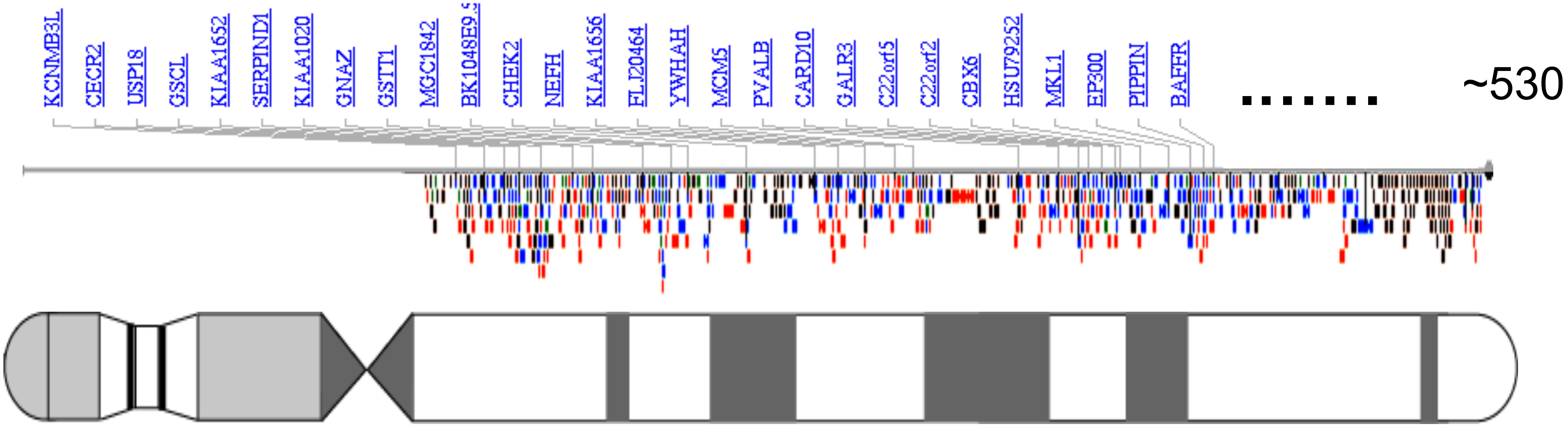
Understanding Protein Function on a Genome-scale through the Analysis of Molecular Networks

Mark B Gerstein
Yale

slides at
Lectures.GersteinLab.org

(See Last Slide for References & More Info.)

The problem: Grappling with Function on a Genome Scale?



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

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

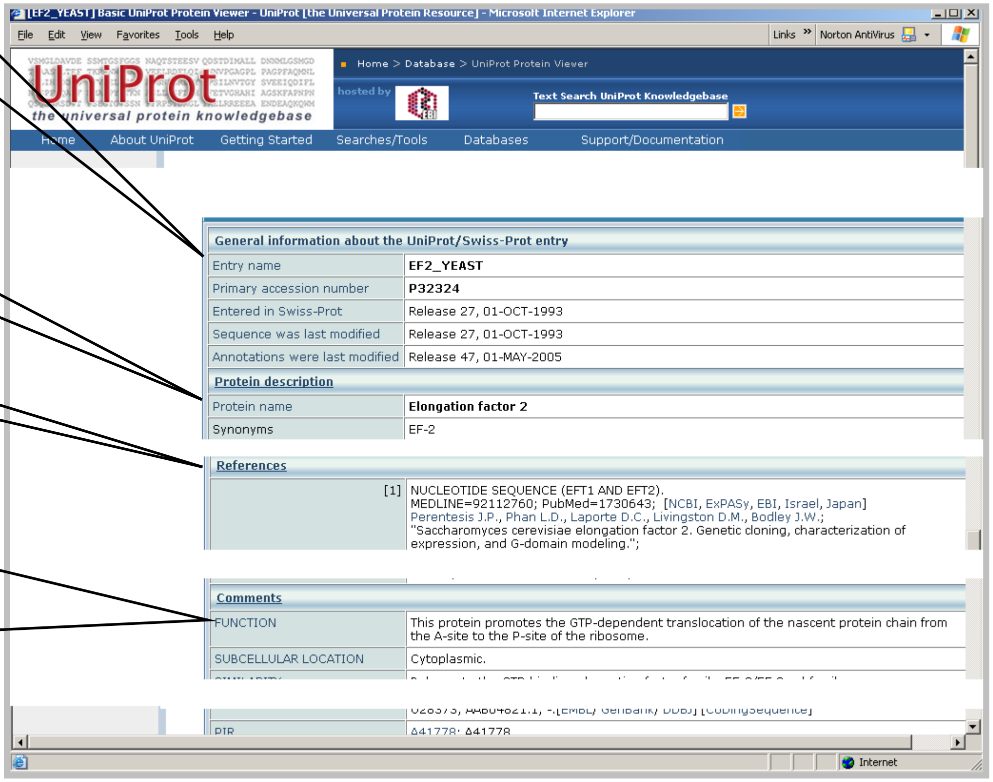
Traditional single molecule way to integrate evidence & describe function

