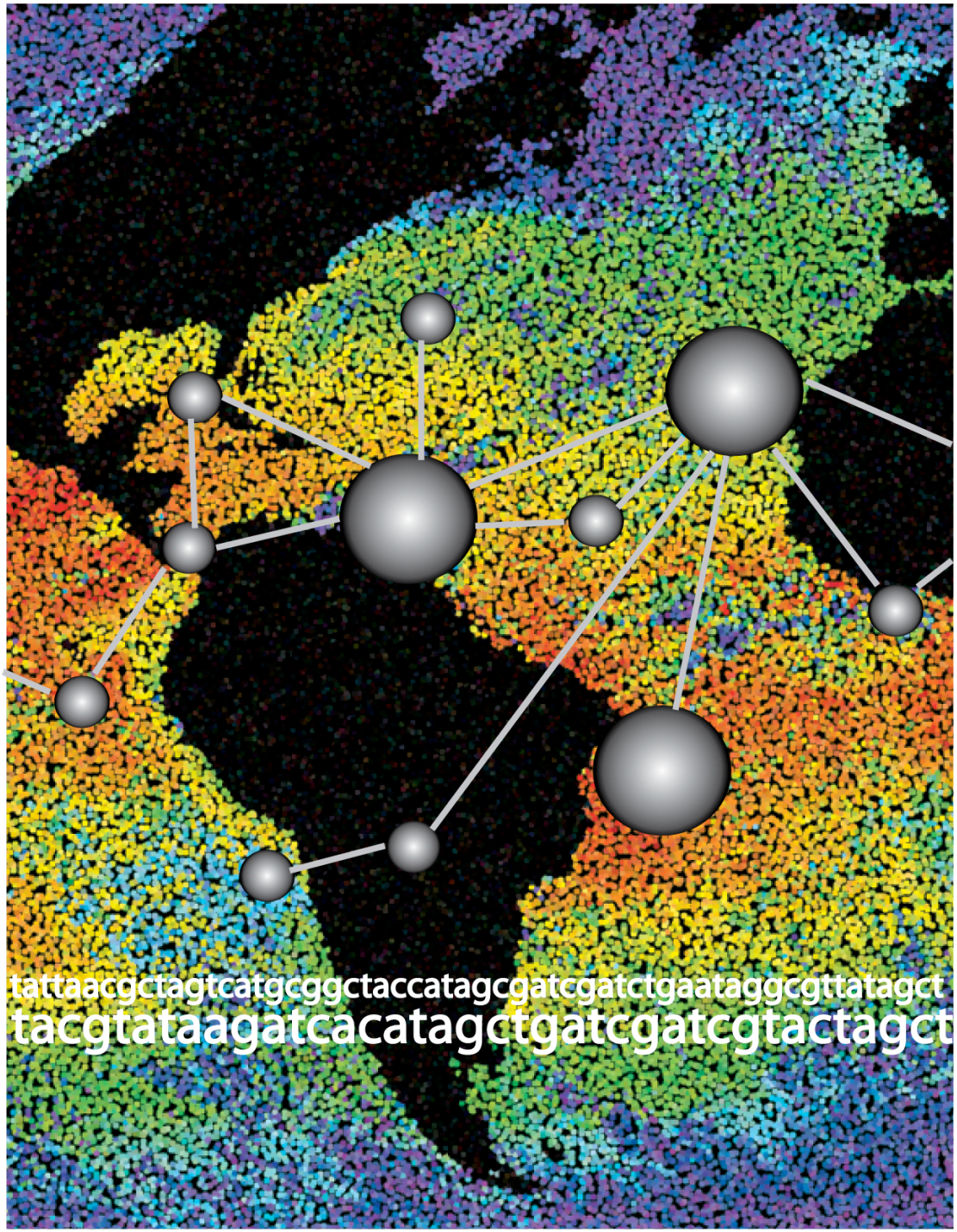


Biological Network Analysis

Mark B Gerstein

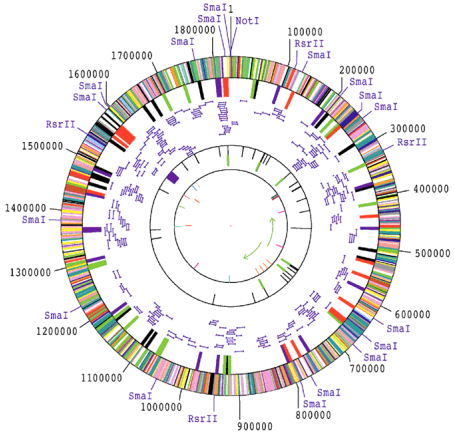
Yale

slides at

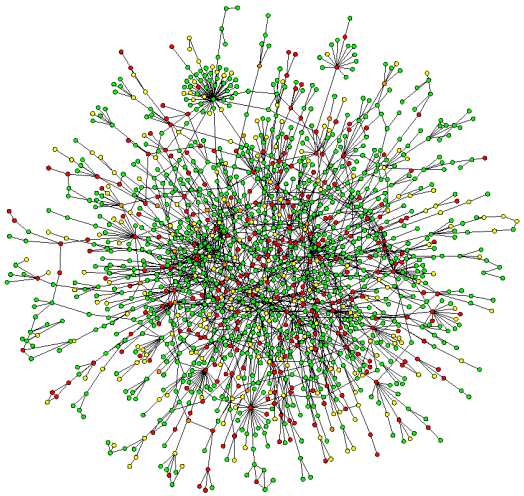
Lectures.GersteinLab.org

(See Last Slide for References & More Info.)

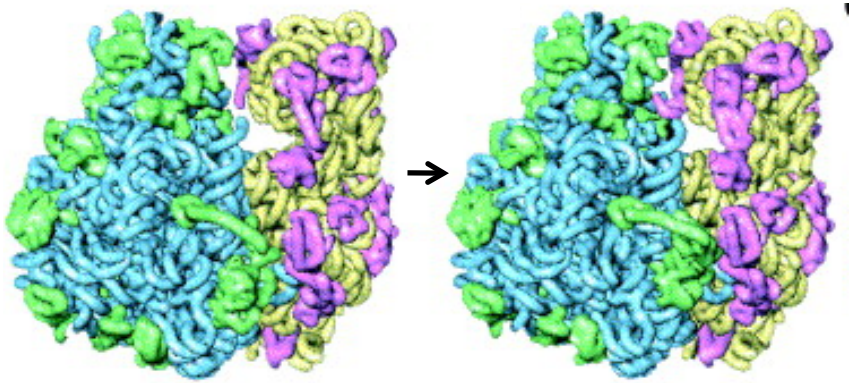
Networks occupy a midway point in terms of level of understanding



1D: Complete Genetic Partslist



~2D: Bio-molecular Network Wiring Diagram



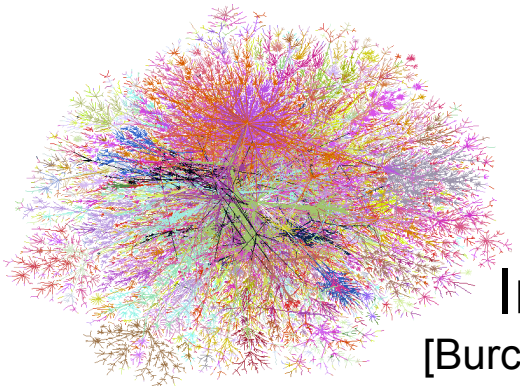
3D and 4D: Detailed structural understanding of cellular machinery (e.g. ribosome in different functional states)

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

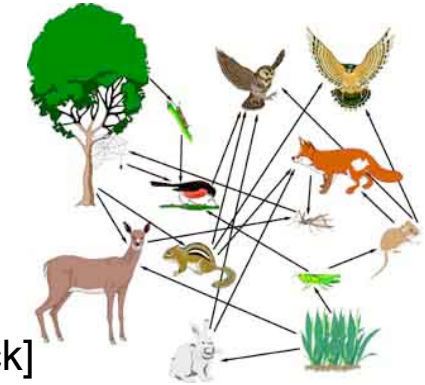
[Jeong et al. Nature, 41:411]

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

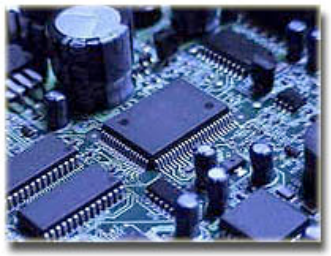
Networks as a universal language



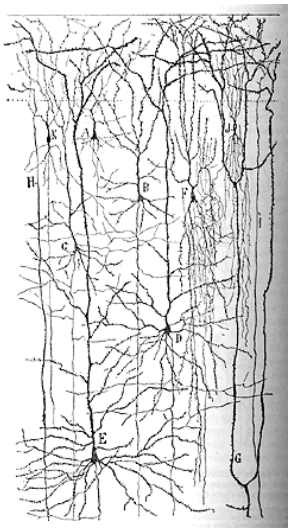
Internet
[Burch & Cheswick]



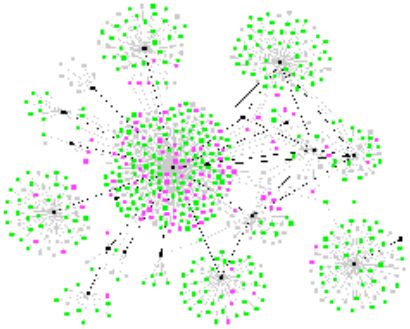
Food Web



Electronic Circuit

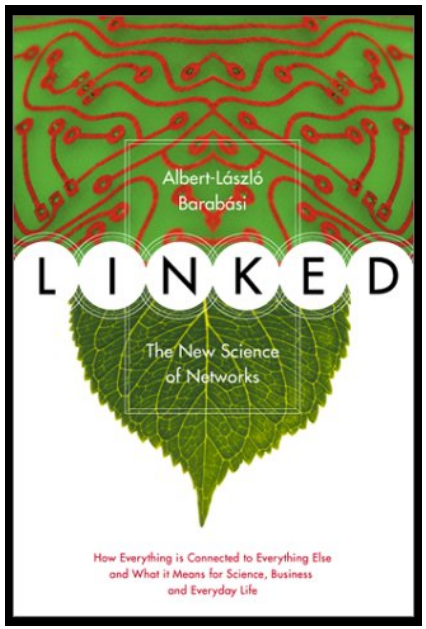


Neural Network
[Cajal]



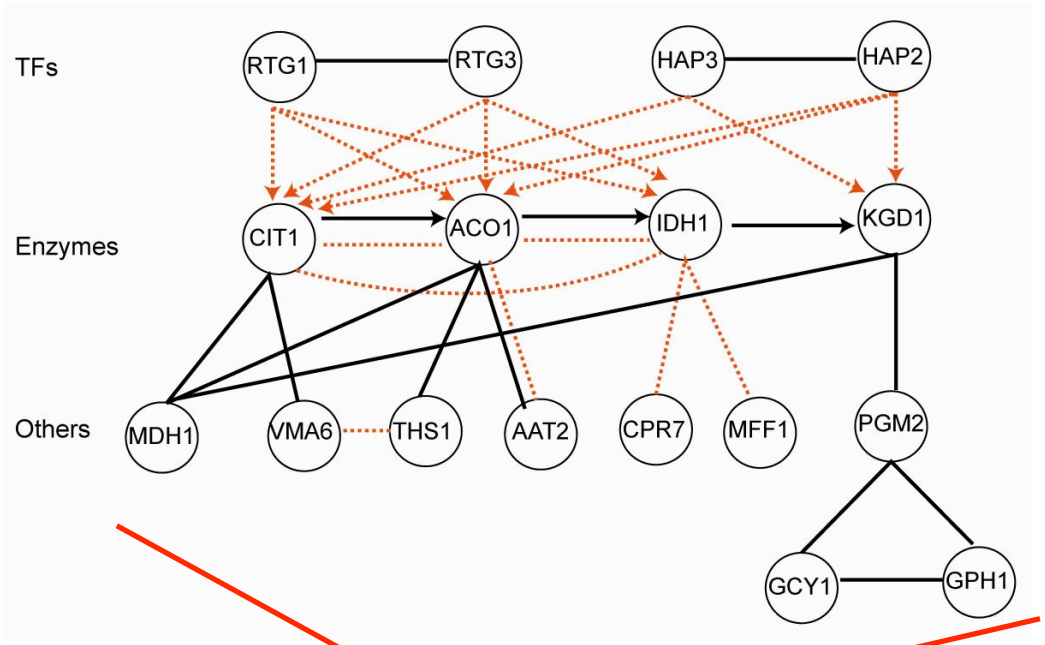
Disease Spread
[Krebs]

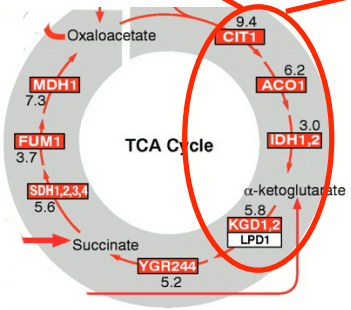
Protein Interactions
[Barabasi]



Social Network

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

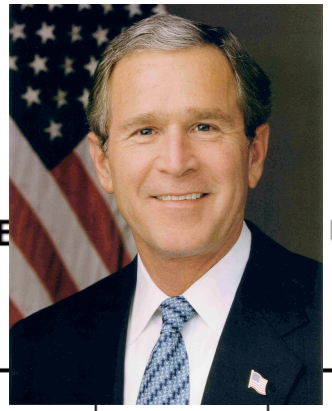
Genetic interaction (synthetic lethal)
Signaling pathways

Outline: Molecular Networks

- Why Networks?
- Network Comparisons
  (reg. net. in many organisms)
  - in rel. to social hierarchy
  - scaling in rel. to partnerships
  - Computer OS Comparisons
- 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
OLICY DEVELOPMENT
IENCE AND TECHNOLOGY POLICY
E U.S. TRADE REPRESENTATIVE

JUDICIAL BRANCH

THE SUPREME COURT OF THE UNITED STATES

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















INDEPENDENT ESTABLISHMENTS AND GOVERNMENT CORPORATIONS

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

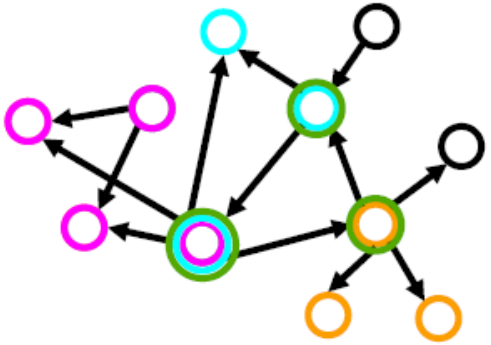
FEDERAL LABOR RELATIONS AUTHORITY
FEDERAL MARITIME COMMISSION
FEDERAL MEDIATION AND CONCILIATION SERVICE
FEDERAL MINE SAFETY AND HEALTH REVIEW COMMISSION
FEDERAL RESERVE SYSTEM
FEDERAL RETIREMENT THRIFT INVESTMENT BOARD
FEDERAL TRADE COMMISSION
GENERAL SERVICES ADMINISTRATION
INTER-AMERICAN FOUNDATION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

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

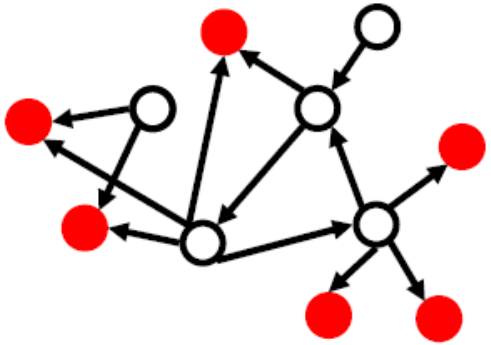
POSTAL RATE COMMISSION
RAILROAD RETIREMENT BOARD
SECURITIES AND EXCHANGE COMMISSION
SELECTIVE SERVICE SYSTEM
SMALL BUSINESS ADMINISTRATION
SOCIAL SECURITY ADMINISTRATION
TENNESSEE VALLEY AUTHORITY
TRADE AND DEVELOPMENT AGENCY
U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT
U.S. COMMISSION ON CIVIL RIGHTS
U.S.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

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

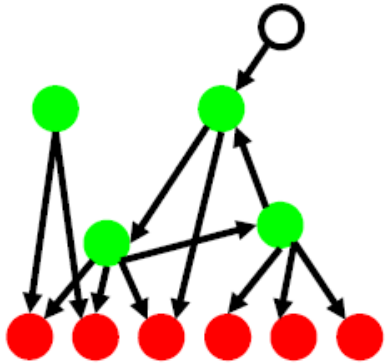
I. Example network with all 4 motifs



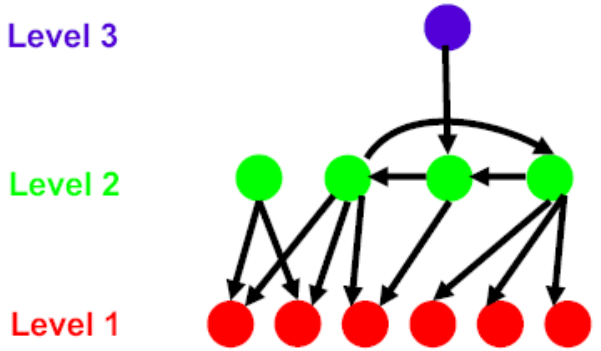
II. Finding terminal nodes (Red)



