Large-scale Analysis of Molecular Networks

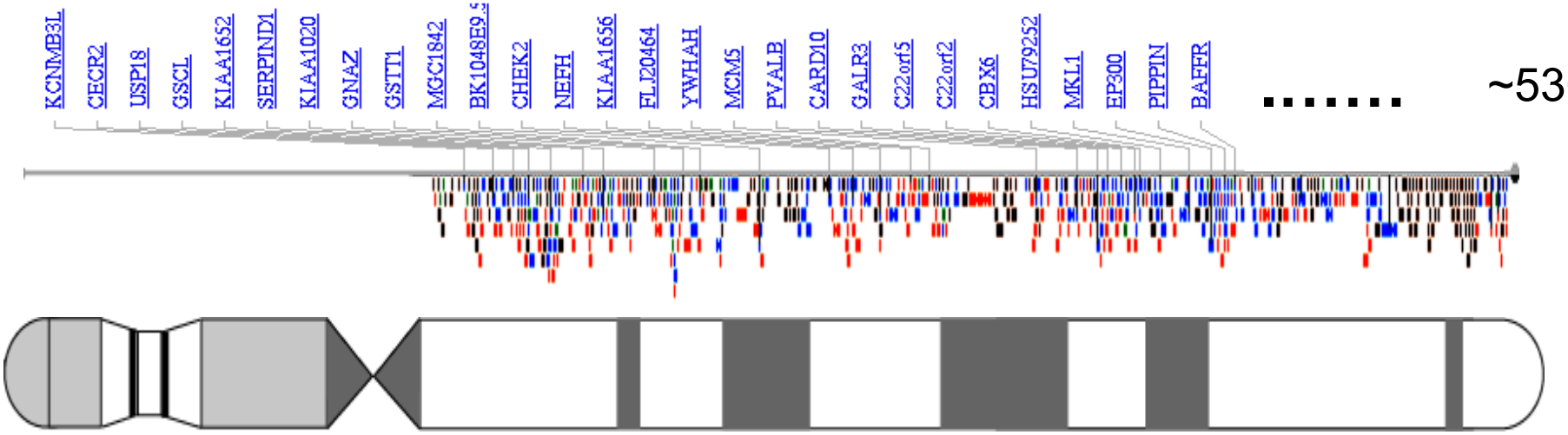
Mark B Gerstein
Yale

slides at

Lectures.GersteinLab.org

(See Last Slide for References & More Info.)

The problem: Grappling with Function on a Genome Scale?



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

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

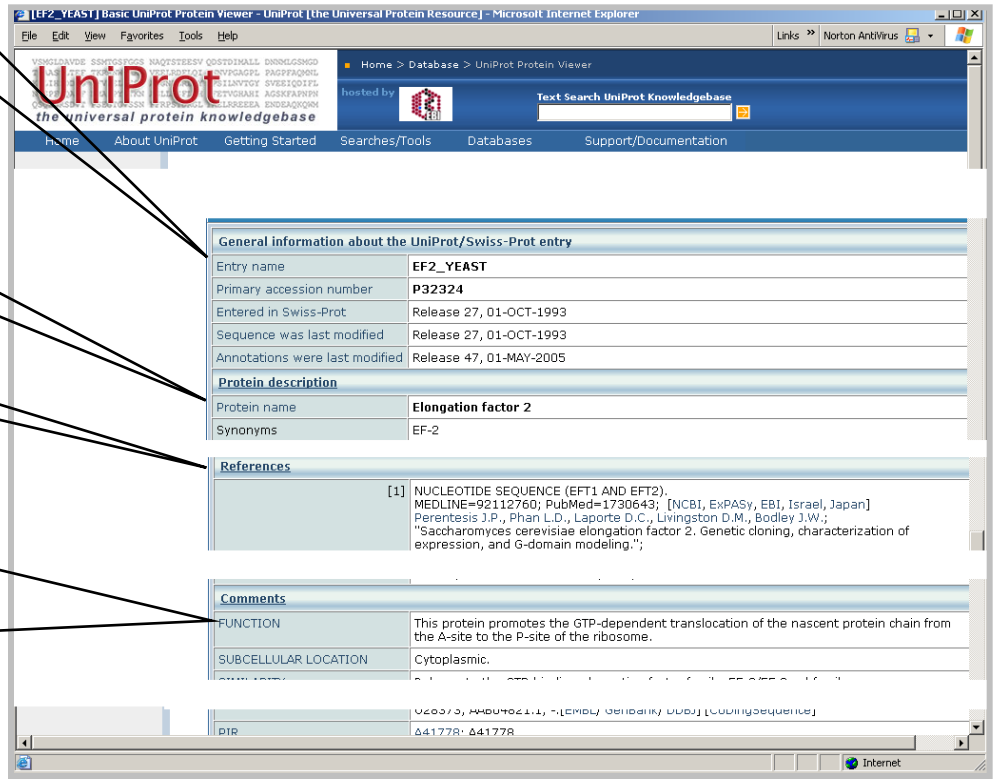
Traditional single molecule way to integrate evidence & describe function

EF2_YEAST

Descriptive Name:
Elongation Factor 2

Lots of references
to papers

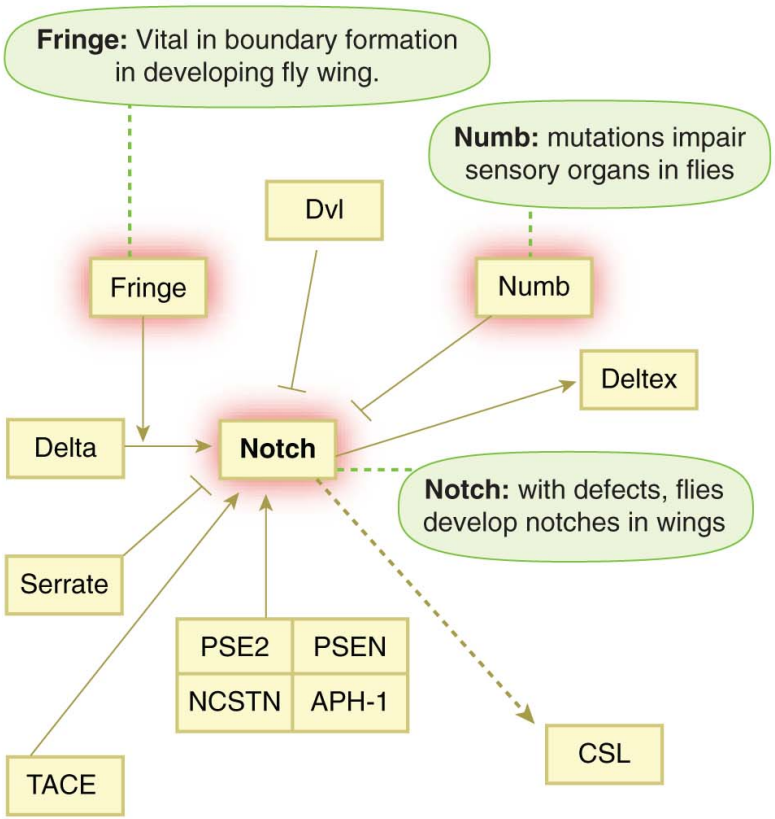
Summary sentence describing function:
This protein promotes the GTP-dependent translocation of the nascent protein chain from the A-site to the P-site of the ribosome.



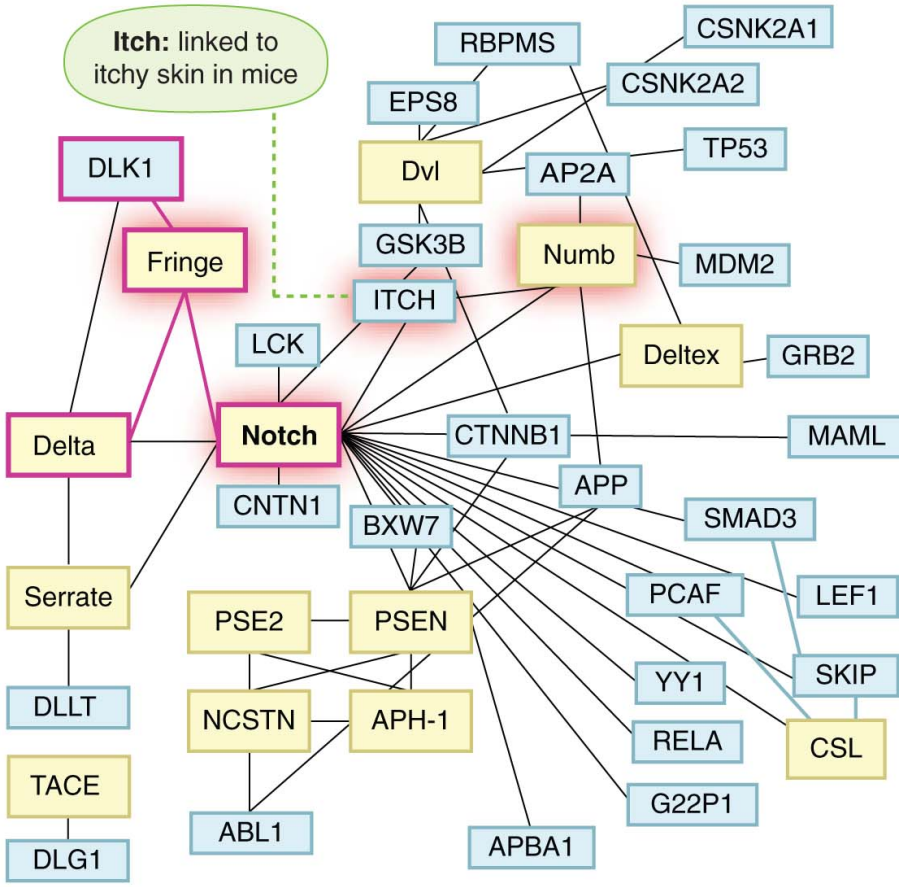
Some obvious issues in scaling single molecule definition to a genomic scale

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

Networks (Old & New)



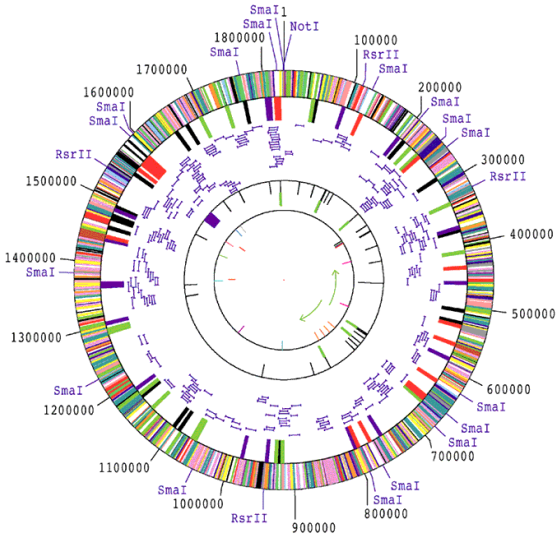
Classical KEGG pathway

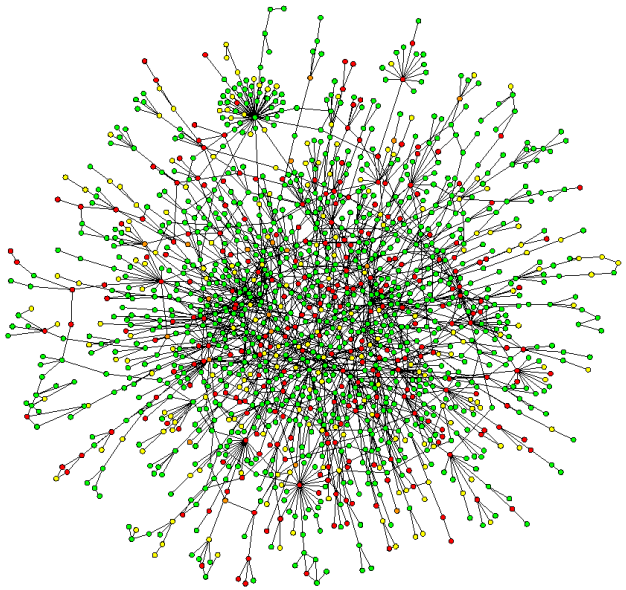


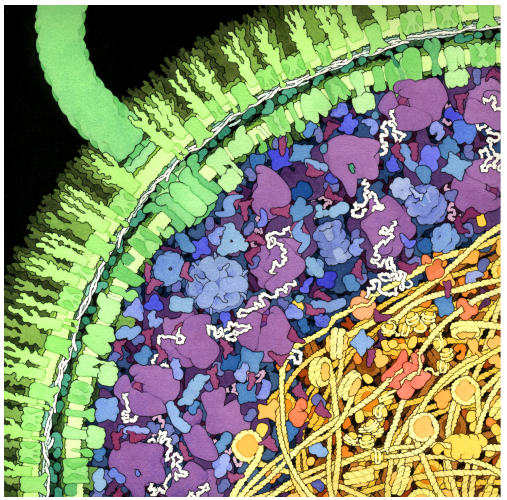
Same Genes in High-throughput Network

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

Networks occupy a midway point in terms of level of understanding







1D: Complete Genetic Partslist

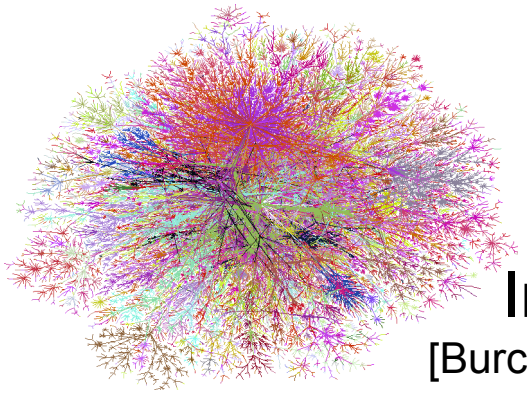
~2D: Bio-molecular Network Wiring Diagram

3D: Detailed structural understanding of cellular machinery

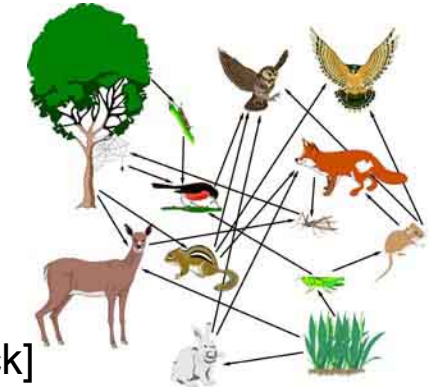
[Fleischmann et al., Science, 269 :496]

[Jeong et al. Nature, 41:411]

Networks as a universal language

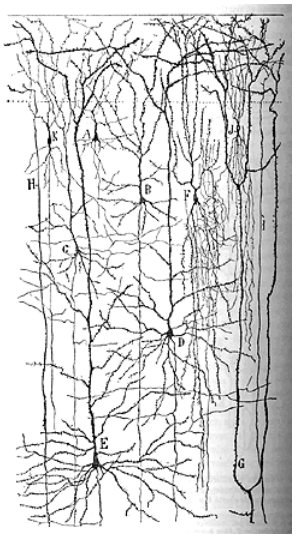


Internet
[Burch & Cheswick]

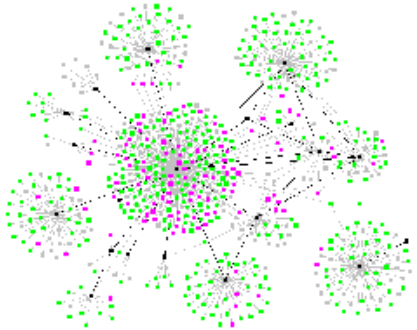


Food Web

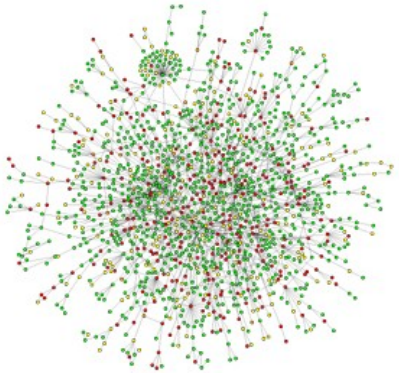
Electronic Circuit



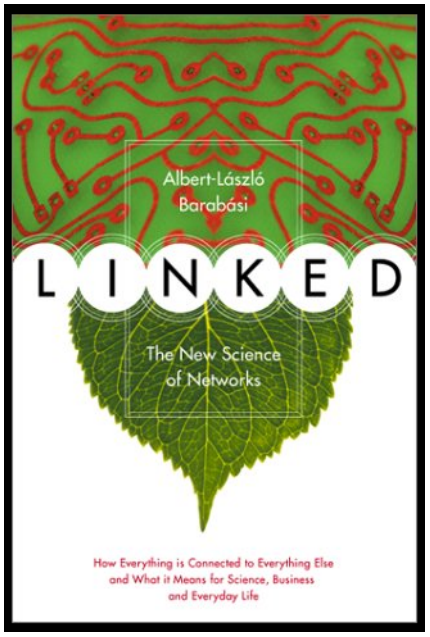
Neural Network
[Cajal]

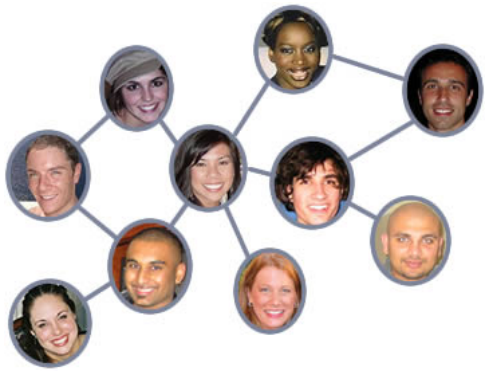


Disease Spread
[Krebs]