EF2_YEAST

Descriptive Name:
Elongation Factor 2

Lots of references
to papers

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



[EF2_YEAST] Basic UniProt Protein Viewer - UniProt [the Universal Protein Resource] - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Links
Norton AntiVirus
UniProt
the universal protein knowledgebase
Home > Database > UniProt Protein Viewer
hosted by
Text Search UniProt Knowledgebase
Home
About UniProt
Getting Started
Searches/Tools
Databases
Support/Documentation
Internet

General information about the UniProt/Swiss-Prot entry	
Entry name	**EF2_YEAST**
Primary accession number	**P32324**
Entered in Swiss-Prot	Release 27, 01-OCT-1993
Sequence was last modified	Release 27, 01-OCT-1993
Annotations were last modified	Release 47, 01-MAY-2005
Protein description	
Protein name	**Elongation factor 2**
Synonyms	EF-2

References	
[1]	NUCLEOTIDE SEQUENCE (EFT1 AND EFT2). MEDLINE=92112760; PubMed=1730643; [NCBI, ExPASy, EBI, Israel, Japan] Perentesis J.P., Phan L.D., Laporte D.C., Livingston D.M., Bodley J.W.; "Saccharomyces cerevisiae elongation factor 2. Genetic cloning, characterization of expression, and G-domain modeling.";

Comments	
FUNCTION	This protein promotes the GTP-dependent translocation of the nascent protein chain from the A-site to the P-site of the ribosome.
SUBCELLULAR LOCATION	Cytoplasmic.

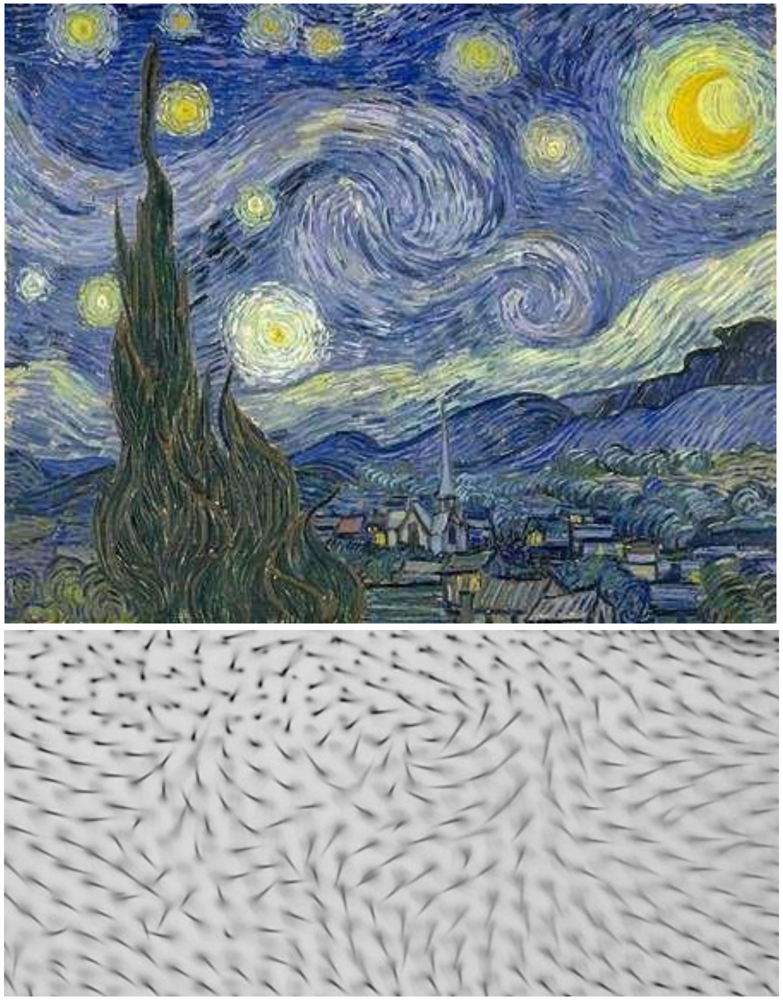
	U28373, AAB04821.1, -.[EMBL/ GenBank/ DDBJ] [CodingSequence]
PIR	A41778; A41778

