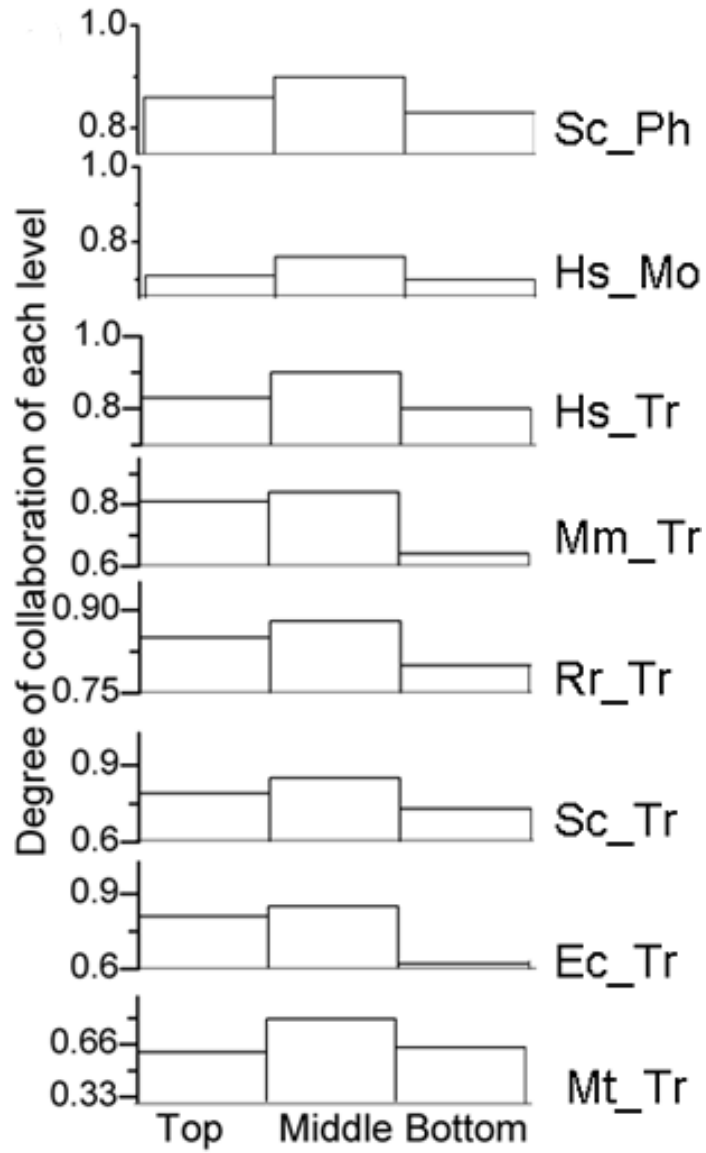


[Bhardwaj et al., PNAS (2010), in press]

Collaborative Nature of the Levels

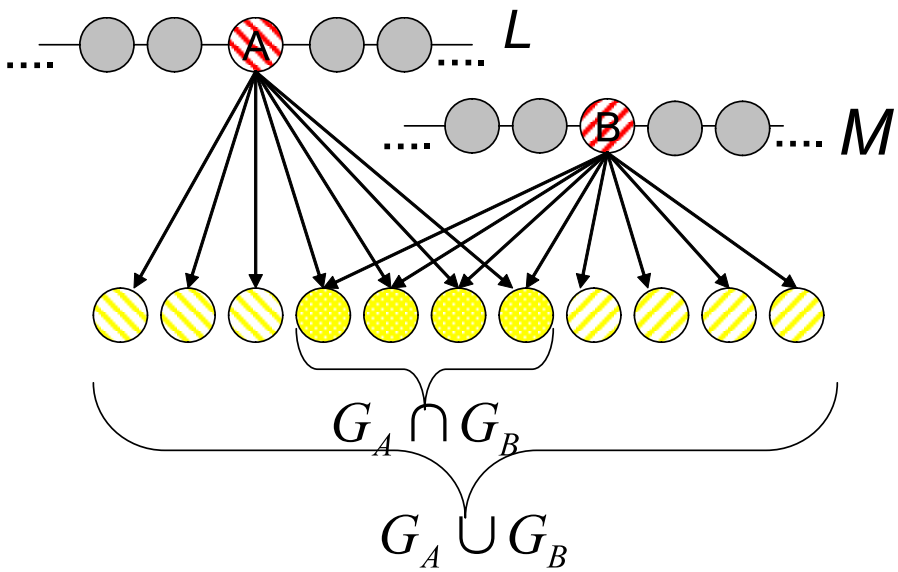


$$D^{L}_{Level-collab} = \frac{0.1+0.3+0.2+0.4}{4} = 0.25$$

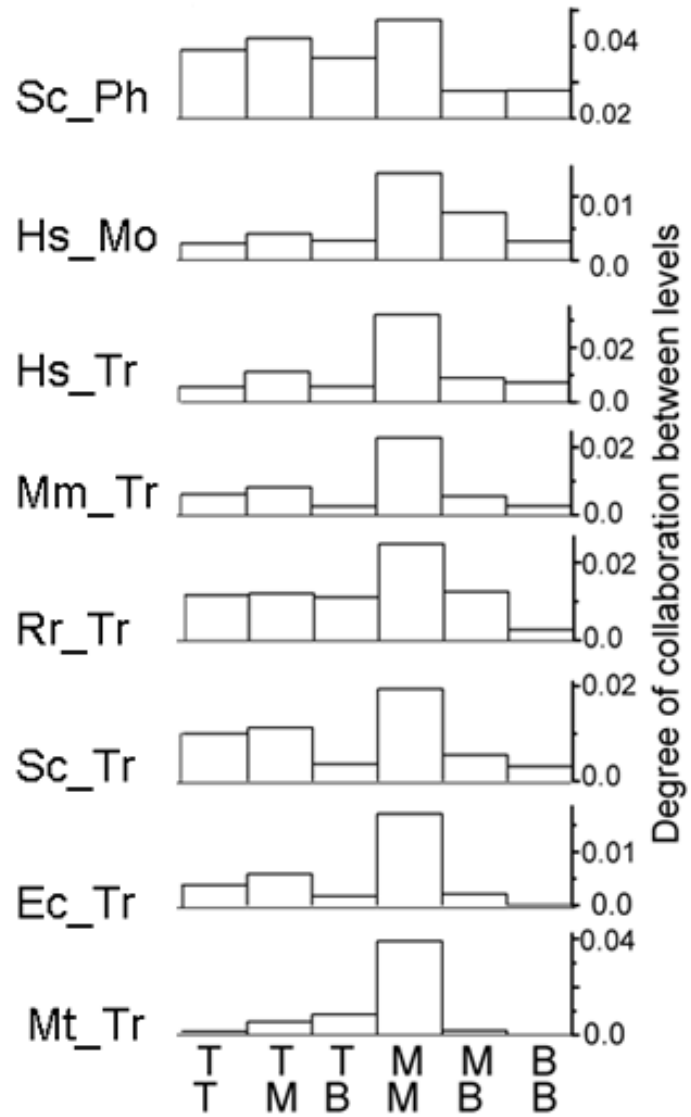


[Bhardwaj et al., PNAS (2010), in press]

Collaboration Between Levels



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



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

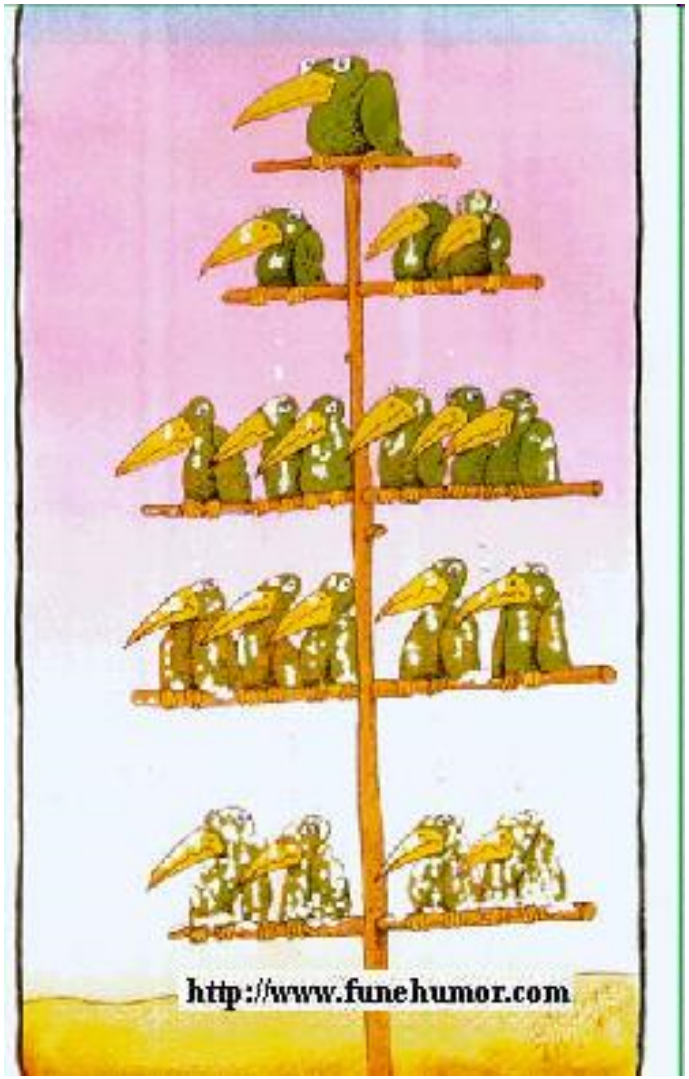
Middle Managers Interact the Most in Efficient Corporate Settings

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

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

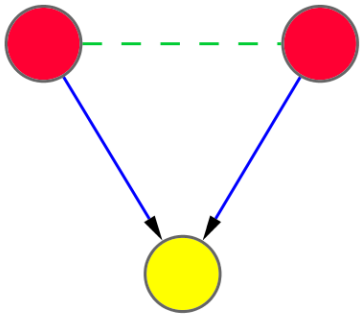
- Floyd, S. W. et al (1993)
  Dinosaurs or Dynamos? Recognizing Middle Management's Strategic Role
  The Academy of Management Executive 8, 47-57.

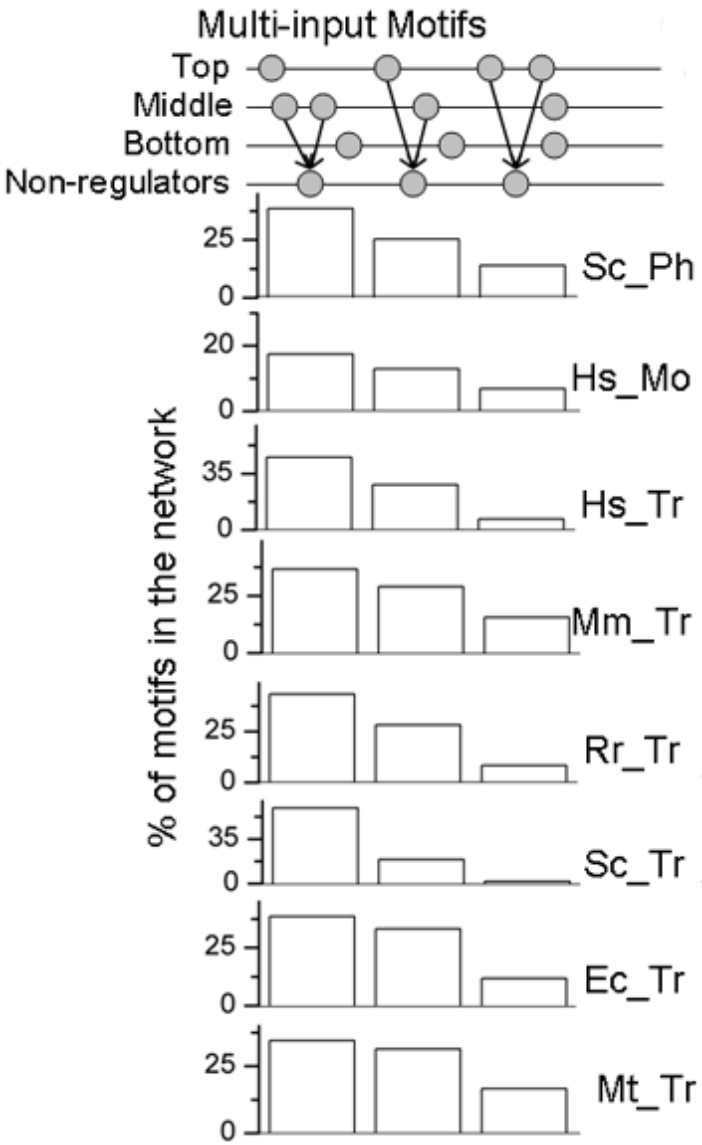
- Floyd, S. W. et al (1997)
  Middle management's strategic influence and organizational performance
  Journal of Management Studies 34, 465-485.



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

Co-regulation Instantiates a Multi-Input Motif

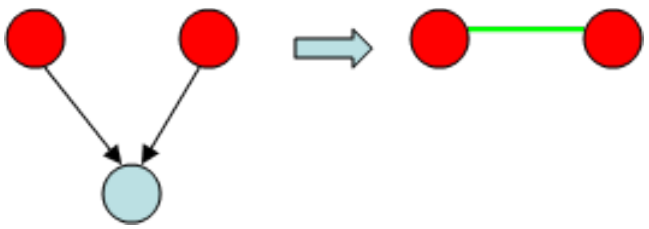




[Bhardwaj et al., PNAS (2010), in press]

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

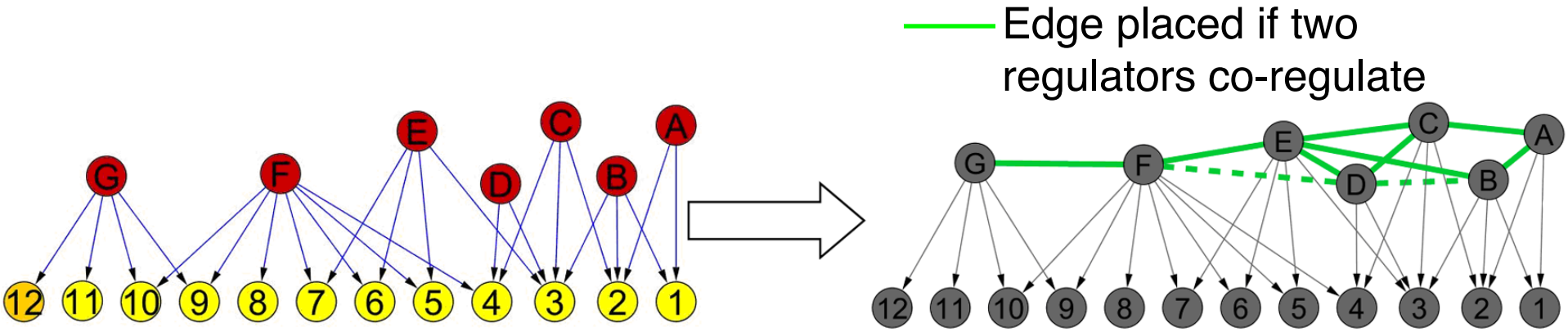
Co-regulation Partnerships



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

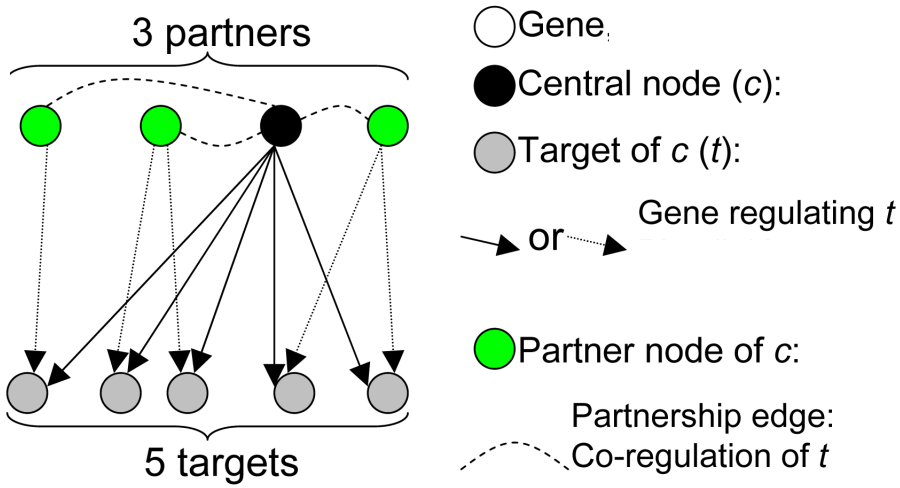
[Bhardwaj et al., PLoS Comp Biol (2010), in press]