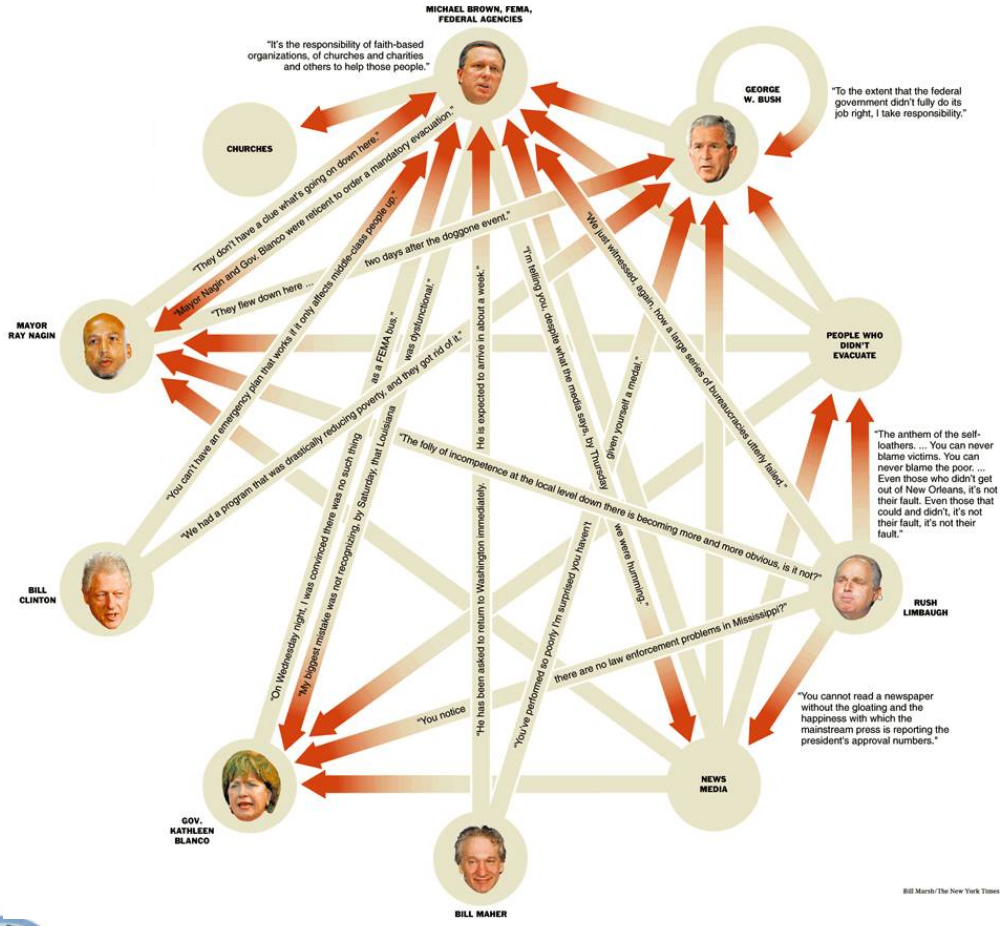
Protein Interactions
[Barabasi]



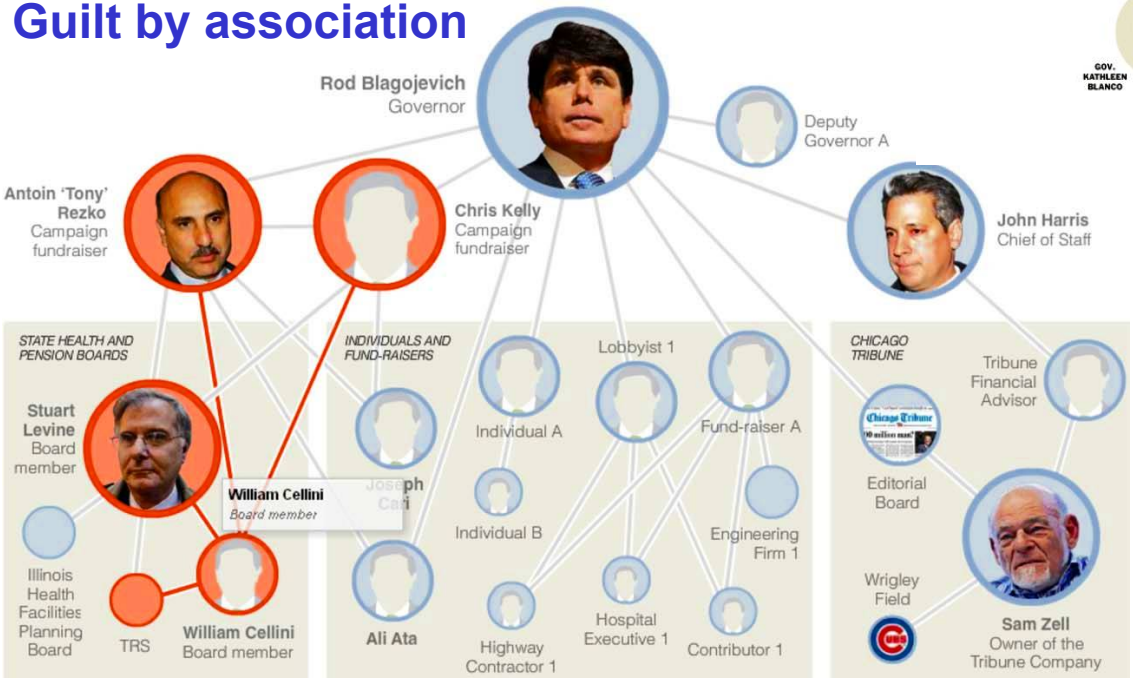


Social Network

Using the position in networks to describe function



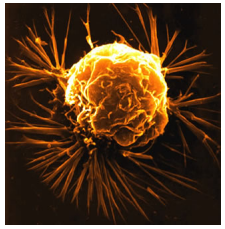
Guilt by association



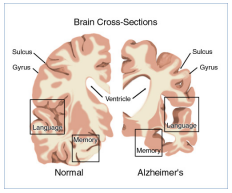
Signal Transduction (the Katrina "Blame Game")

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

Network pathology & pharmacology

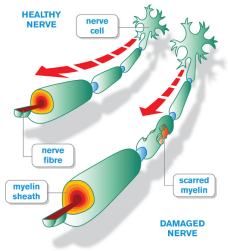


Breast Cancer

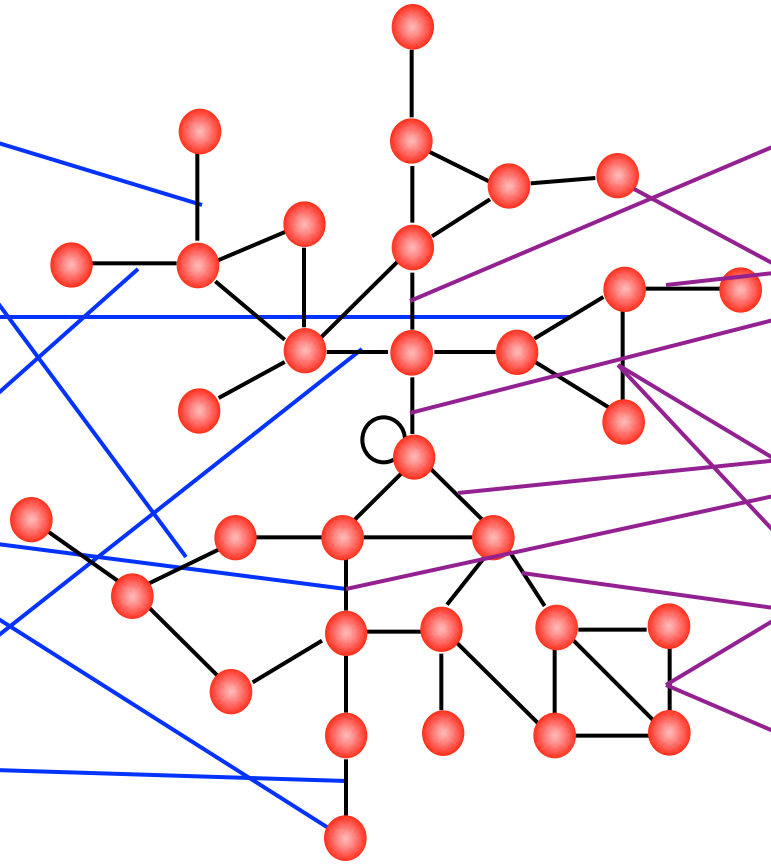


Alzheimer's Disease

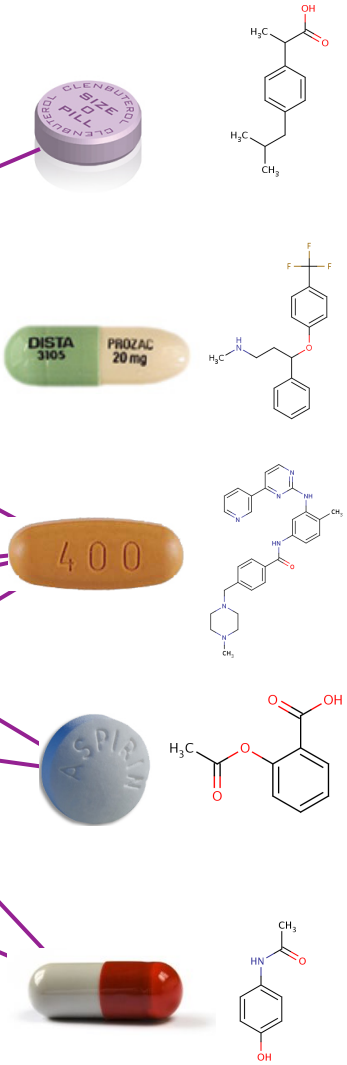
Parkinson's Disease



Multiple Sclerosis



Interactome networks

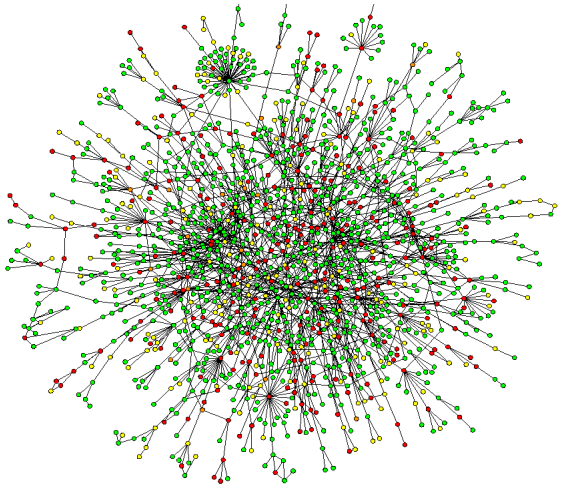


[Adapted from H Yu]

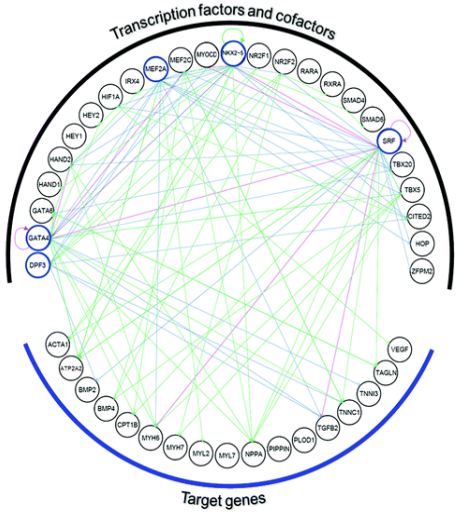
Outline: Molecular Networks

- Why Networks?
- Network Structure: Key Positions
  - ◊ Hubs & Bottlenecks
  - ◊ Tops of a Hierarchy
  - ◊ RE-score nodes
- Networks, Variation & the Environment
  - ◊ Which pathways change most with the environment

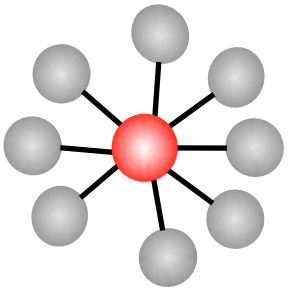
Different Types of Molecular Networks



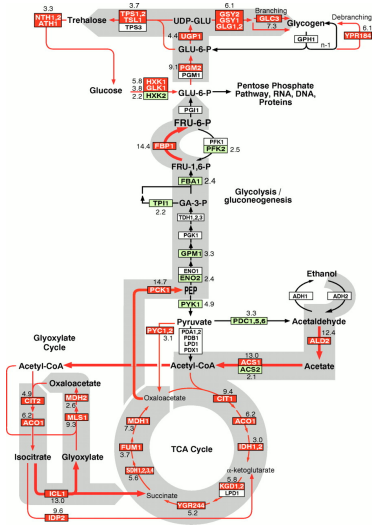
Protein-protein Interaction networks



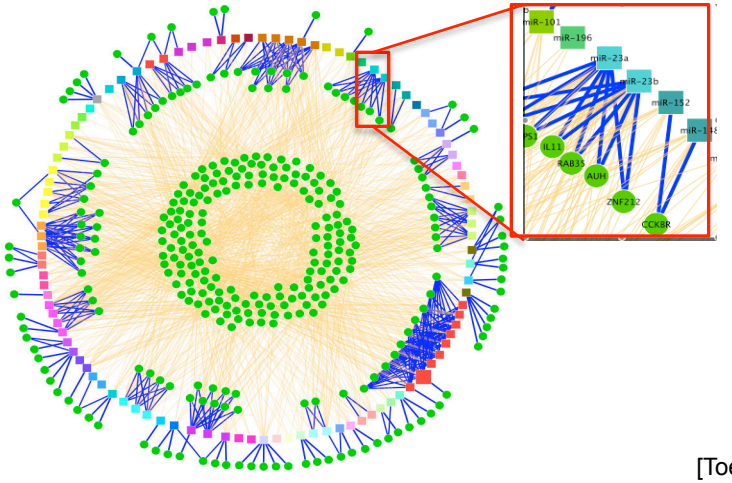
TF-target-gene Regulatory networks



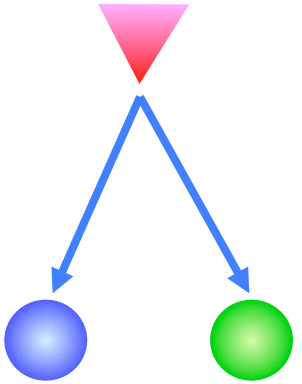
Undirected



Metabolic pathway networks



miRNA-target networks

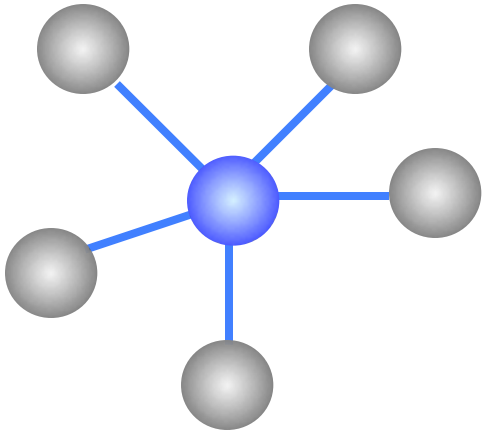


Directed

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

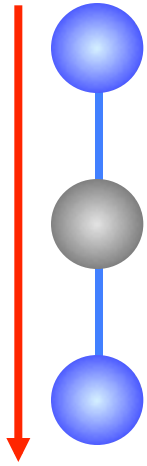
Global topological measures

Indicate the gross topological structure of the network



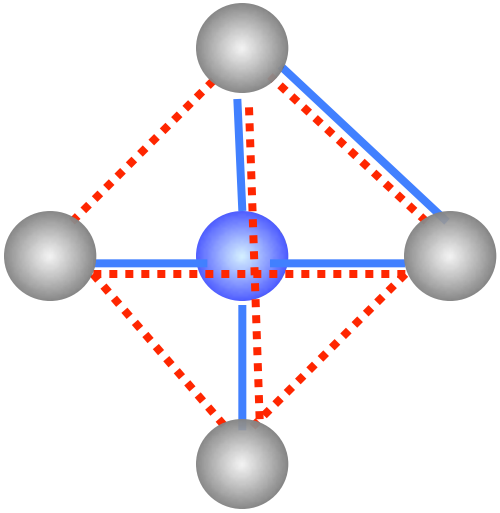
Degree (K)

5



Path length (L)

2



Clustering coefficient (C)