III. Finding mid-level nodes (Green)

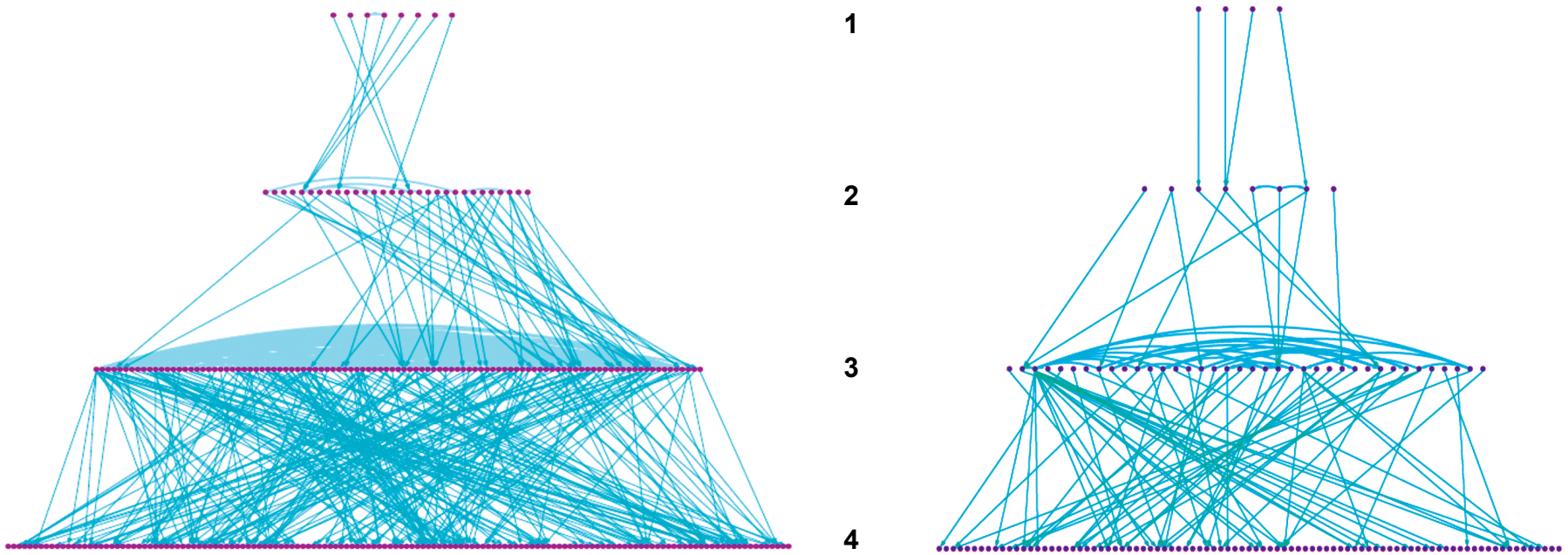


IV. Finding top-most nodes (Blue)



[Yu et al., PNAS (2006)]

Regulatory Networks have similar hierarchical structures

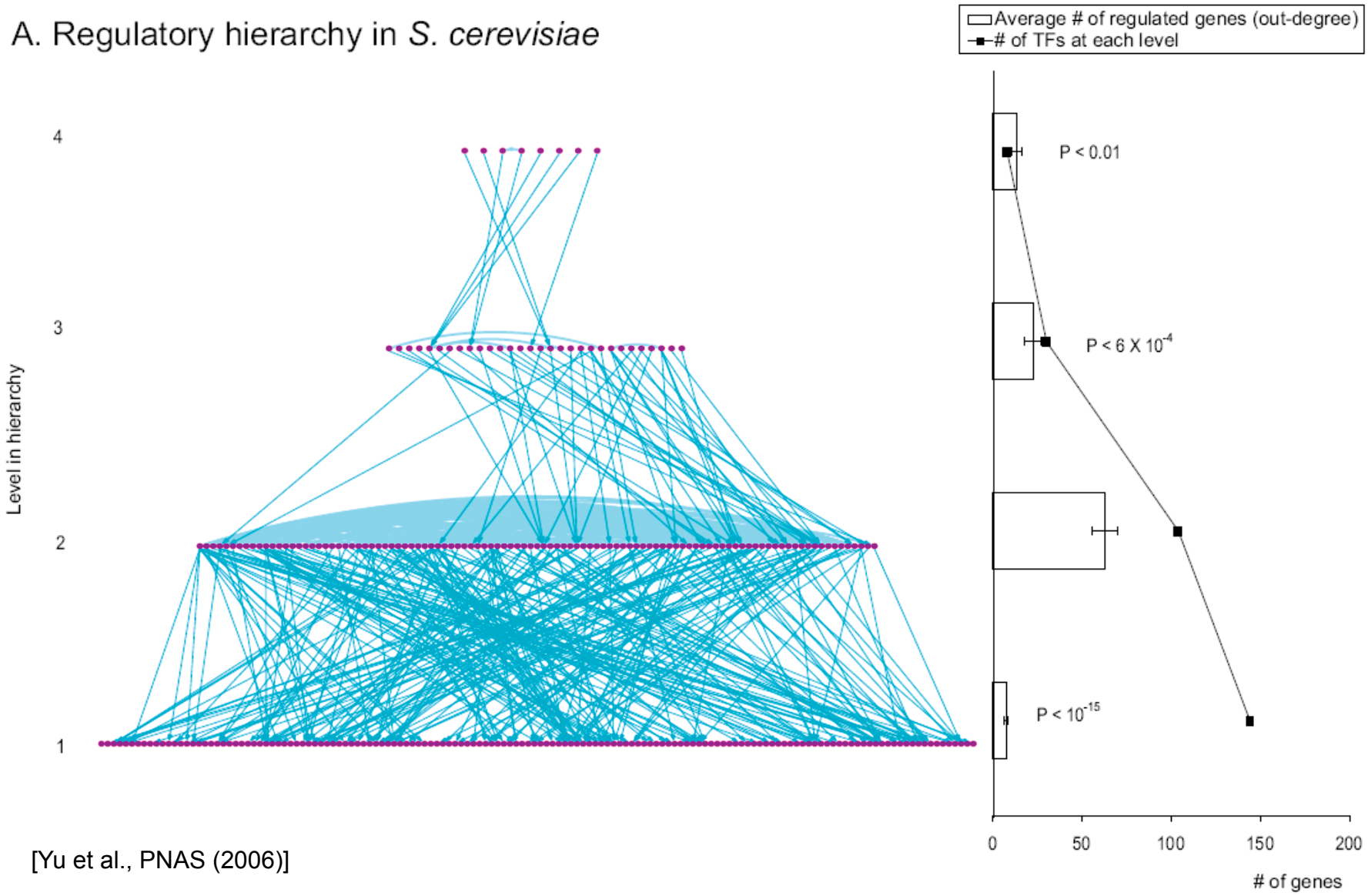


S. cerevisiae

E. coli

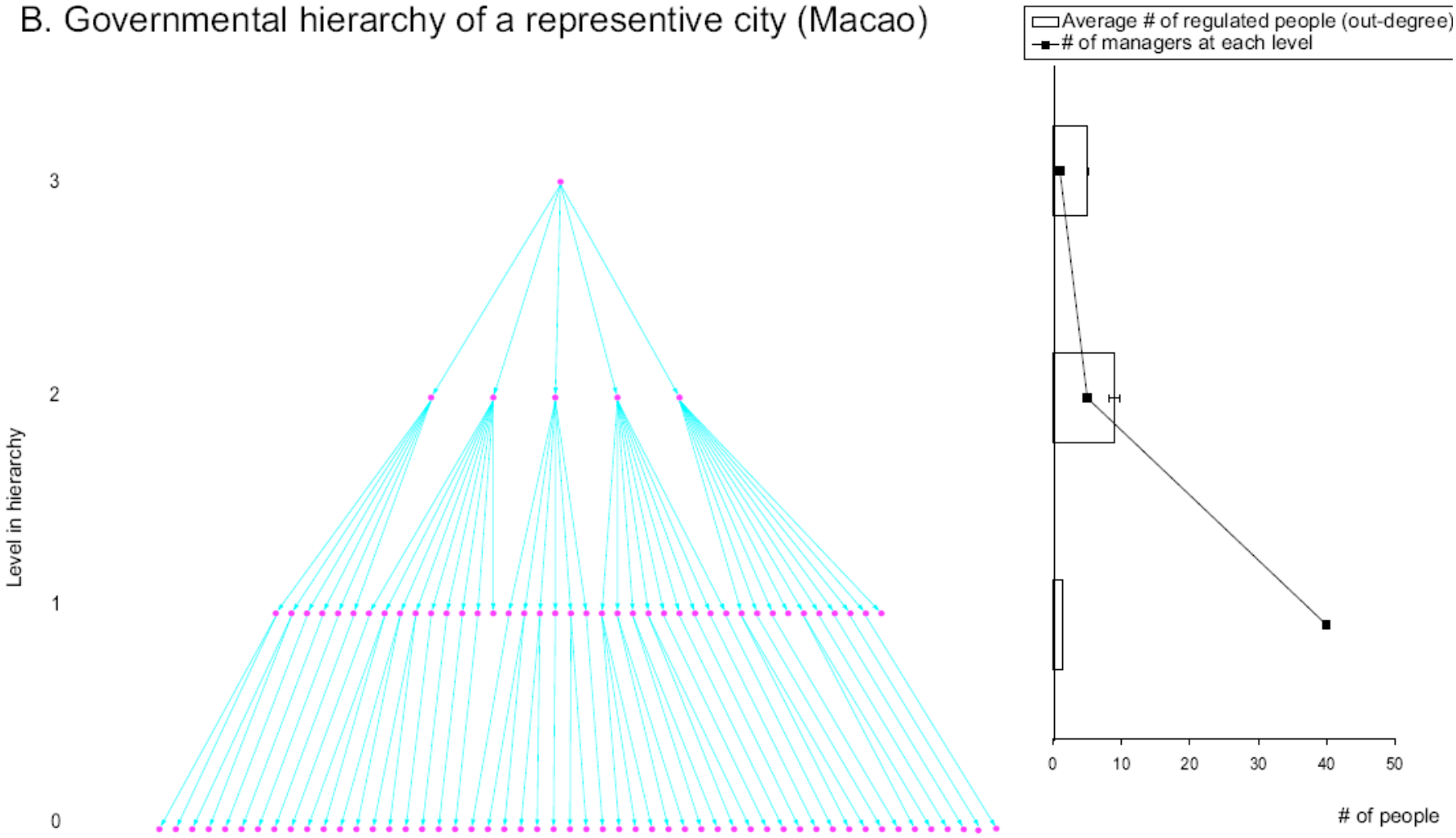
[Yu *et al.*, *Proc Natl Acad Sci U S A* (2006)]

Yeast Regulatory Hierarchy: the Middle-managers Rule

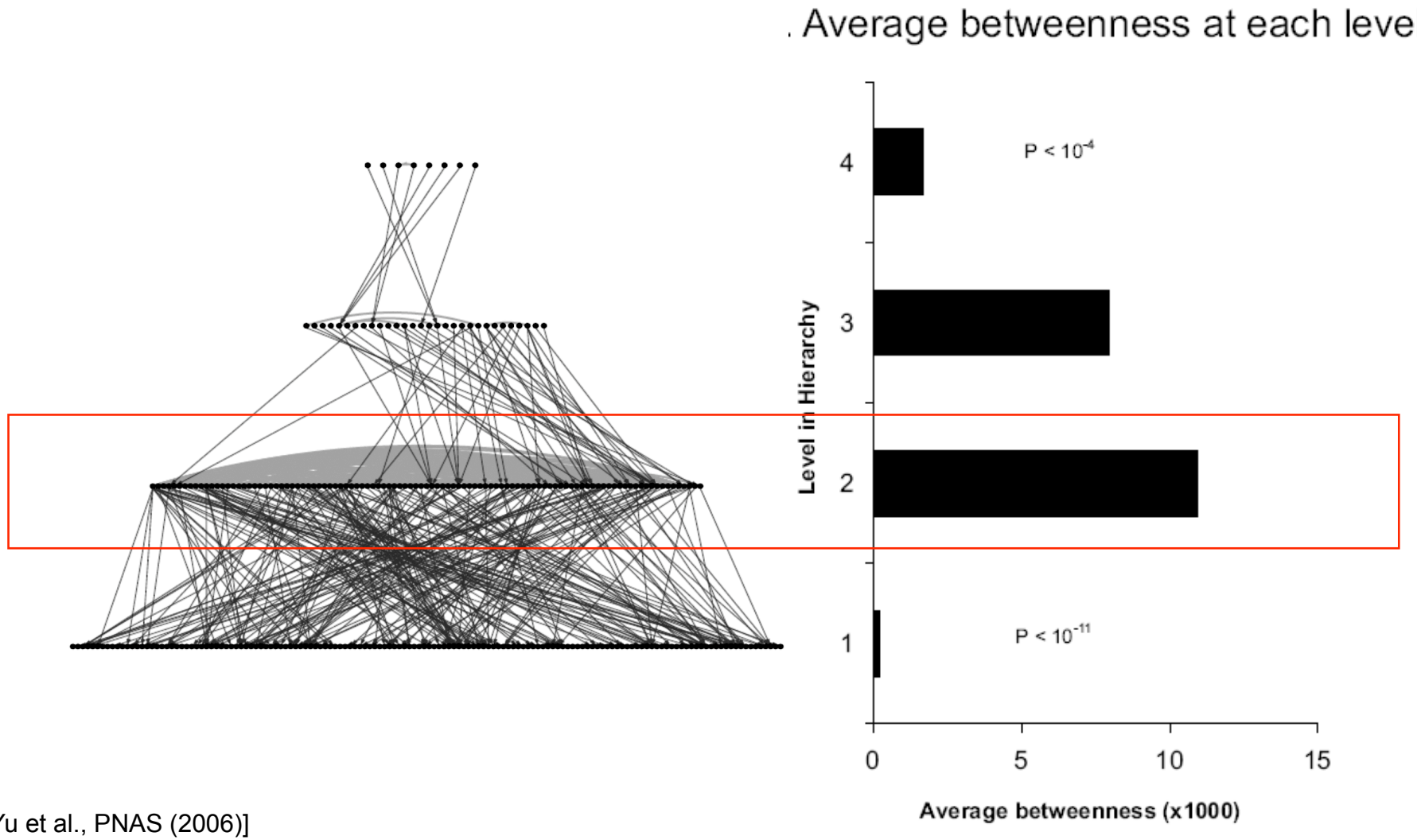


[Yu et al., PNAS (2006)]

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



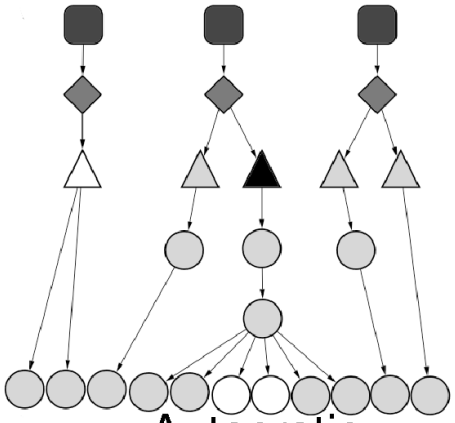
Characteristics of Regulatory Hierarchy: Middle Managers are Information Flow Bottlenecks



[Yu et al., PNAS (2006)]

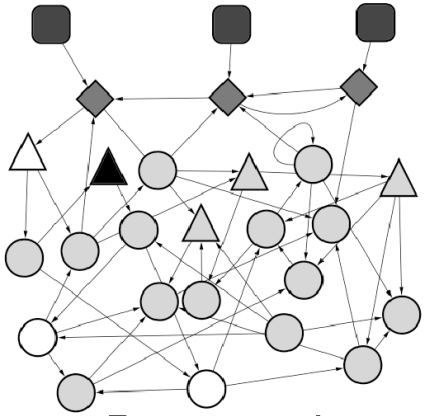
Network Comparison #2
Broadening the comparison to different types of hierarchies & different types of biological networks

Different kinds of Hierarchies



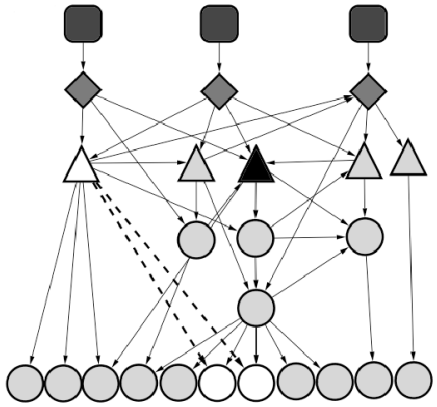
Autocratic

- 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

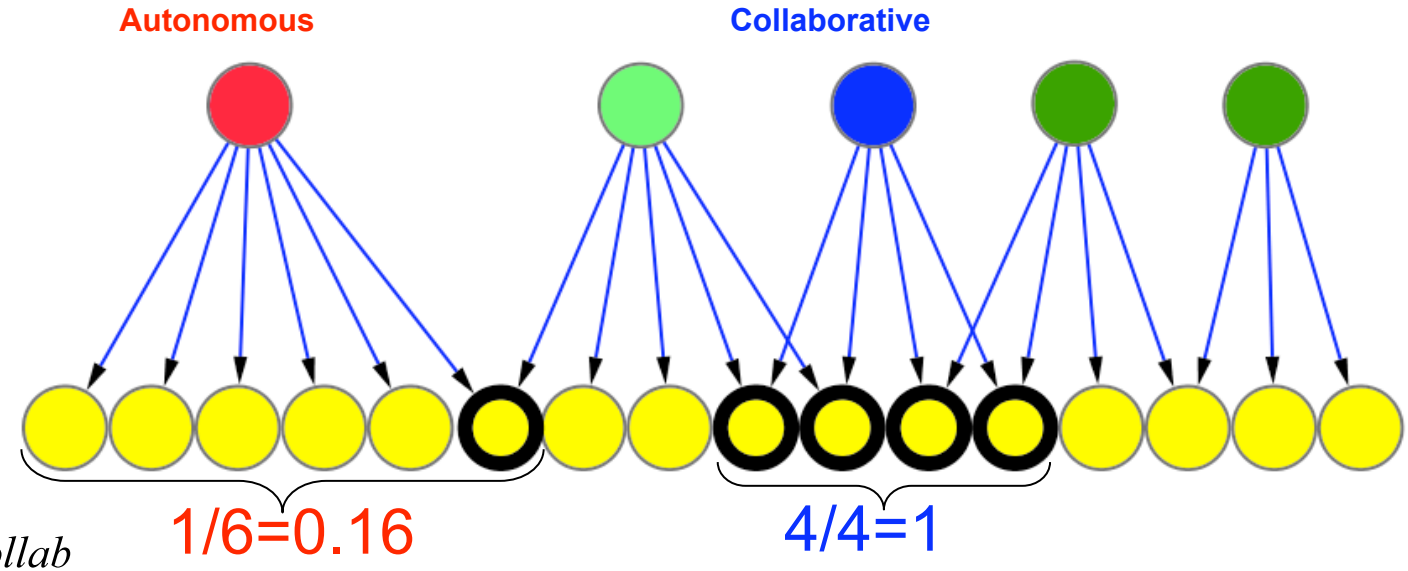
	Autocratic	Democratic	Intermediate
Betweenness (△)	1.03	3.6	3.3
Betweenness (▲)	4.1	1.08	3.4
Var. Betw. (triangles)	2.1	0.58	1.74
Var. Betw. (all)	2.9	1.4	1.9
$D_{Net\text{-}collab}$	0	0.91	0.71