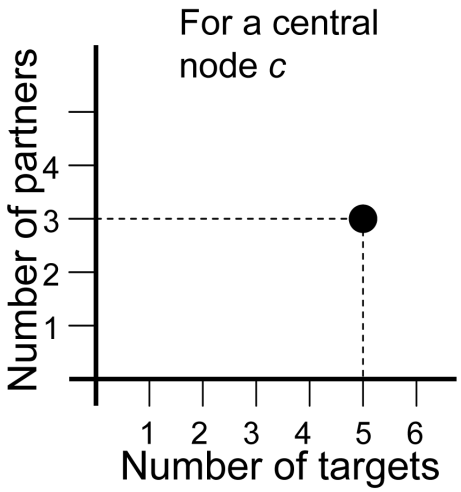
Building and Analysis of Networks

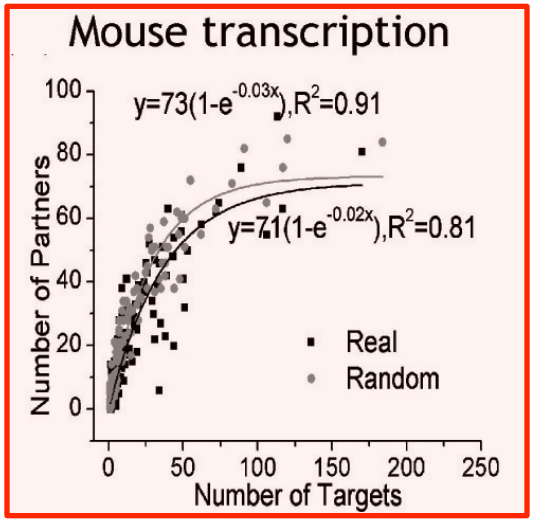


Network type	Species	Number of regulators	Number of targets	Number of interactions
Transcription	*E. coli*	160	1,420	3,123
Transcription	Yeast	157	4,410	12,873
Transcription	Mouse	144	1,092	2,403
Transcription	Rat	91	461	1,092
Transcription	Human	156	3,032	6,896
Phosphorylation	Yeast	87	1,337	4,083
Modification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

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

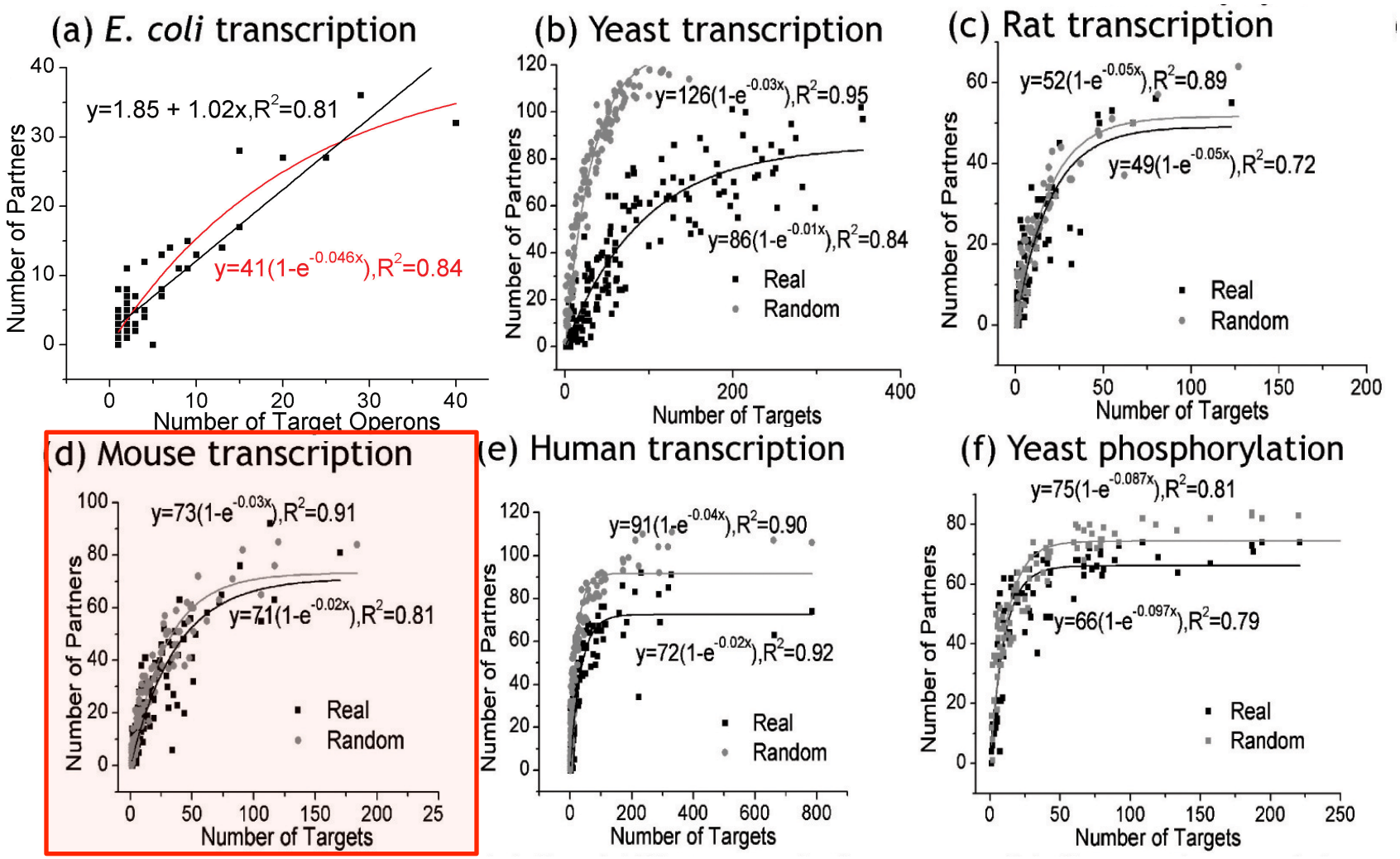






Scaling of Regulatory Partnerships with Targets

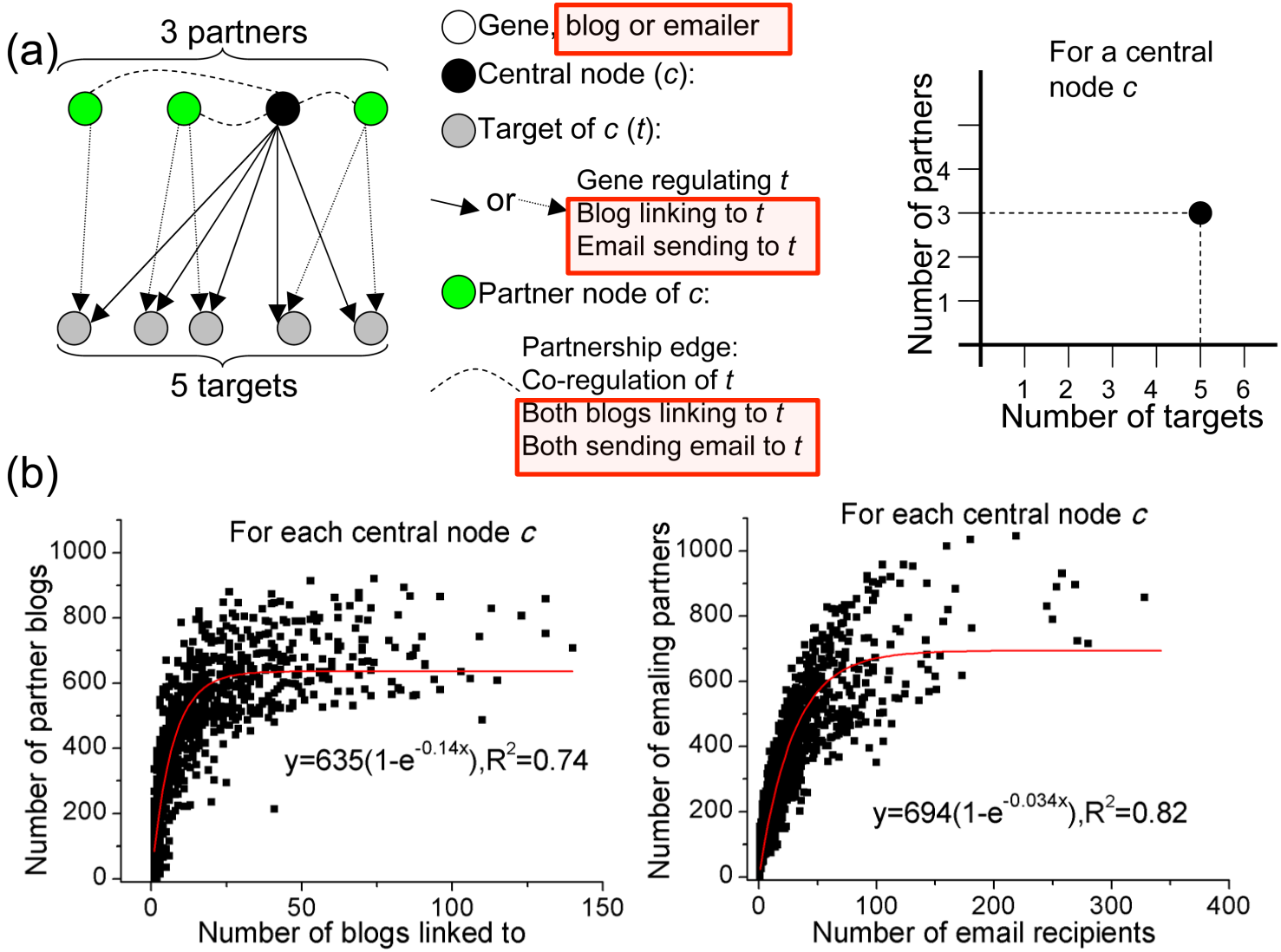
Scaling across many networks



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

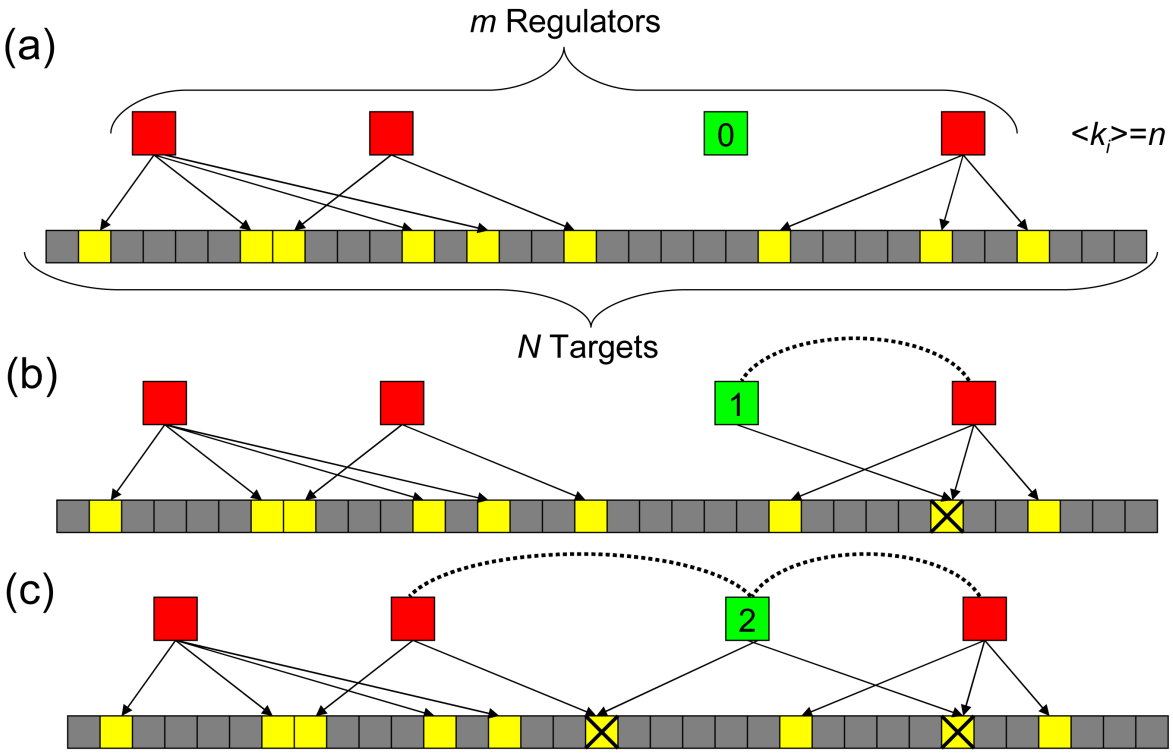
[Bhardwaj et al., PLoS Comp Biol (2010), in press]

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



A Simple Theoretical Model

On average, each regulator has *n* targets
f_i : the number of partners
k_i :the number of targets



$<k_i>=n$

When $f=0$,

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

When $f=1$,

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

Generalizing,

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

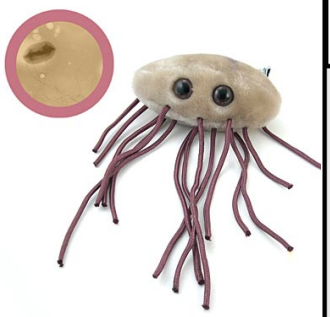
Integrating this, we get:

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

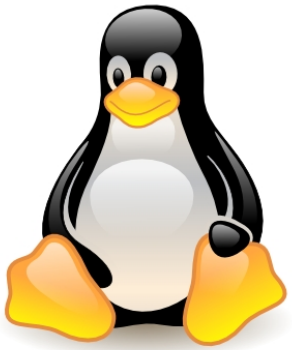
[Bhardwaj et al., PLoS Comp Biol (2010), in press]

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

E. Coli Transcriptional regulatory network vs Linux kernel call graph



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



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

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

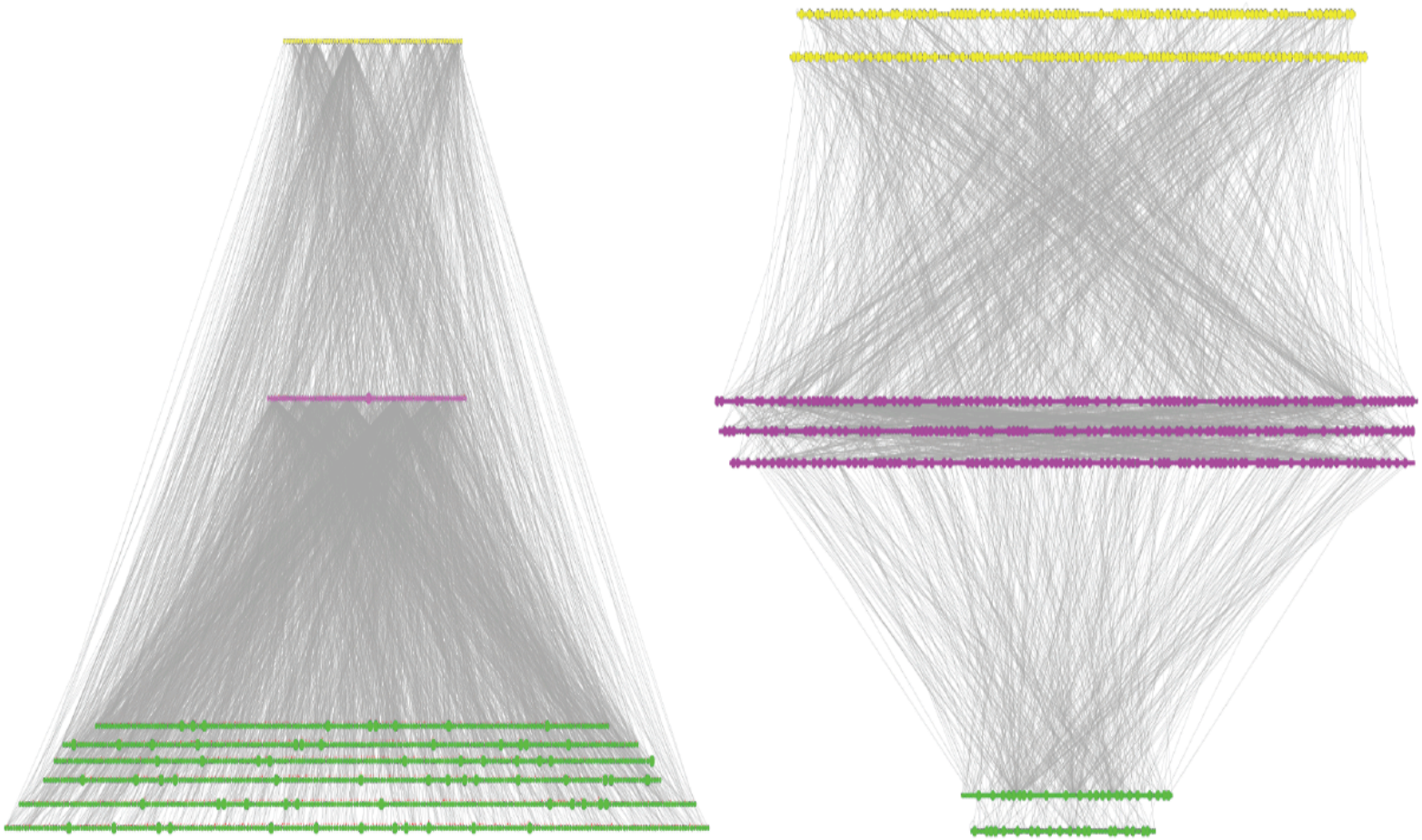
E. coli transcriptional regulatory network