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

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

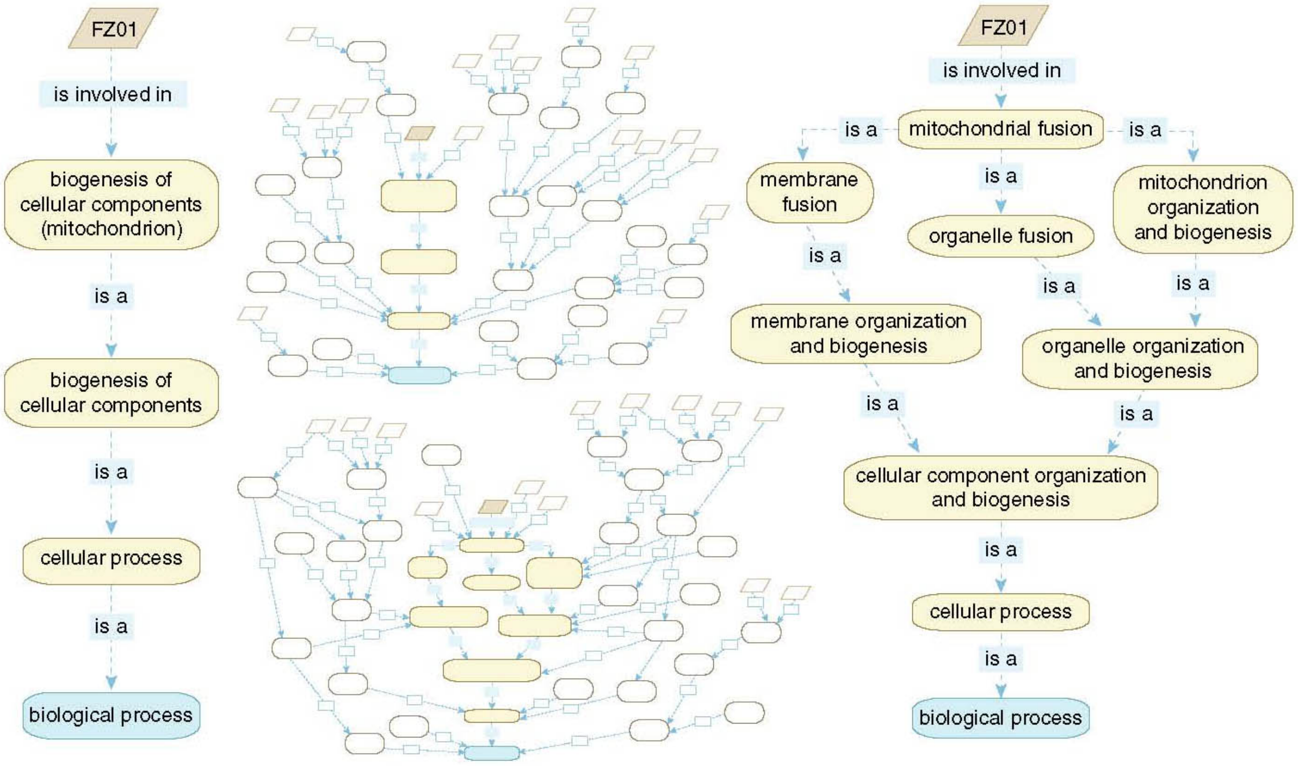
Some obvious issues in scaling single molecule definition to a genomic scale

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



[Seringhaus et al. GenomeBiology (2008)]

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


FZ01
is involved in
biogenesis of cellular components (mitochondrion)
is a
biogenesis of cellular components
is a
cellular process
is a
biological process
FZ01
is involved in
mitochondrial fusion
is a
membrane fusion
is a
organelle fusion
is a
mitochondrion organization and biogenesis
is a
membrane organization and biogenesis
is a
organelle organization and biogenesis
is a
cellular component organization and biogenesis
is a
cellular process
is a
biological process

MIPS (Mewes et al.)

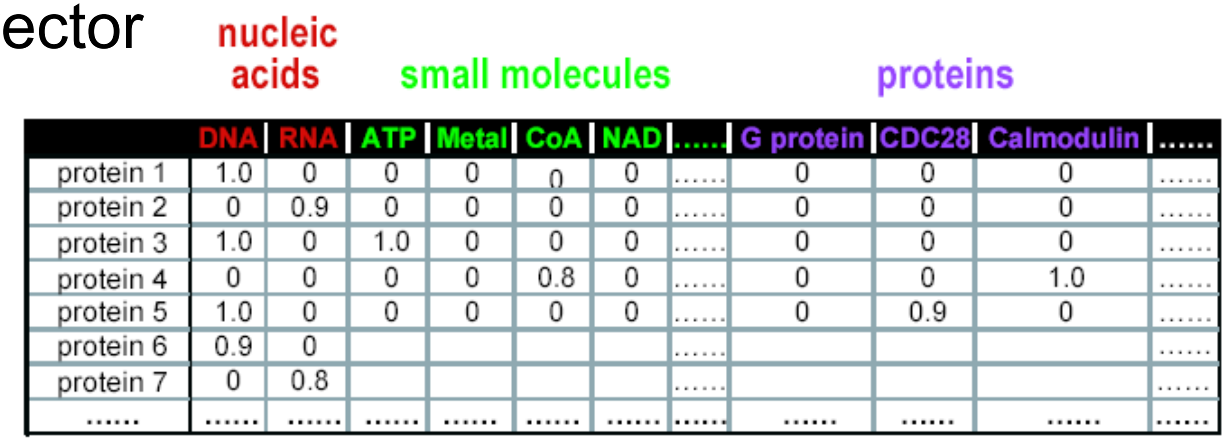
GO (Ashburner et al.)