Decreases σ^2 of 'stress'

Increases 'collaboration'

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

Collaborative Nature of the Nodes

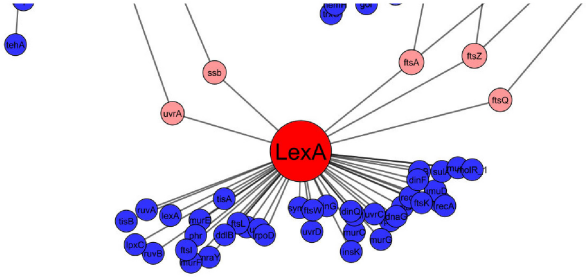


Degree of Collaboration D^i_{collab}

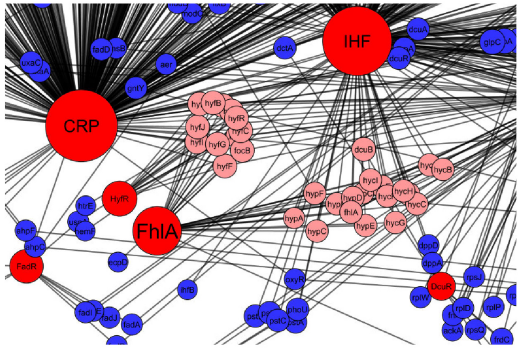
More Collaborative: Democratic

More Autonomous: Autocratic

Autonomous

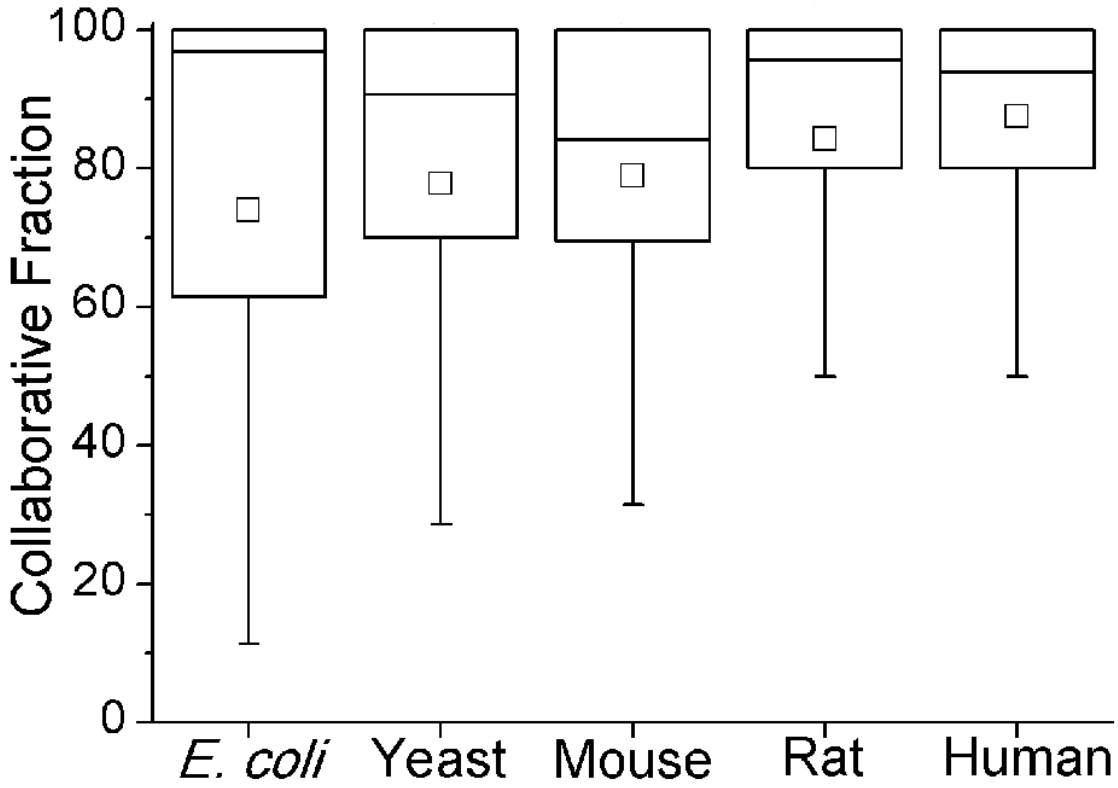


Collaborative



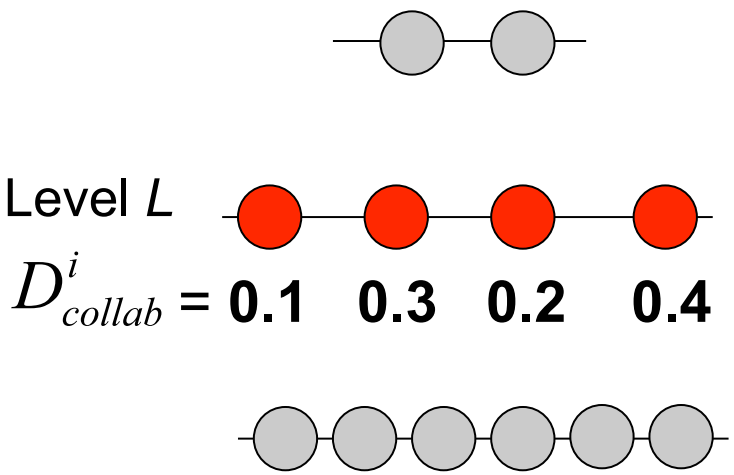
[Bhardwaj et al., PNAS (2010), in press]

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



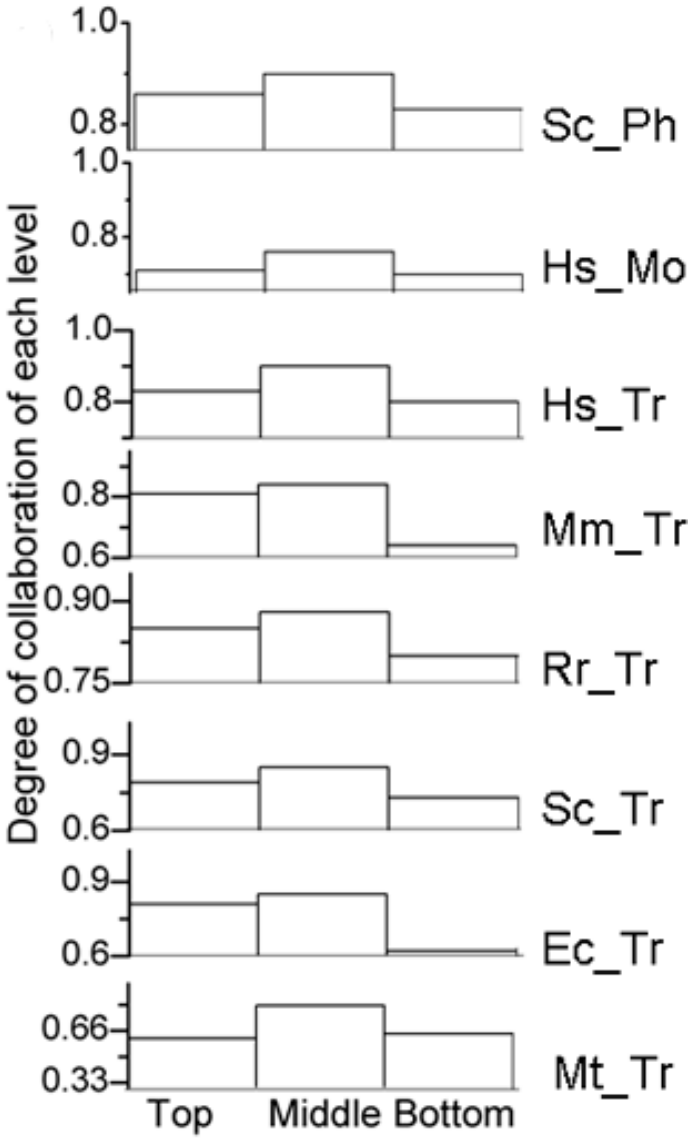
[Bhardwaj et al., PNAS (2010), in press]

Collaborative Nature of the Levels

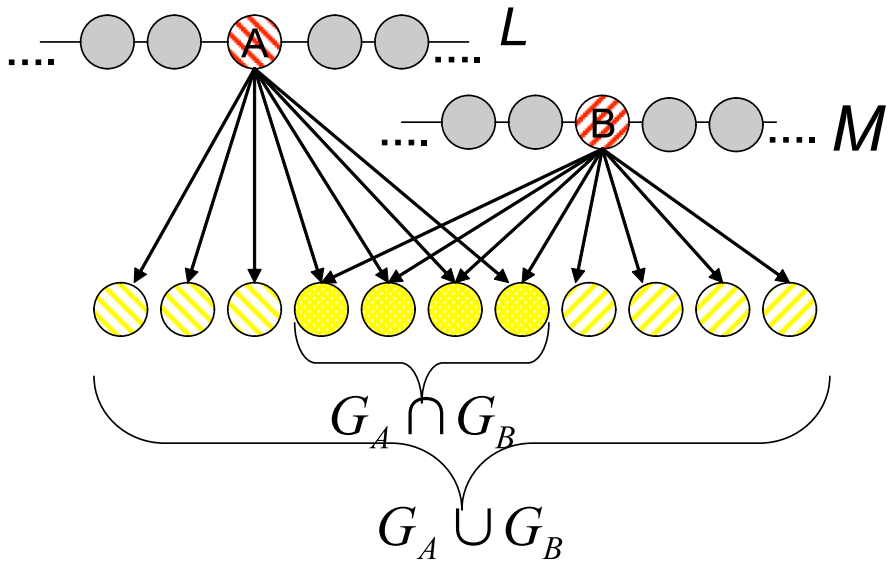


$$D^{L}_{Level-collab} = \frac{0.1+0.3+0.2+0.4}{4} = 0.25$$

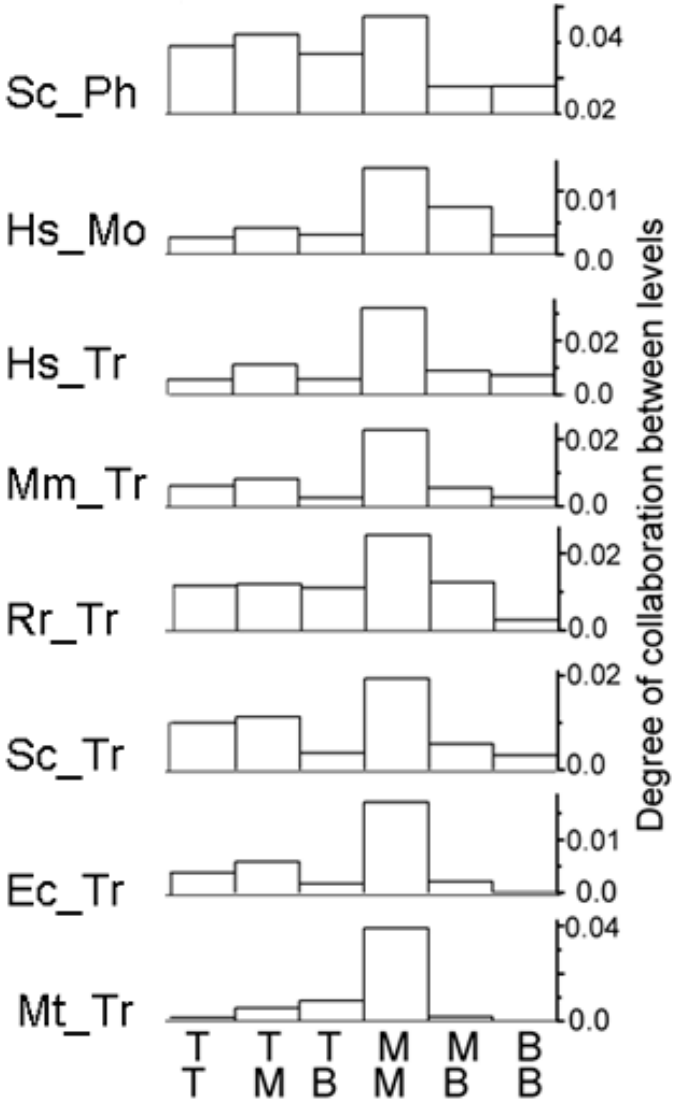
[Bhardwaj et al., PNAS (2010), in press]



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]

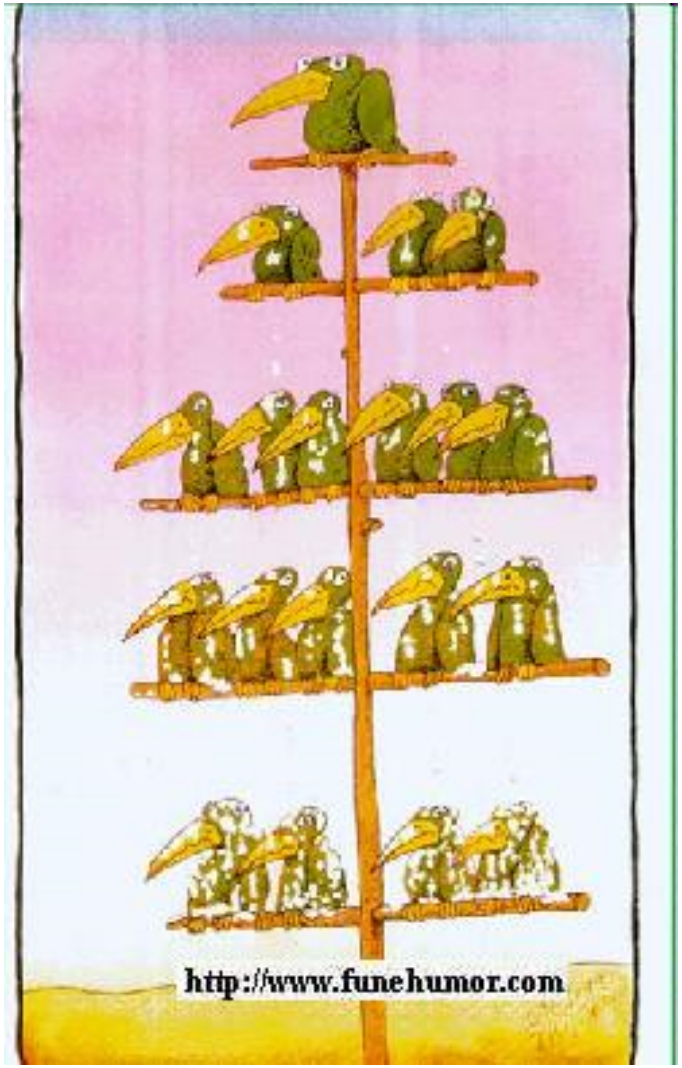
Middle Managers Interact the Most in Efficient Corporate Settings

- 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]

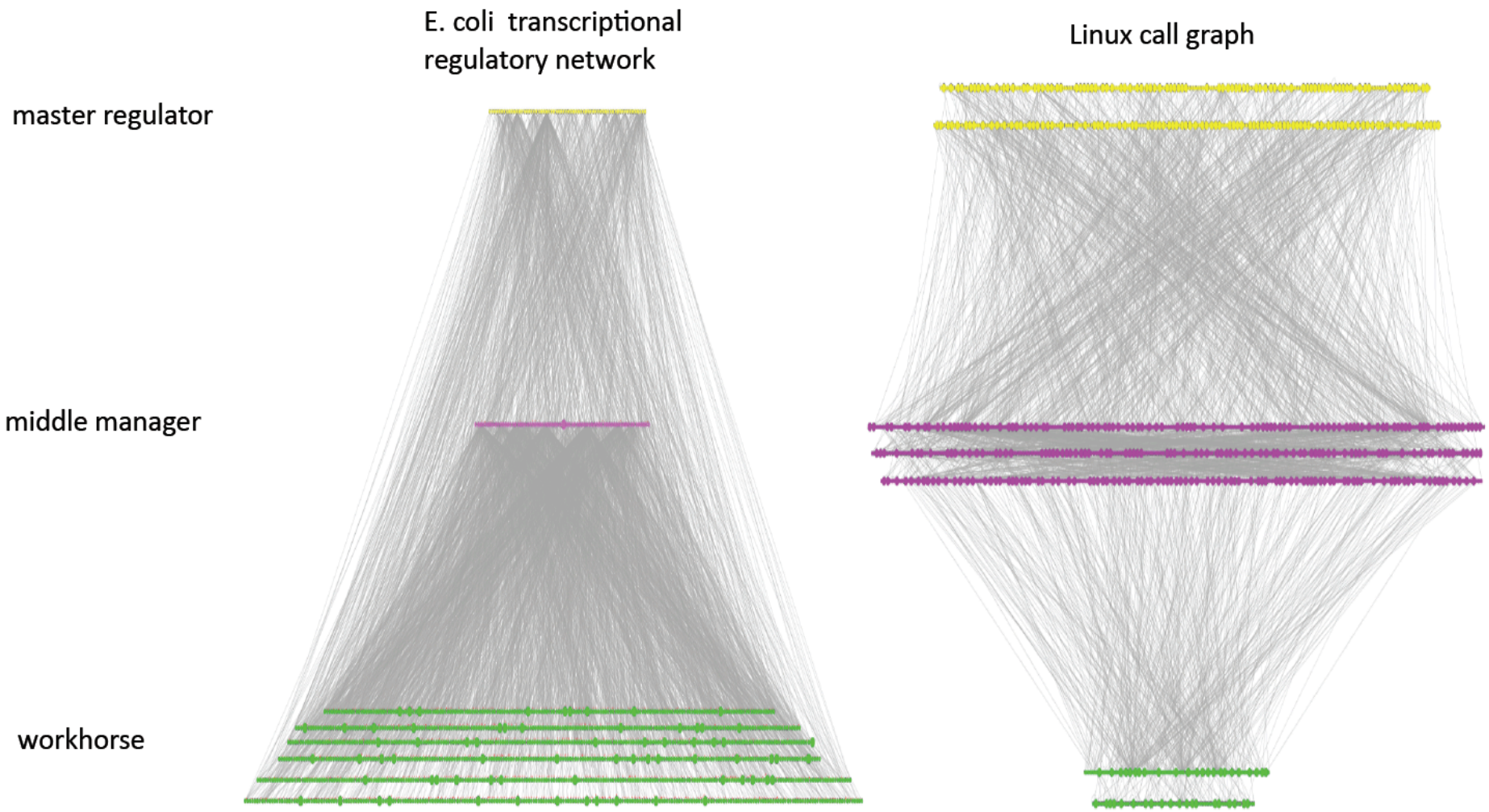
Network Comparison #3: Comparing the structure and evolution of biological regulatory networks and software call graphs