Linux call graph

master regulator

middle manager

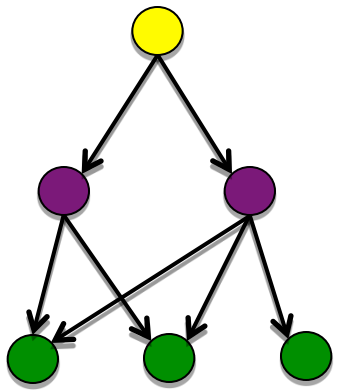
workhorse



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

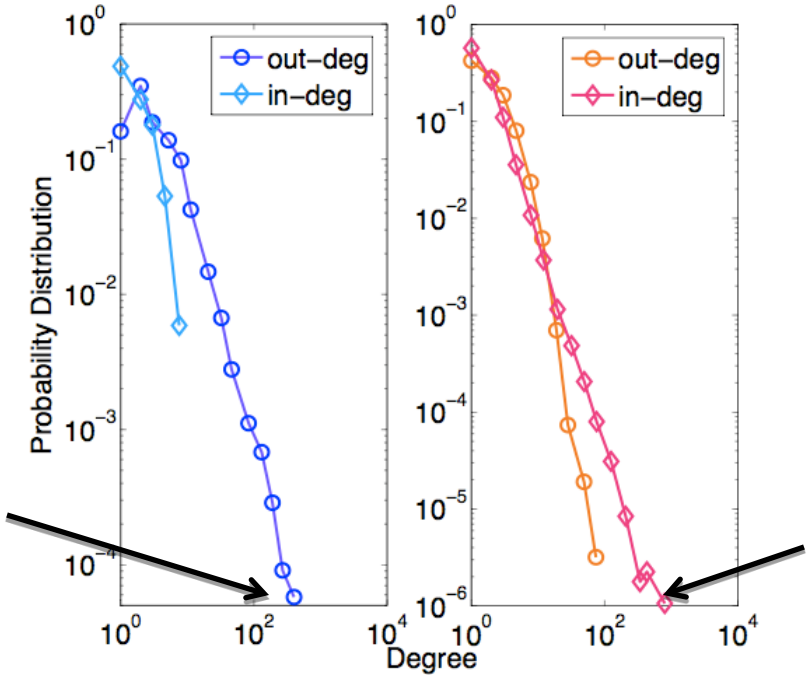
Comparison: hierarchical organization

Pyramidal vs
Top-heavy



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

Degree distribution
Roles of hubs



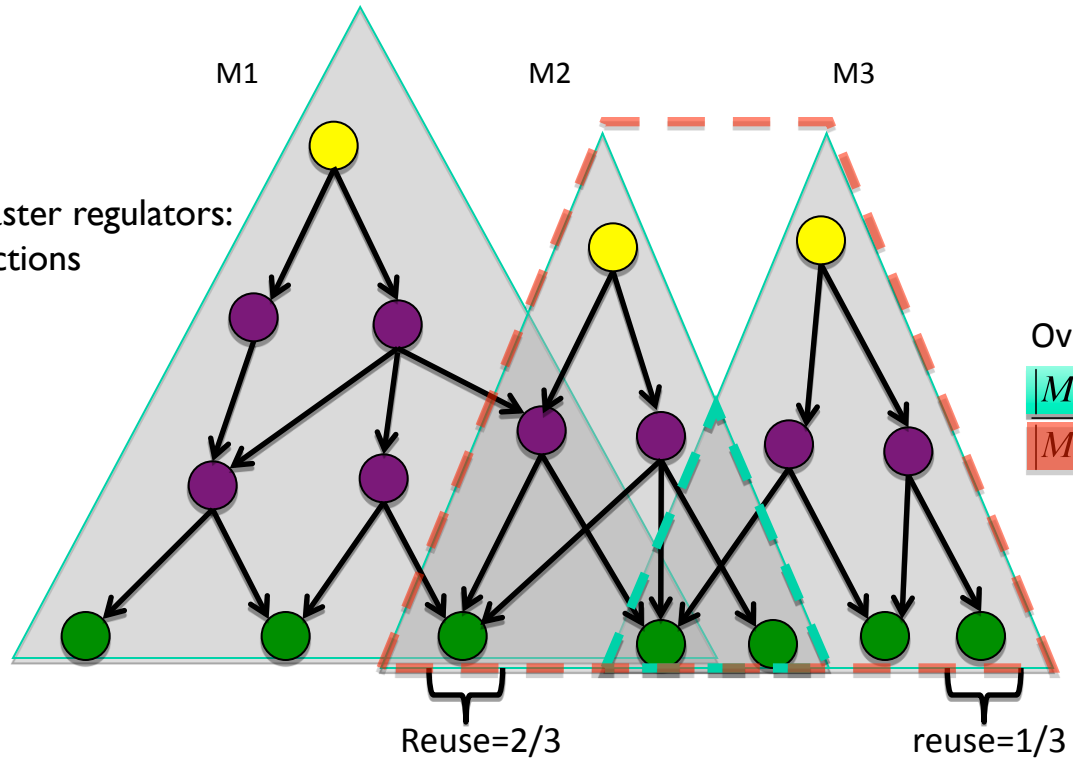
out-degree hubs
e.g. “crp”

in-degree hubs
e.g. “printk”

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

Comparison: organization of modules

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



Overlap(M2,M3)=

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

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%

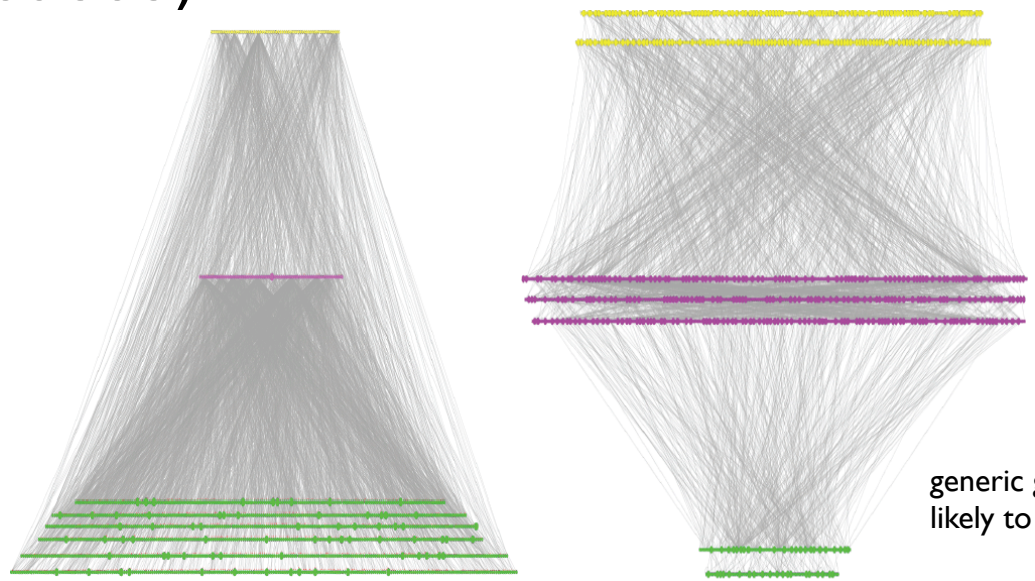
Call graph:
modules overlap,
Functions are highly
reused (generic):
"printk"

[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

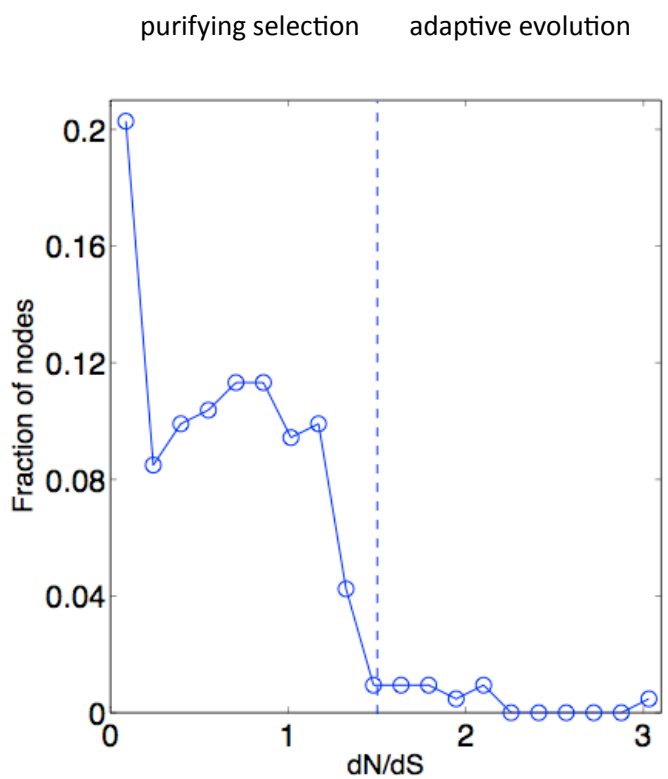


generic guys are more
likely to be persistent

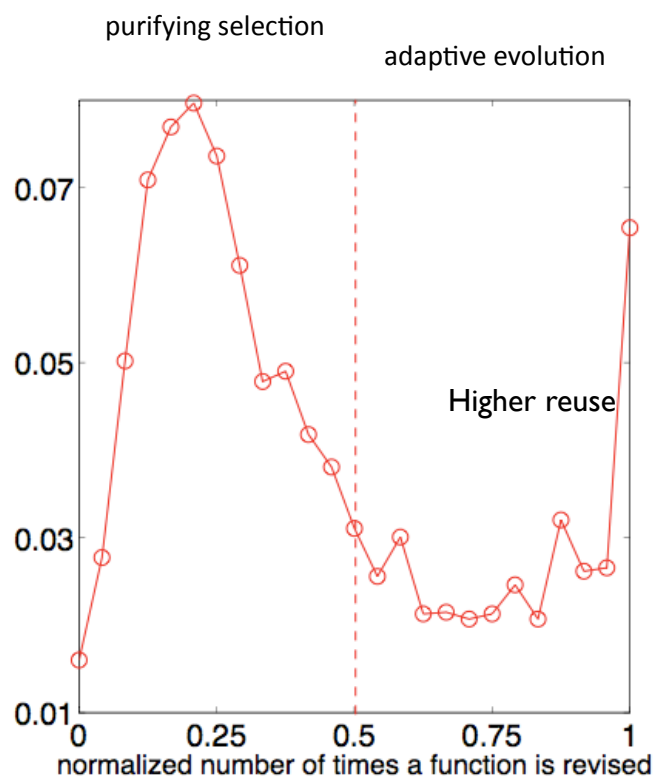
- Building of the hierarchy:
  - ◊ TRN: Bottom up. Regulatory changes are the main driving forces of evolution
  - ◊ Call graph: top down

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

Evolutionary rate of persistent functions



Persistent genes evolve slowly

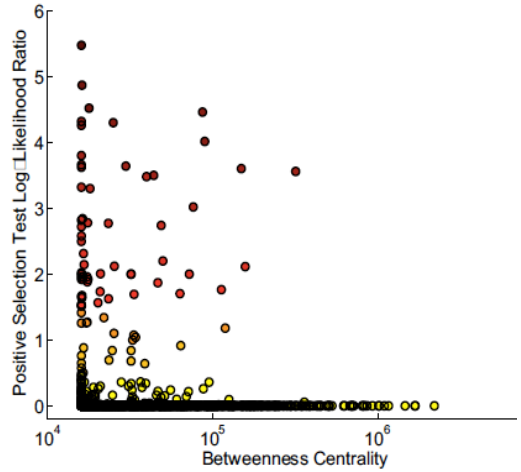


Two classes of persistent functions

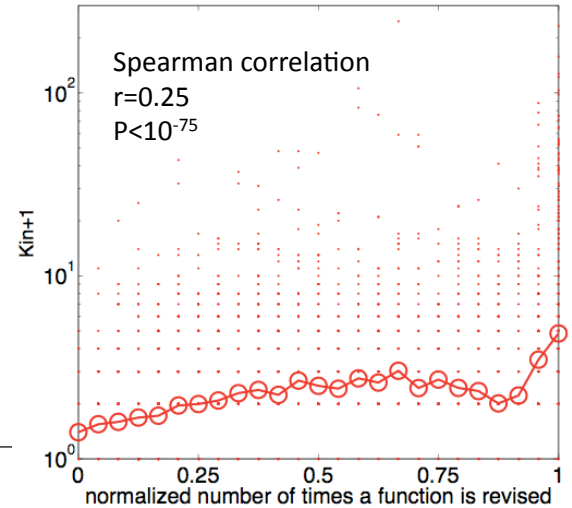
[Yan et al., PNAS (2010), in press]

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]

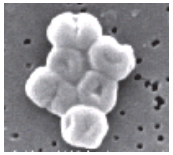
Network Dynamics Across Environments: Metabolic Pathways

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

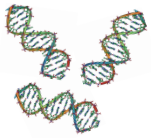
What is Metagenomics?

Traditional Genomics

Select organism and culture

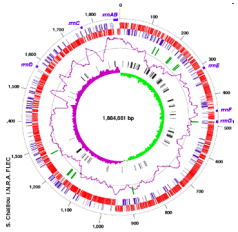


Extract DNA and sequence



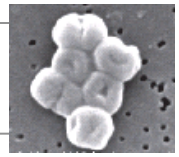
atgctcgatctcg
atcgatctcgctg
atgccgatctaa

Assemble and annotate

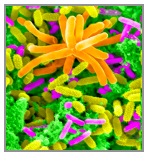


Contig 1
Contig 2
. . .
. . .

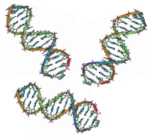
Estimated that less than 1% of microbes can be cultured



Metagenomics

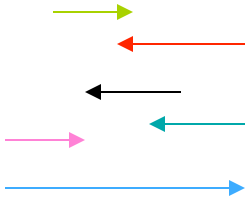


Collect sample from environment



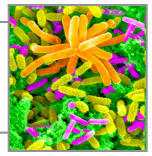
atgctcgatctcg
atcgatctcgctg
atgccgatctaa