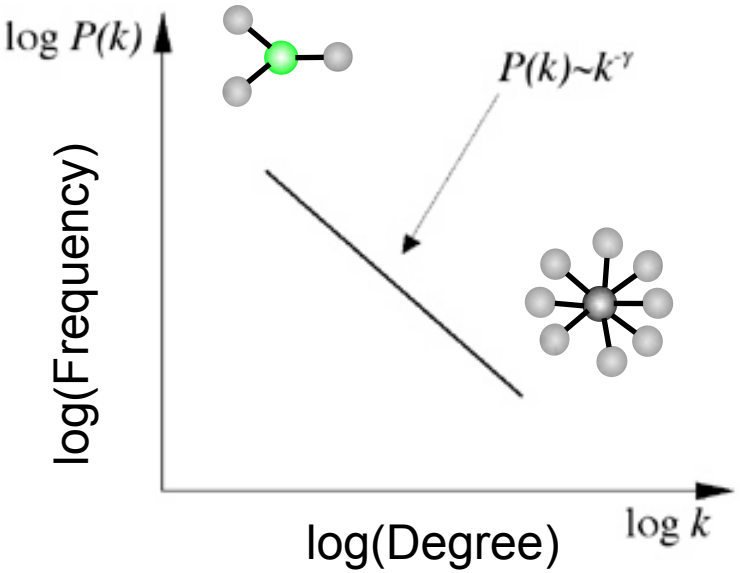
1/6

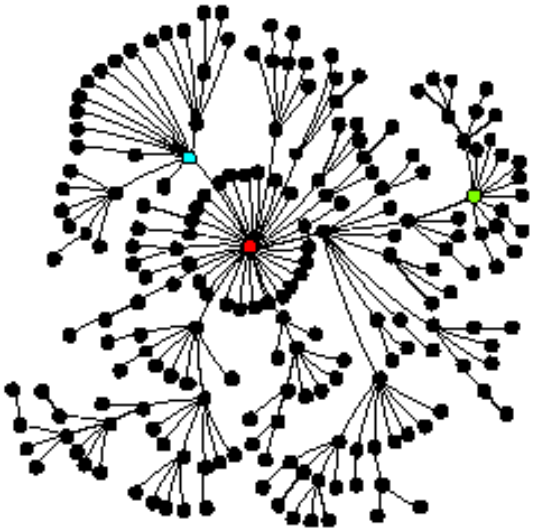
Interaction and expression networks are ***undirected***

[Barabasi]

Scale-free networks

Power-law distribution





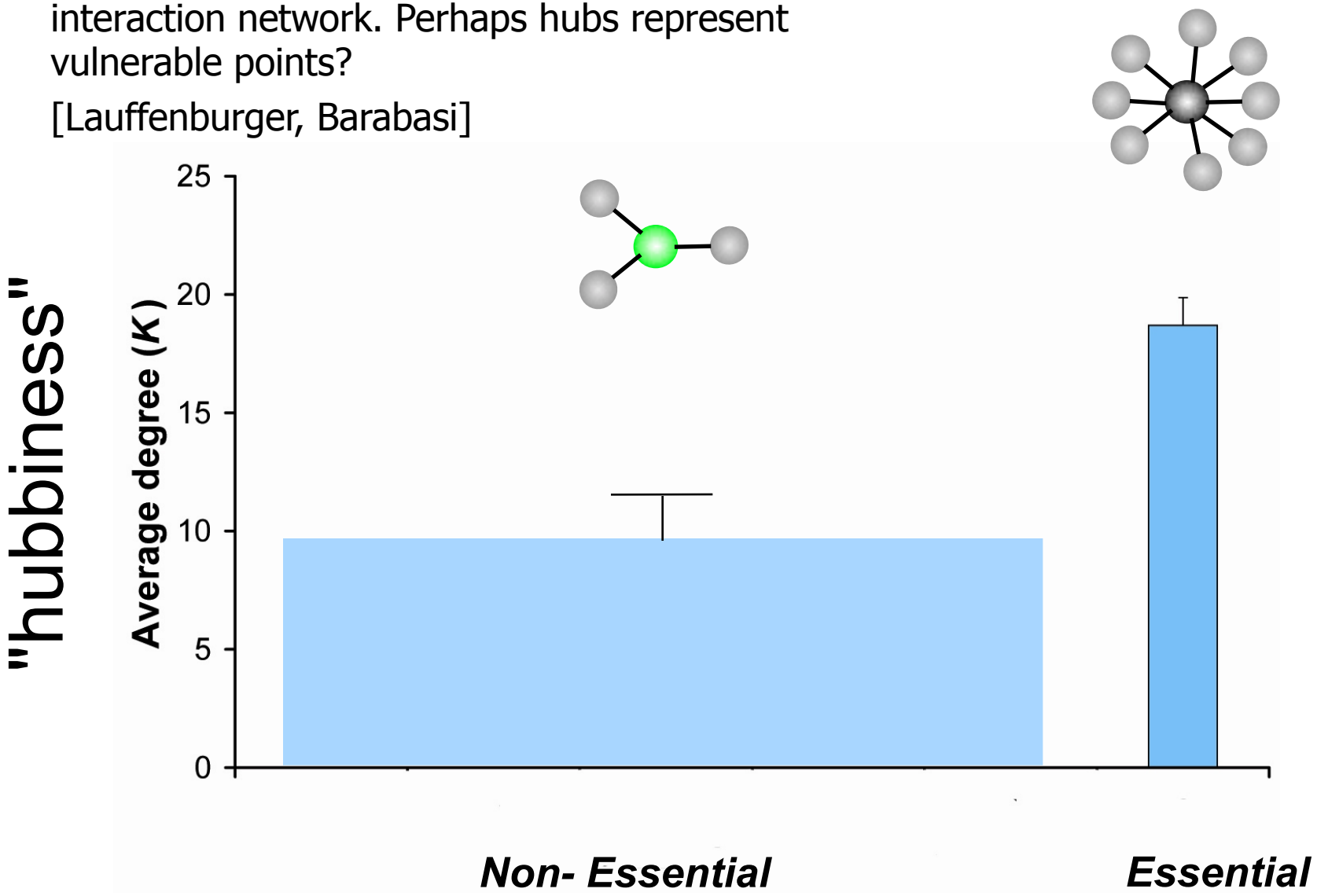
Hubs dictate the structure of the network

[Barabasi]

Hubs tend to be Essential

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

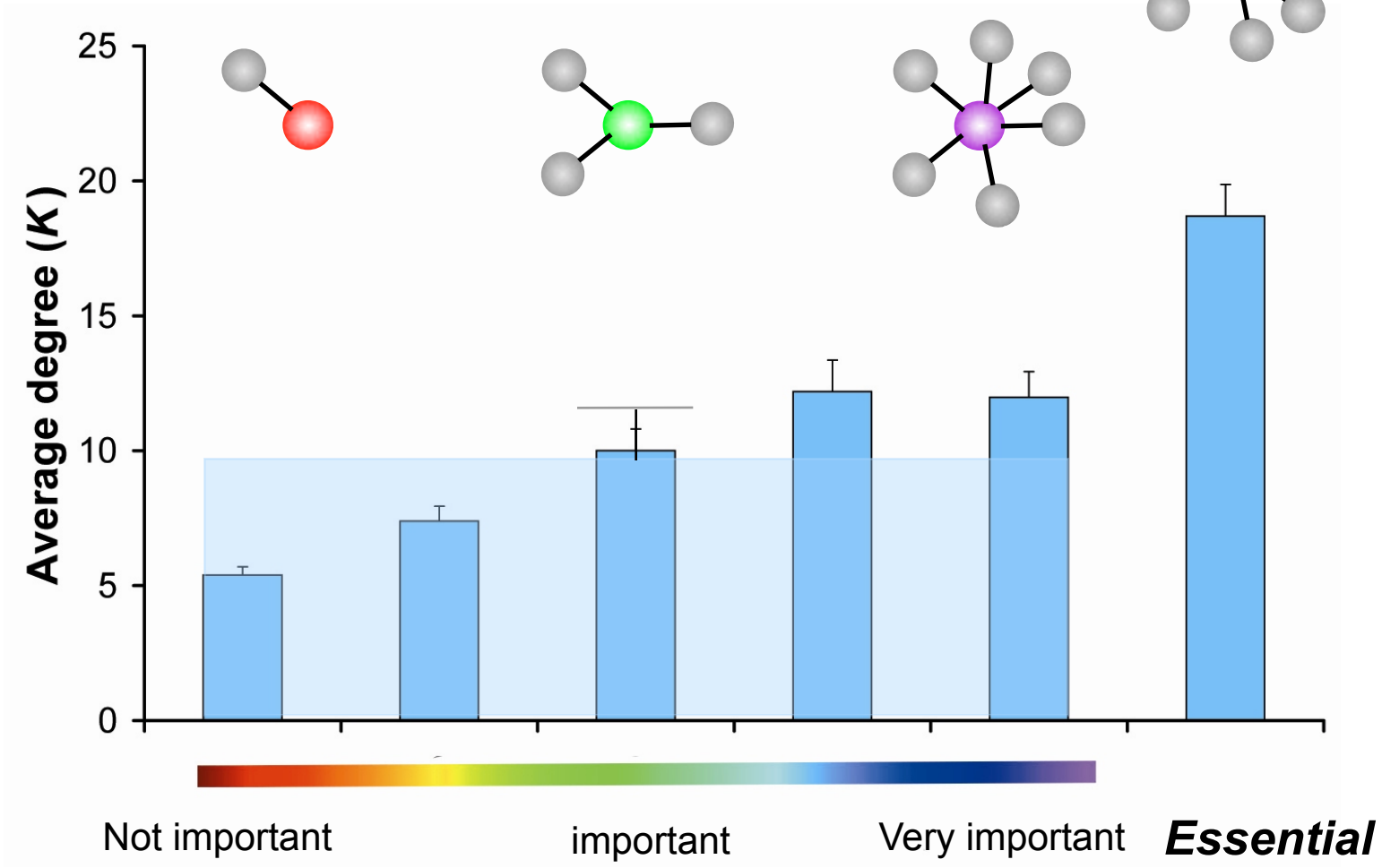
[Lauffenburger, Barabasi]



[Yu et al., 2003, TIG]

Relationships extends to "Marginal Essentiality"

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



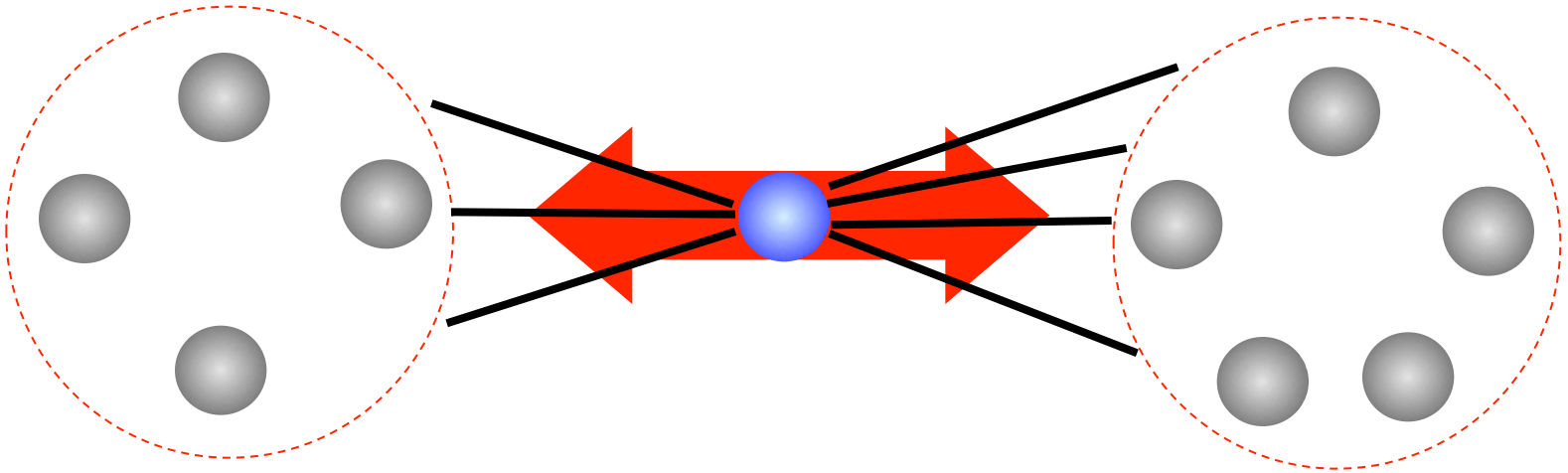
[Yu et al., 2003, TIG]

Another measure of Centrality: Betweenness centrality

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

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

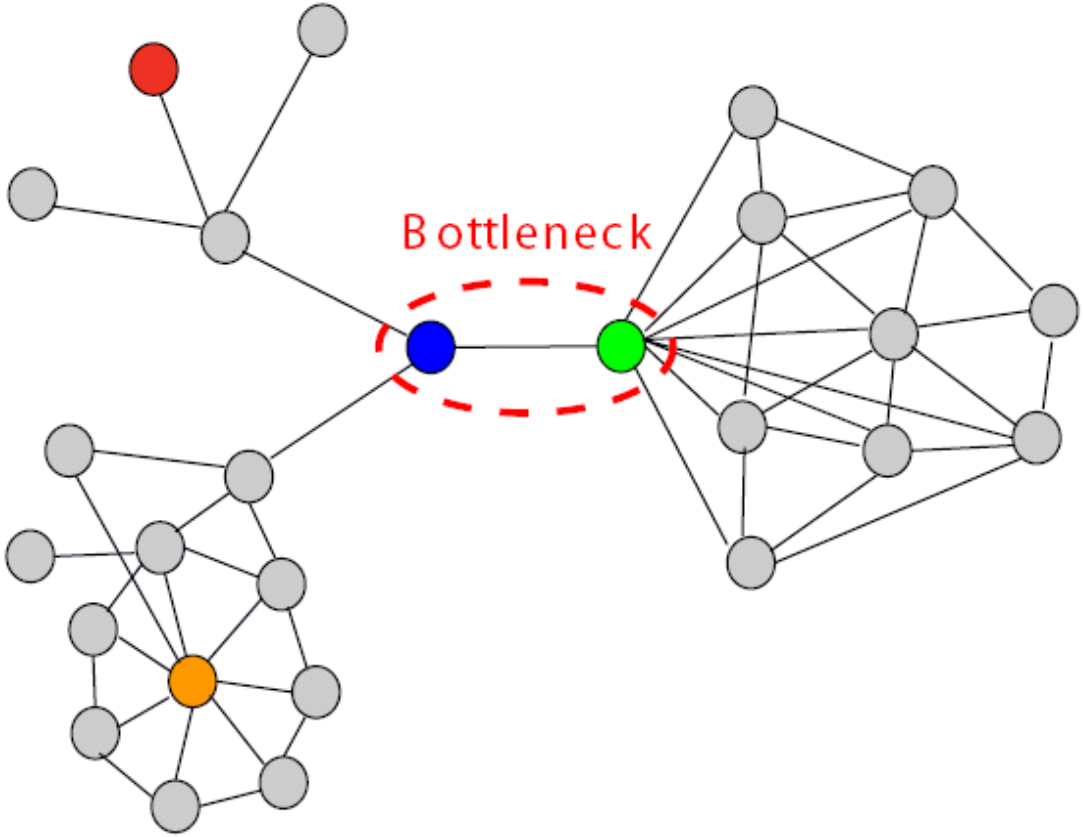
Girvan & Newman (2002) PNAS 99: 7821.