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

Towards Developing Standardized Descriptions of Function

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

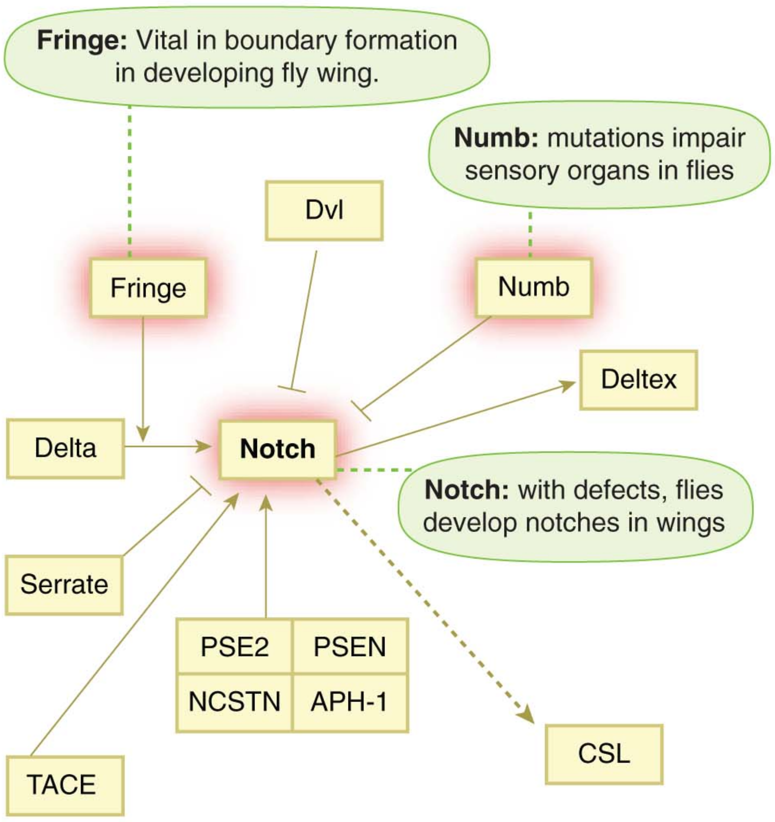
- Function as a vector



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

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

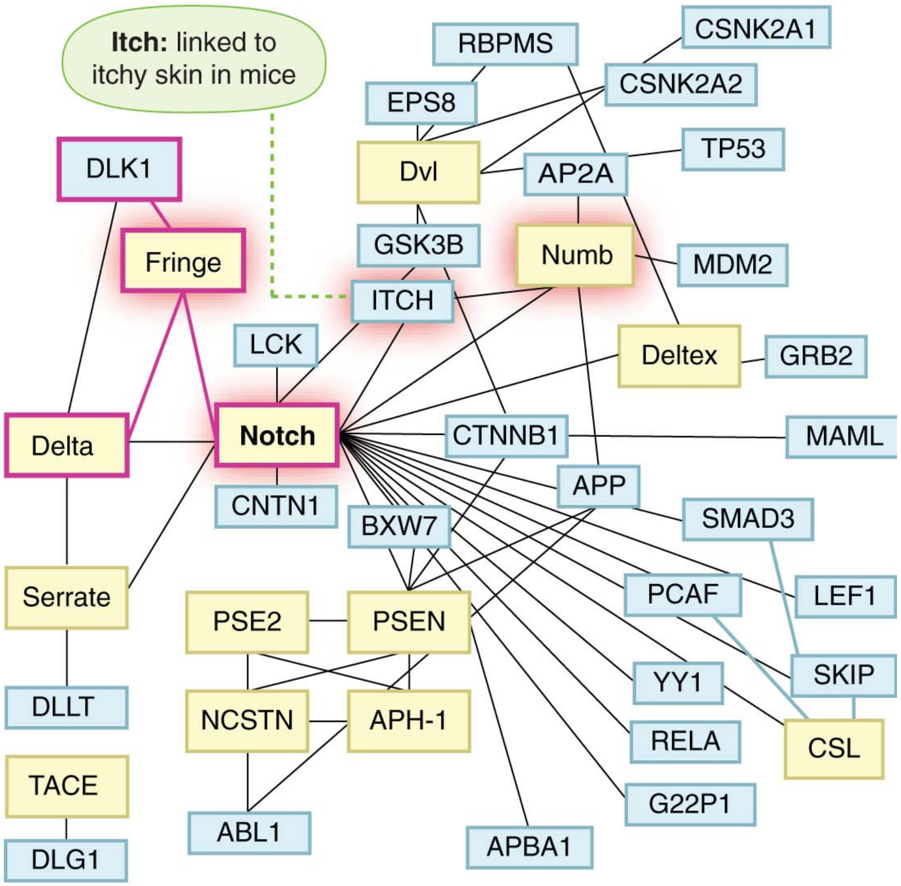
Networks (Old & New)


Fringe: Vital in boundary formation in developing fly wing.