Extract DNA and sequence

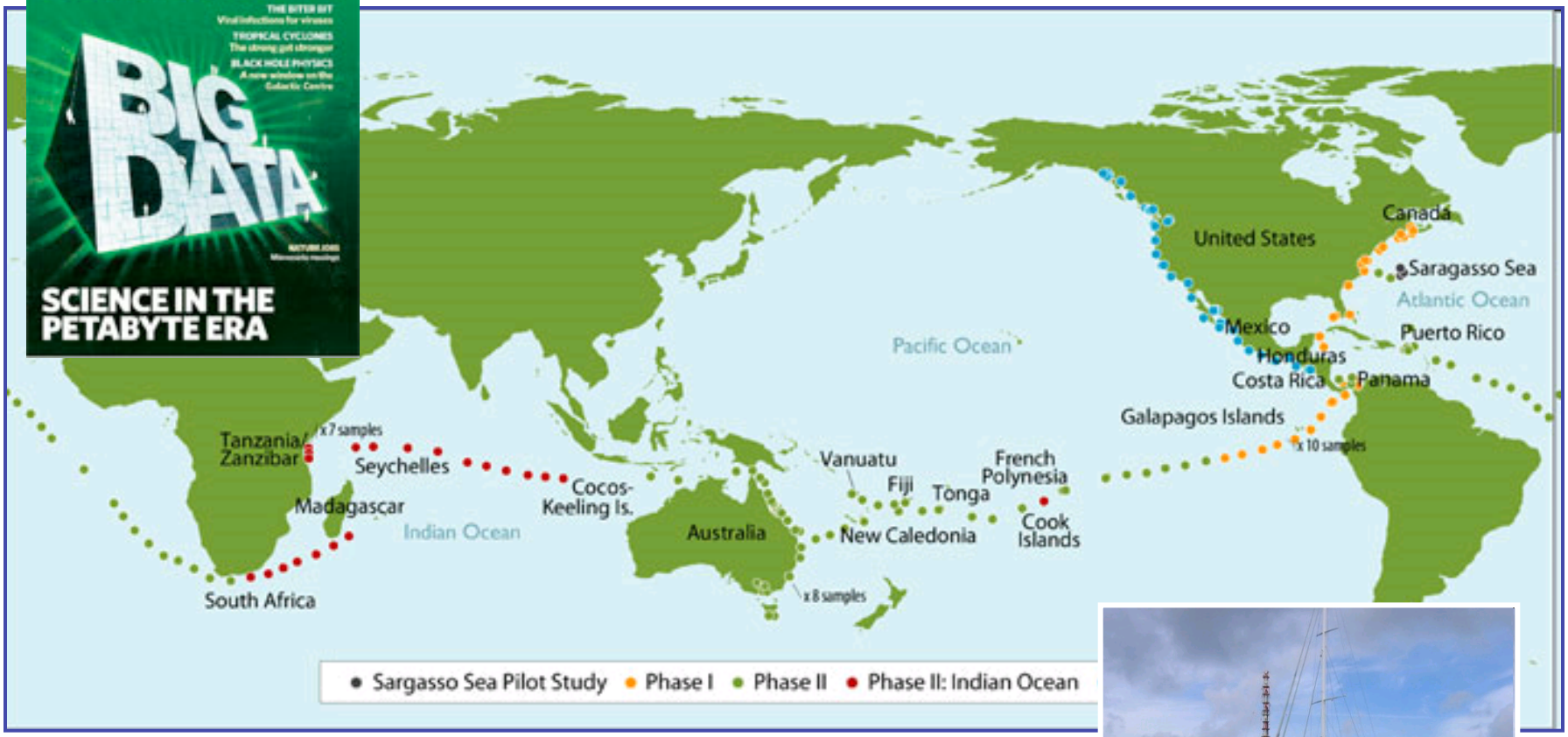


Assemble and annotate

Lose information about which gene belongs to which microbe



Sorcerer II Global Ocean Survey

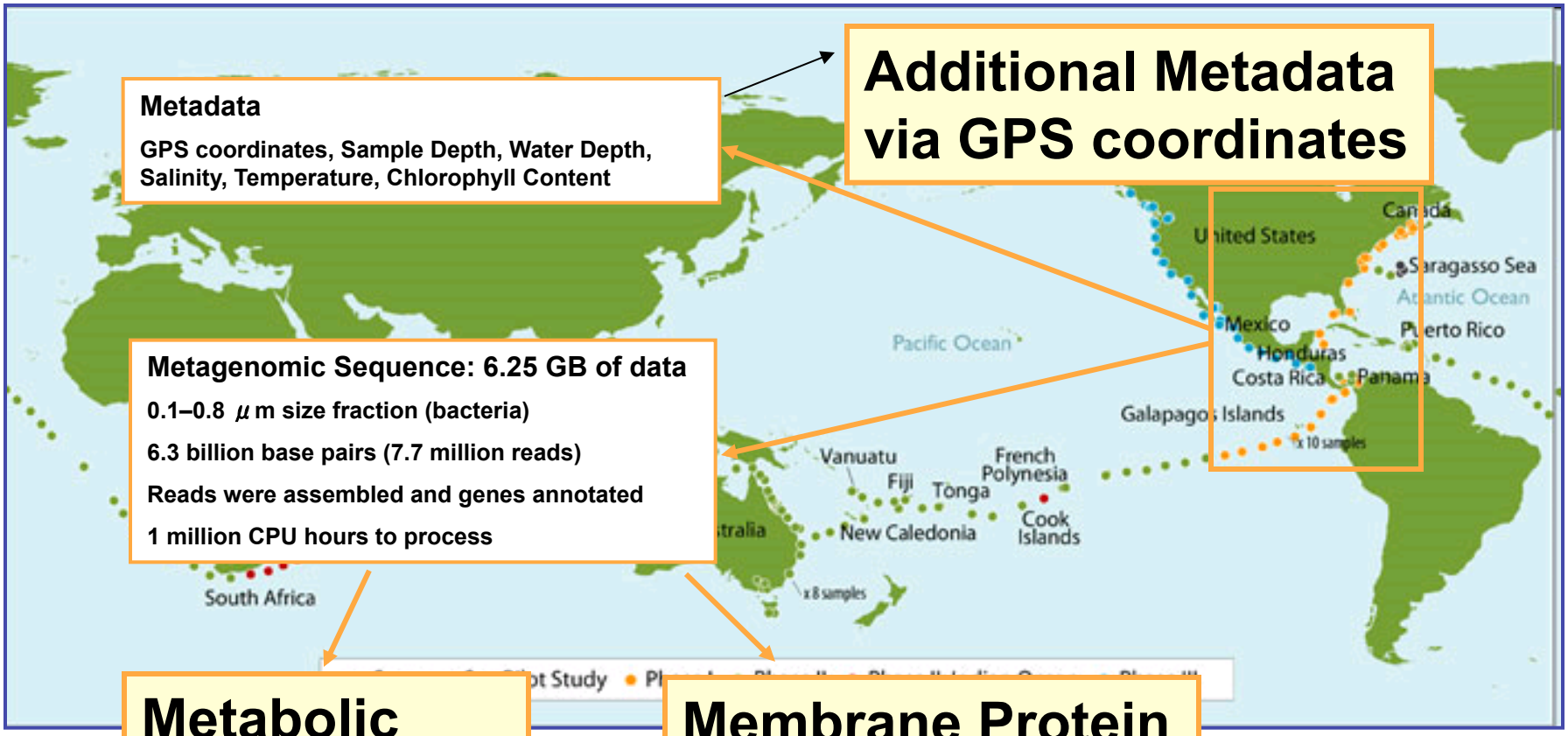


Sorcerer II **journey August 2003- January 2006**

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

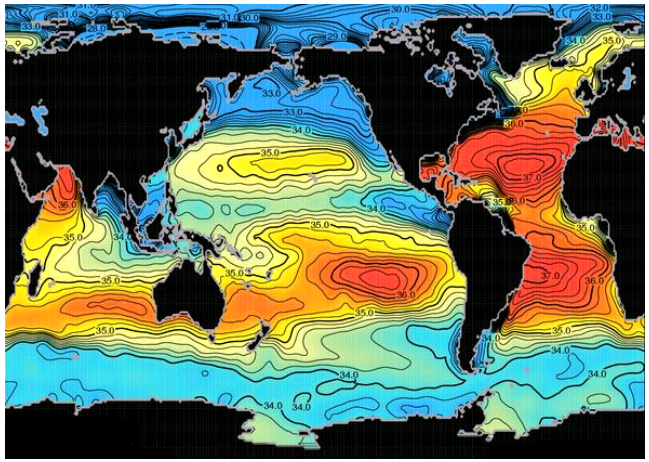
Rusch, et al., *PLOS Biology* 2007

Extracting environmental data using GPS Coordinates

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

GPS coordinates allow us to extract information from other sources:



*** World Ocean Atlas**

*** National Center for Ecological Analysis and Synthesis**

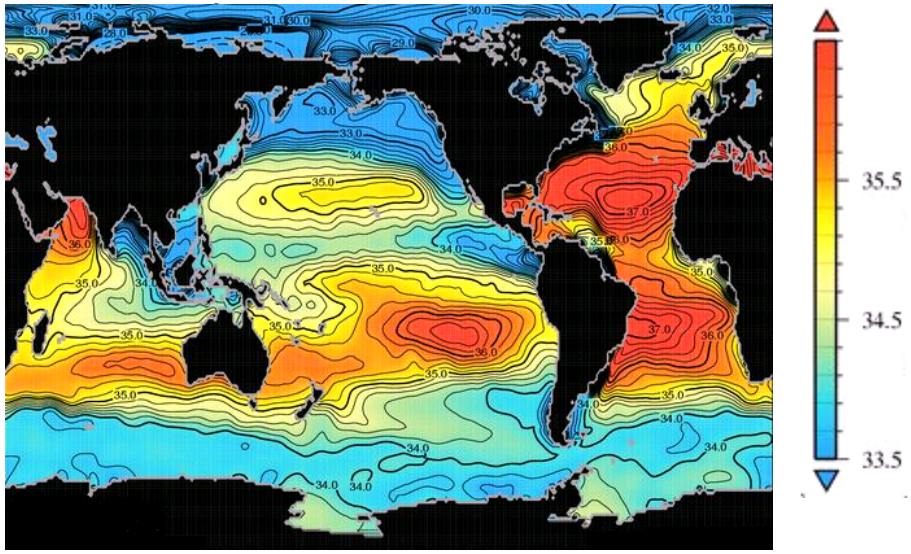
Patel et. al., *Genome Research* 2010

World Ocean Atlas 2005

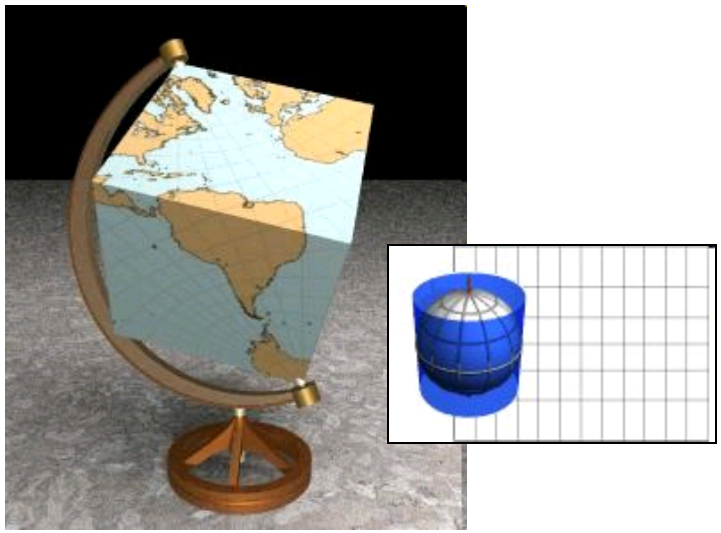
NOAA (National Oceanic and Atmospheric Administration) and
NODC (National Oceanographic Data Center)

*** Cumulative annual data at the ocean surface**
*** Resolution is 1 degree latitude/longitude**

Annual Phosphate [umol/l] at the



. . . no simple geometric shape matches the Earth



Nutrient Features Extracted:

Phosphate

Silicate

Nitrate

Apparent Oxygen Utilization

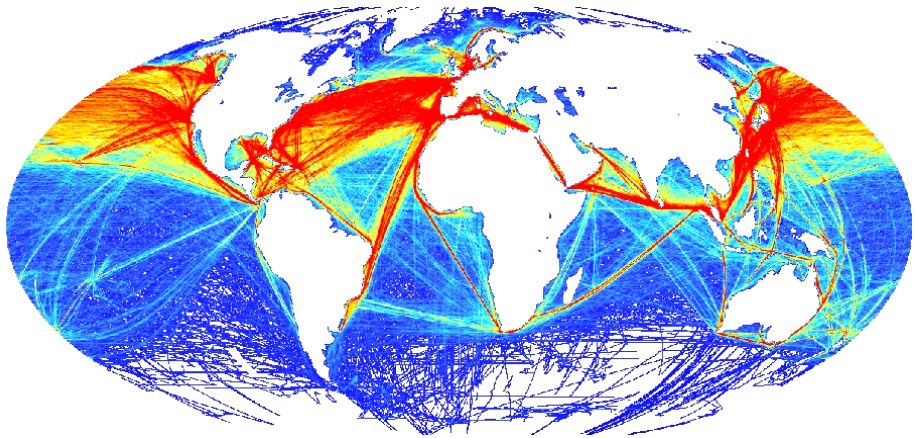
Dissolved Oxygen

National Center for Ecological Analysis and Synthesis (NCEAS)