Betweenness centrality -- Bottlenecks

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

Bottlenecks & Hubs



Hub-bottleneck **node**

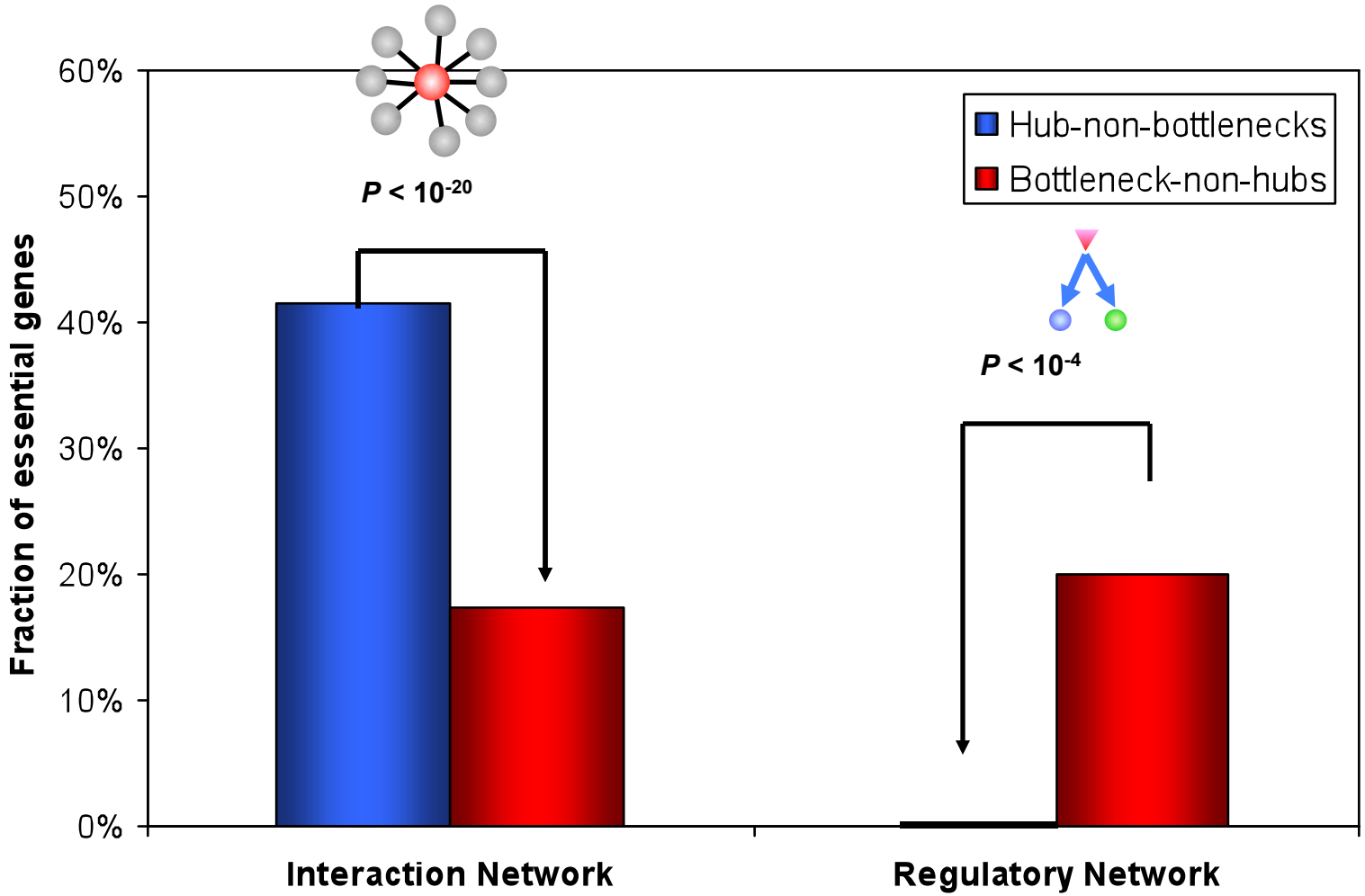
Non-hub-bottleneck **node**

Hub-non-bottleneck **node**

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

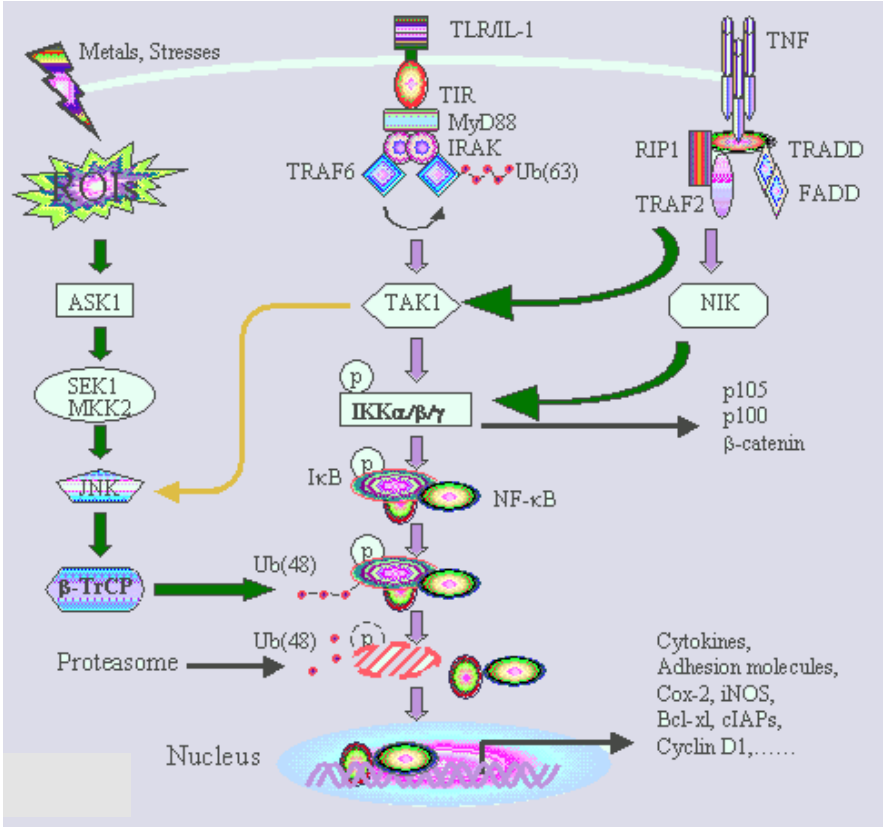
[Yu et al., PLOS CB (2007)]

Bottlenecks are what matters in regulatory networks



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

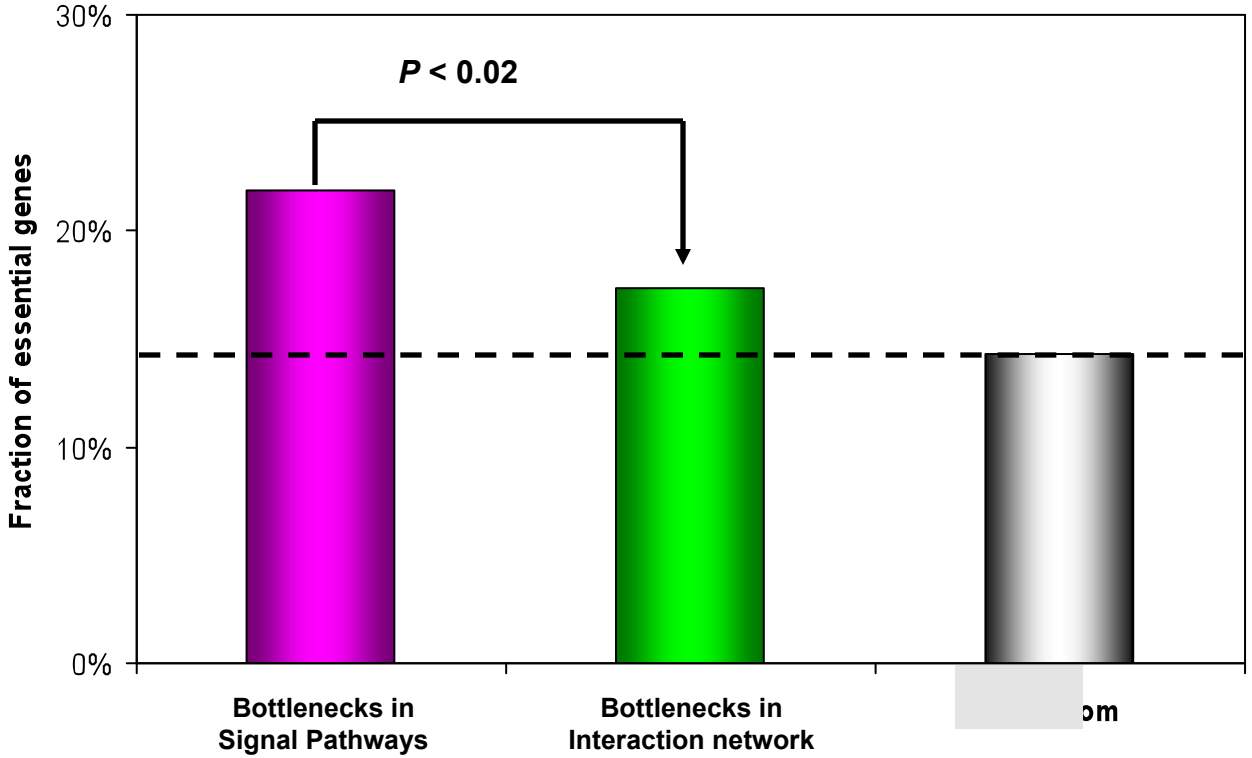
Signaling transduction pathways are directed

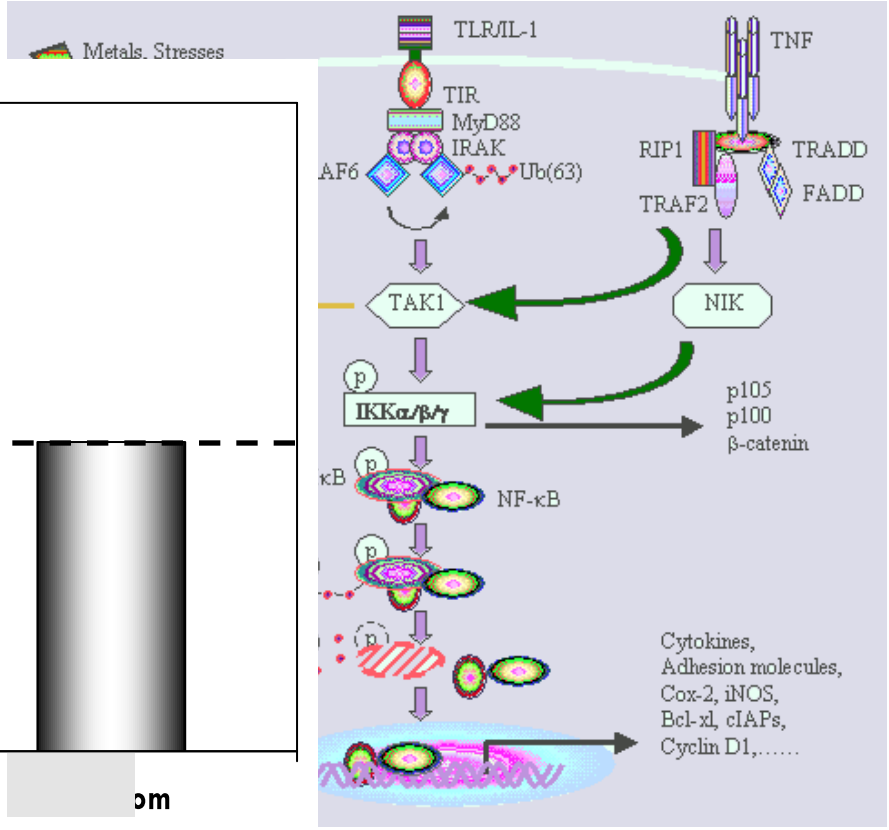


[Xianglin Shi]

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

Bottlenecks in signaling pathways are important





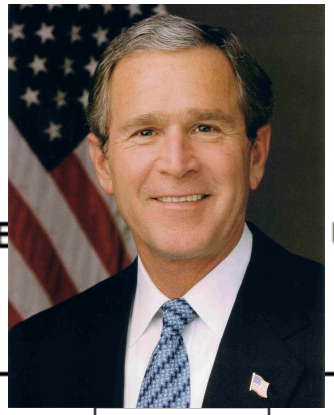
[Xianglin Shi]

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

Outline: Molecular Networks

- Why Networks?
- Network Structure: Key Positions
  - ◊ Hubs & Bottlenecks
  - ◊ Tops of a Hierarchy
  - ◊ RE-score nodes
- Networks, Variation & the Environment
  - ◊ Which pathways change most with the environment

Social Hierarchy



THE GOVE UNITED STATES

LEGISLATIVE BRANCH

THE CONGRESS

SENATE HOUSE

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

EX SIDENT

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

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

JUDICIAL BRANCH

THE SUPREME COURT OF THE UNITED STATES

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














INDEPENDENT ESTABLISHMENTS AND GOVERNMENT CORPORATIONS

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

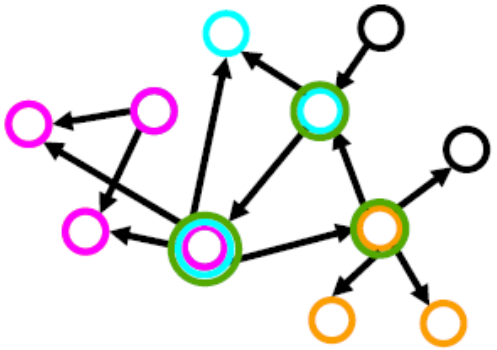
FEDERAL LABOR RELATIONS AUTHORITY
FEDERAL MARITIME COMMISSION
FEDERAL MEDIATION AND CONCILIATION SERVICE
FEDERAL MINE SAFETY AND HEALTH REVIEW COMMISSION
FEDERAL RESERVE SYSTEM
FEDERAL RETIREMENT THRIFT INVESTMENT BOARD
FEDERAL TRADE COMMISSION
GENERAL SERVICES ADMINISTRATION
INTER-AMERICAN FOUNDATION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

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

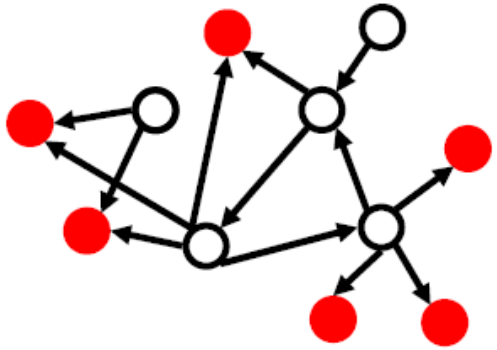
POSTAL RATE COMMISSION
RAILROAD RETIREMENT BOARD
SECURITIES AND EXCHANGE COMMISSION
SELECTIVE SERVICE SYSTEM
SMALL BUSINESS ADMINISTRATION
SOCIAL SECURITY ADMINISTRATION
TENNESSEE VALLEY AUTHORITY
TRADE AND DEVELOPMENT AGENCY
U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT
U.S. COMMISSION ON CIVIL RIGHTS
U.S.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

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

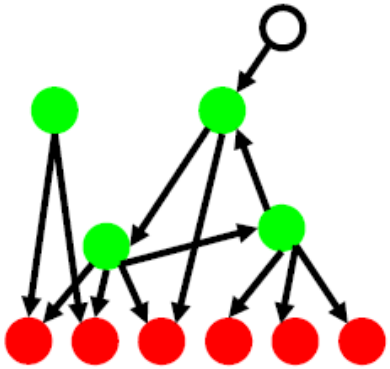
I. Example network with all 4 motifs



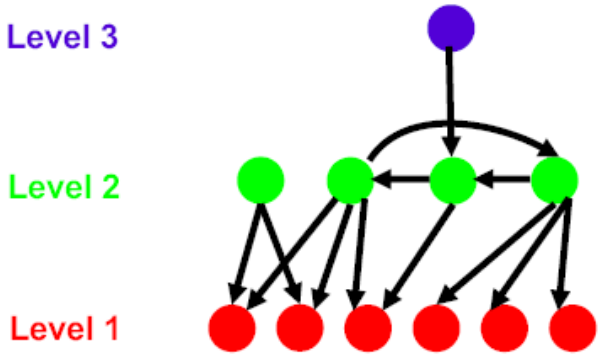
II. Finding terminal nodes (Red)



III. Finding mid-level nodes (Green)

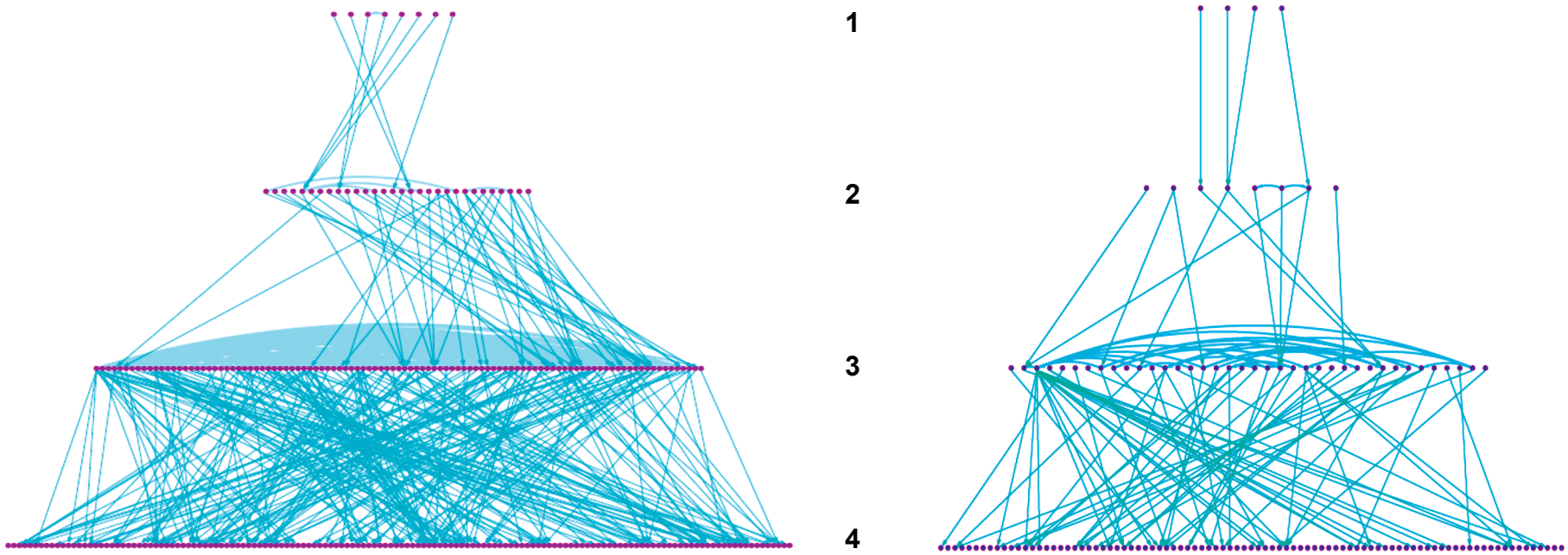


IV. Finding top-most nodes (Blue)



[Yu et al., PNAS (2006)]