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

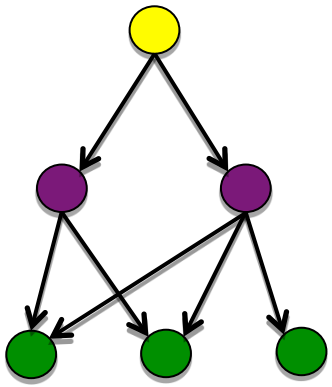
[Yan et al., PNAS (2010), in press]



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

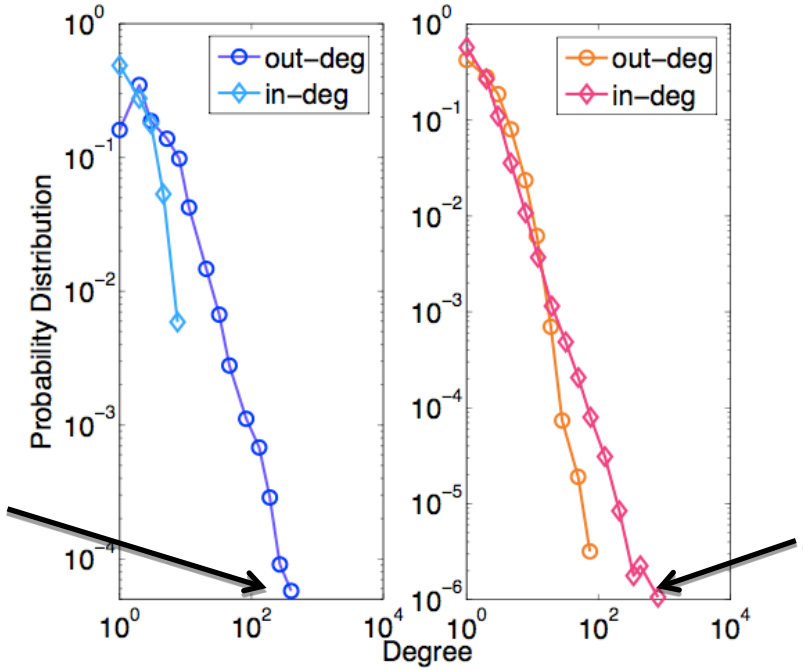
Comparison: hierarchical organization

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



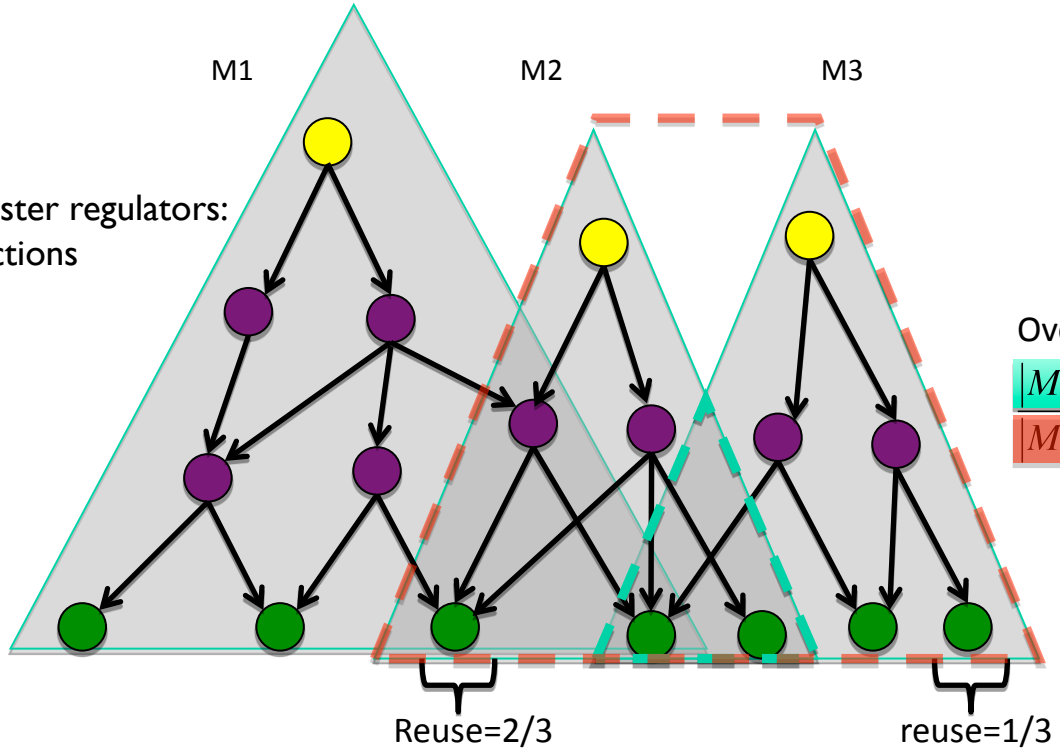
out-degree hubs
e.g. “crp”

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%

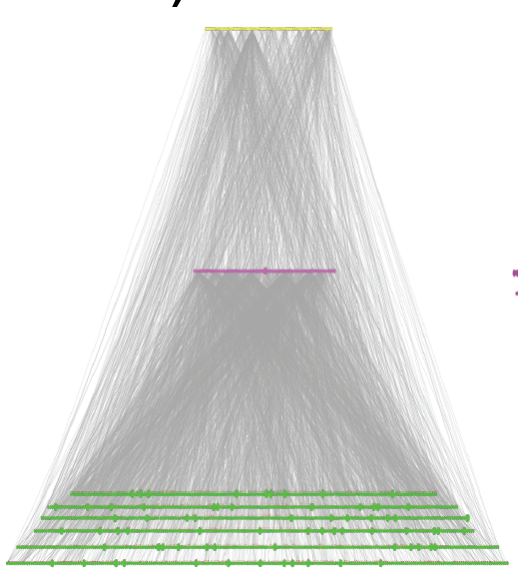
Call graph:
modules overlap,
Functions are highly
reused (generic):
"printk"

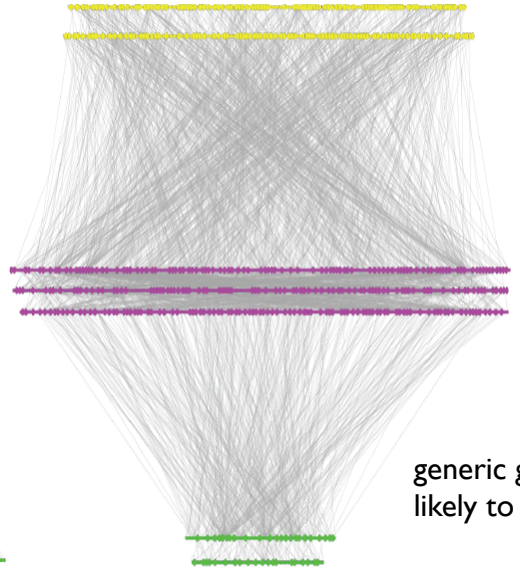
[Yan et al., PNAS (2010), in press]

Comparison of persistent components

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

specialized proteins
are preserved
across genomes



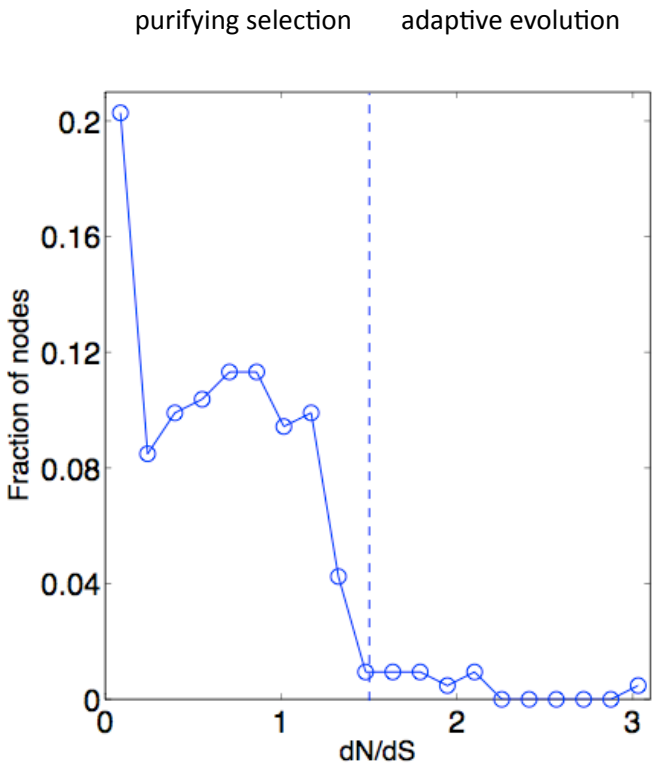


generic guys are more
likely to be persistent

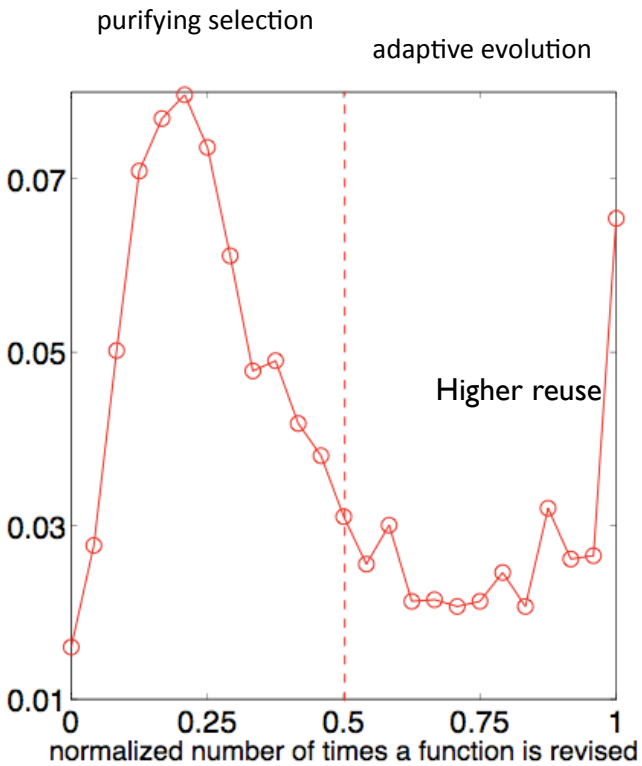
- Building of the hierarchy:
  - ◊ TRN: Bottom up. Regulatory changes are the main driving forces of evolution
  - ◊ Call graph: top down

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

Evolutionary rate of persistent functions



Persistent genes evolve slowly

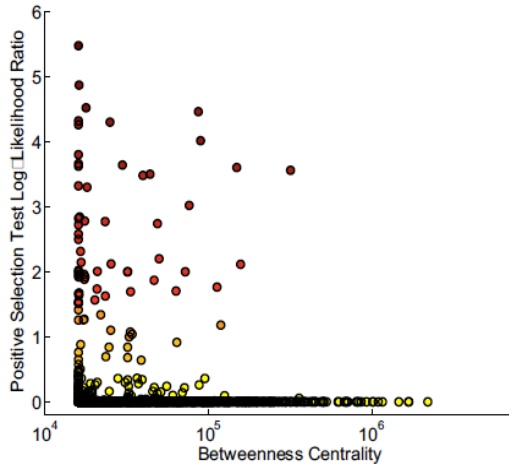


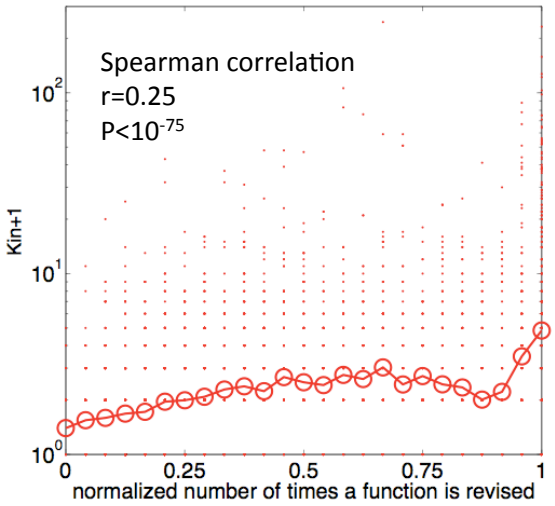
Two classes of persistent functions

[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]

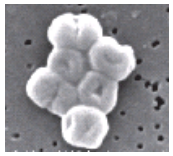
Network Dynamics Across Environments: Metabolic Pathways

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

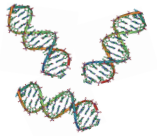
What is Metagenomics?

Traditional Genomics

Select organism and culture

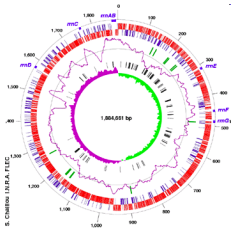


Extract DNA and sequence



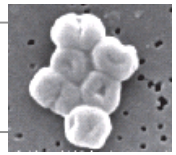
atgctcgatctcg
atcgatctcgctg
atgccgatctaa

Assemble and annotate

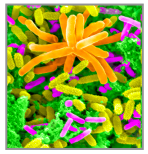


Contig 1
Contig 2
...
...

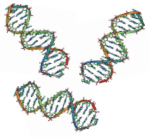
Estimated that less than 1% of microbes can be cultured



Metagenomics

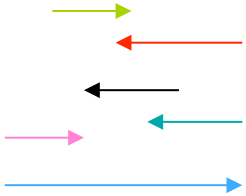


Collect sample from environment



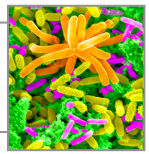
atgctcgatctcg
atcgatctcgctg
atgccgatctaa