Numb: mutations impair sensory organs in flies

Notch: with defects, flies develop notches in wings

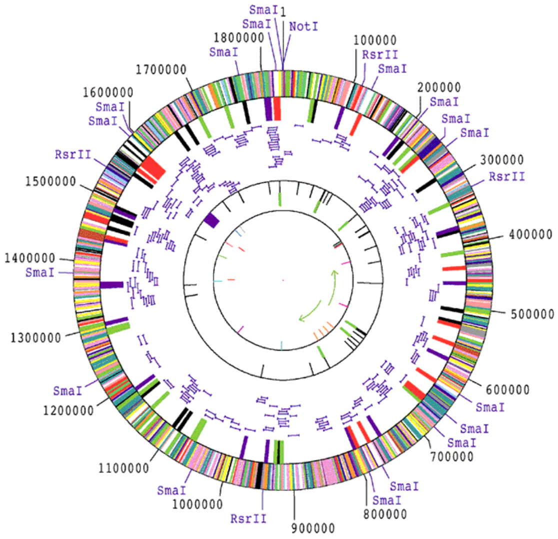
Classical KEGG pathway

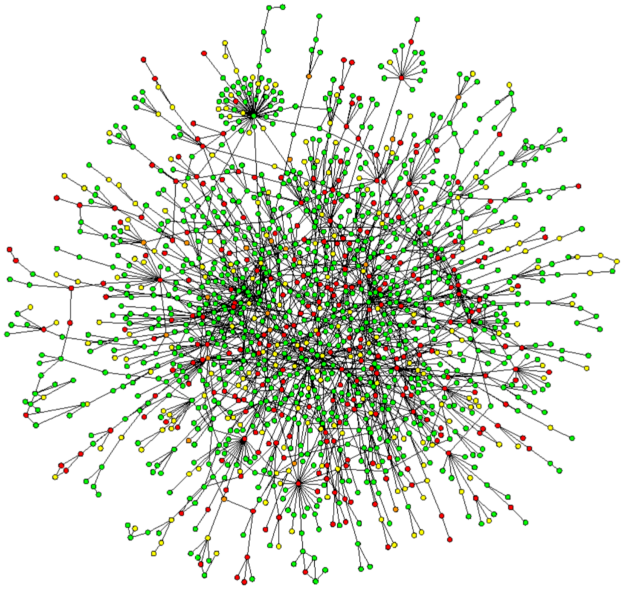

Itch: linked to itchy skin in mice

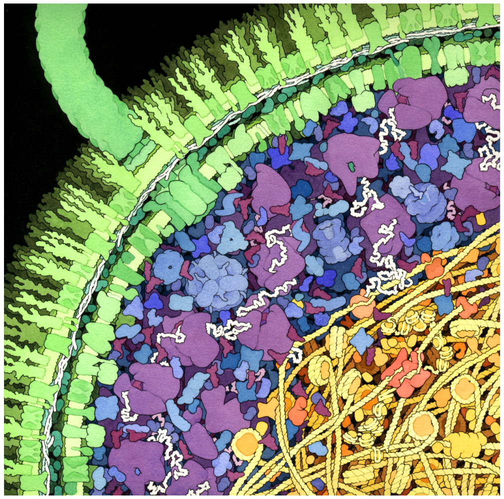
Same Genes in High-throughput Network

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

Networks occupy a midway point in terms of level of understanding







1D: Complete Genetic Partslist

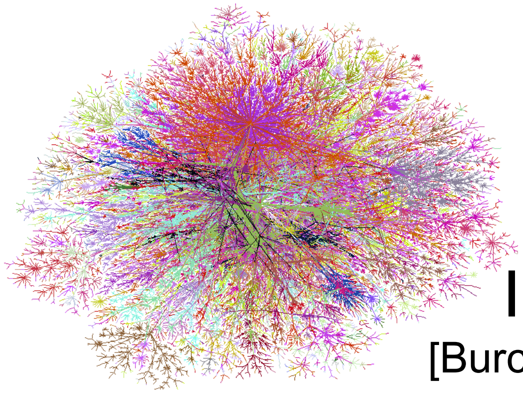
~2D: Bio-molecular Network Wiring Diagram

3D: Detailed structural understanding of cellular machinery

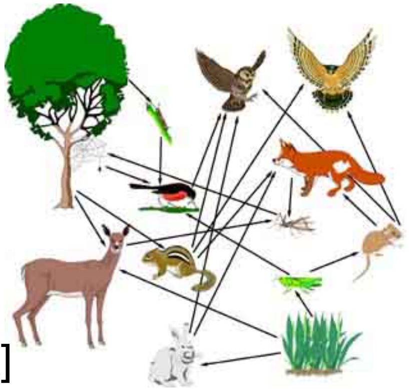
[Fleischmann et al., Science, 269 :496]

[Jeong et al. Nature, 41:411]

Networks as a universal language

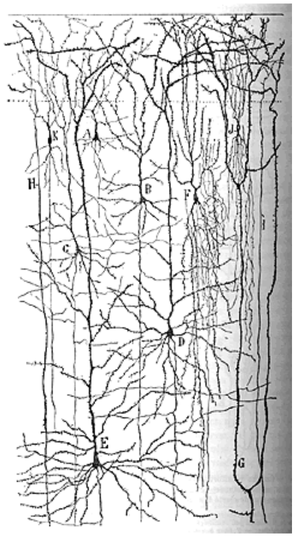


Internet
[Burch & Cheswick]