Regulatory Networks have similar hierarchical structures

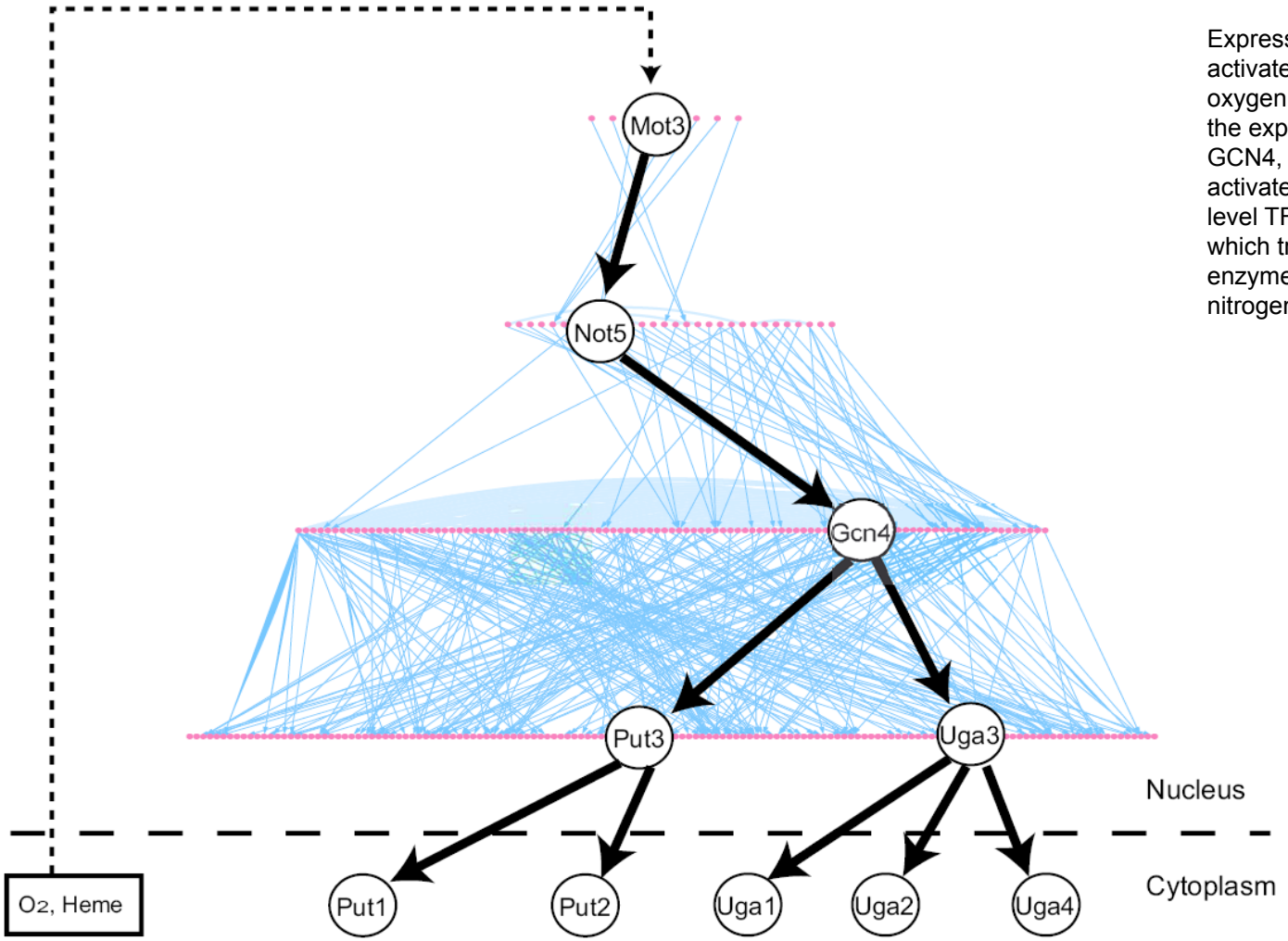


S. cerevisiae

E. coli

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

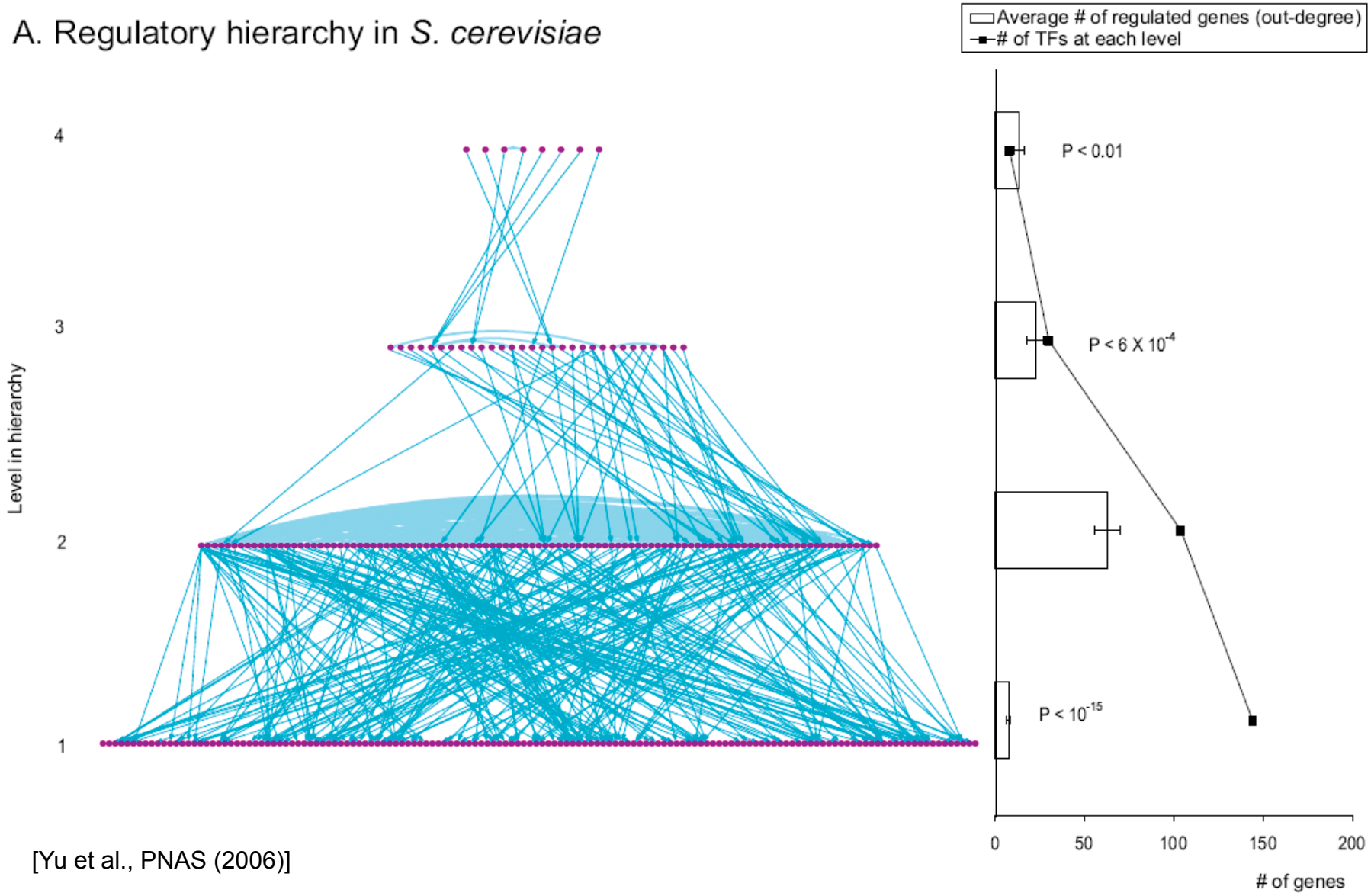
Example of Path Through Regulatory Network



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

[Yu et al., PNAS (2006)]

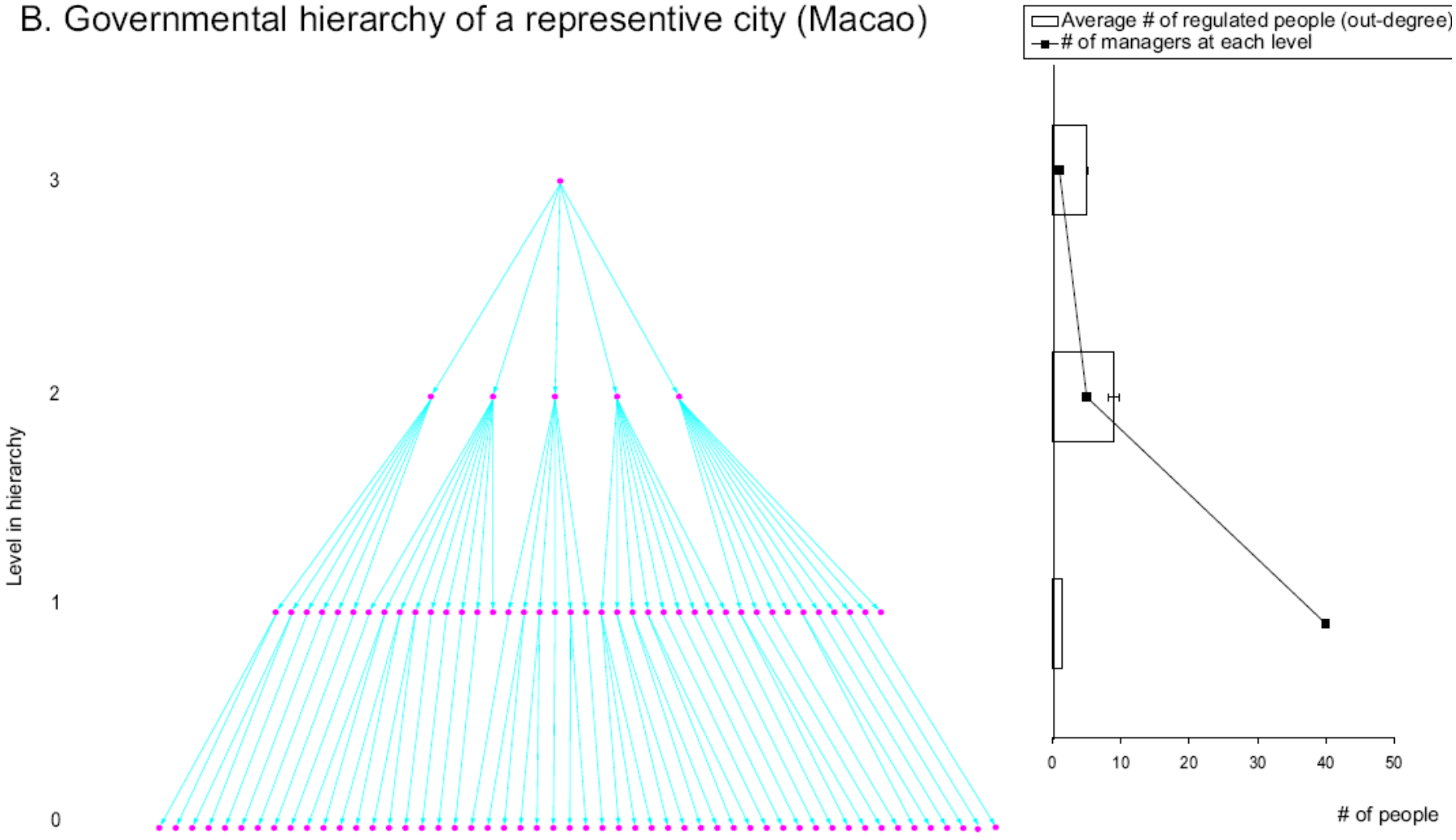
Yeast Regulatory Hierarchy: the Middle-managers Rule



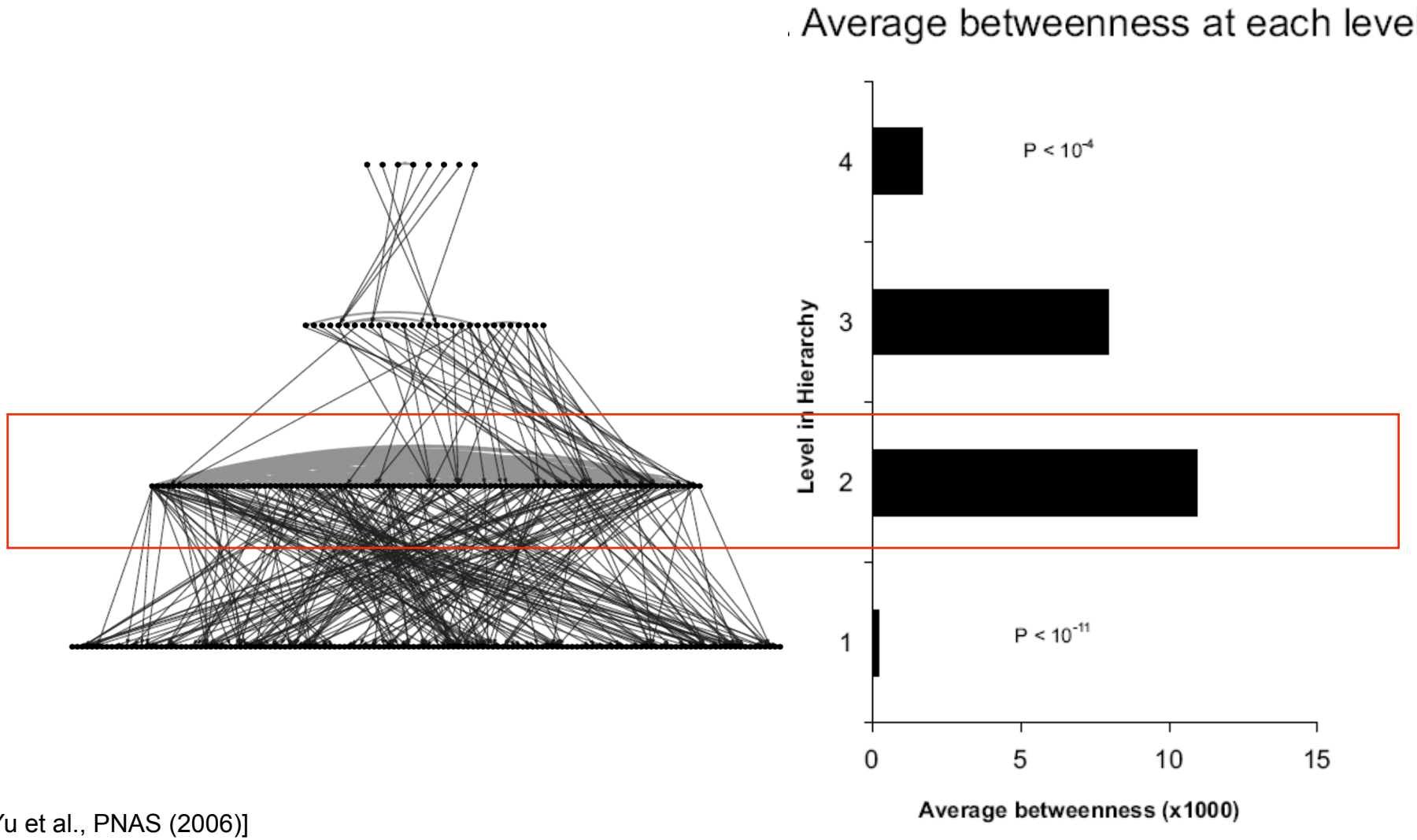
[Yu et al., PNAS (2006)]

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

B. Governmental hierarchy of a representive city (Macao)

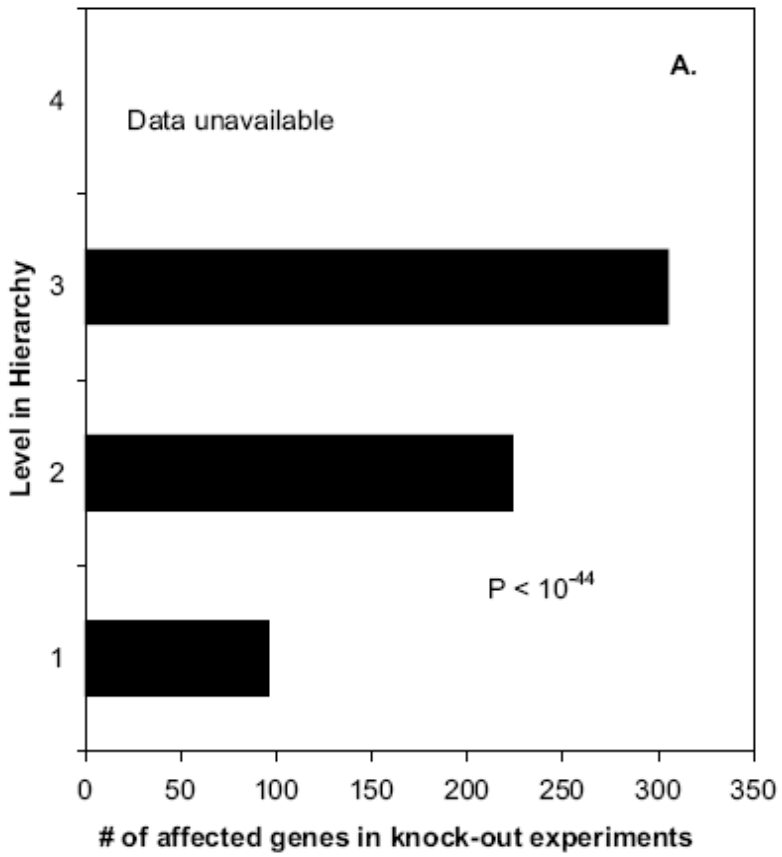


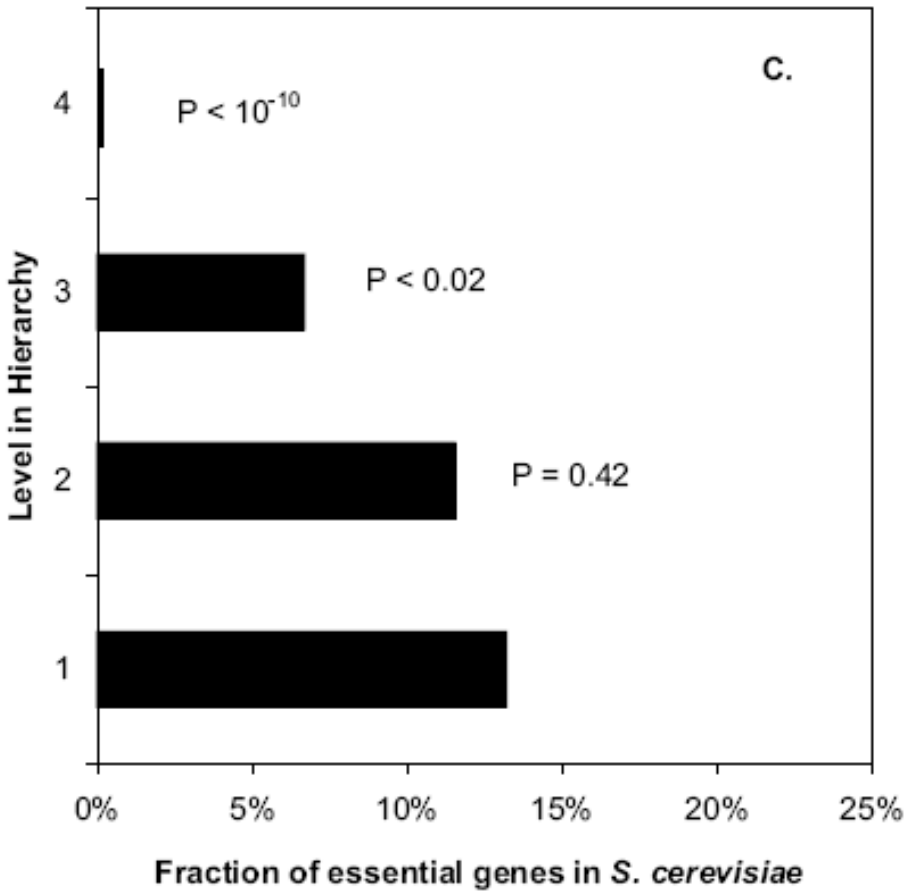
Characteristics of Regulatory Hierarchy: Middle Managers are Information Flow Bottlenecks