Extract DNA and sequence



Assemble and annotate

Lose information about which gene belongs to which microbe



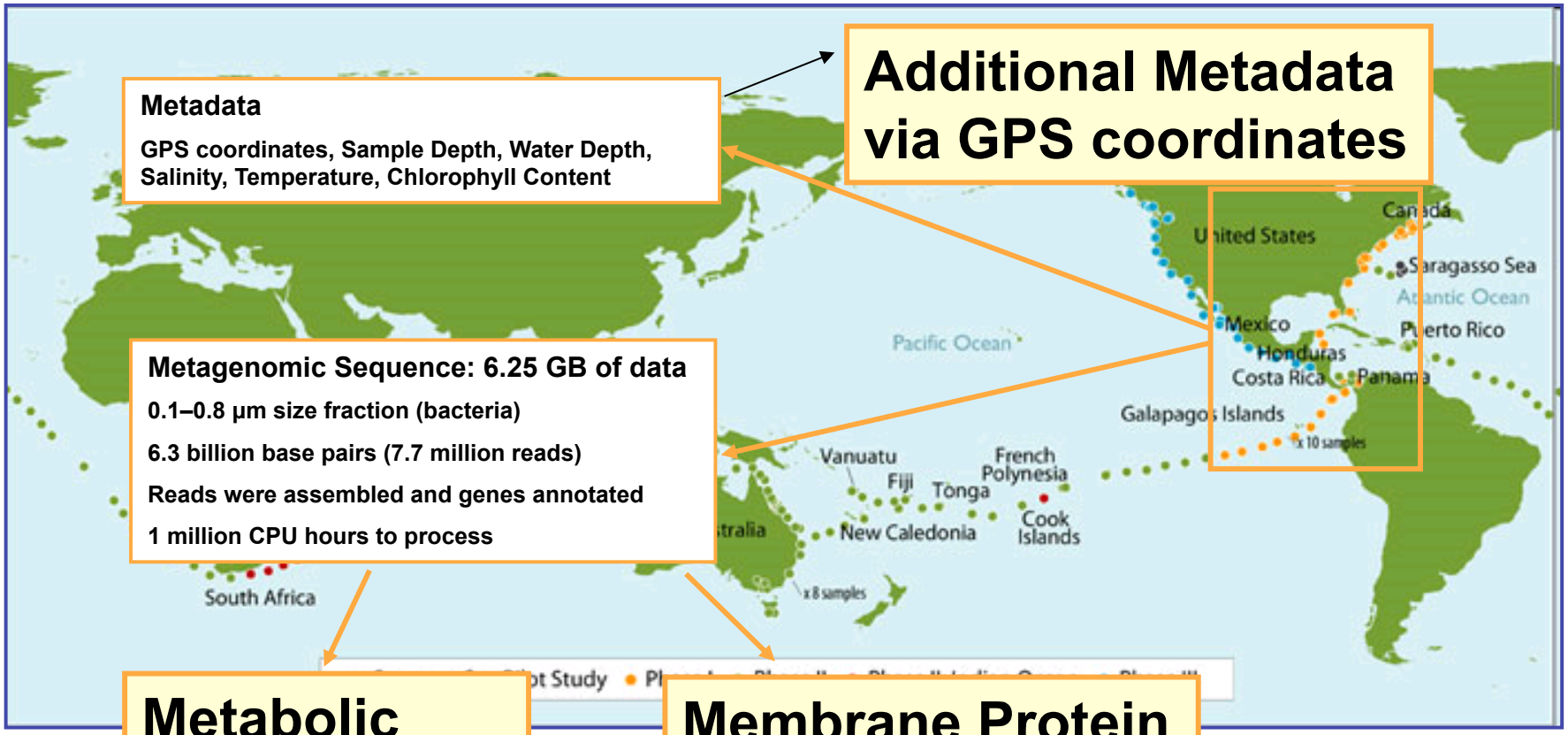
Sorcerer II Global Ocean Survey

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

Sample approximately every 200 miles

Rusch, et al., *PLOS Biology* 2007

Sorcerer II Global Ocean Survey

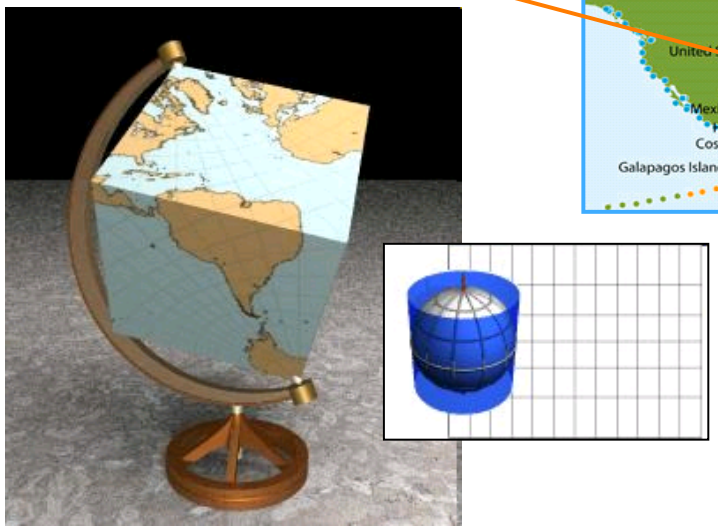


Rusch, et al., *PLOS Biology* 2007

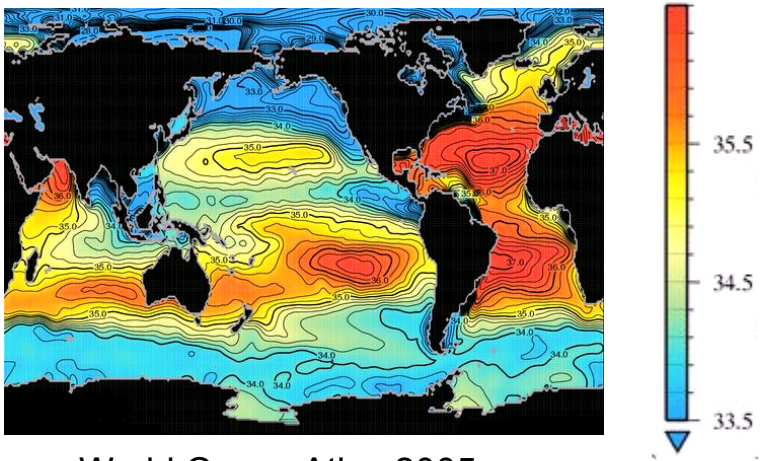
Extracting Environmental Data from Other Sources

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

Anthropogenic Features Extracted:

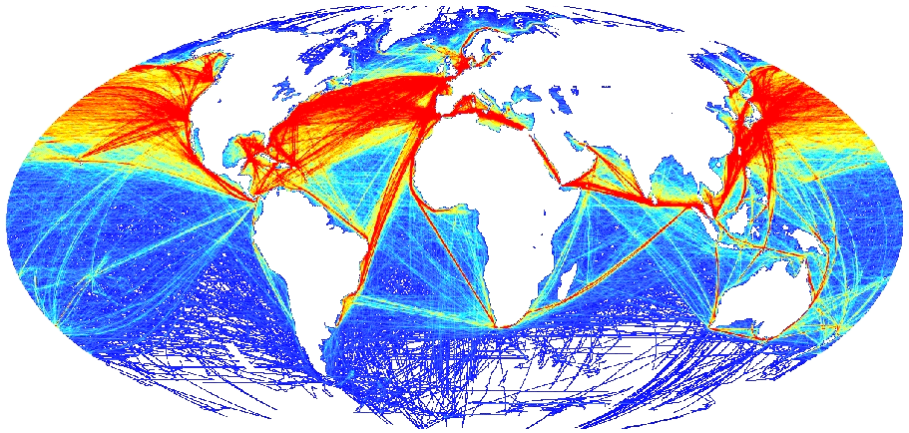
Ultraviolet radiation

Shipping

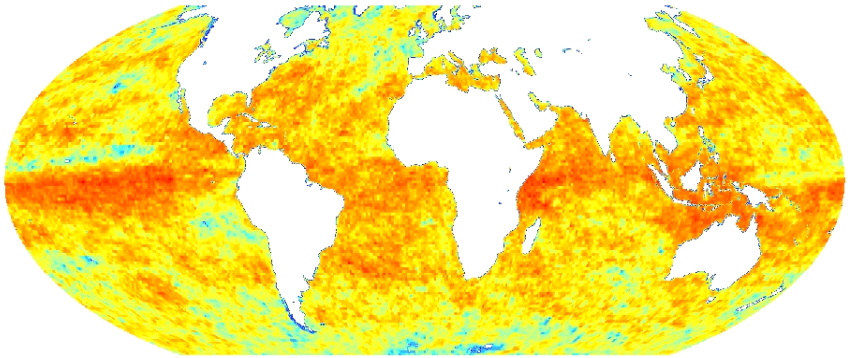
Pollution

Climate Change

Ocean Acidification

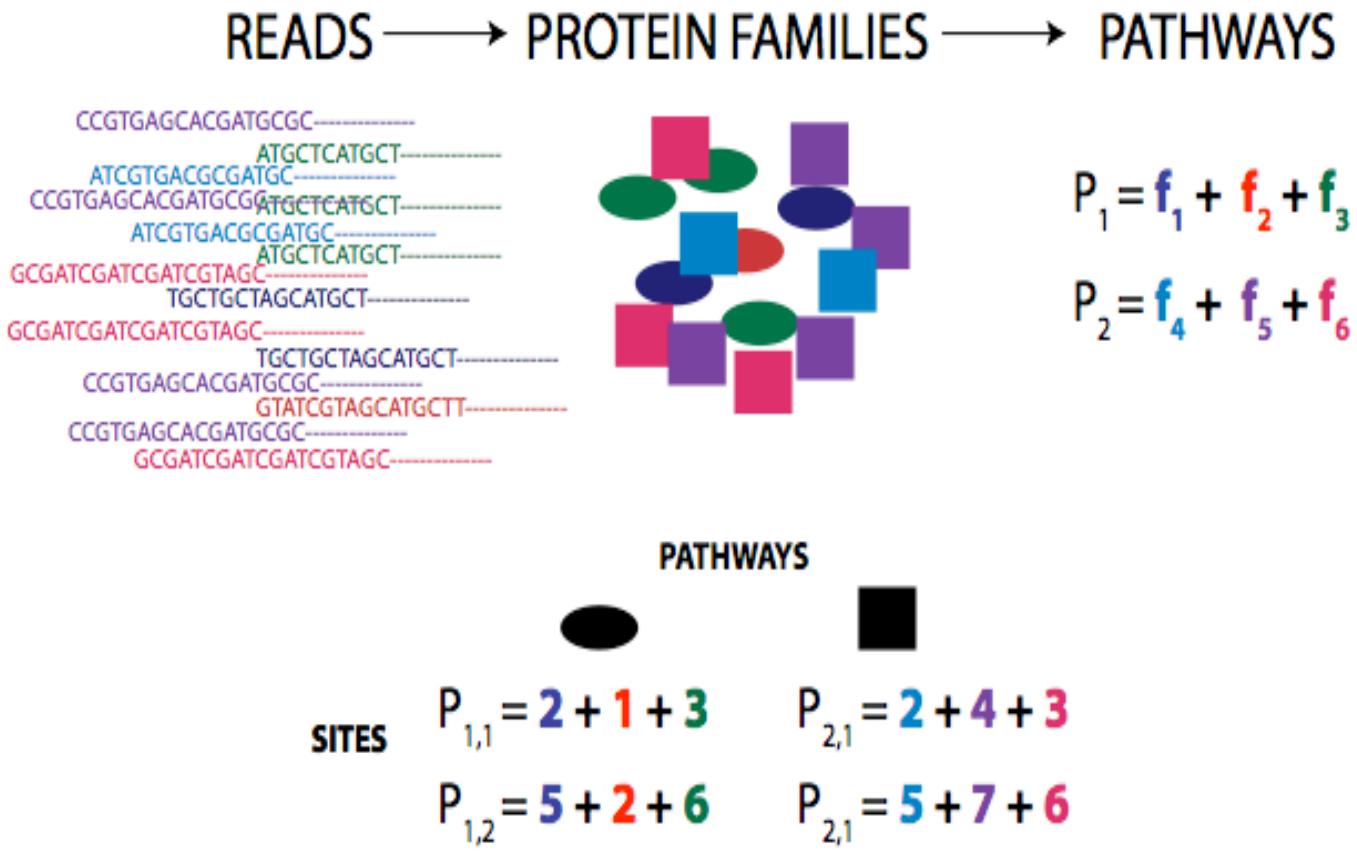


Shipping

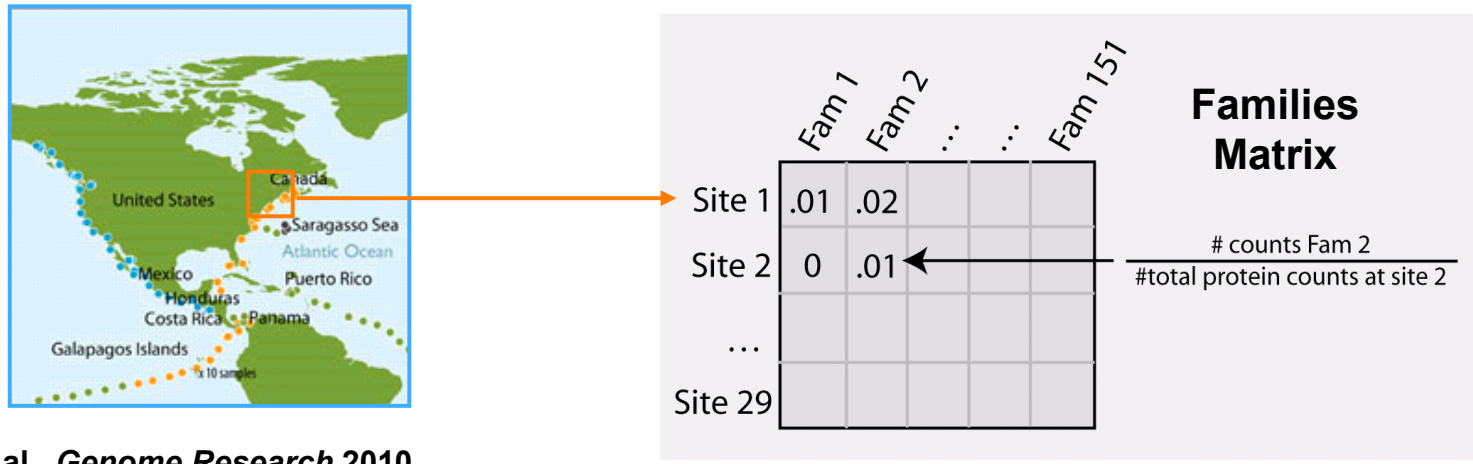


Climate Change

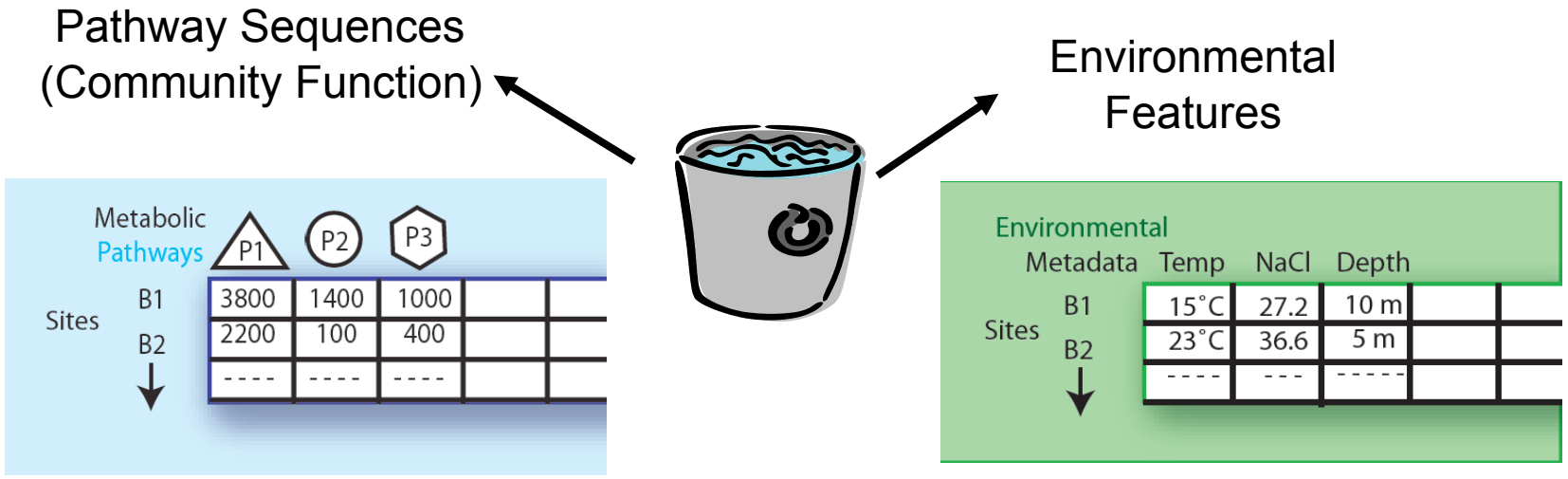
Halperin et al. *Science* 2008



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



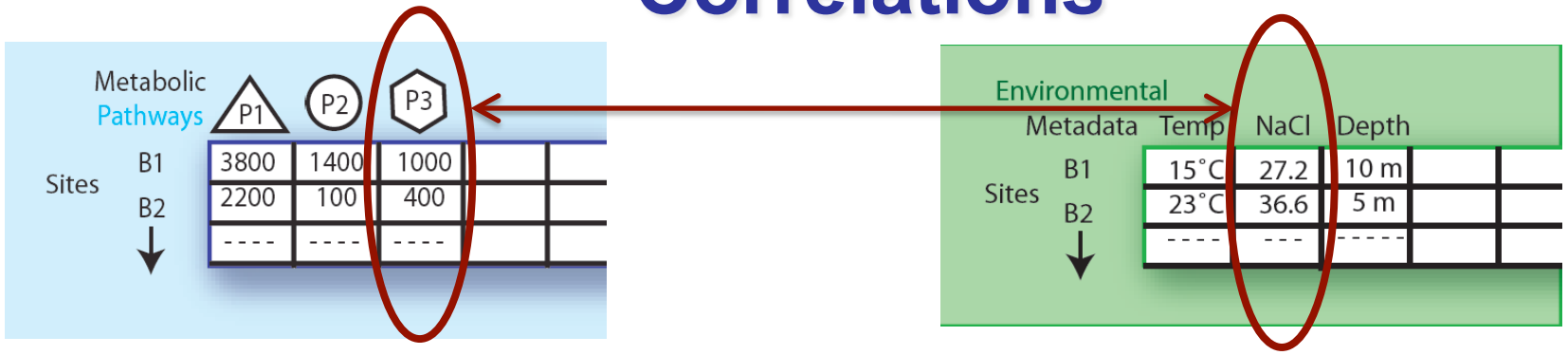
Patel et. al., *Genome Research* 2010

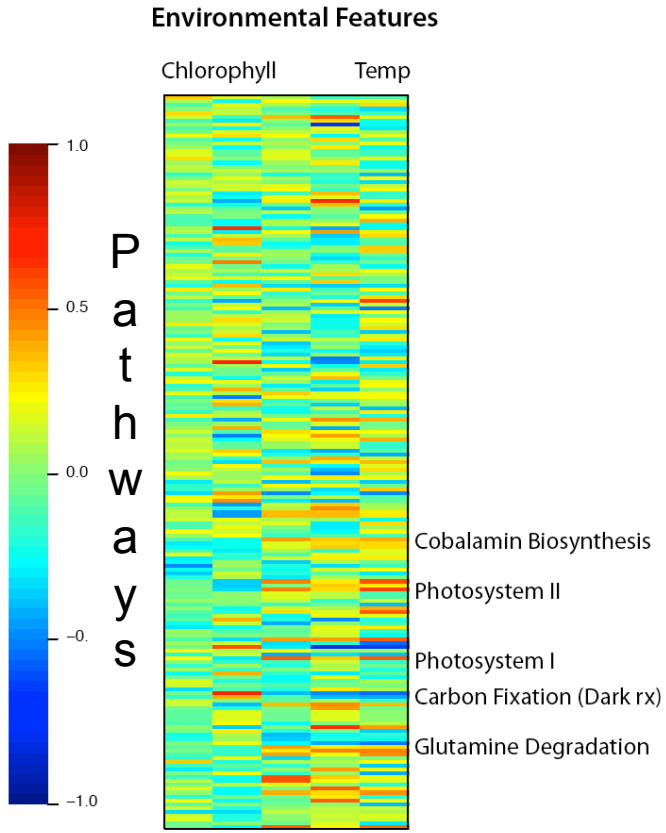


Expressing data as matrices indexed by site, env. var., and pathway usage

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

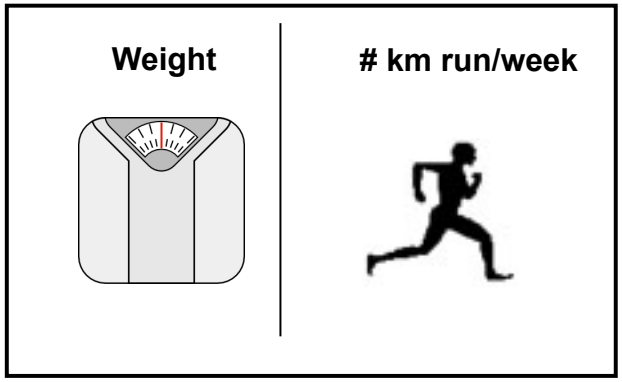
Simple Relationships: Pairwise Correlations