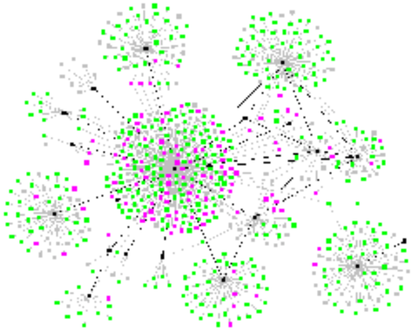


Food Web

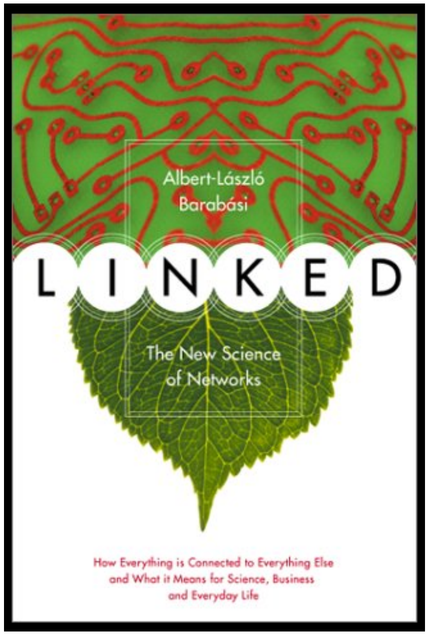
Electronic
Circuit

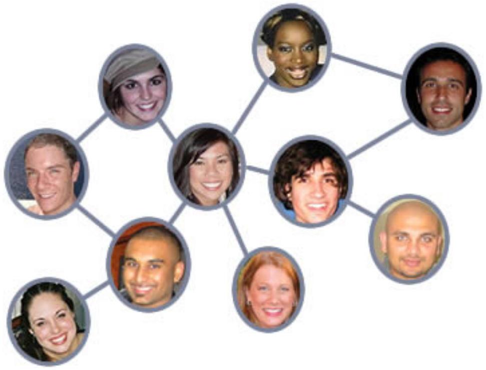


Neural Network
[Cajal]



Disease
Spread
[Krebs]

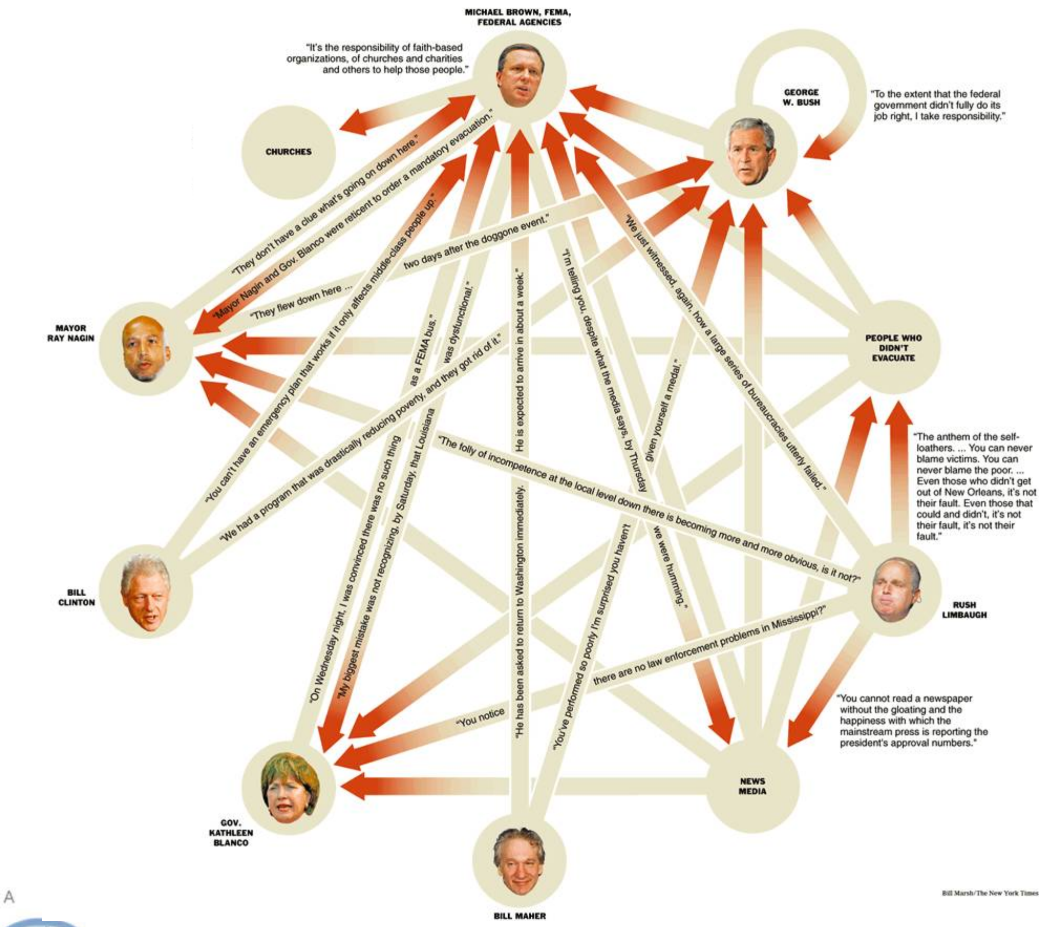

Albert-László
Barabási
LINKED
The New Science
of Networks
How Everything is Connected to Everything Else
and What it Means for Science, Business
and Everyday Life