[Yu et al., PNAS (2006)]

Characteristics of Regulatory Hierarchy: The Paradox of Influence and Essentiality





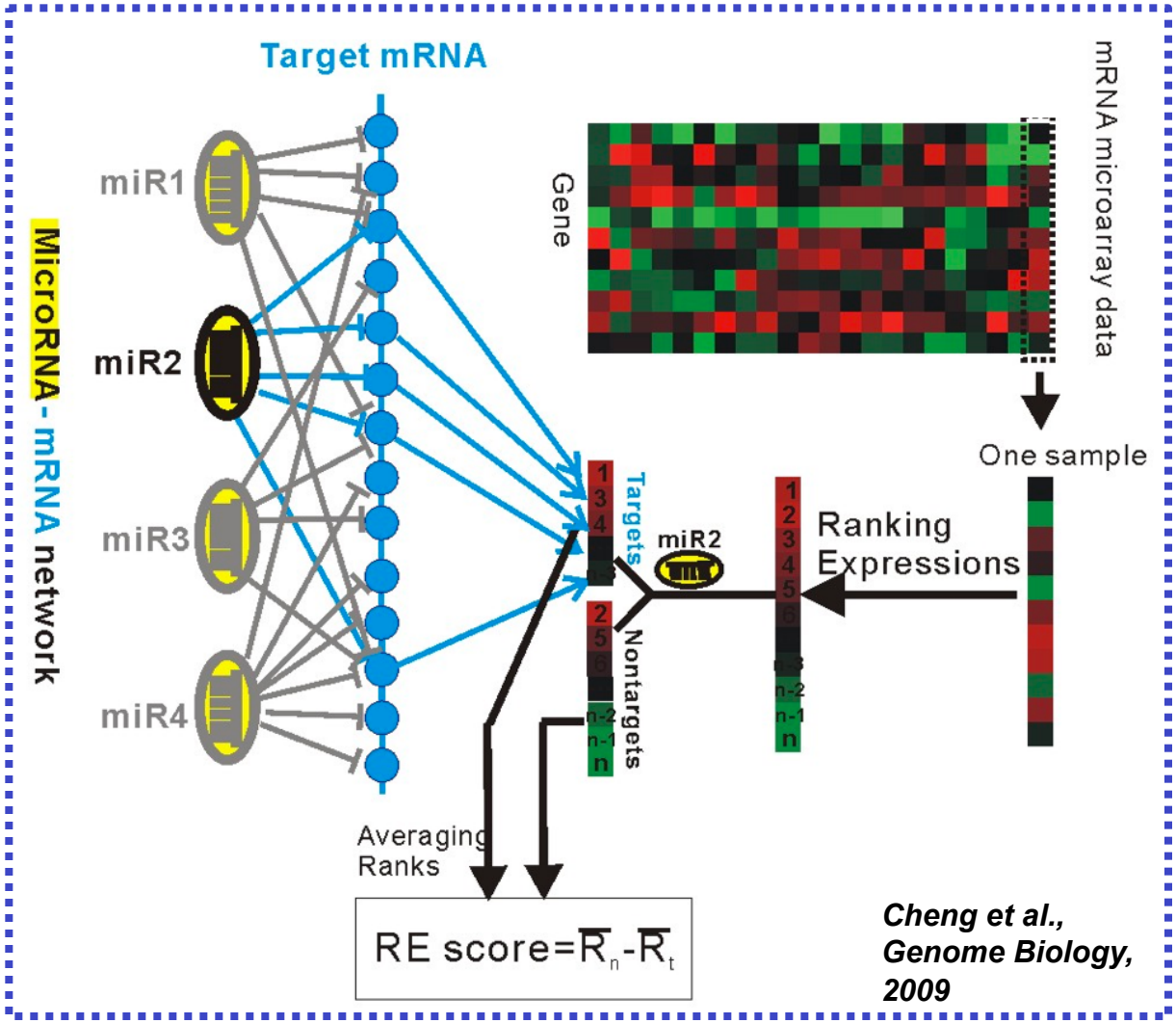
[Yu et al., PNAS (2006)]

Outline: Molecular Networks

- Why Networks?
- Network Structure: Key Positions
  - ◊ Hubs & Bottlenecks
  - ◊ Tops of a Hierarchy
  - ◊ RE-score nodes
- Networks, Variation & the Environment
  - ◊ Which pathways change most with the environment

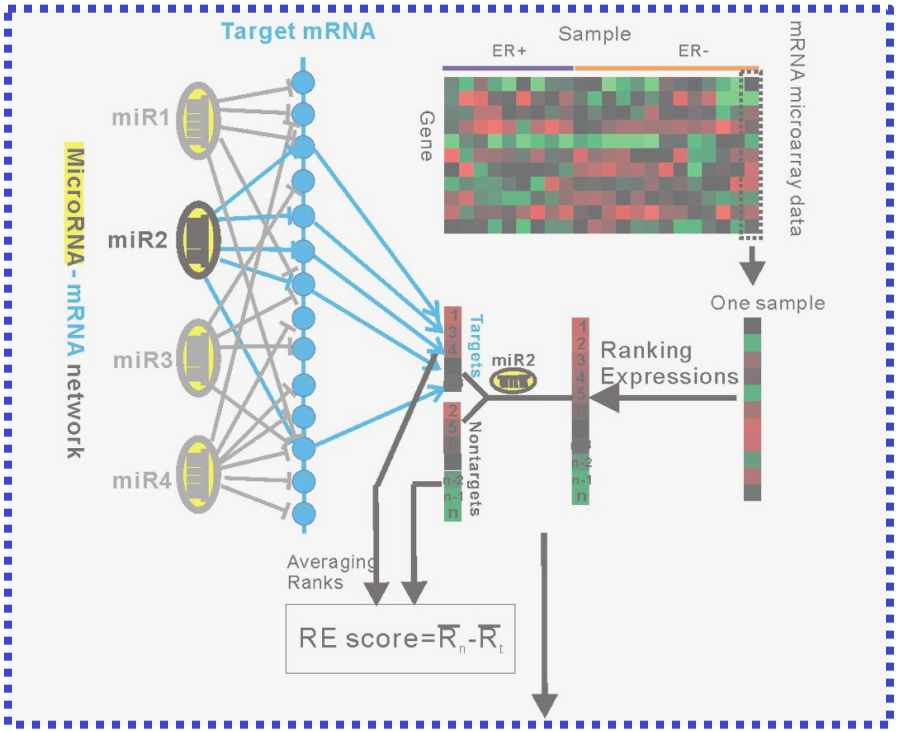
- How much does a regulator influence its targets?
- For miRNA-target networks easy to calculate, as all influence is down-regulation
  - ◊ target prediction via: TargetScan, PITA, PicTar, miRanda, …
- Look at down-reg. genes in a sample & compare with targets of a specific micro-RNA
  - ◊ more down-reg genes => stronger regulatory effect

RE-score: Another way to identify "important" network nodes

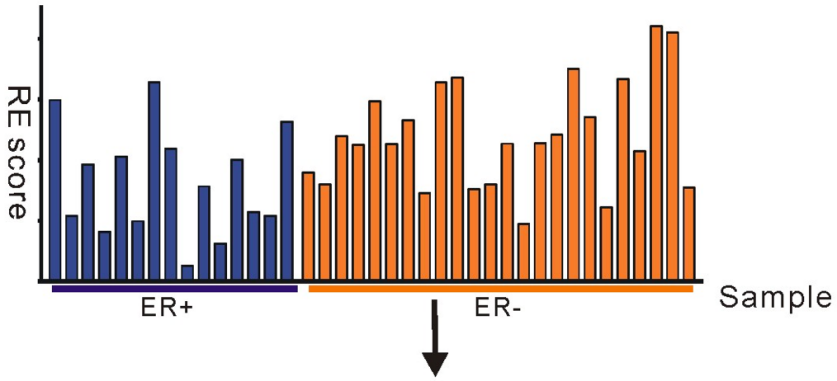


Application of RE-score to measure changing miRNA effect in different conditions

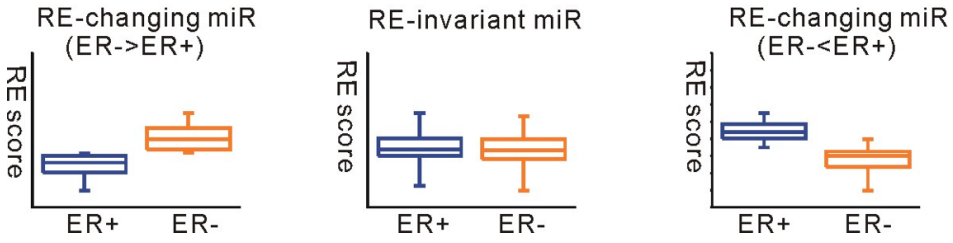
(ER- and ER+ breast cancer)



Calculating RE scores of a miRNA in each sample



Comparing the RE scores between ER+ and ER-

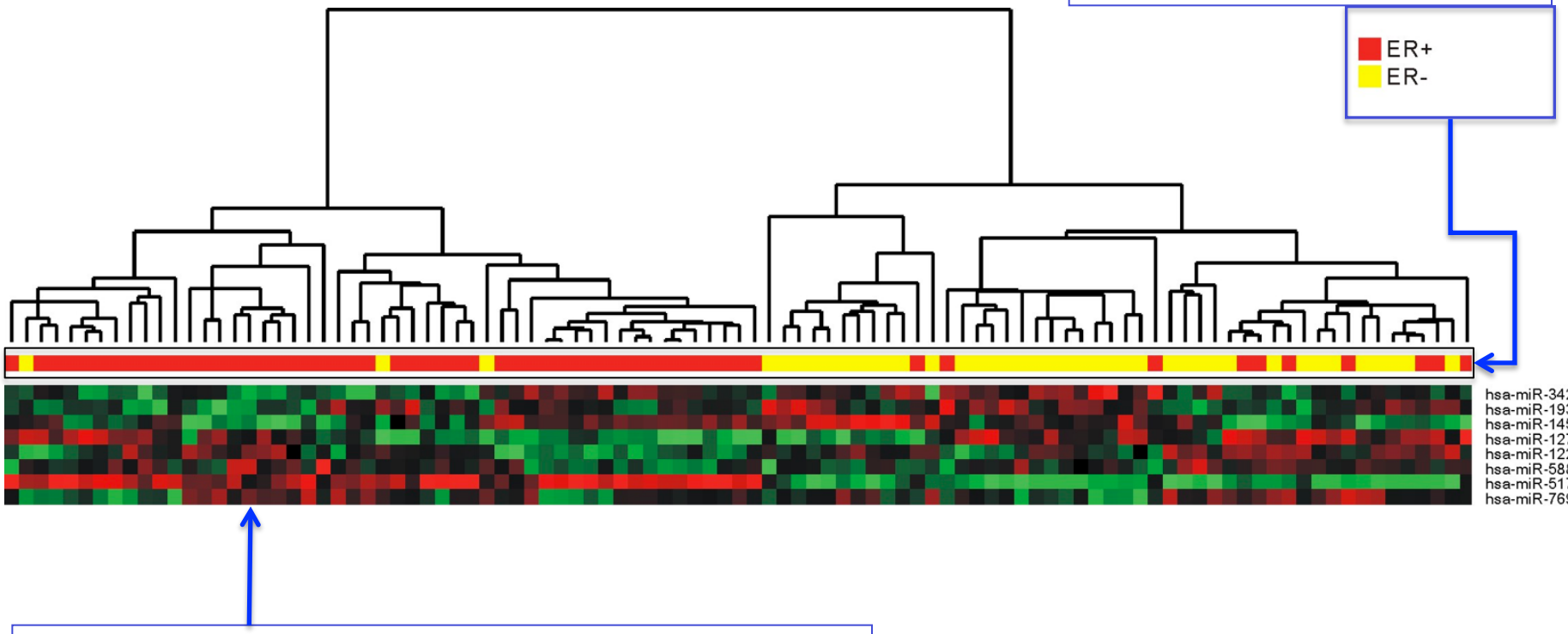


Cheng et al., Genome Biology, 2009

RE-score can be used to classify cancers

(3) Clustering based on RE score divides samples into 2 main types of cancer

(4) Clustering better than based on indiv. gene expression levels



(1) RE-score profile for diff. miRNA in 1 cancer sample.
(2) Tabulate over many different breast cancer samples

Cheng et al., Genome Biology, 2009

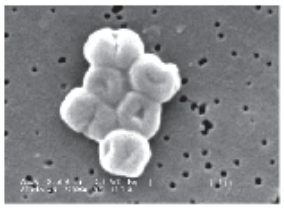
Outline: Molecular Networks

- Why Networks?
- Network Structure: Key Positions
  - ◊ Hubs & Bottlenecks
  - ◊ Tops of a Hierarchy
  - ◊ RE-score nodes
- Networks, Variation & the Environment
  - ◊ Which pathways change most with the environment

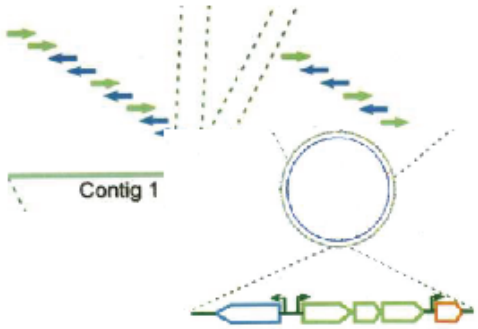
What is metagenomics?