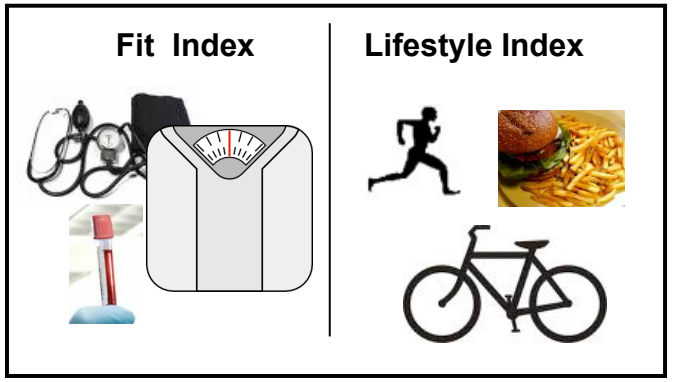




[Gianoulis et al., PNAS (in press, 2009)]

Canonical Correlation Analysis: Simultaneous weighting

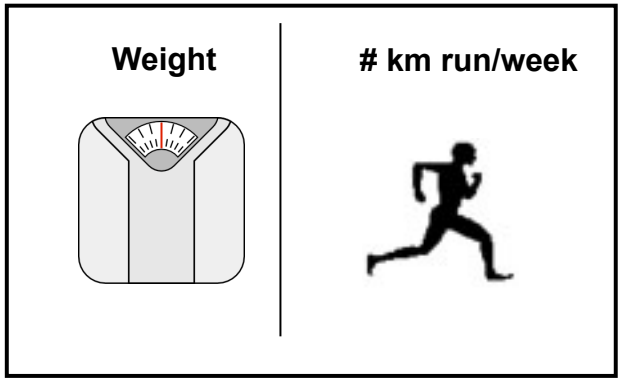


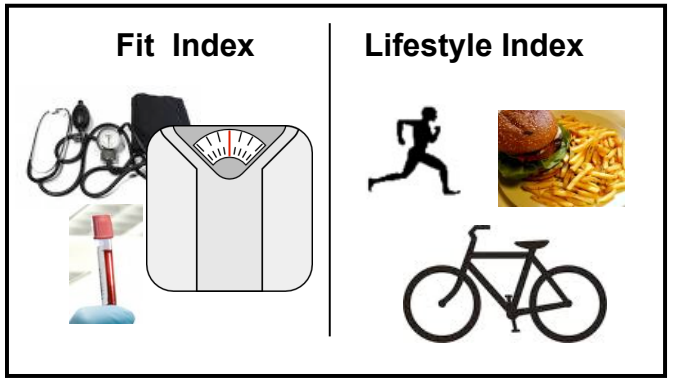


Fit Index = a + b +c

Lifestyle Index = a +b + c

Canonical Correlation Analysis: Simultaneous weighting



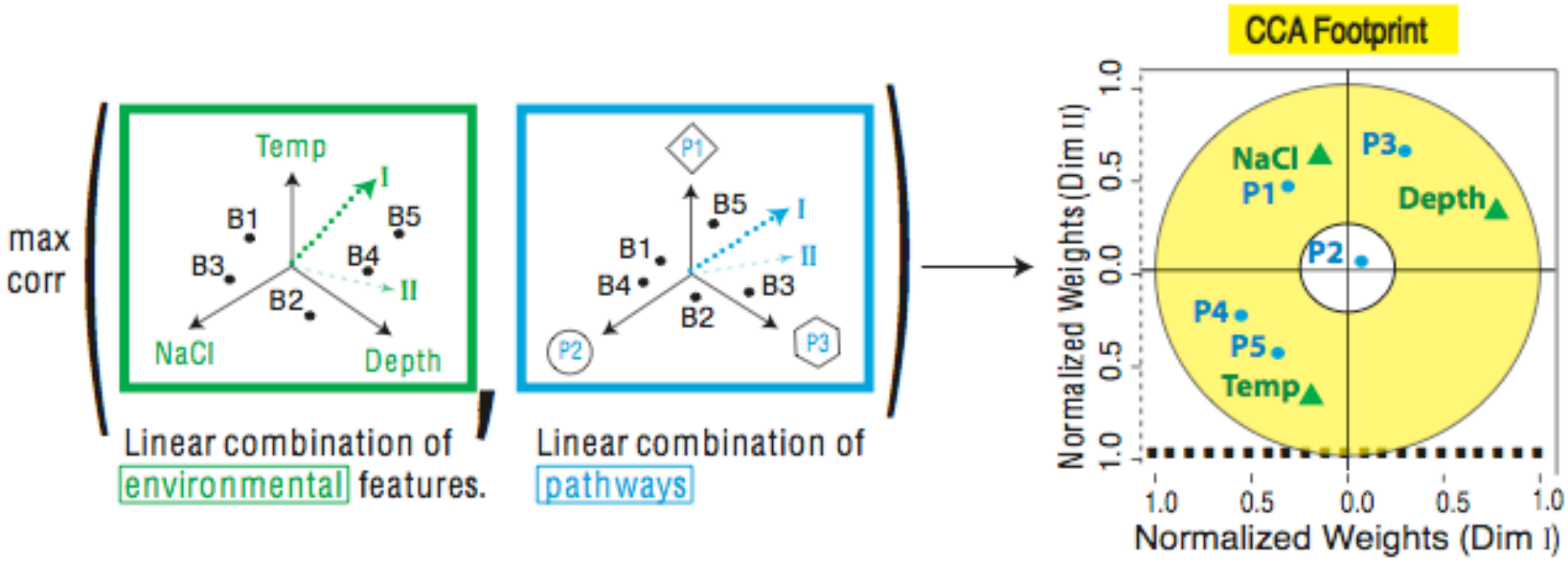


Fi

Lif

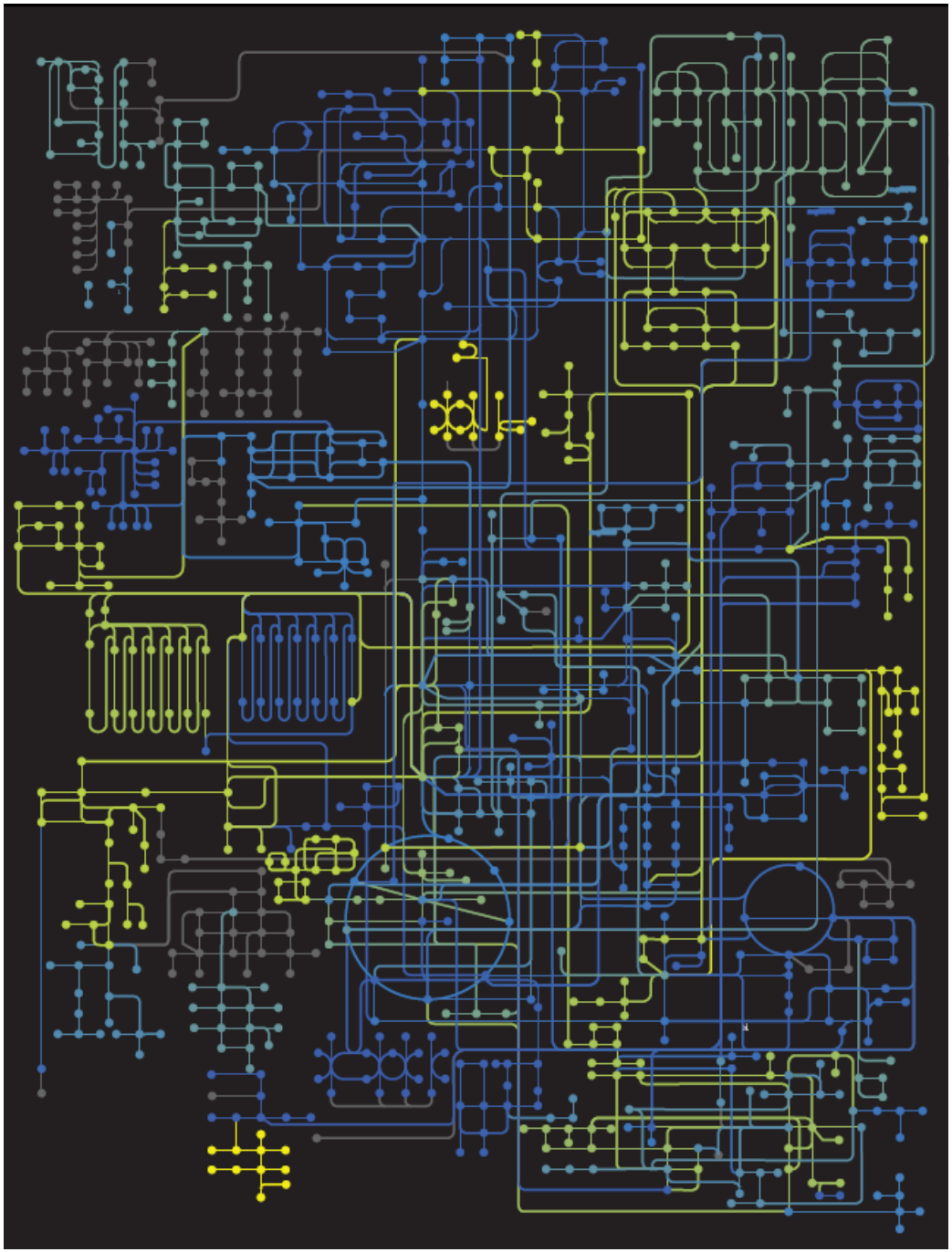
Environmental Features	Metabolic Pathways/ Protein Families
Temp etc	Photosynthesis etc
Chlorophyll	Lipid Metabolism

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

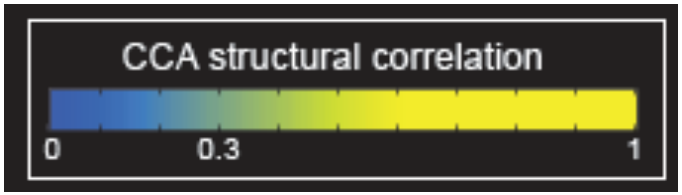


The goal of this technique is to interpret cross-variance matrices
We do this by defining a change of basis.

Gianoulis et al., PNAS 2009

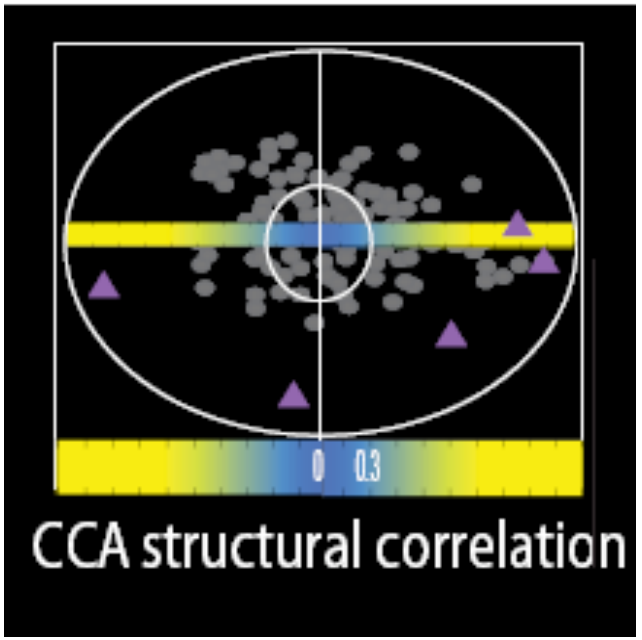


Strength of Pathway co-variation with environment



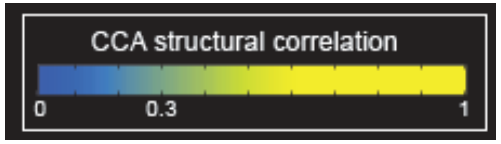
Environmentally invariant

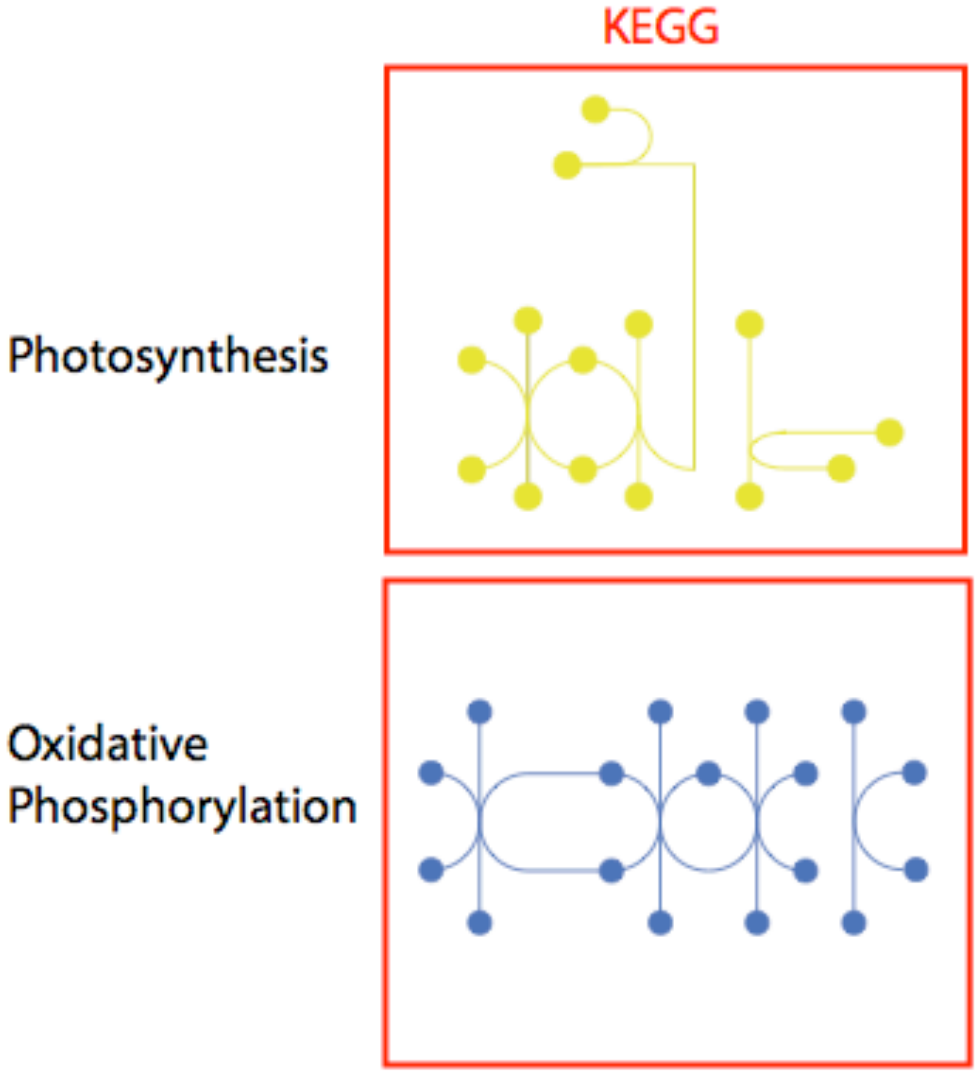
Environmentally variant



Gianoulis et al., *PNAS* 2009

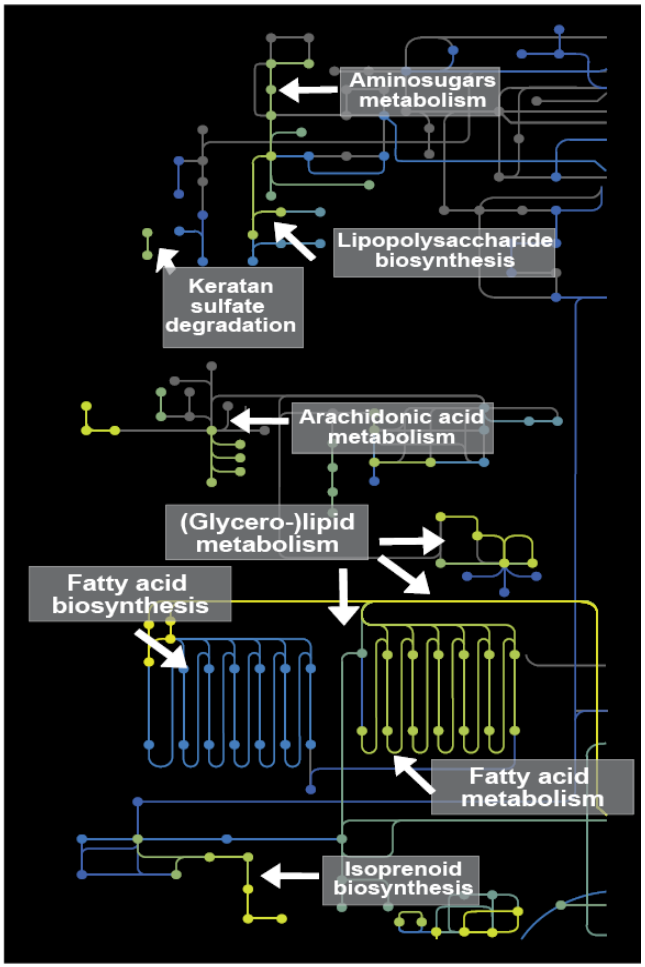
Conclusion #1: energy conversion strategy, temp and depth