*** 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:

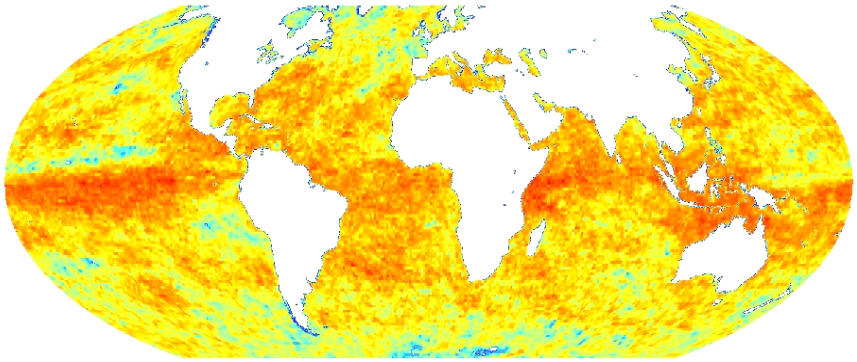
Ultraviolet radiation

Shipping

Pollution

Climate Change

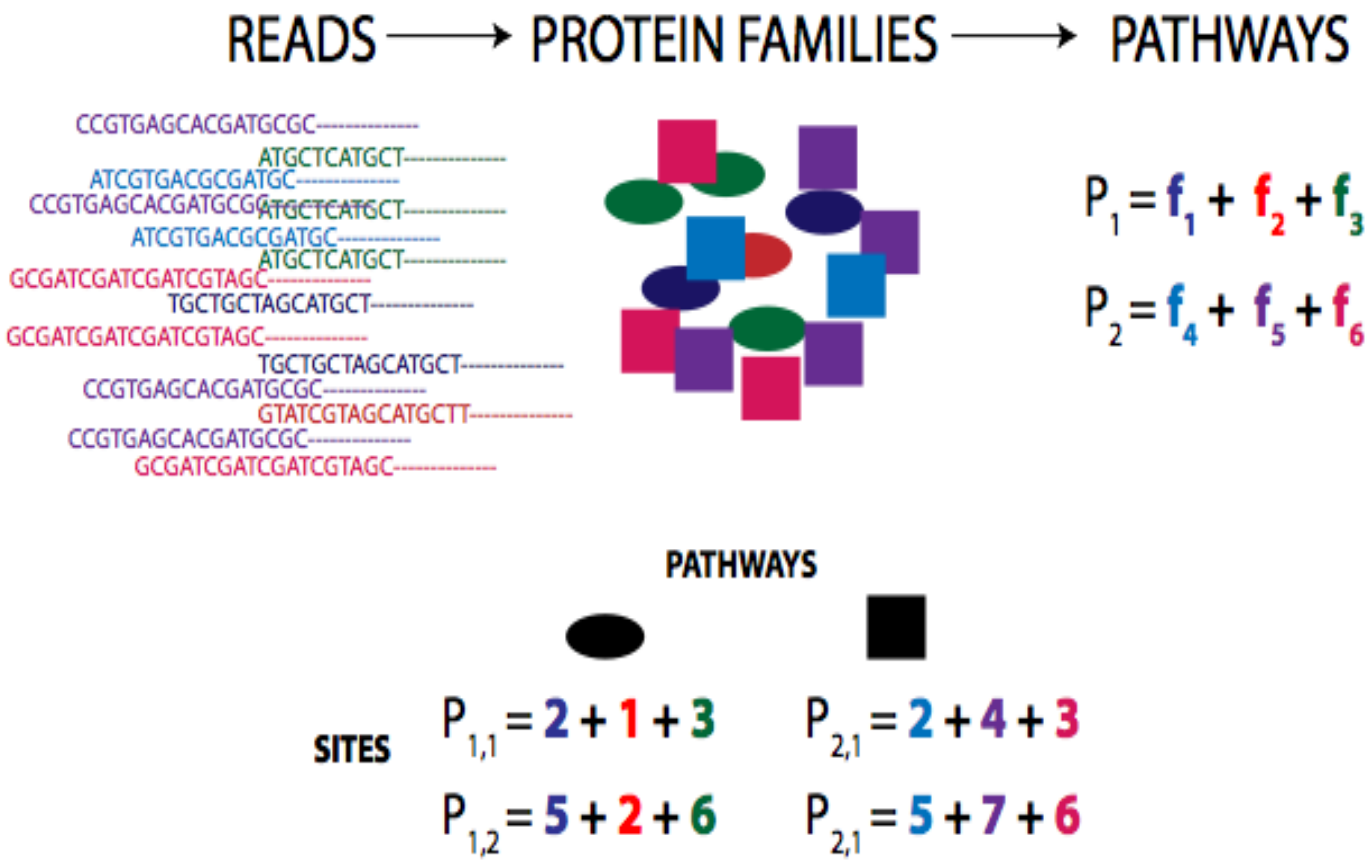
Ocean Acidification

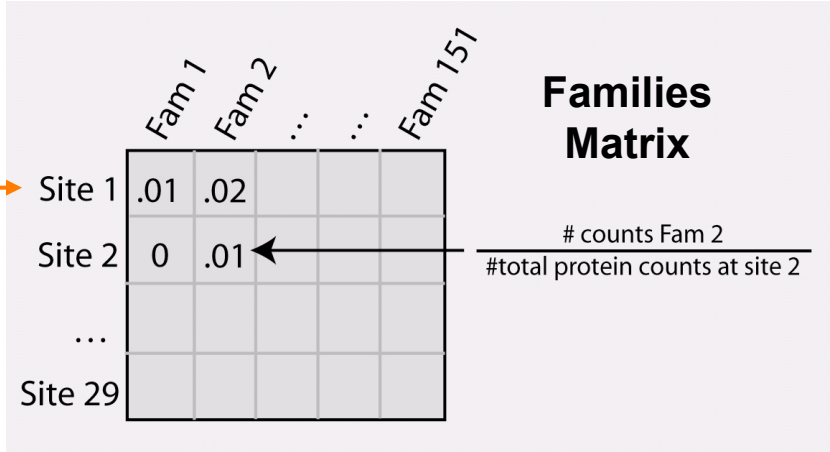


Climate Change

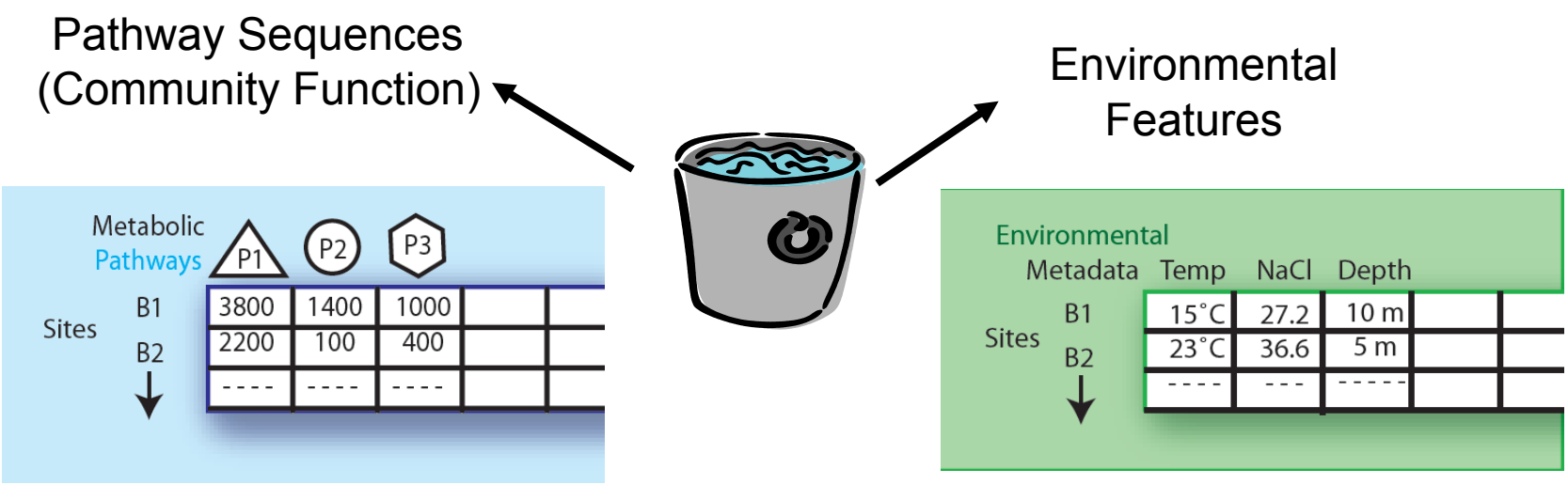
Halperin et al. *Science* 2008

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





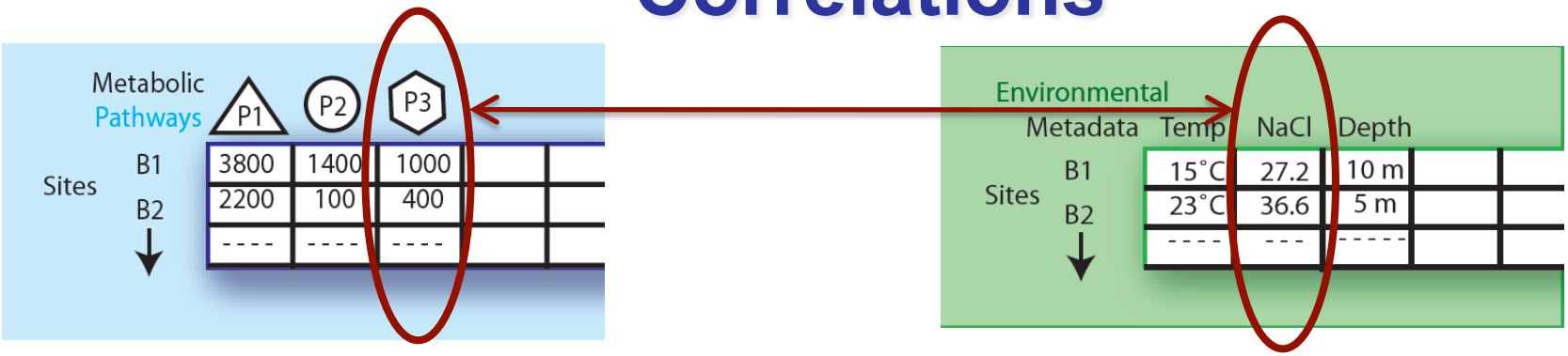
Patel et. al., *Genome Research* 2010

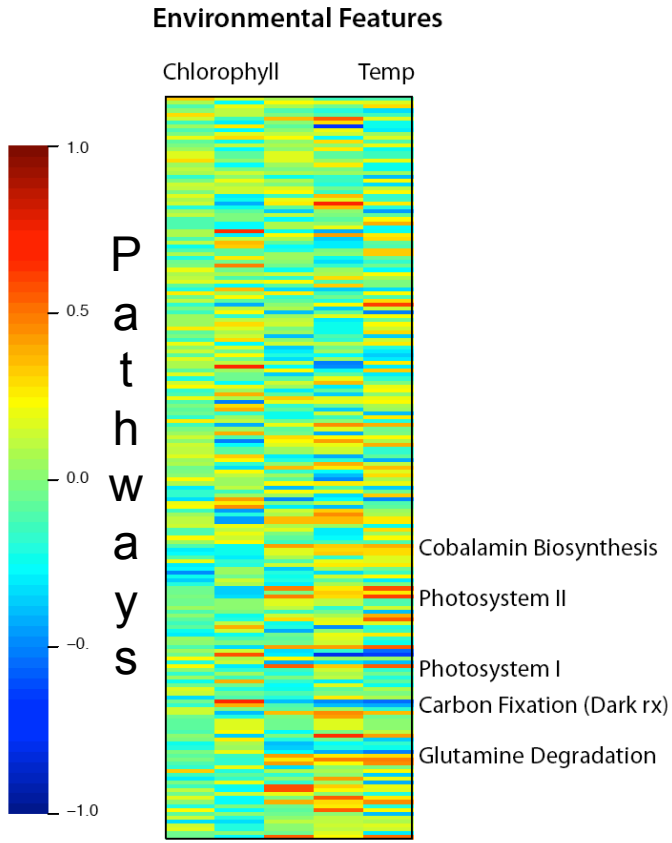


Expressing data as matrices indexed by site, env. var., and pathway usage

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

Simple Relationships: Pairwise Correlations

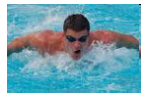
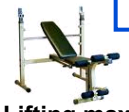




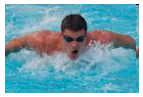
[Gianoulis et al., PNAS (in press, 2009)]

Canonical Correlation Analysis: Simultaneous weighting

Fitness Test	# km run/week
5K run	

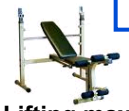
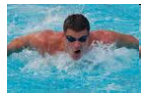
Fitness Text Index	Exercise Index
5K run	 
Heart Rate	 
 Lifting max	

FTI = a 5K run + b  Lifting max + c Heart Rate

EI = a  + b  + c  + d 

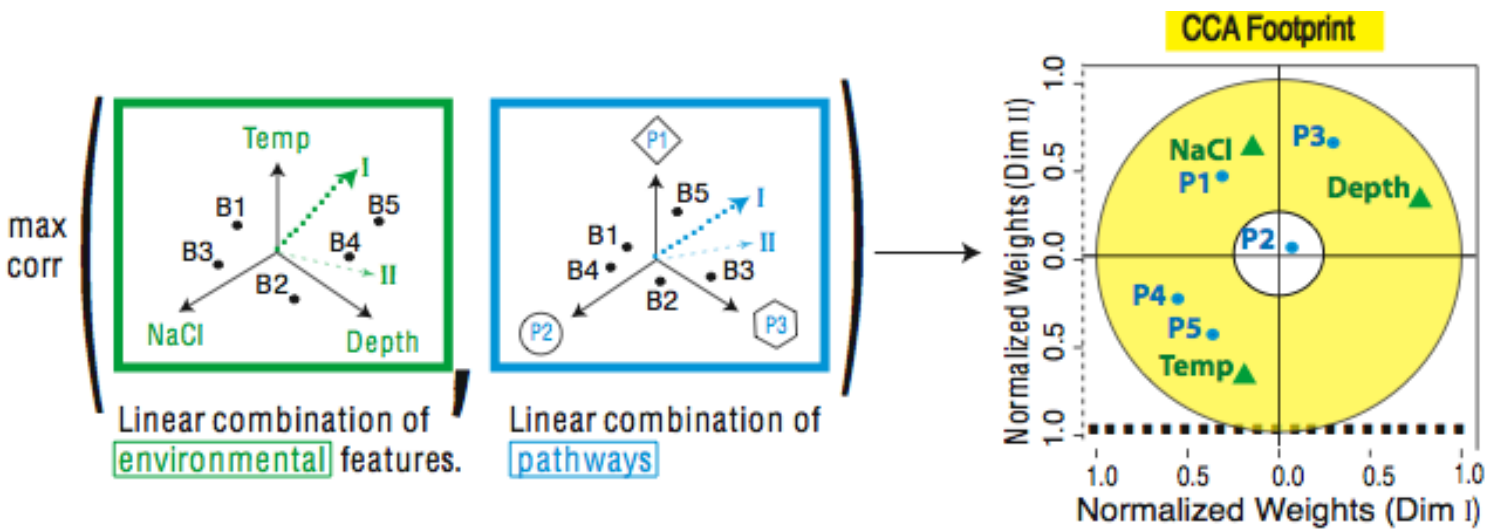
Canonical Correlation Analysis: Simultaneous weighting

Fitness Test	# km run/week
5K run	

Fitness Text Index	Exercise Index
5K run Heart Rate  Lifting max	   

Environmental Features	Metabolic Pathways/ Protein Families
Temp etc	Photosynthesis etc
Chlorophyll	Lipid Metabolism

Environmental-Metabolic Space



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

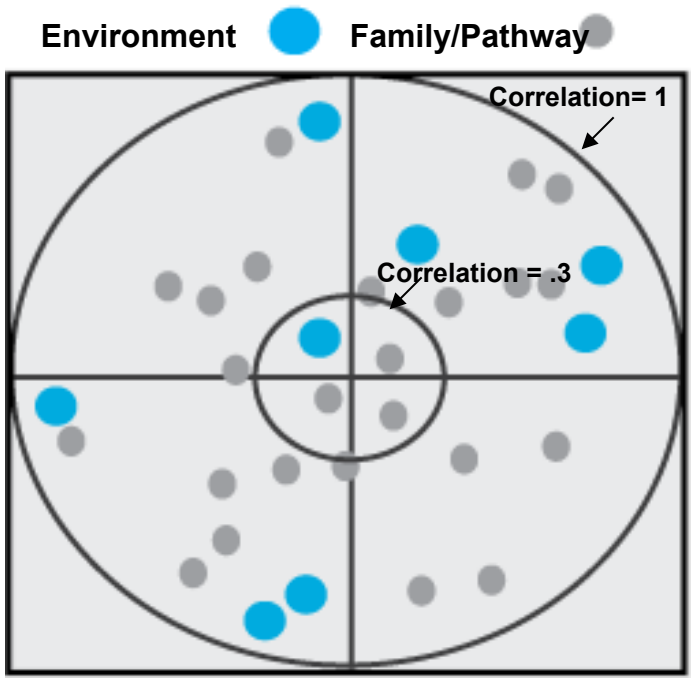
Given $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$

$$C = \begin{matrix} \Sigma_X & \Sigma_{X,Y} \\ \Sigma_Y & \Sigma_{Y,X} \end{matrix}$$

$$\max_{a,b} Corr(U,V) = \frac{a' \Sigma_{12} b}{\sqrt{a' \Sigma_{11} a} \sqrt{b' \Sigma_{22} b}}$$

Gianoulis et al., PNAS 2009

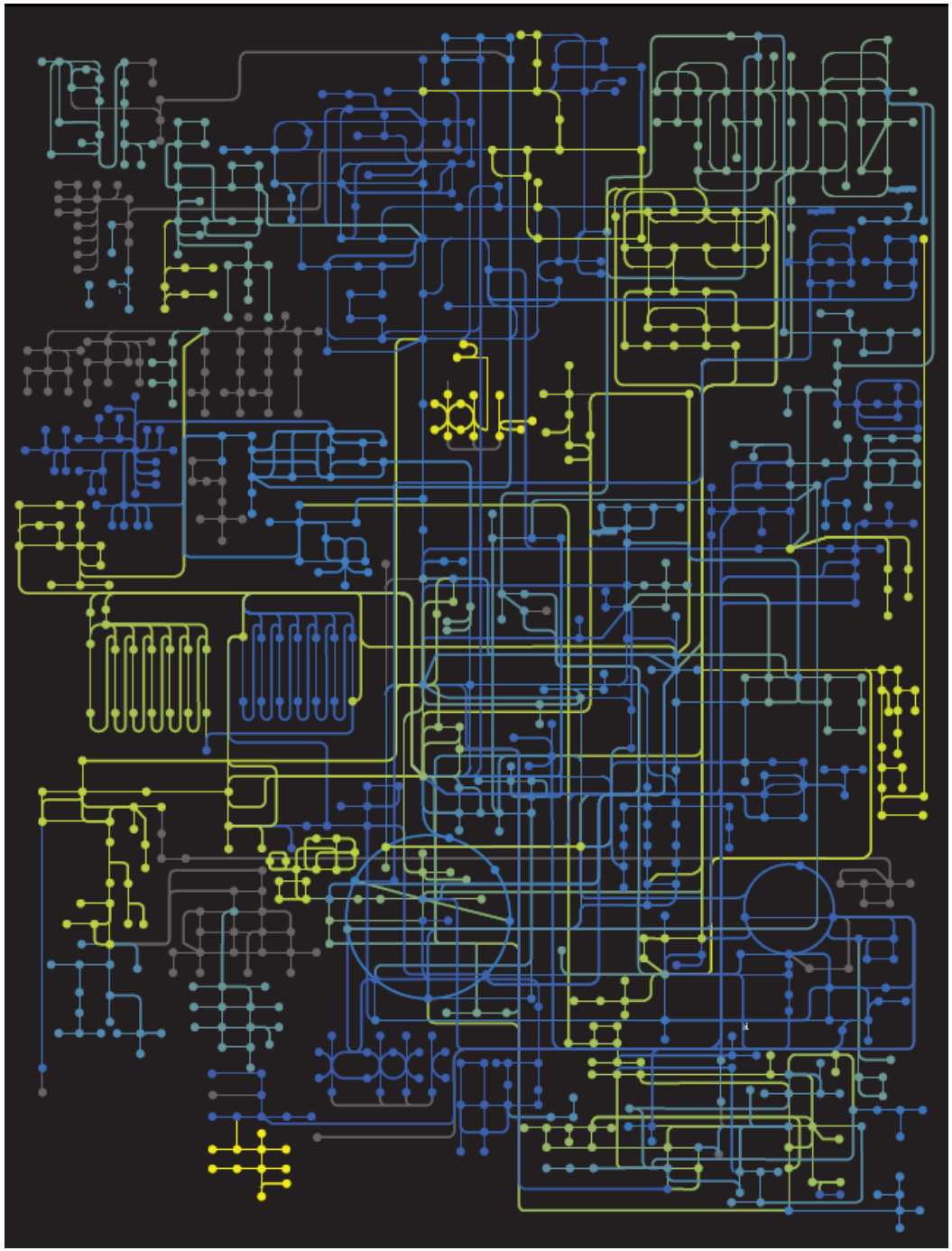
CCA Example



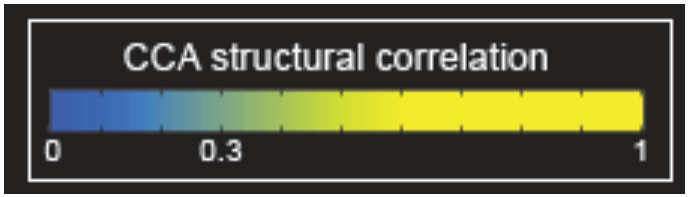
This plot shows the correlations in the first and second dimensions

Correlation Circle: The closer the point is to the outer circle, the higher the correlation