Genomics Approach

Culture Microbes → Extract DNA → Sequence → Assemble and Annotate



ATCGTATA
CGCGAAG
ACGTCTGA
AGTGCTGCT

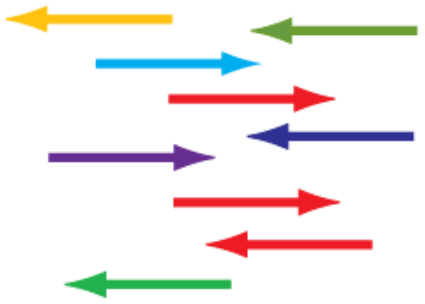


PROBLEM: Estimated that less than 1% can be cultured in the lab

Metagenomics Approach

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

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

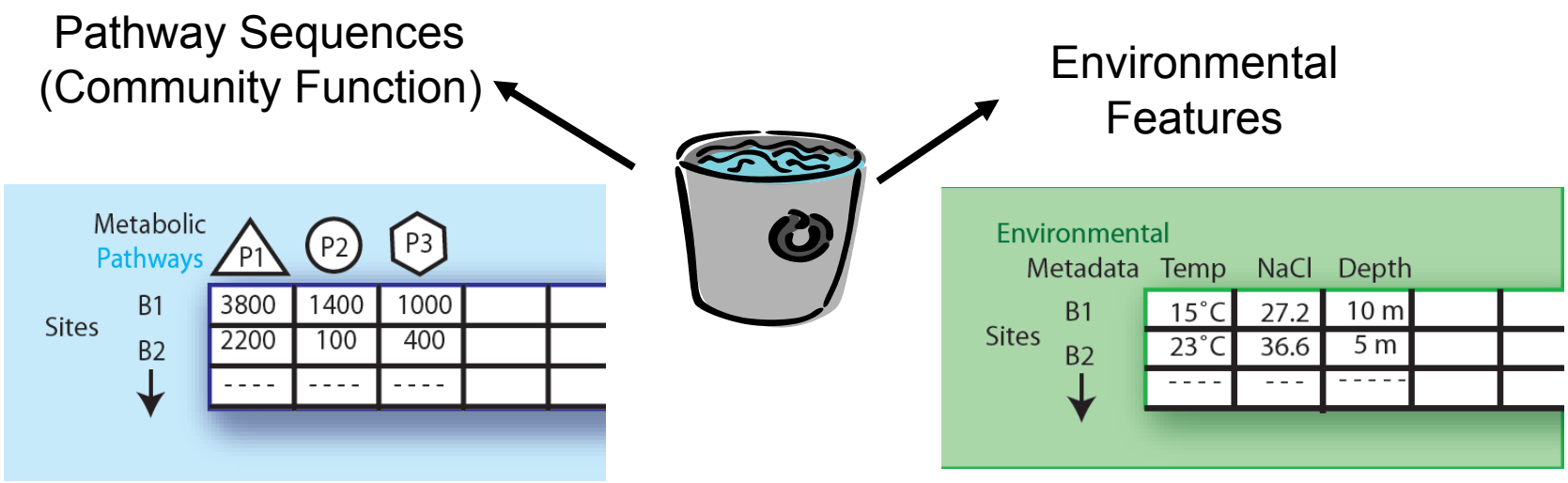


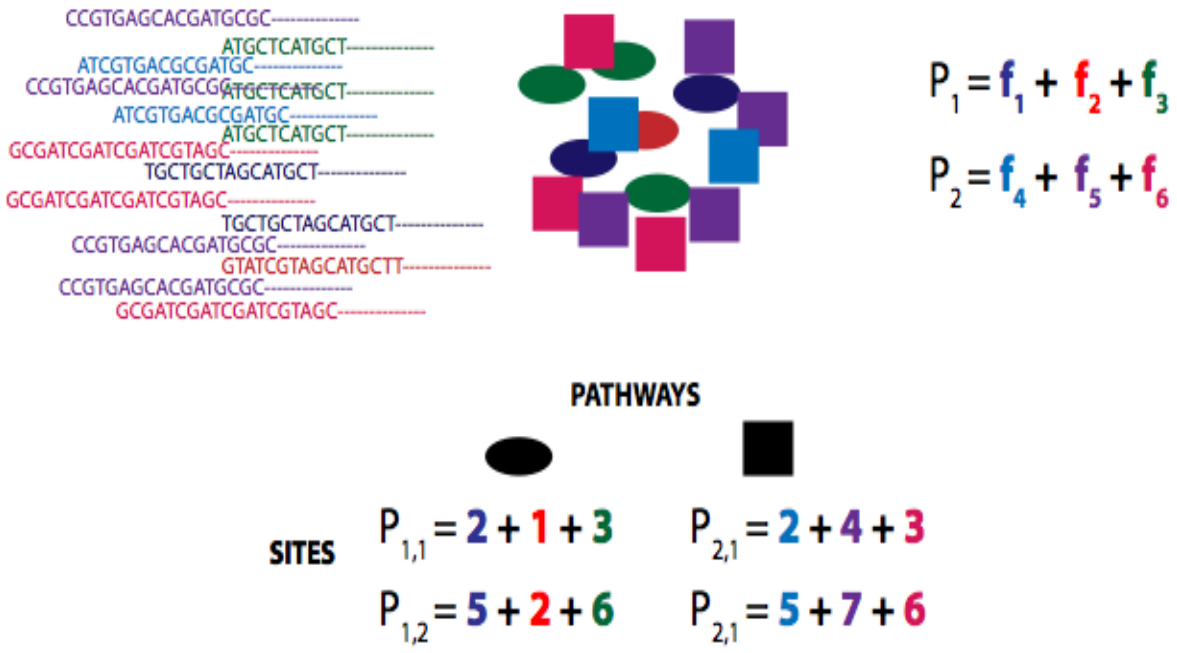
PROBLEM: Lose information about which gene belongs to which microbe.

Global Ocean Survey Statistics (GOS)

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

Rusch, et al., PLOS Biology 2007



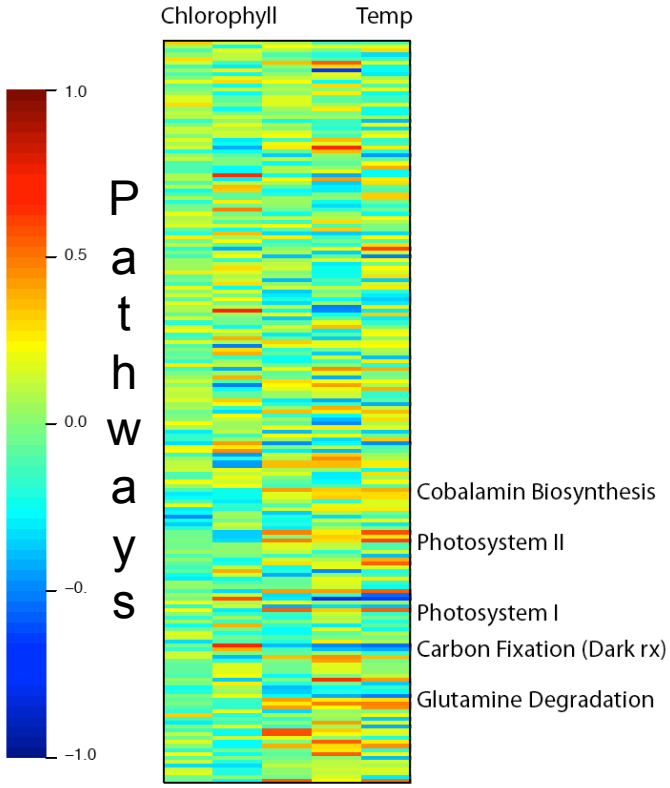


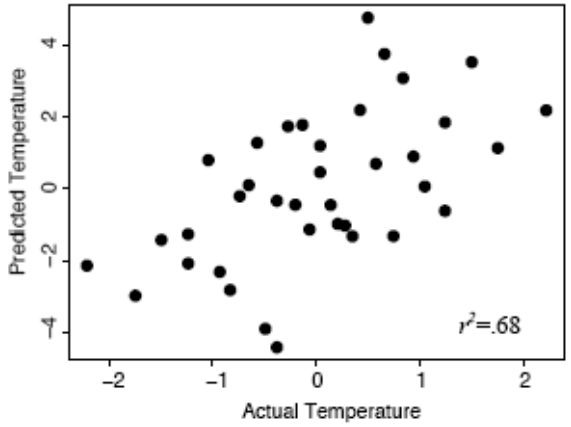
Expressing data as matrices indexed by site, env. var., and pathway usage

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

Simple Relationships: Pairwise Correlations

Environmental Features





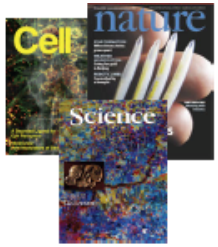
[Gianoulis et al., PNAS (in press, 2009)]

Canonical Correlation Analysis: Simultaneous weighting

Score	# of papers published
GRE	

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

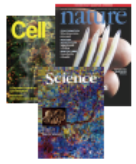
UPI = a GRE + b  + c GPA

GPI = a'  + b'  + c' 

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

Canonical Correlation Analysis: Simultaneous weighting

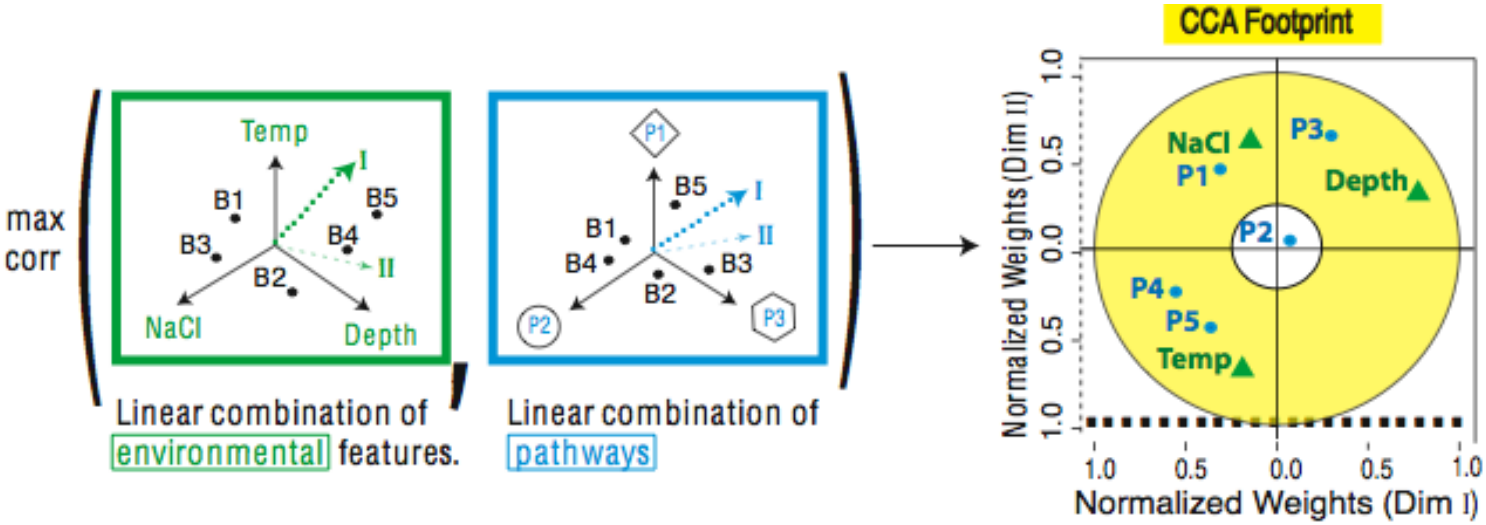
Score	# of papers published
GRE	

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

Environmental Features	Metabolic Pathways
Temp etc	Photosynthesis etc
Chlorophyll	Lipid Metabolism

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

Environmental-Metabolic Space

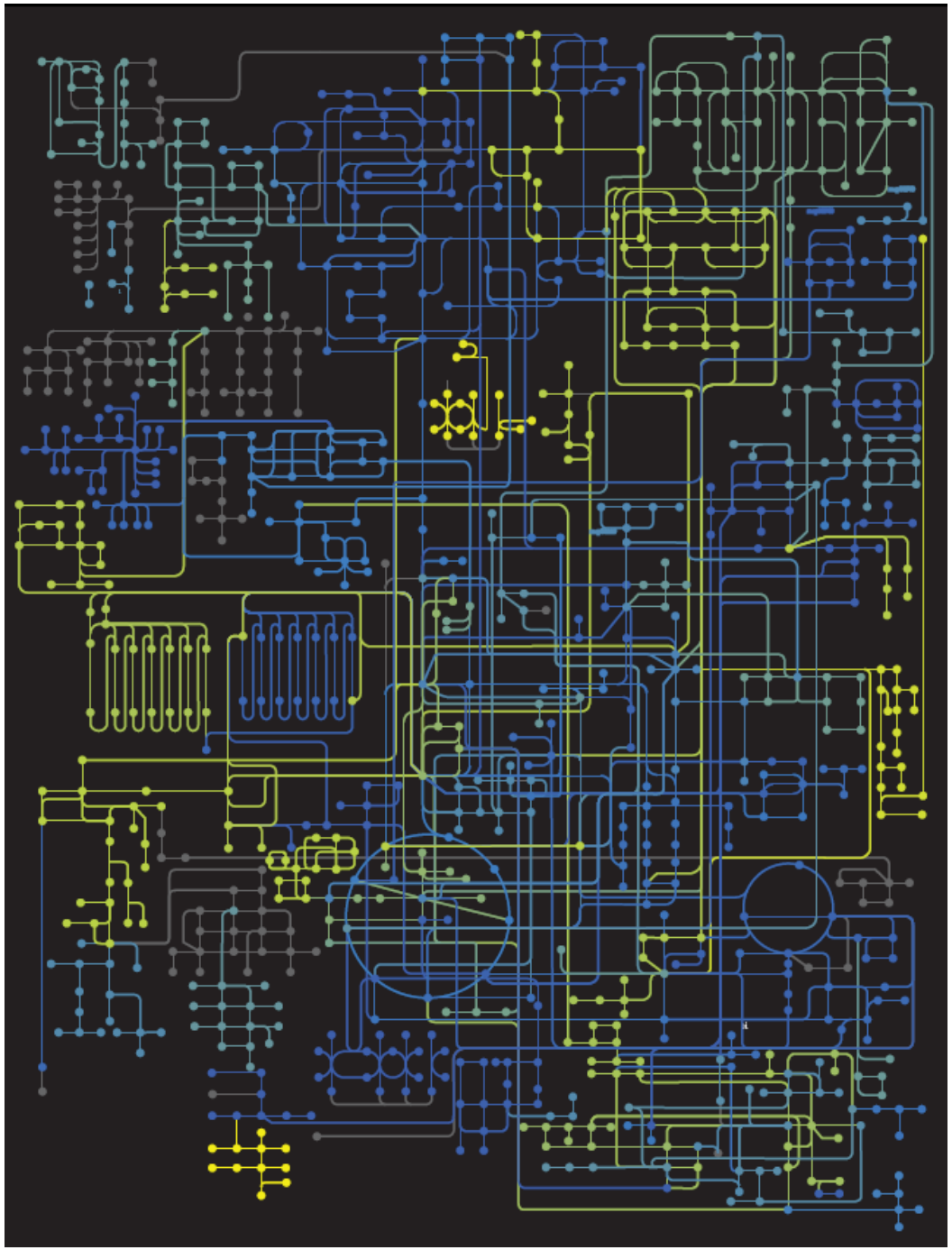


The goal of this technique is to interpret cross-variance matrices
We do this by defining a change of basis.

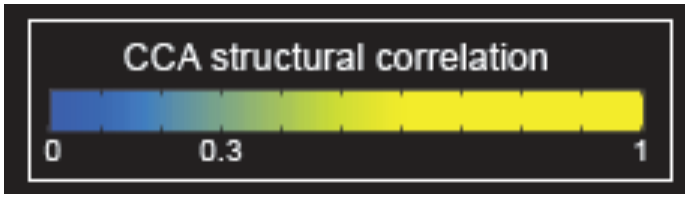
$$\text{Given } X=\{x_1,x_2,\ldots,x_n\} \text{ and } Y=\{y_1,y_2,\ldots,y_m\}$$