Social Network

Protein
Interactions
[Barabasi]

Using the position in networks to describe function


MICHAEL BROWN, FEMA, FEDERAL AGENCIES
CHURCHES
GEORGE W. BUSH
MAYOR RAY NAGIN
PEOPLE WHO DIDN'T EVACUATE
BILL CLINTON
RUSH LIMBAUGH
NEWS MEDIA
GOV. KATHLEEN BLANCO
BILL MAHER
"They don't have a clue what's going on down here."
"Mayor Nagin and Gov. Blanco were reticent to order a mandatory evacuation."
"They flew down here ... two days after the doggone event."
"You can't have an emergency plan that works if it only affects middle-class people up."
"We had a program that was drastically reducing poverty, and they got rid of it."
"On Wednesday night, I was convinced there was no such thing as a FEMA bus."
"My biggest mistake was not recognizing, by Saturday, that Louisiana was dysfunctional."
"He has been asked to return to Washington immediately. He is expected to arrive in about a week."
"I'm telling you, despite what the media says, by Thursday we were humming."
"You've performed so poorly I'm surprised you haven't given yourself a medal."
"We just witnessed, again, how a large series of bureaucracies utterly failed."
"The folly of incompetence at the local level down there is becoming more and more obvious, is it not?"
"You notice there are no law enforcement problems in Mississippi?"

"It's the responsibility of faith-based organizations, of churches and charities and others to help those people."

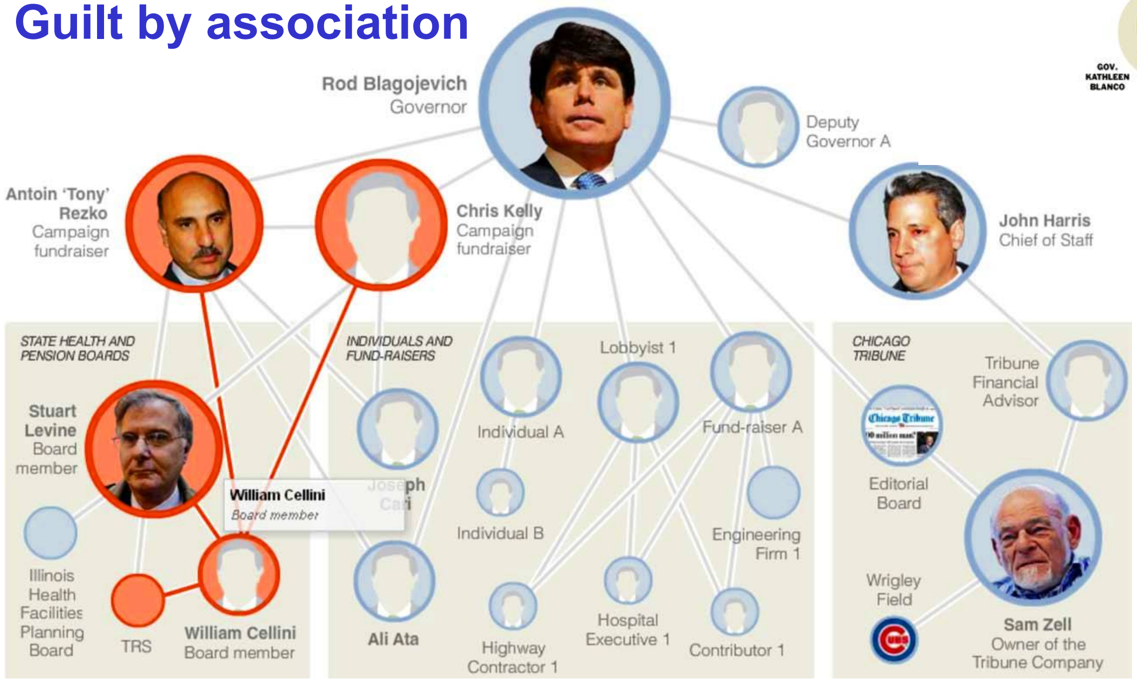
"To the extent that the federal government didn't fully do its job right, I take responsibility."

"The anthem of the self-loathers. ... You can never blame victims. You can never blame the poor. ... Even those who didn't get out of New Orleans, it's not their fault. Even those that could and didn't, it's not their fault, it's not their fault."

"You cannot read a newspaper without the gloating and the happiness with which the mainstream press is reporting the president's approval numbers."