Variables projected in the same direction are correlated

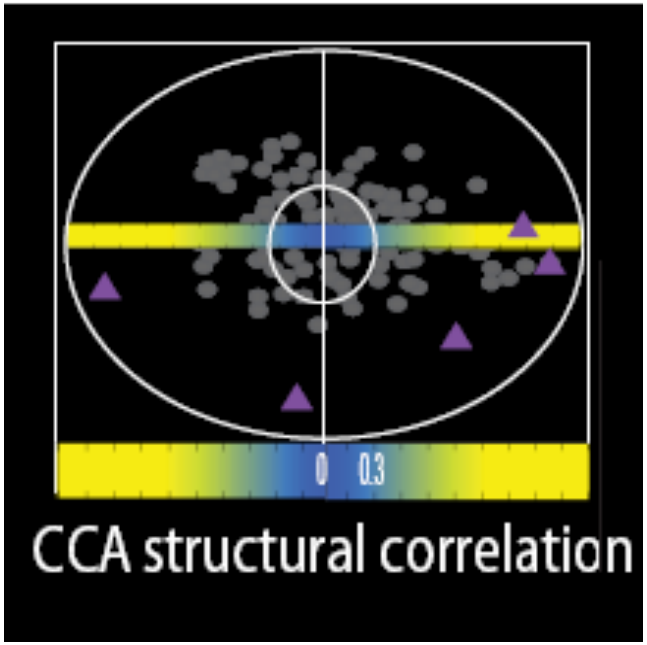


Strength of Pathway co-variation with environment



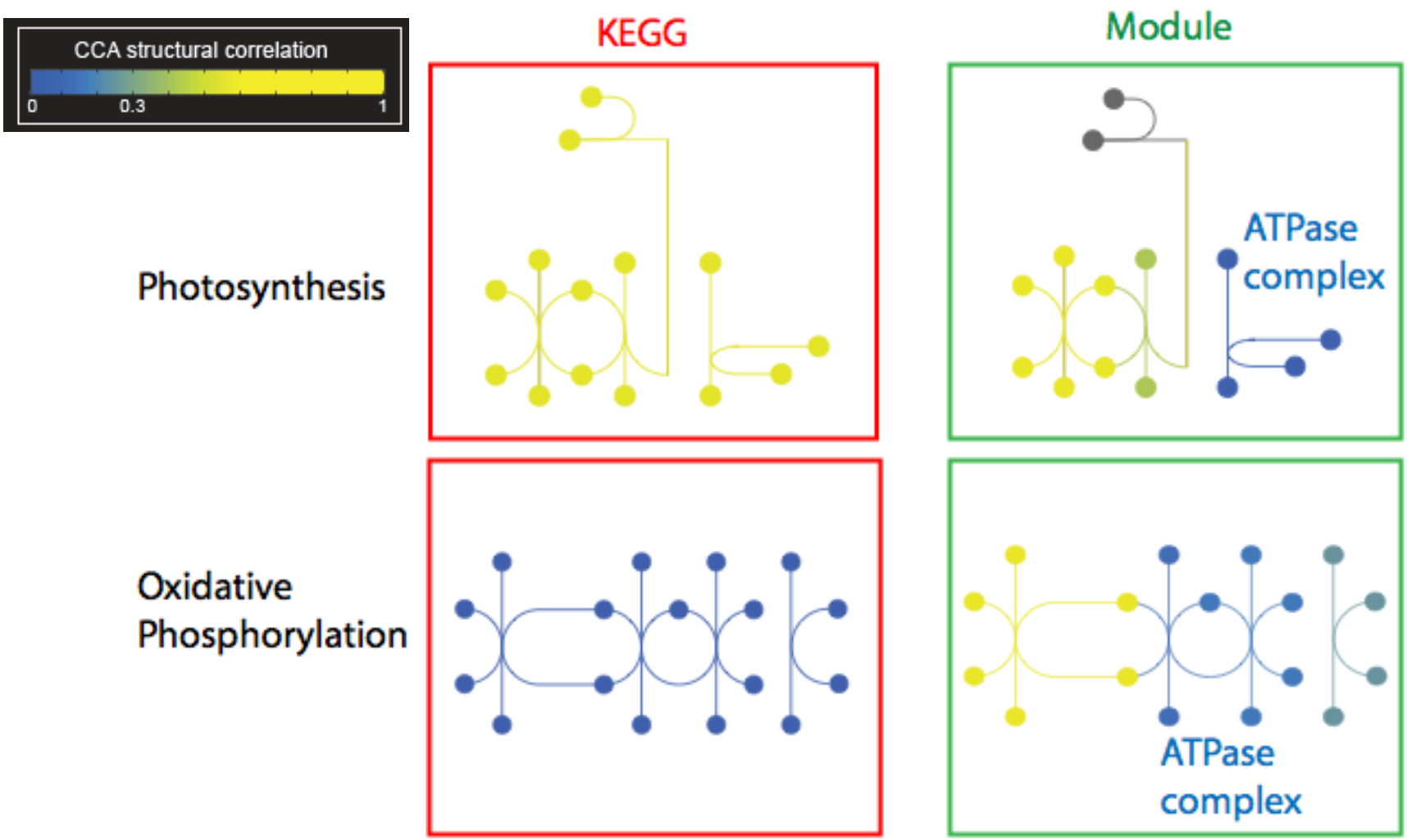
Environmentally invariant

Environmentally variant



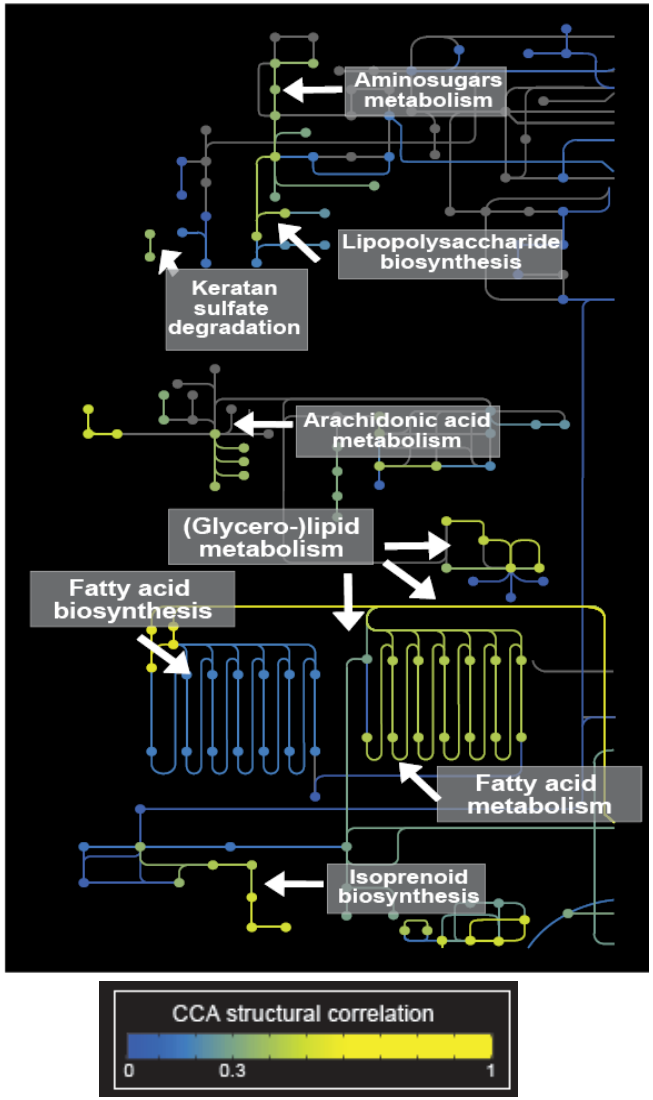
Gianoulis et al., *PNAS* 2009

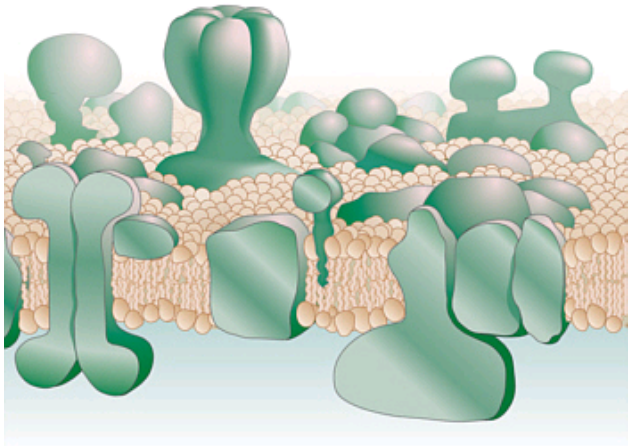
Conclusion #1: energy conversion strategy, temp and depth



Gianoulis et al., *PNAS* 2009

Conclusion #2: Outer Membrane components vary with the environment



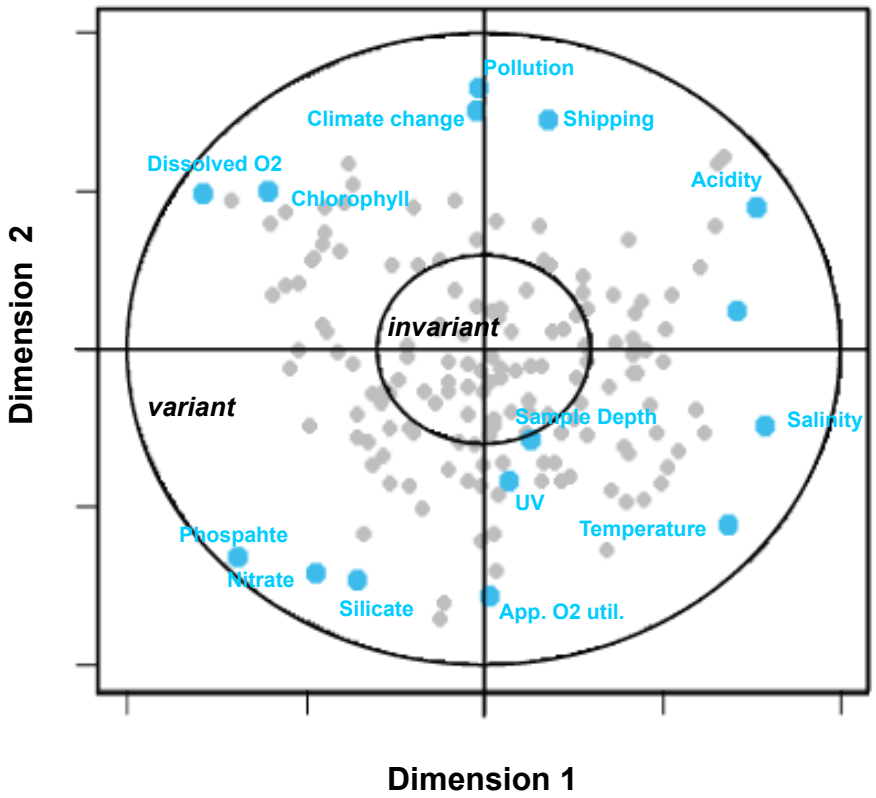


Membrane proteins interact with the environment, transporting available nutrients, sensing environmental signals, and responding to changes

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

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

CCA results for Membrane Proteins



107 variant membrane protein families

44 invariant membrane protein families

Difficult to see the strength and directionality of a relationship

Weights of the features are difficult to visualize and compare

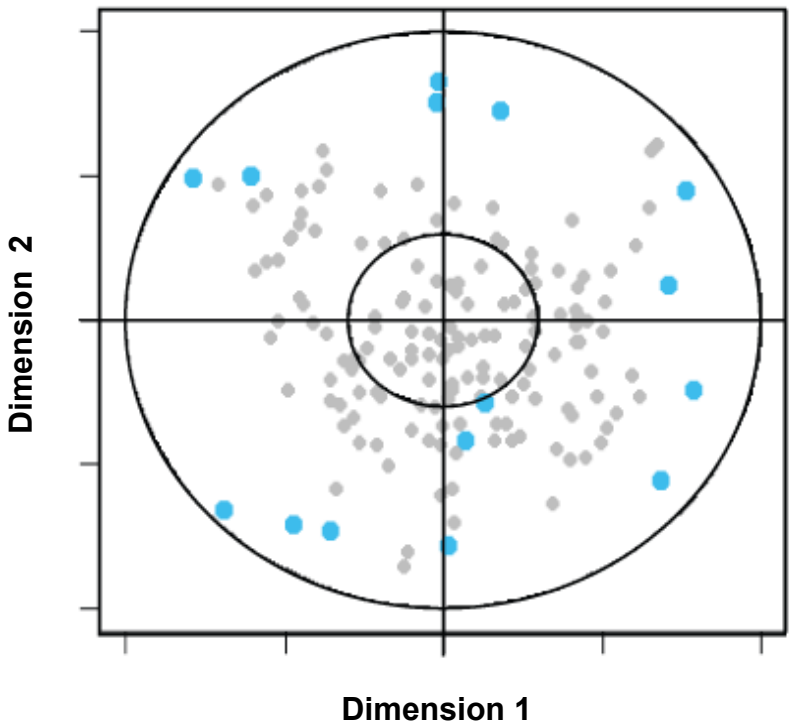
There is no means of quantifying the variation between sets of features

Patel et. al., *Genome Research* 2010

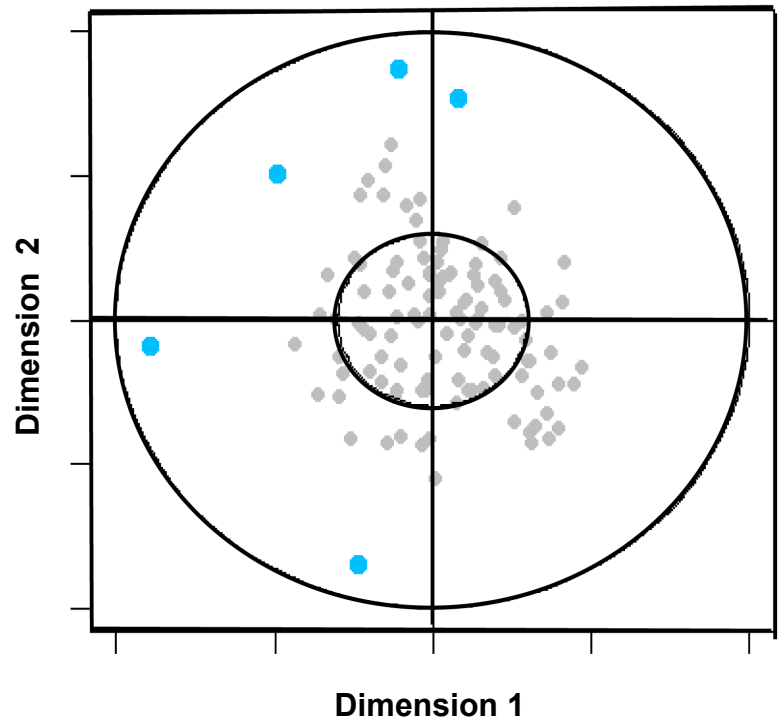
Membrane Proteins vary more than Metabolic Pathways

Median absolute structural Correlation Coefficient

Membrane Proteins = 0.3

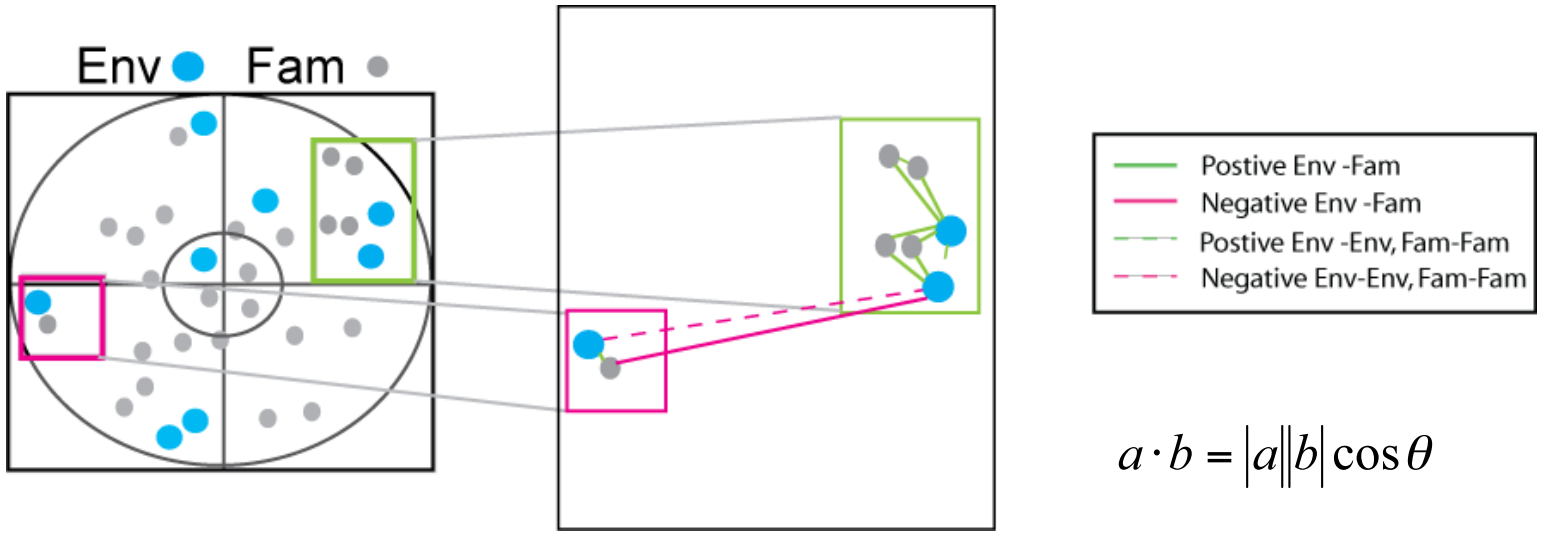


Metabolic Pathways = 0.17



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

Protein Families and Environmental Features Network (PEN)

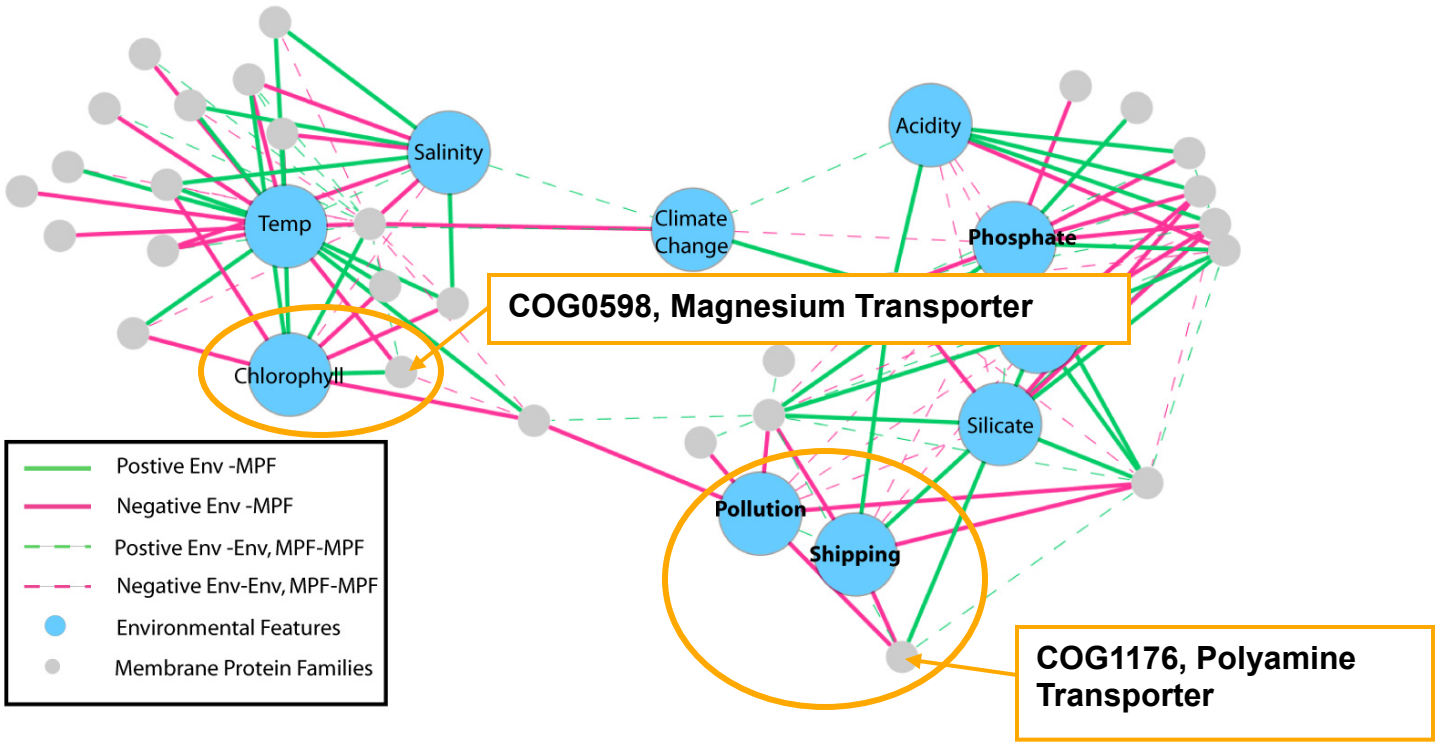


$$a \cdot b = |a||b|\cos\theta$$

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

Patel et. al., *Genome Research* 2010

Protein Families and Environmental Features Network (PEN)