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

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

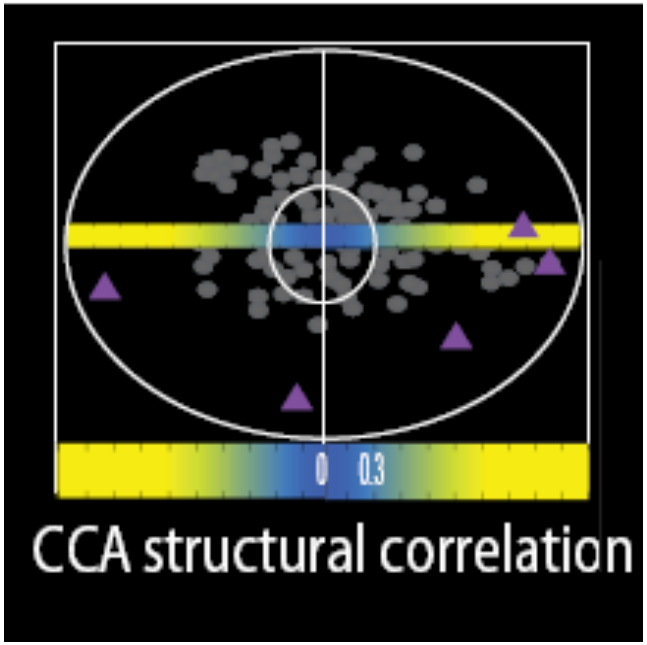


Strength of Pathway co-variation with environment



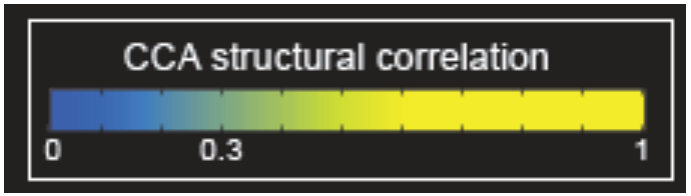
Environmentally invariant

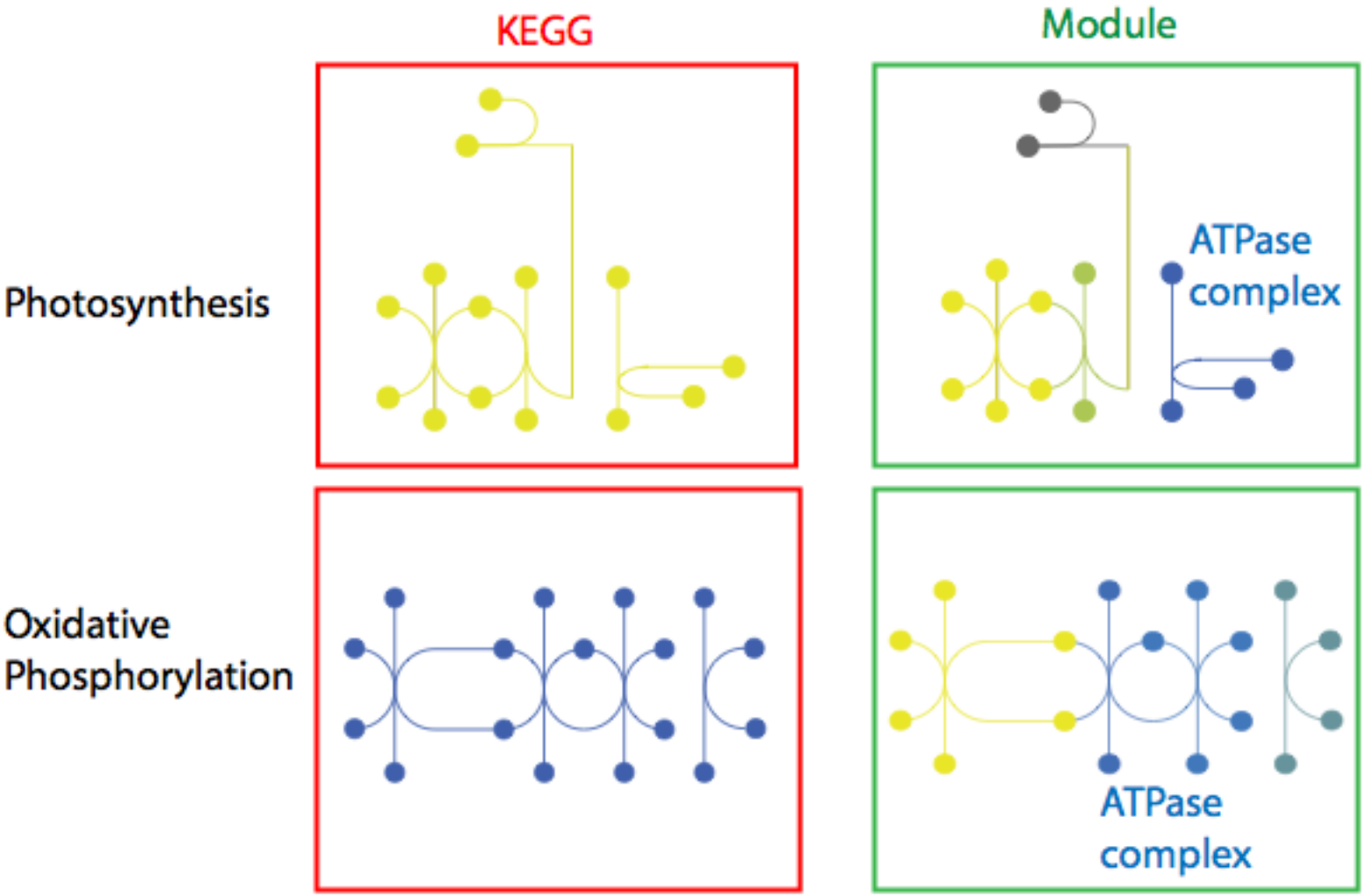
Environmentally variant



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

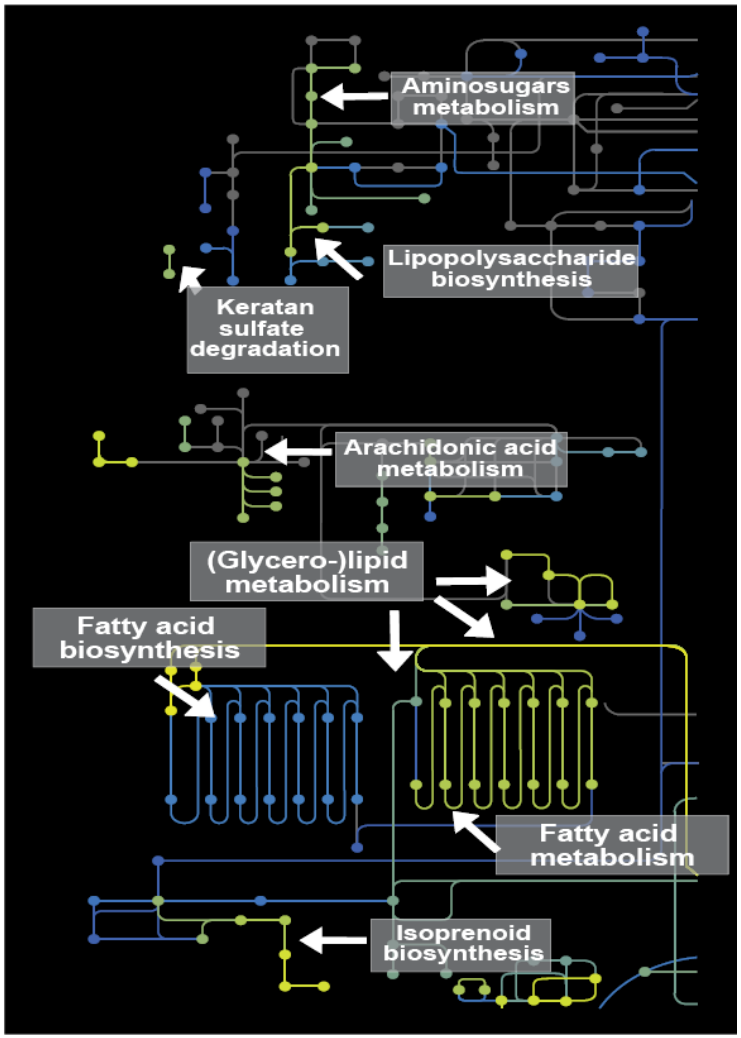
Conclusion #1: energy conversion strategy, temp and depth

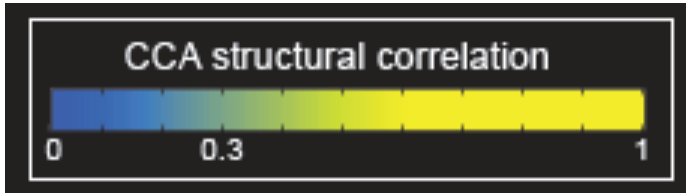




[Gianoulis et al., PNAS (in press, 2009)]

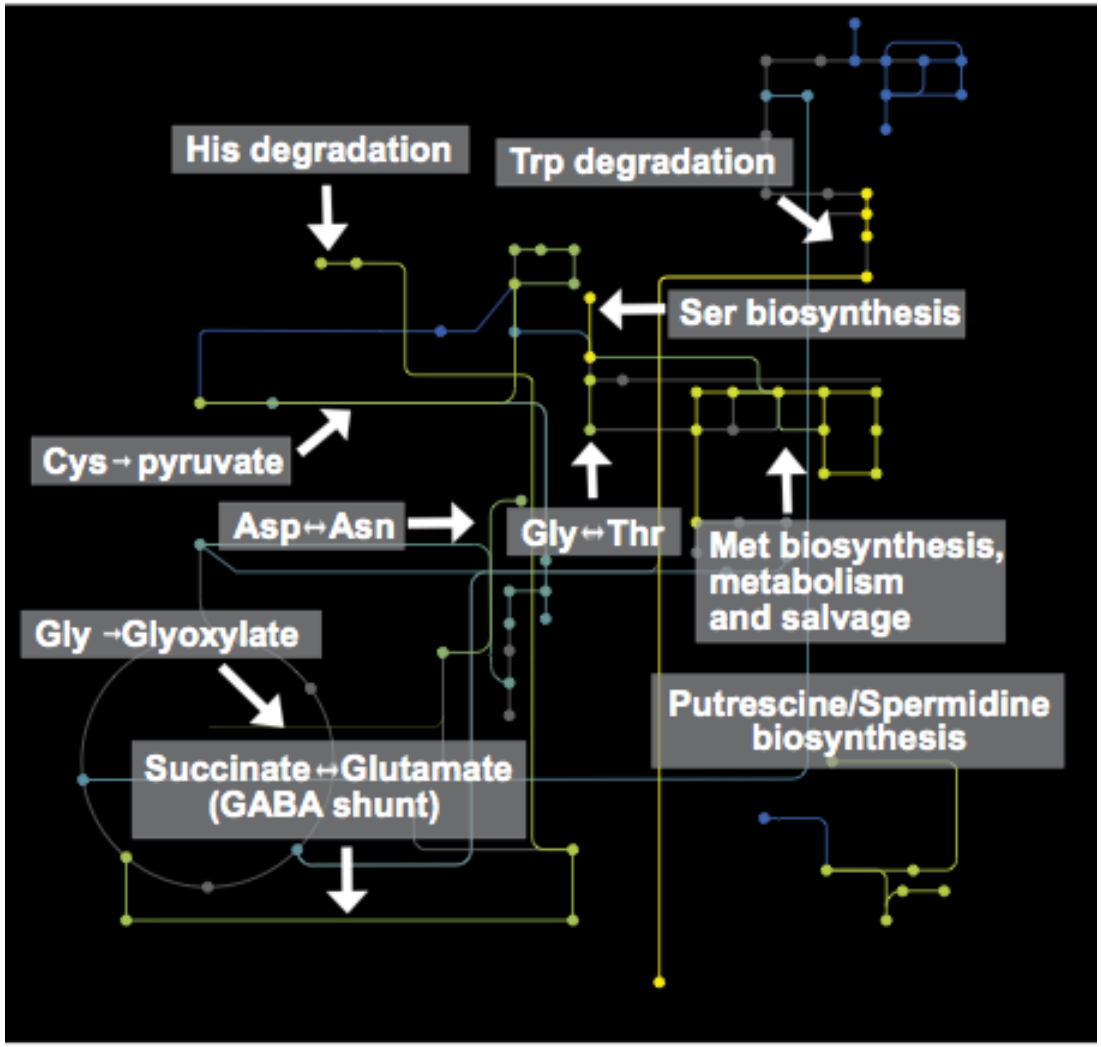
Conclusion #2: Outer Membrane components vary the environment





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

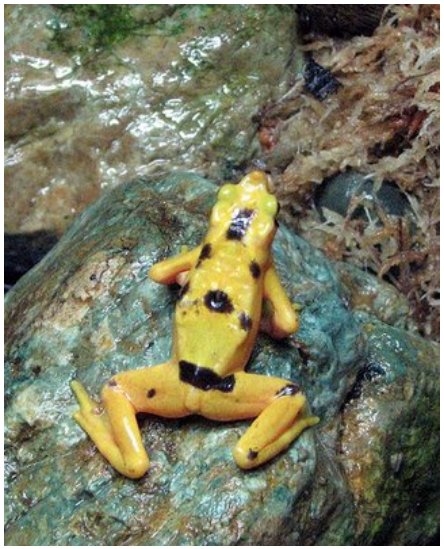
Conclusion #3: Covariation of AA biosynthesis and Import



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

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

Biosensors:
Beyond Canaries in a Coal Mine



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

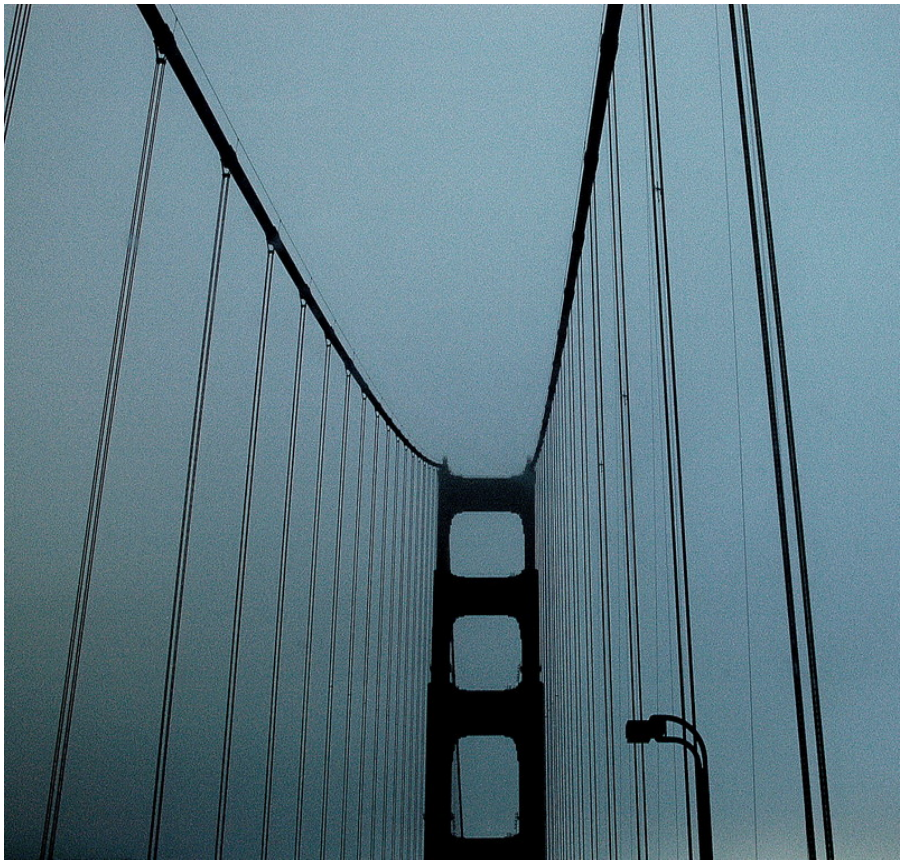
Outline: Molecular Networks

- Why Networks?
- Network Structure: Key Positions
  - ◊ Hubs & Bottlenecks
  - ◊ Tops of a Hierarchy
  - ◊ RE-score nodes
- Networks, Variation & the Environment
  - ◊ Which pathways change most with the environment

Conclusions:
Analysis of Network Structure

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

Conclusions:
Points of Network Centrality



- RE-score measures degree of (down) regulation of targets vs. non-targets
- Application to miRNA network
- Different RE-score of miRNAs can be used in cancer classification

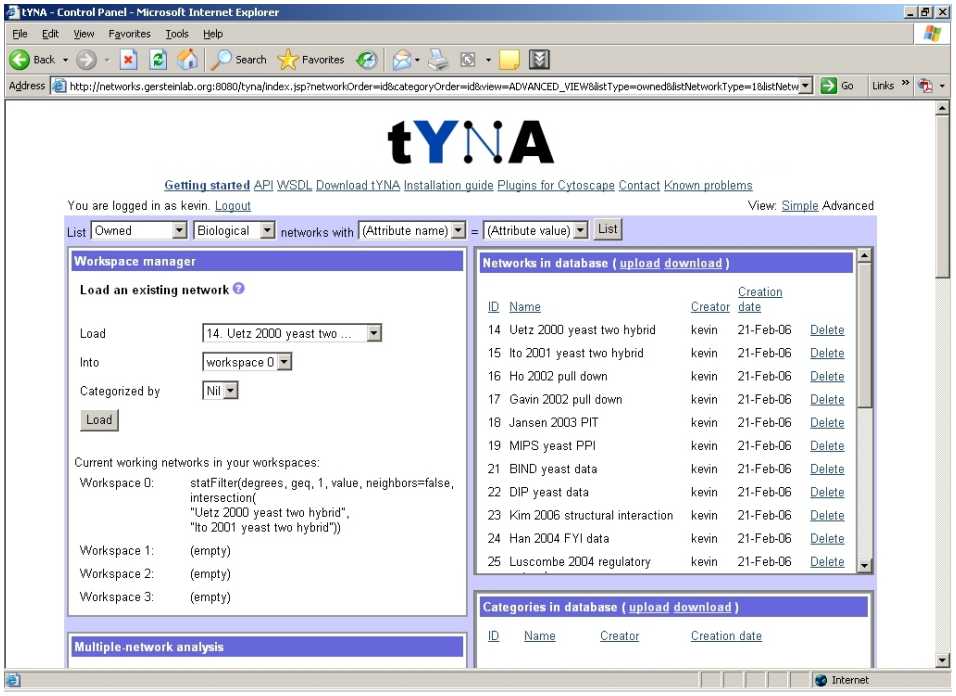
Conclusions: Networks Dynamics across Environments

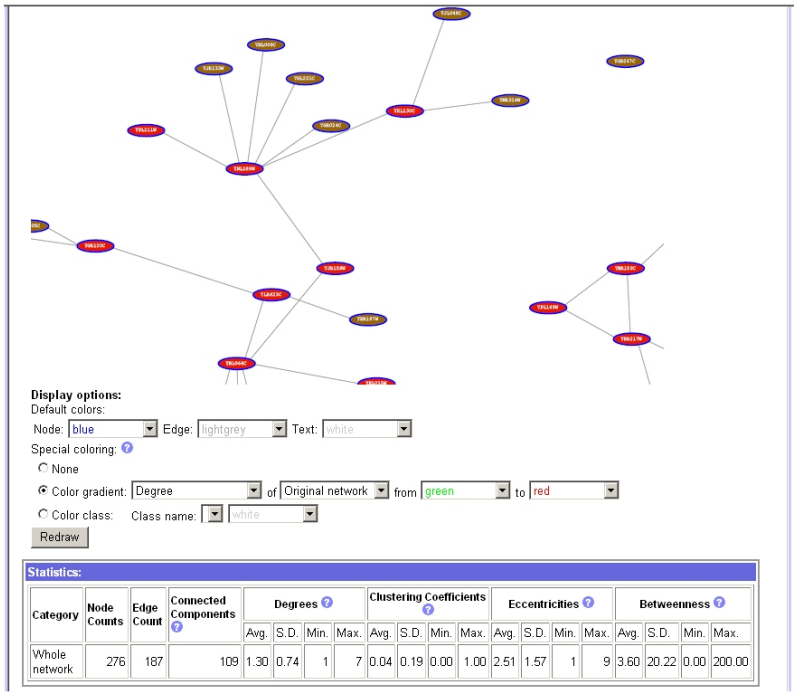
- Developed and adapted techniques to connect quantitative features of environment to metabolism.
- Applied to available aquatic datasets, we identified footprints that were predictive of their environment (potentially could be used as biosensor).
- Strong correlation exists between a community's energy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

- an automated web tool

tYNA

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





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

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

Acknowledgements

H Yu
P Kim
K Yip
T Gianoulis
C Cheng

A Paccanaro
P Alves
T Emonet
P Cayting
M Seringhaus
Y Xia
J Korbel
A Sboner
P Patel
P Bork
J Raes
E Franzosa
M Snyder
N Bhardwaj
R Alexander

Networks.GersteinLab.org

Job opportunities currently for postdocs & students

More Information on this Talk

TITLE: Large-scale Analysis of Molecular Networks

SUBJECT: Networks

DESCRIPTION: Laufer Center Inaugural Symposium, Stony Brook University, NY; 2009.09.25, 09:00-09:30; [I:**STONEYBROOK**] (Short networks talk. Aiming to fit into 27′ w. 3′ questions. Modifications of **rescore*** compared to previous networks talk.)

NOTES: Orange background slides are hidden and not meant to me shown. 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** .