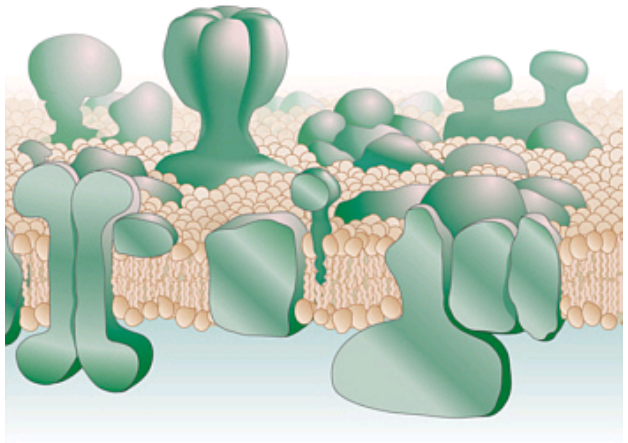




Gianoulis et al., *PNAS* 2009

Conclusion #2: Outer Membrane components vary with the environment



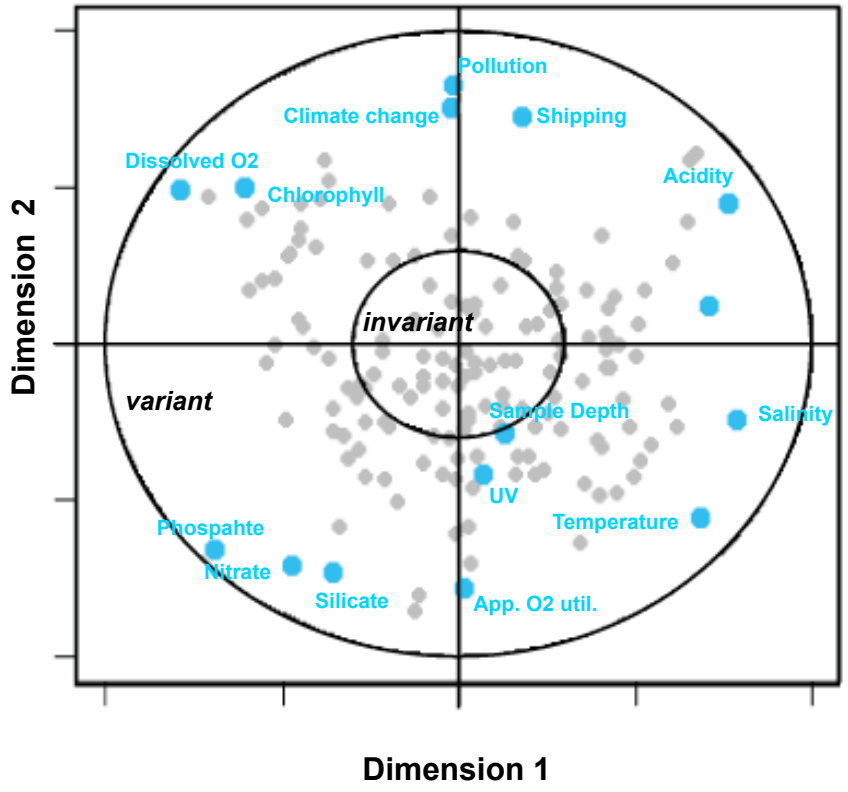


Membrane proteins interact with the environment, transporting available nutrients, sensing environmental signals, and responding to changes

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

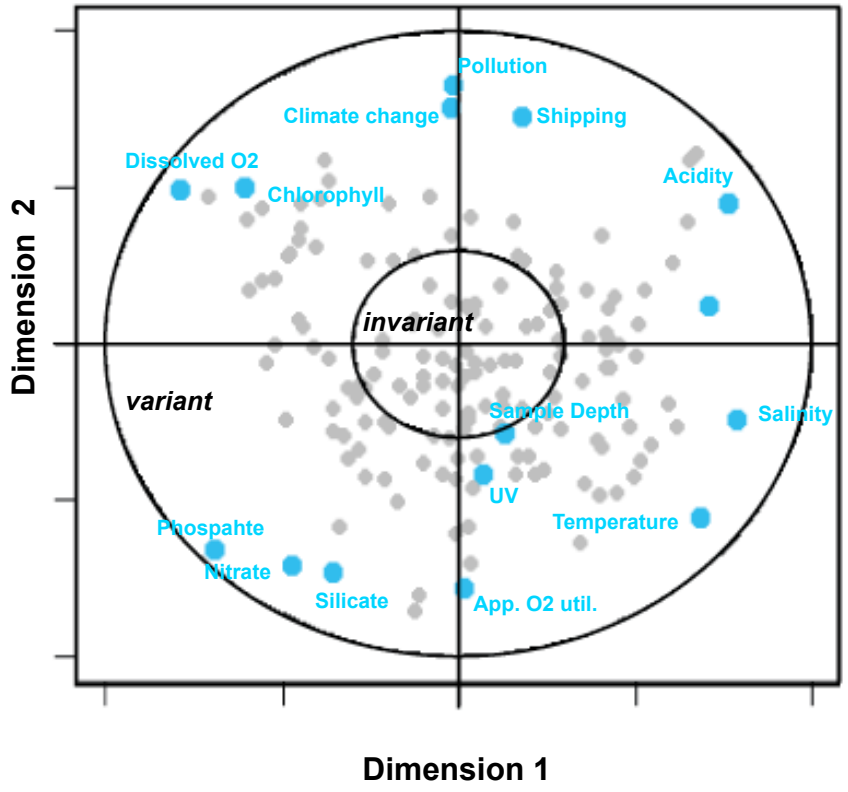
20% have NO KNOWN FUNCTION

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

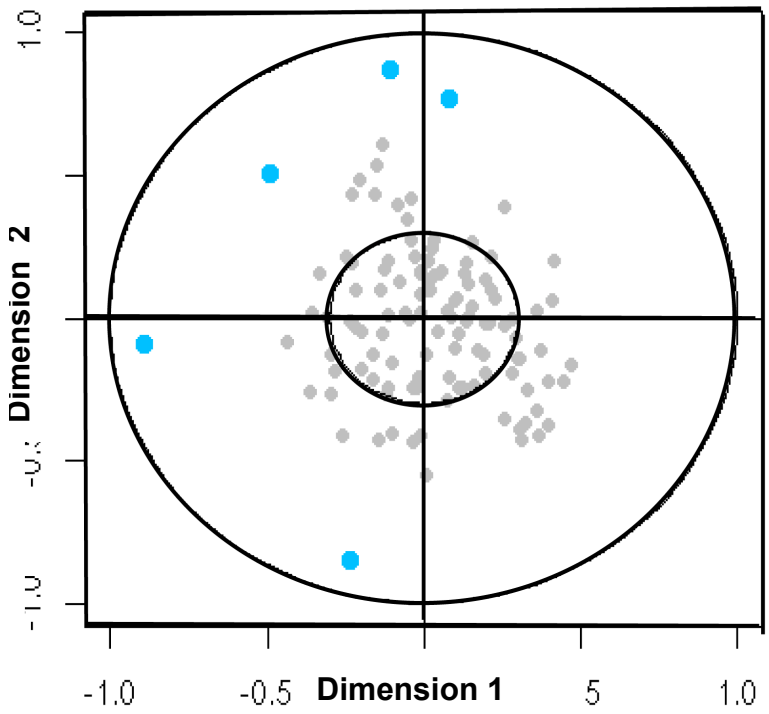
Membrane Proteins co-vary more than Metabolic Pathways

Median absolute structural Correlation Coefficient

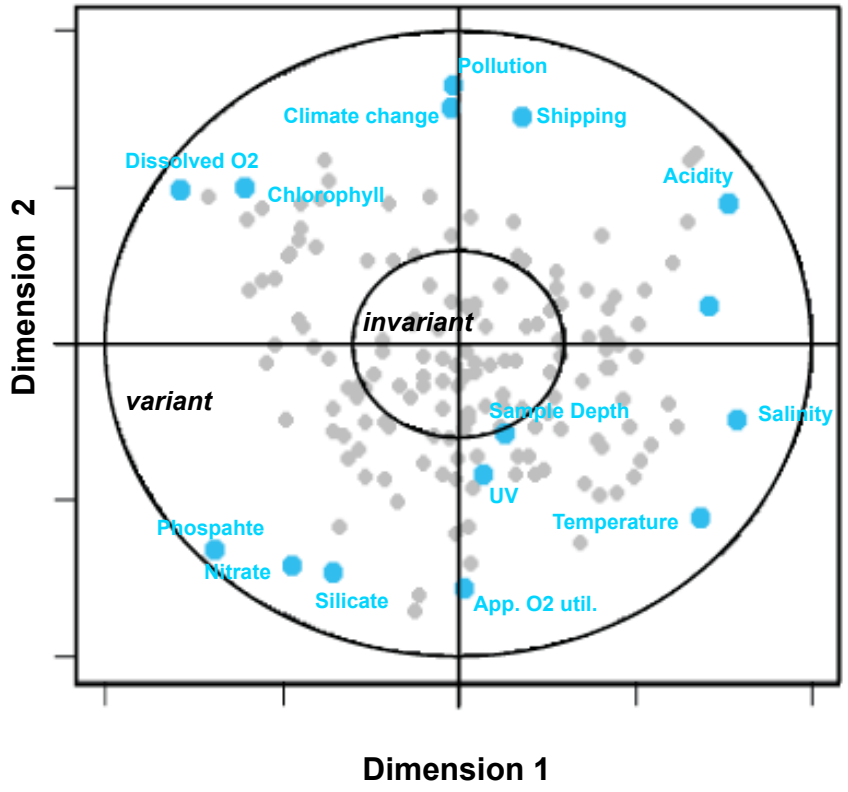
Membrane Proteins = 0.3



Metabolic Pathways = 0.17



CCA Limitations



Four Major Obstacles

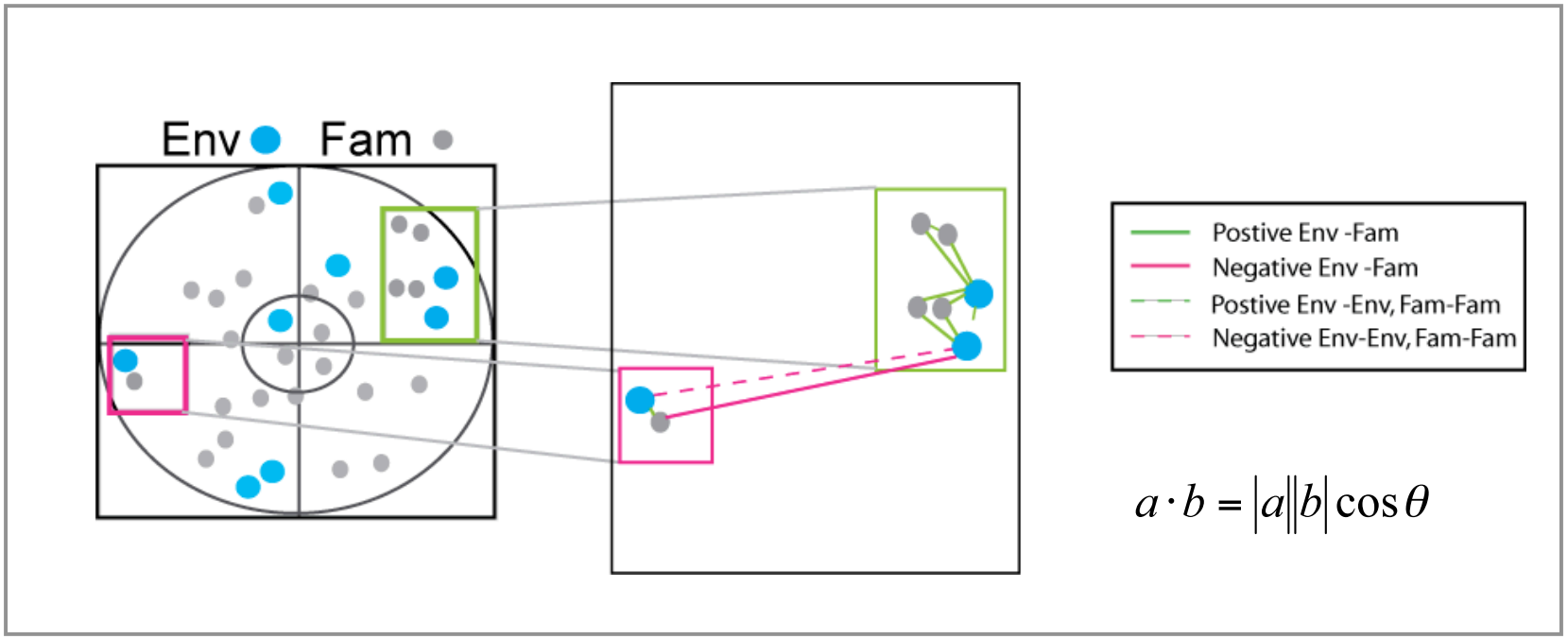
(1) Strength and directionality of relationships not intuitive

(2) Relative weights of features are difficult to visualize and compare.

(3) No real means of quantifying covariation between specific sets of features.

(4) Difficult to visualize or compare results in more than 2 dimensions.

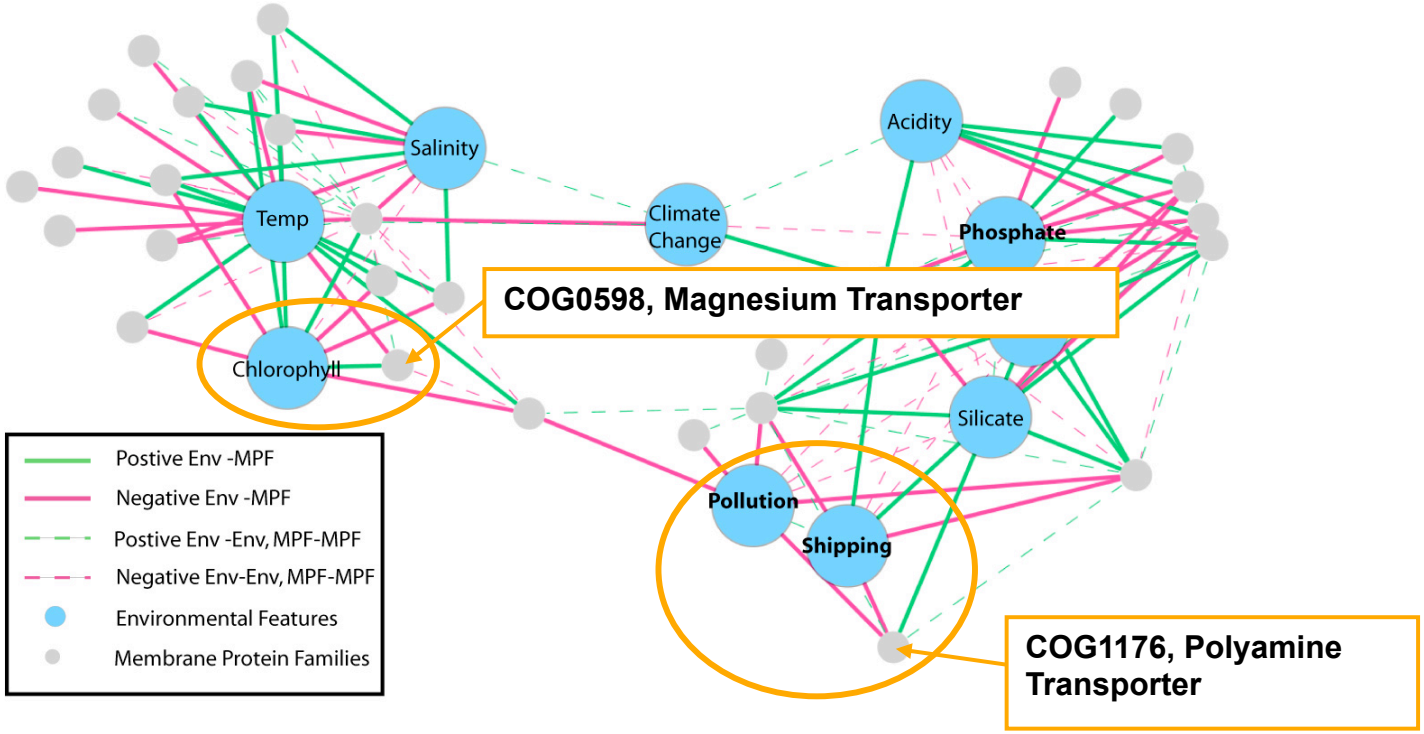
Protein Families and Environmental Features Network (PEN)



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

Patel et. al., *Genome Research* 2010

Protein Families and Environmental Features Network (PEN)