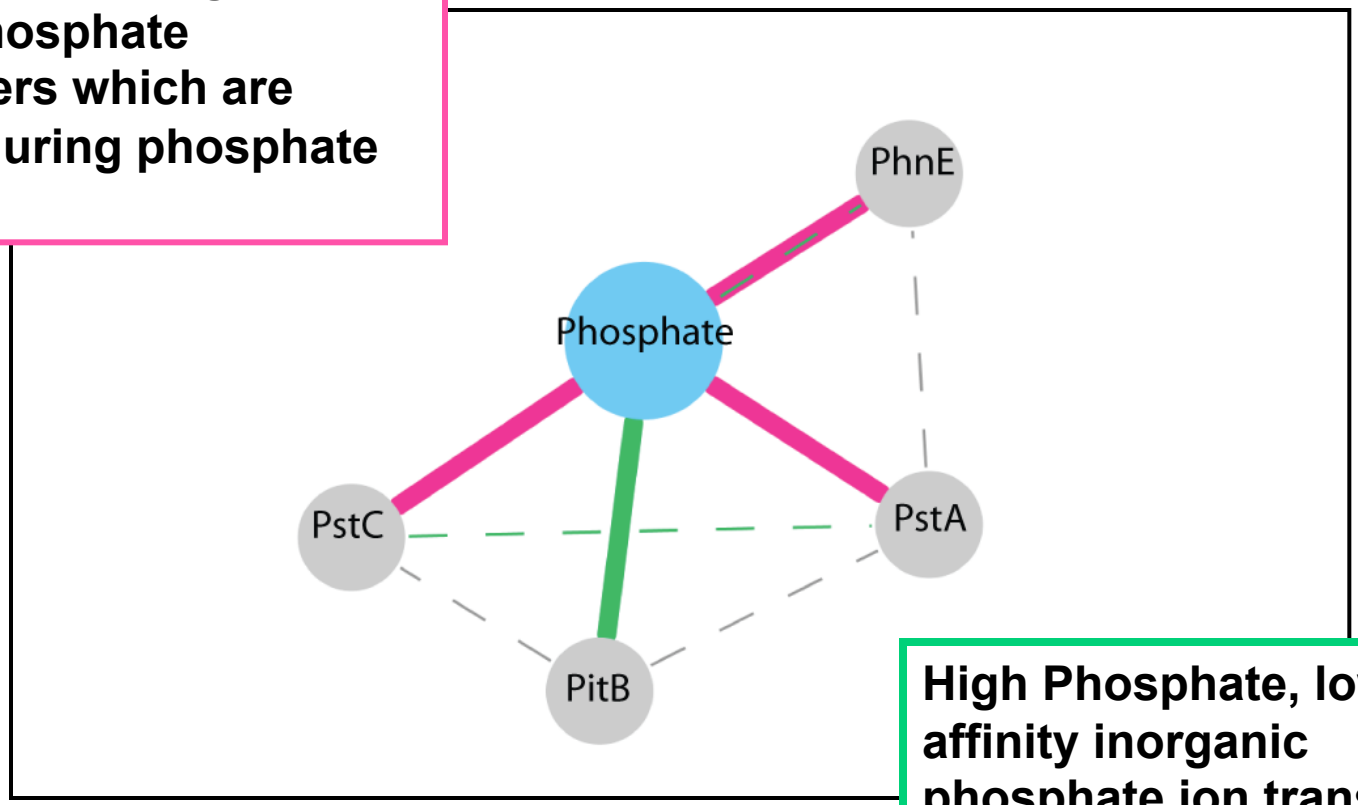


"Bi-modules": groups of environmental features and membrane proteins families that are associated

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

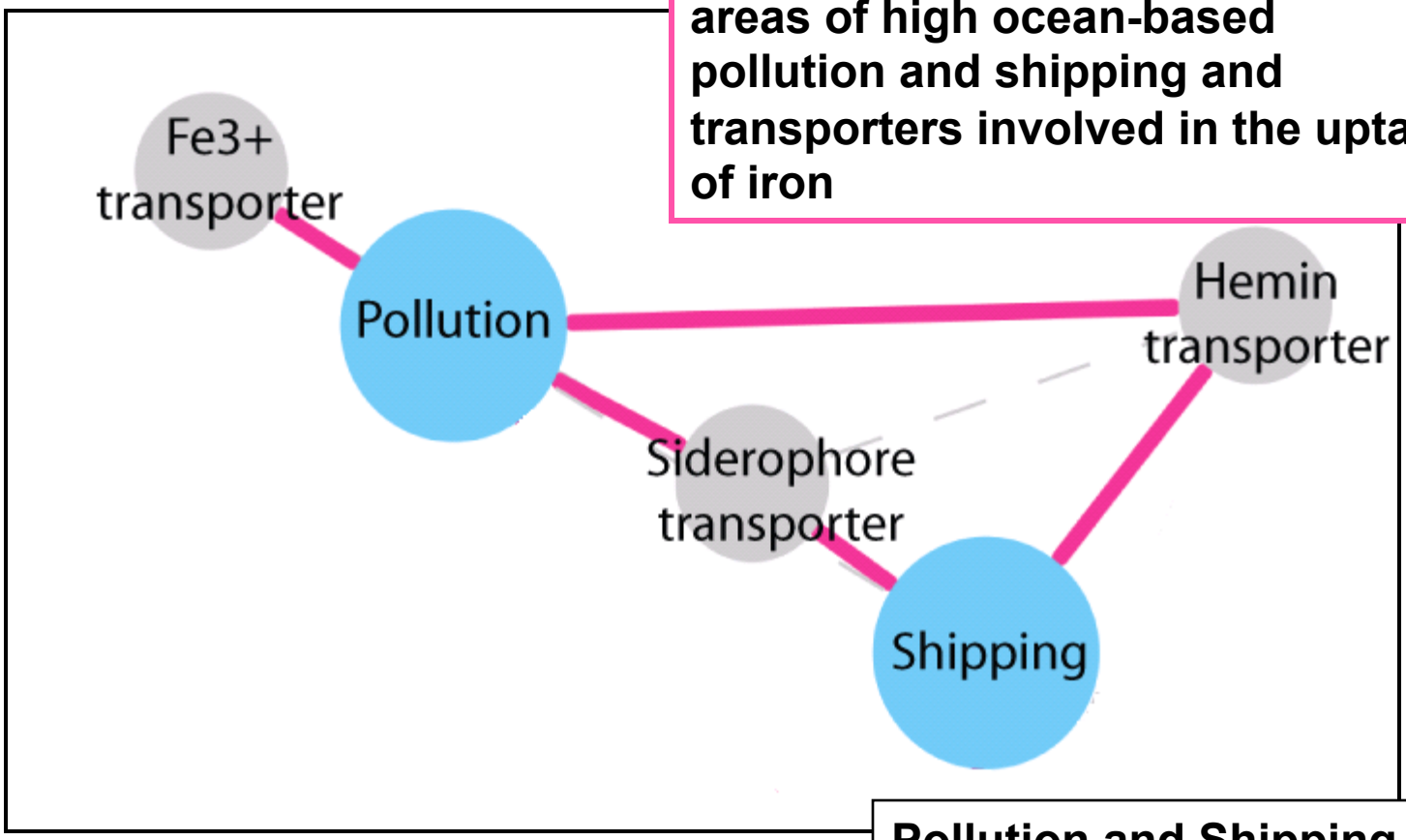


High Phosphate, low affinity inorganic phosphate ion transporter which are constitutively expressed

Patel et. al., *Genome Research* 2010

Bi-module 2: Iron Transporters/Pollution/Shipping

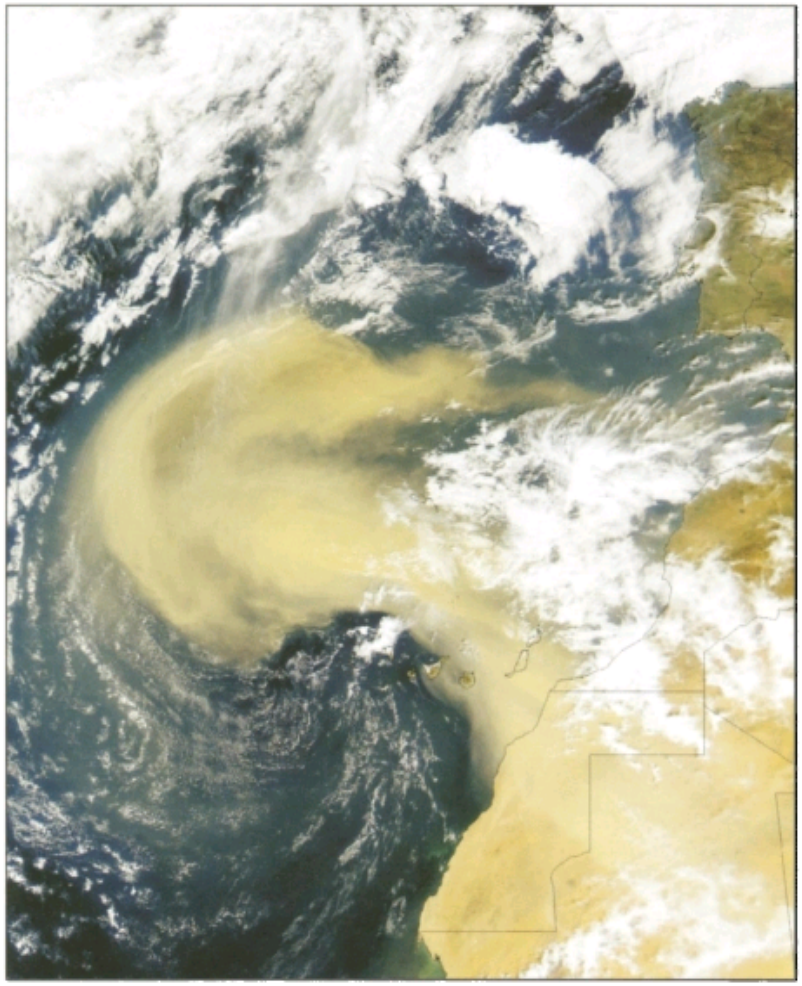
Negative relationship between areas of high ocean-based pollution and shipping and transporters involved in the uptake of iron



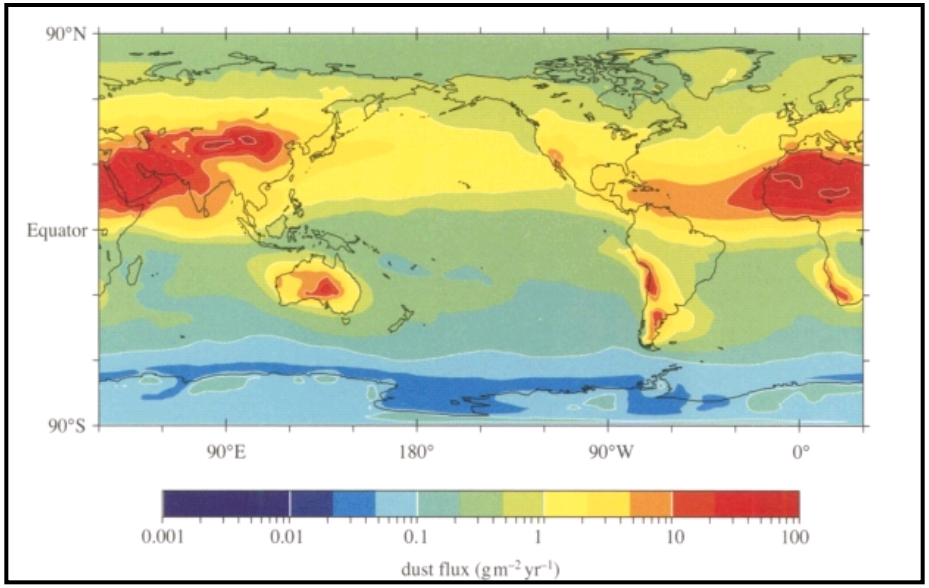
Pollution and Shipping may be a proxy for iron concentrations →

Patel et. al., *Genome Research* 2010

Bi-module 2: Iron Transporters/Pollution/Shipping



Rigwell A. J. (2002) *Phil. Trans. R. Soc. Lond.*

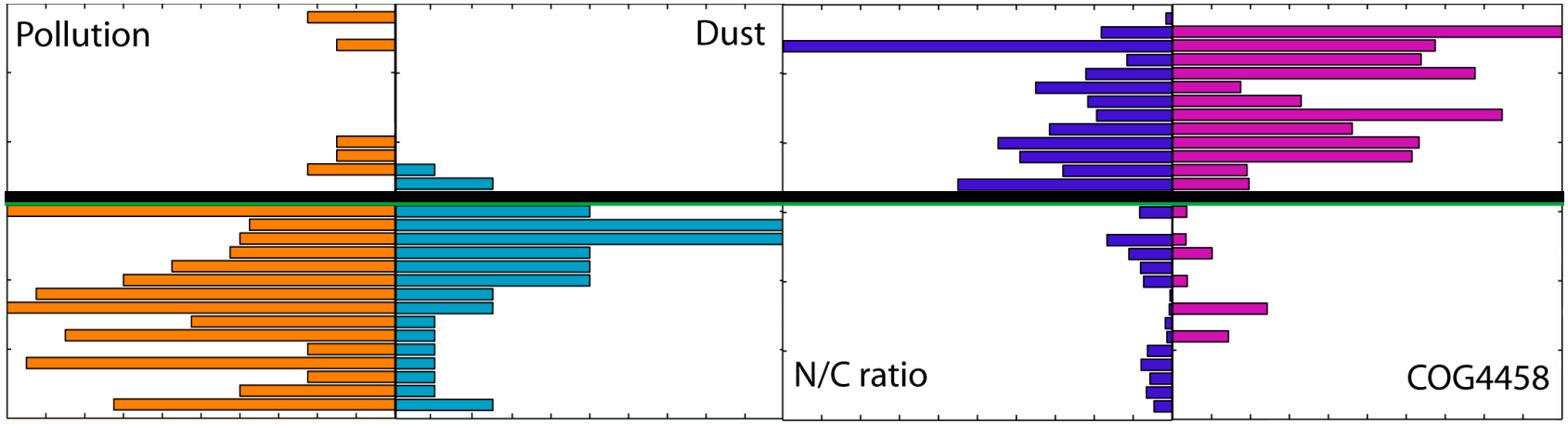


Iron is usually limiting in oceans: High Nitrate-Nutrient/Low Chlorophyll regions
Delivery of iron to is usually by:

- **terrestrial input**
- **fluvial (rivers) input**
- **upwelling from the ocean floor**
- **aeolian dust from land**

Patel et. al., *Genome Research* 2010

Bi-module 2: Iron Transporters/Pollution/Shipping



Pollution and Dust ↑

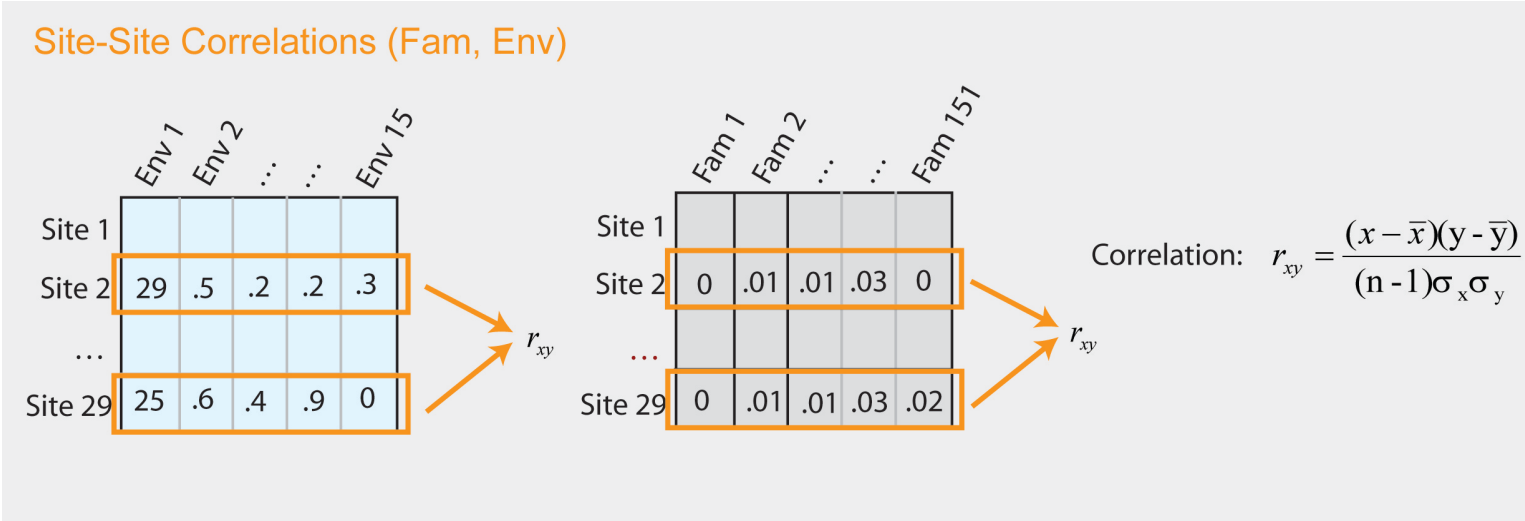
N/C and Iron Transporters ↓

-Negative correlation between COG4558 and COG0609 and dust/pollution values (p-value <0.01)

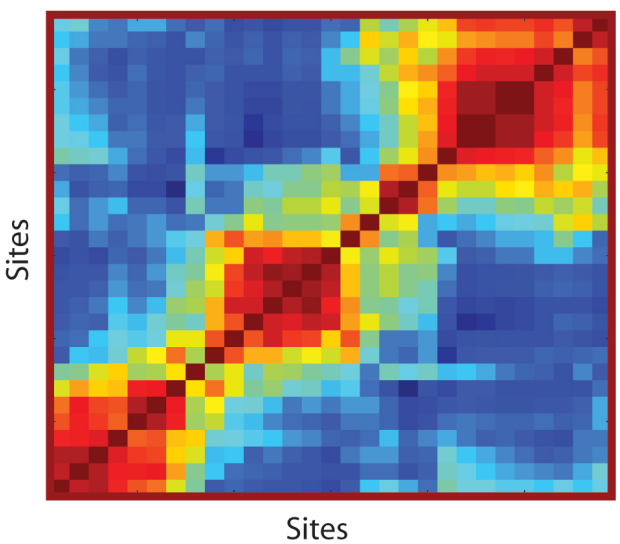
- Searching the BRENDA database for enzymes using iron as a cofactor reveal that an increase in these two COGs negatively correlated to the amount of enzymes present that required iron.

Patel et. al., *Genome Research* 2010

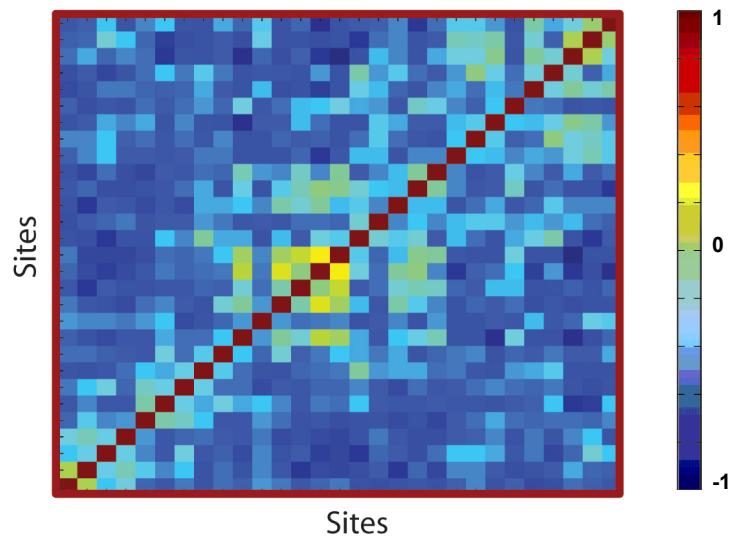
How Similar are the Sites to each other?



Environmental Features Site Correlations



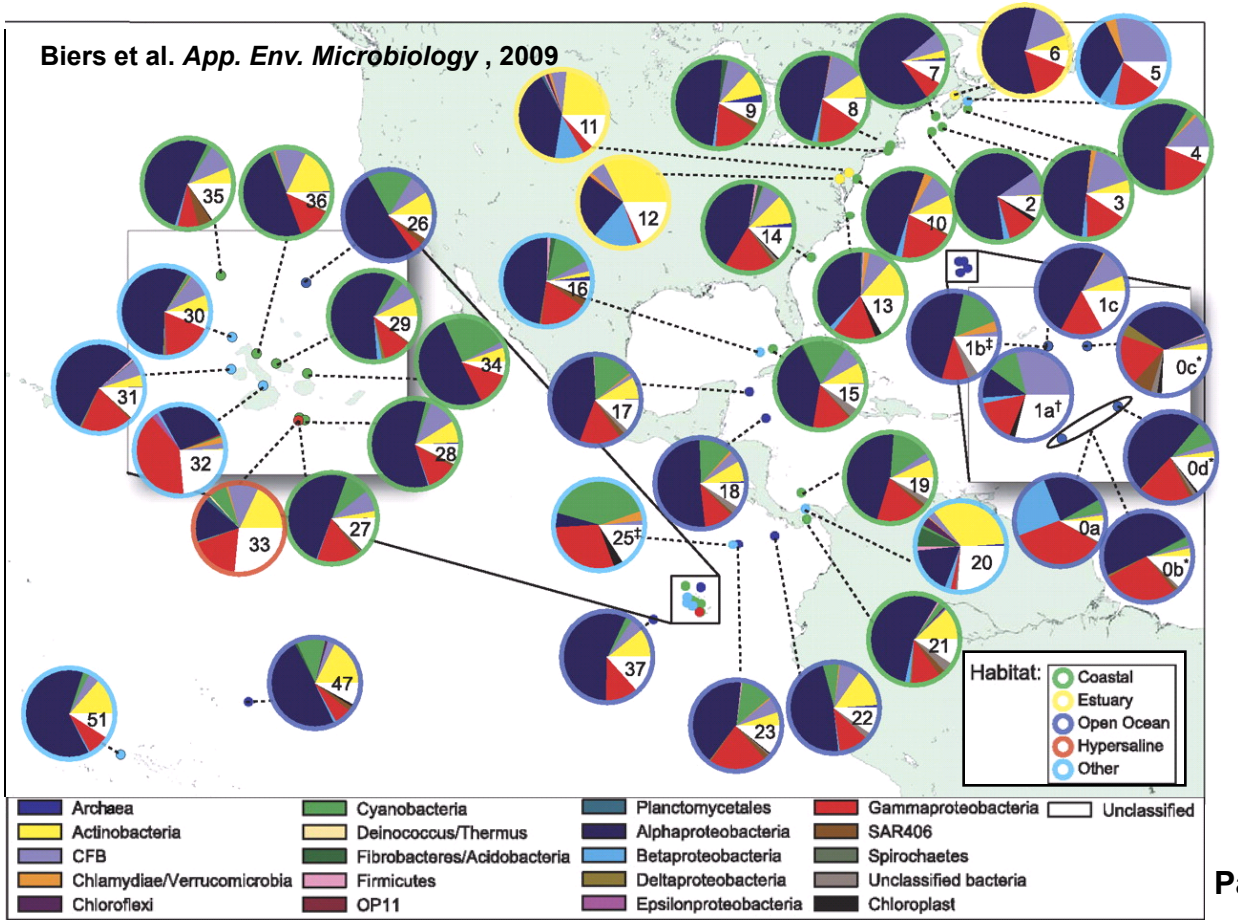
Membrane Protein Families Site Correlations

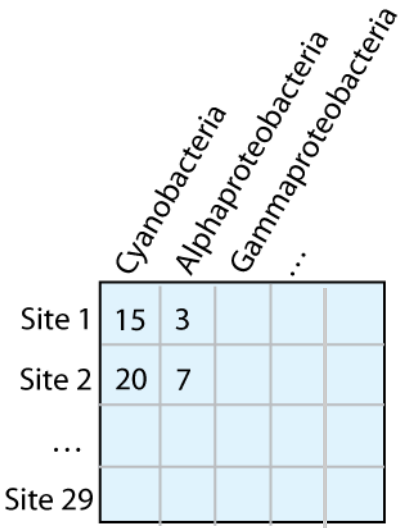


Patel et. al., *Genome Research* 2010

Species Distribution

- The 16S rRNA gene is a component of the small prokaryotic ribosomal subunit
- Bacteria with 16S rRNA gene sequences more similar than 97% are considered the same 'species'
- 10,025 16S genes found and classified

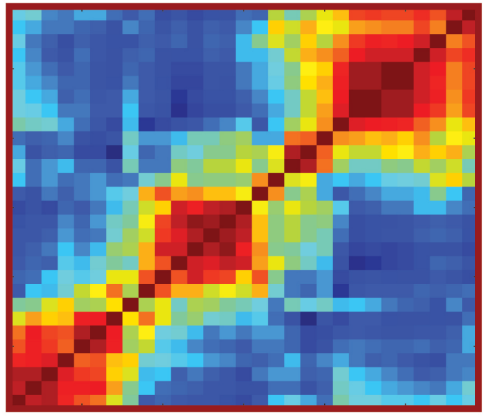




20% level, "phylum"

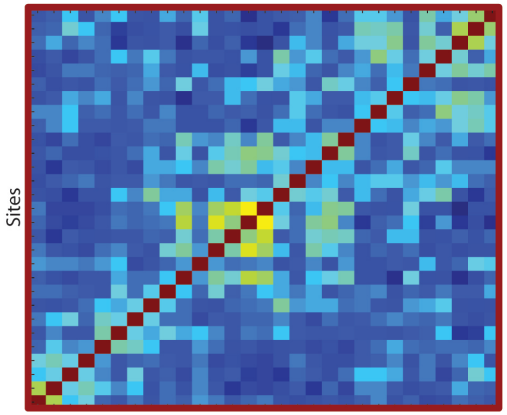
Patel et. al., *Genome Research* 2010

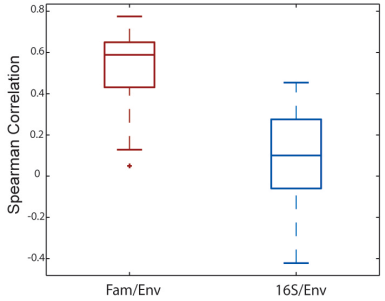
Environmental Features Site Correlations



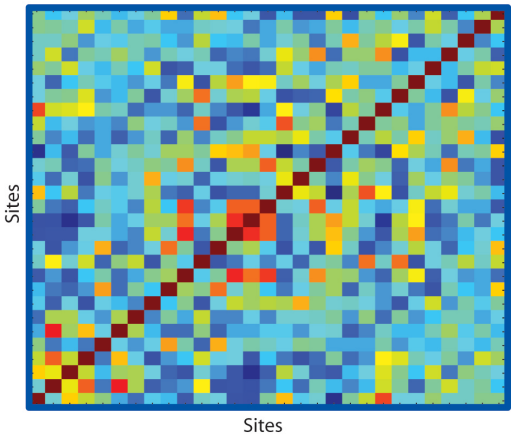
This suggests that the observed membrane protein variation is more a function of the measured environmental features, than phylogenetic diversity.

Membrane Protein Families Site Correlations





16S Site Correlations



Method: For each site, we correlated the EF profile distances and its MPF frequency profile distances and 16S profile distances

Environmental Features Site Correlations

	Site 1	Site 2	Site 3	Site 4	…					
Site 1	1	.8	.5	.7						
Site 2	.8	1	.3	.4						
Site 3	.5	.3	1	.7						
…	.7	.4	.7	1						

Membrane Protein Families Site Correlations

	Site 1	Site 2	Site 3	Site 4	…					
Site 1	1	.2	.6	.3						
Site 2	.2	1	.9	.8						
Site 3	.6	.9	1	.6						
…	.3	.8	.6	1						

$$r_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{(n-1)\sigma_x \sigma_y}$$

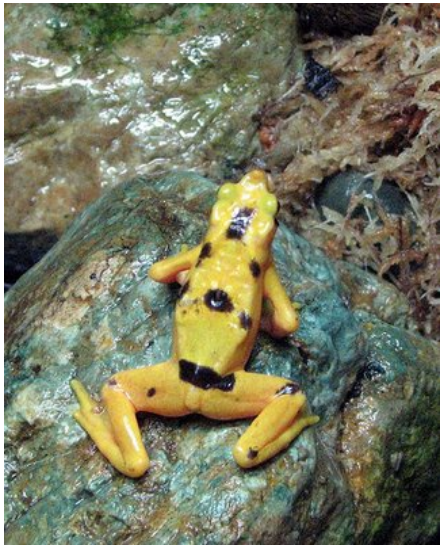
Spearman rank correlation:

1	2	4	3	…					

1	4	2	3	…					

Patel et. al., *Genome Research* 2010

Biosensors:
Beyond Canaries in a Coal Mine



Outline: Molecular Networks

- Why Networks?
- Backround:
  Central Network Points
- Networks & Variation
  (human ppi)
- Social Network Comparisons
  (reg. net. in many organisms)
  - in rel. to social hierarchy
  - scaling in rel. to partnerships
- Computer OS Comparisons
  (E. coli reg. net)
- Network Dynamics Across Environments
  (prokaryote metab. pathways)
  - Metabolic Pathways
  - Entry pts. (Mem. Proteins)

Conclusions: Networks & Variation

- Positive selection (adaptive evolution) at the network periphery
  - On a sequence level, it can be seen as positive selection of peripheral nodes
  - On a structural level, it can be seen as the pattern of SDs that display significantly higher allele frequencies in non-central genes

Conclusions: Comparison to Social and Regulatory Hierarchies

- Regulatory Network 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)
- Number of collaborative partners saturates even while scale of targets governed increases
  - Also seen in social networks

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

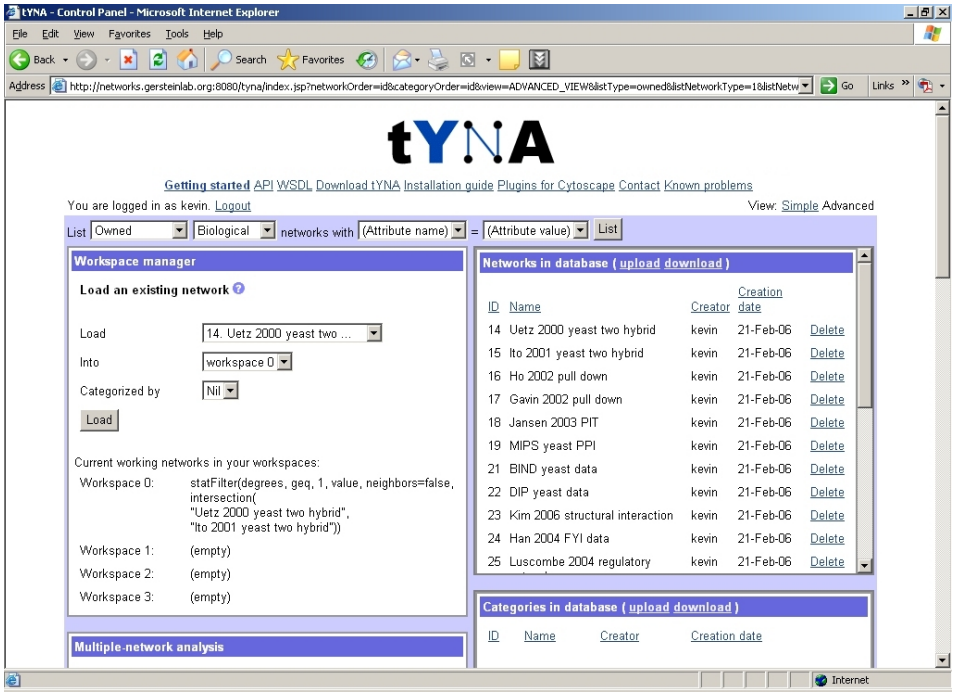
Conclusions: Network Dynamics Across Environments

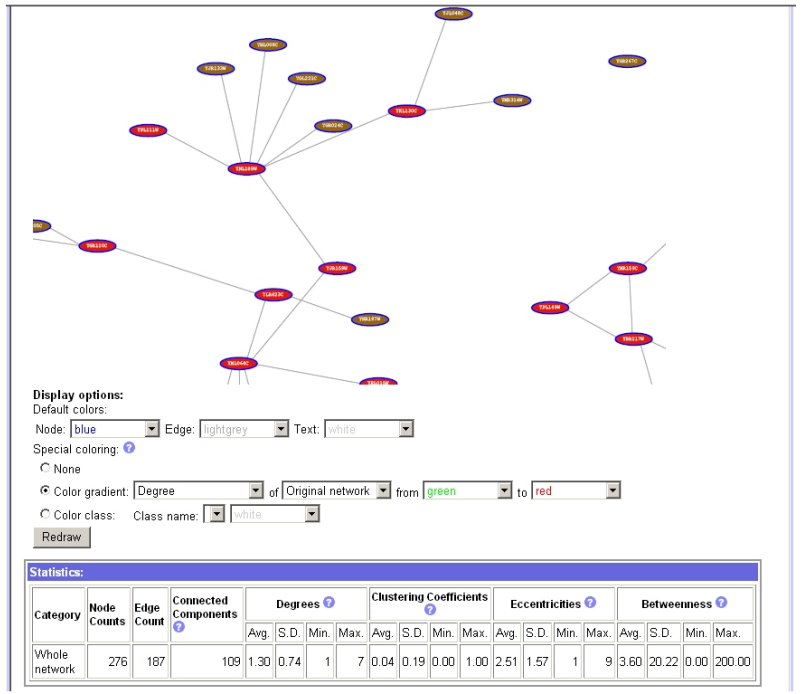
- Developed approach to connect quantitative features of environment to usage of pathways & families
  - CCA + PEN
- Applied to available aquatic datasets, identified footprints predictive of environment (potentially useful as biosensor)
- Integration of geospatial data can highlight unexpected trends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

- an automated web tool

(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

P Kim
N Bhardwaj
K-K Yan
P Patel
T Gianoulis
H Yu

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

Networks.GersteinLab.org

Job opportunities currently for postdocs & students

More Information on this Talk

SUBJECT: Networks

DESCRIPTION:
Institut de recherches cliniques de Montreal (IRCM), Montreal, Quebec; 2010.05.03, 11:30-12:30; [**I:IRCM**] (Long networks talk, derived from [I:**BROWNMATH**], including **metamembrane*** for 1st time. Takes 55' without callgraph sect.)

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

.

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