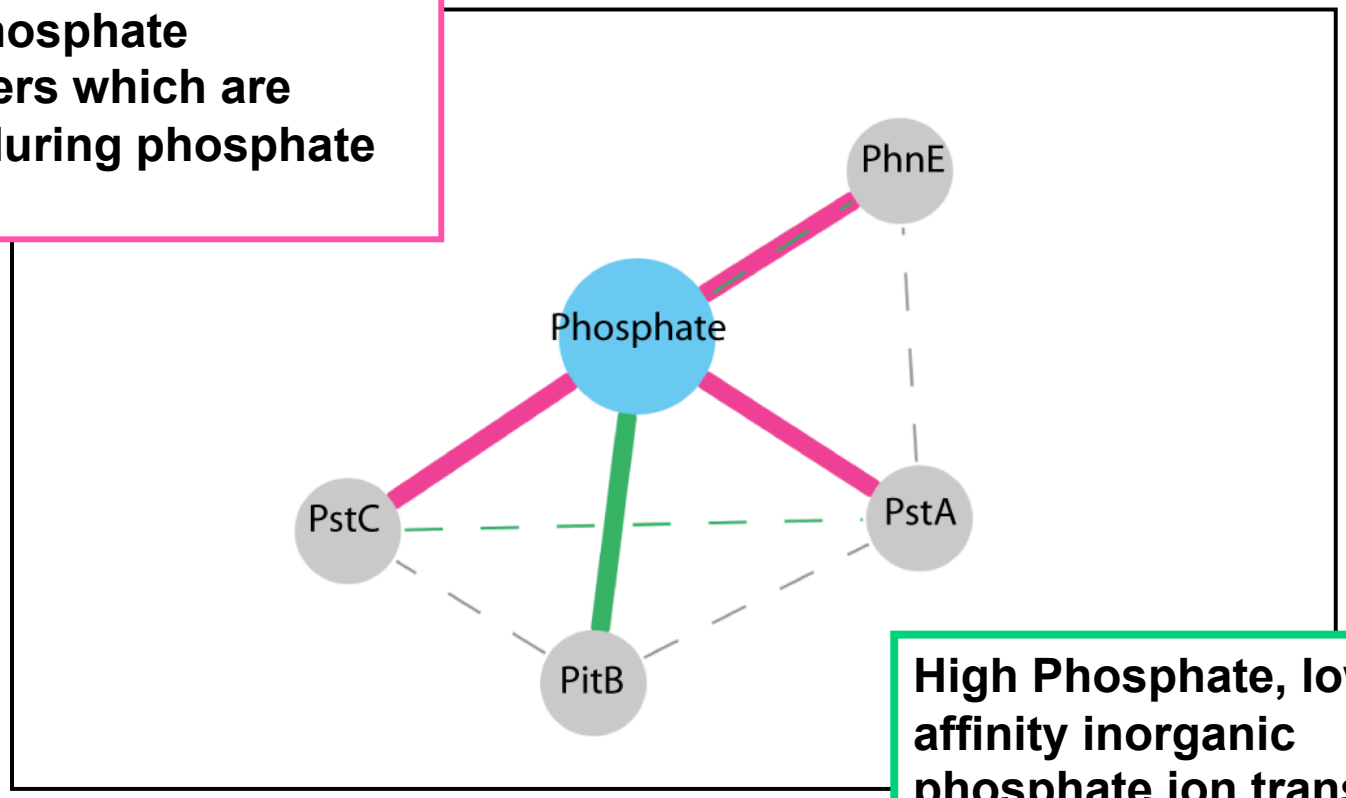


"Bi-modules": groups of environmental features and membrane proteins families that are associated

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

Bi-module 1: Phosphate/Phosphate Transporters

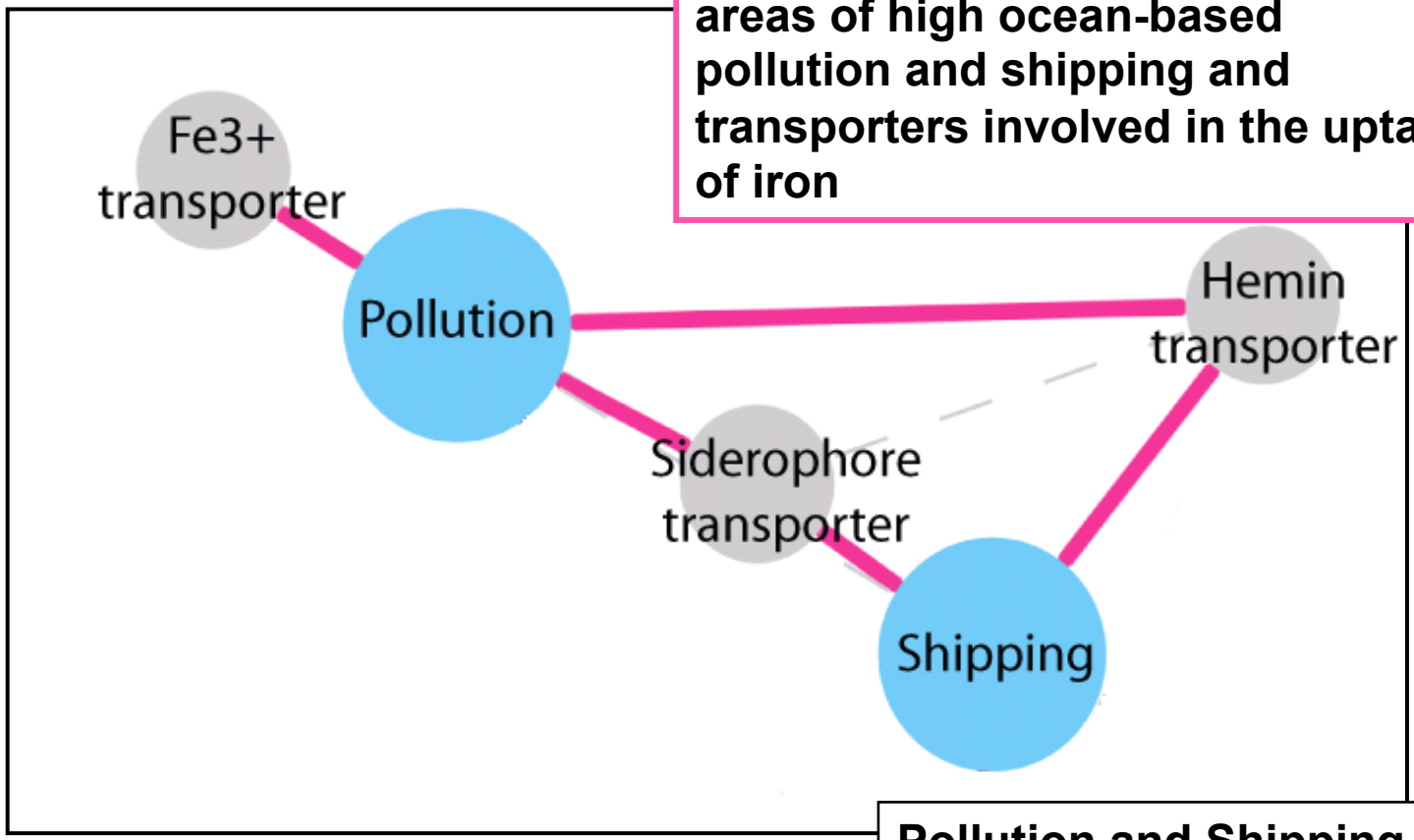
Low Phosphate, high affinity phosphate transporters which are induced during phosphate limitation



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

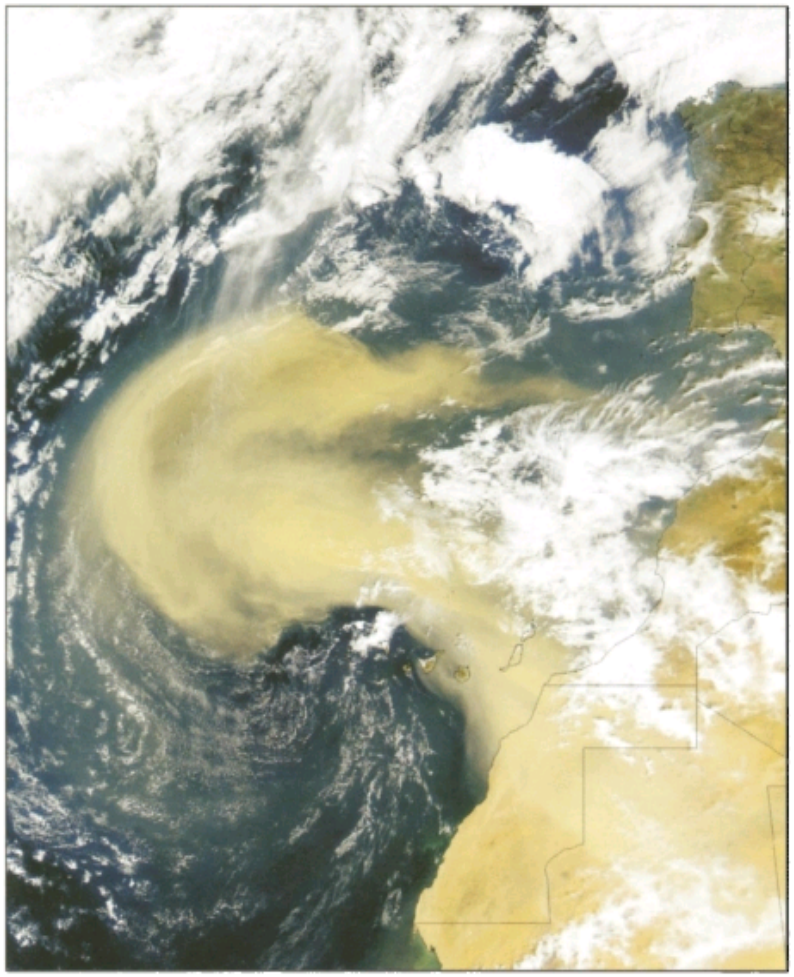


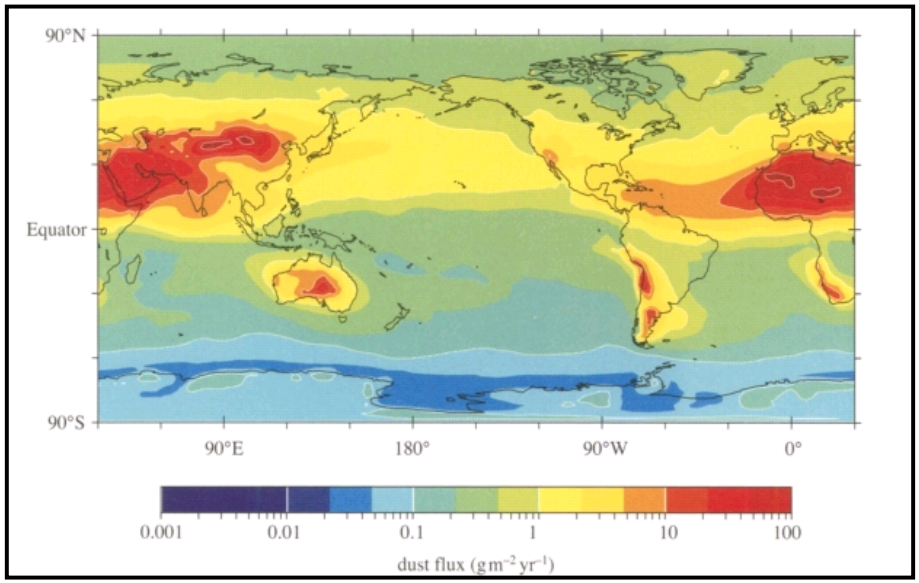
Negative relationship between areas of high ocean-based pollution and shipping and transporters involved in the uptake of iron

Pollution and Shipping may be a proxy for iron concentrations →

Patel et. al., *Genome Research* 2010

Bi-module 2: Iron Transporters/Pollution/Shipping





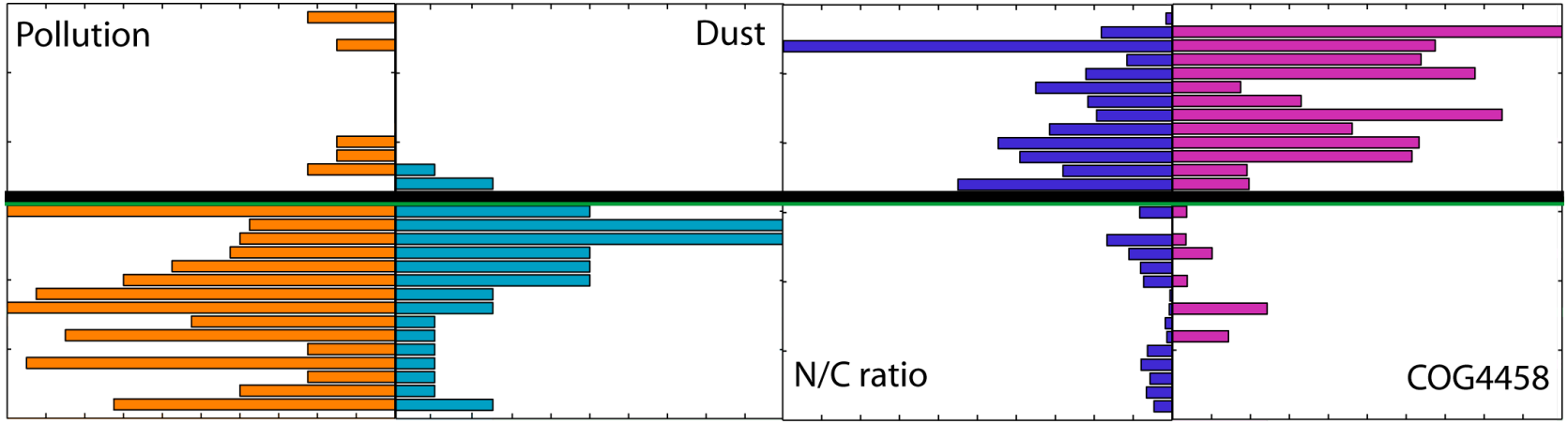
Iron is usually limiting in oceans: High Nitrate-Nutrient/Low Chlorophyll regions
Delivery of iron to is usually by:

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

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

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)

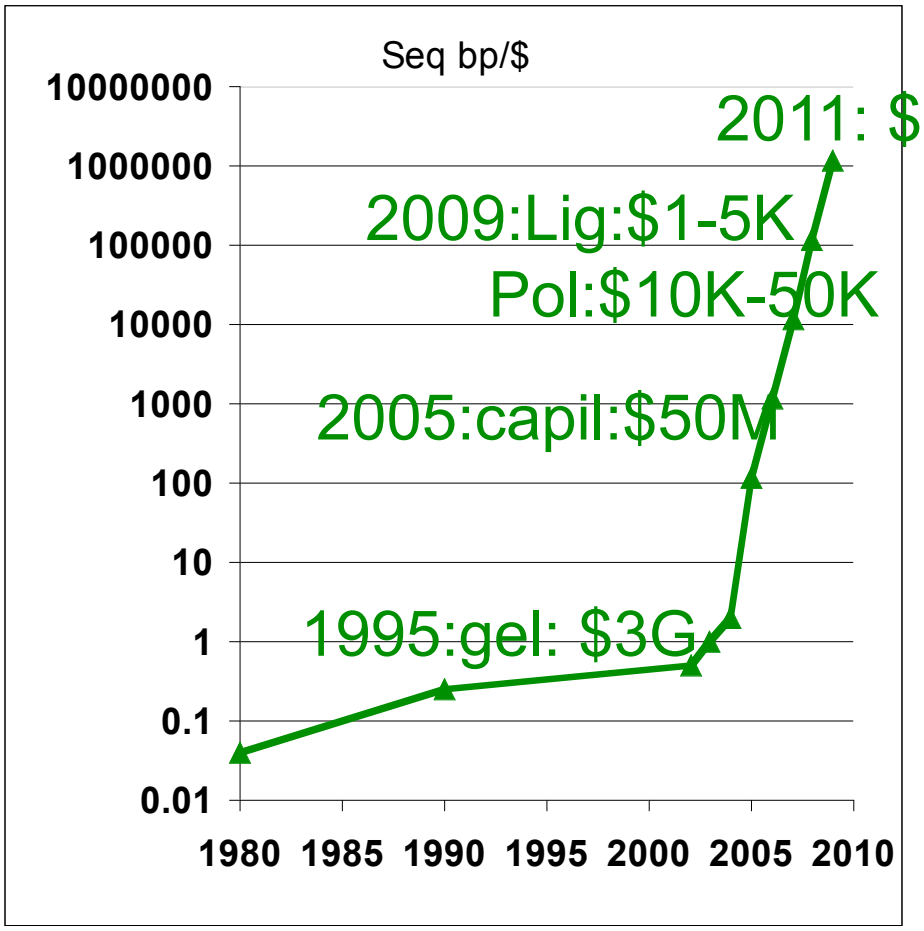
- Searching the BRENDA database for enzymes using iron as a cofactor reveal that an increase in these two COGs negatively correlated to the amount of enzymes present that required iron.

Patel et. al., *Genome Research* 2010

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

(Moore's law)
1.5x/yr for
electronics
vs
10x/yr for DNA
Sequencing

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



Carr and Church, Nat Biotech 2009

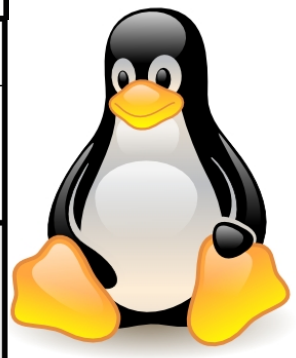
Outline: Molecular Networks

- Why Networks?
- Network Comparisons
  (reg. net. in many organisms)
  - in rel. to social hierarchy
  - scaling in rel. to partnerships
  - Computer OS Comparisons
- 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)

		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

Acknowledge-ments

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

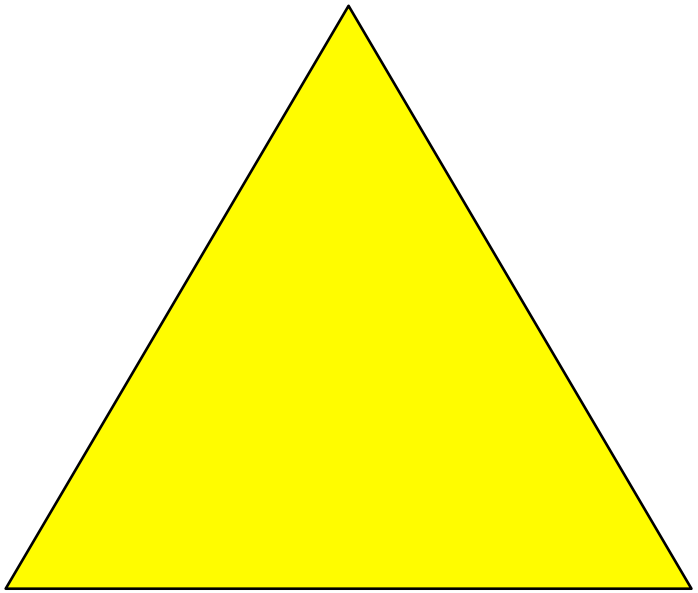
A Paccanaro
K Yip
R Bjornson
G Fang
Y Xia
J Korbel
J Raes
P Bork
D Engelman
M Snyder

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:
New York Academy of Sciences, A Look at the Tools and Comparative Approaches of Systems Biology, 2010.06.10, 17:00-17:40; [I:NYSYSBIO] (Medium-length networks talk, derived from [I:BROWNMATH] with metagenomics updated.)

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 .