Bill Marsh/The New York Times

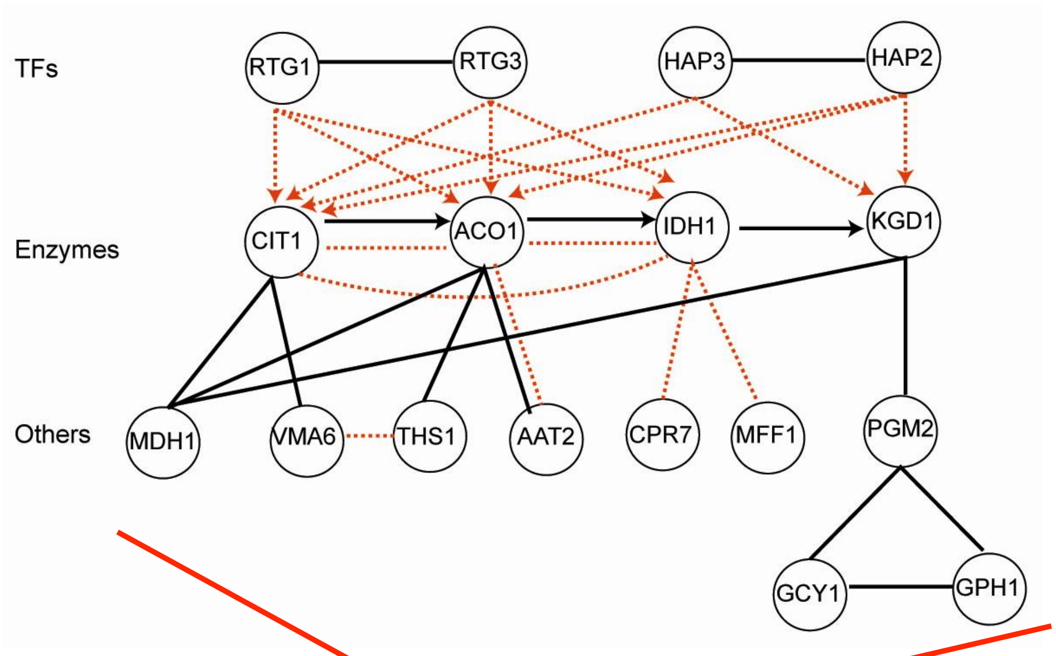
Guilt by association

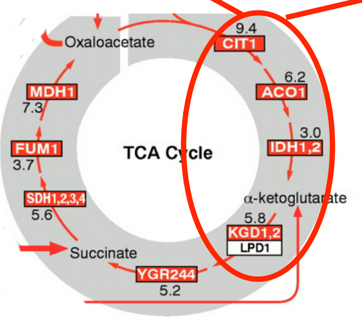


Finding the causal regulator (the "Blame Game")

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

Combining networks forms an ideal way of integrating diverse information





Part of the TCA cycle

Metabolic pathway

Transcriptional regulatory network

Physical protein-protein Interaction

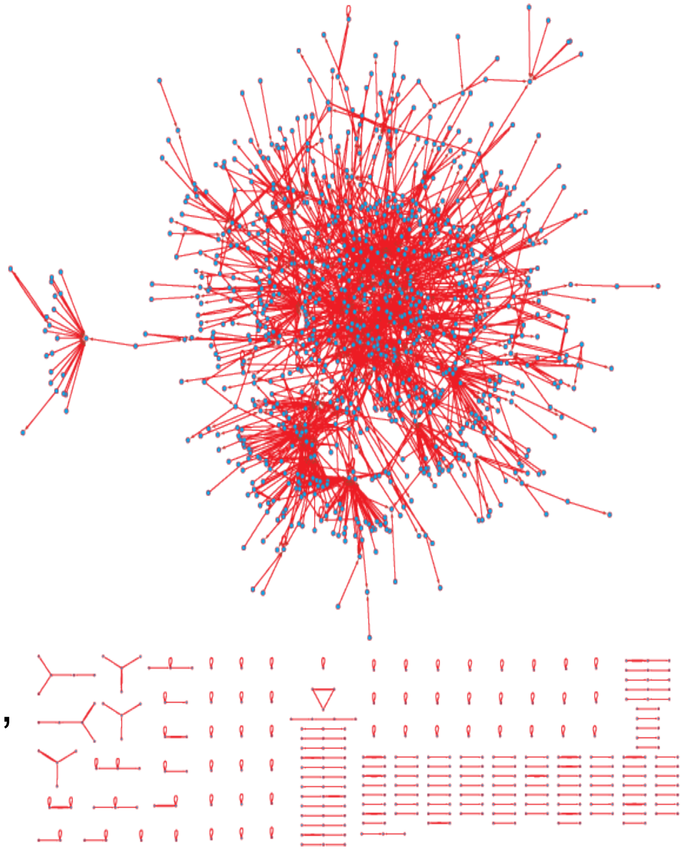
Co-expression Relationship

Genetic interaction (synthetic lethal)
Signaling pathways

Outline: Molecular Networks

- Why Networks?
- Predicting Networks (yeast ppi)
  - ◊ Propagating known information
- Central Points in Networks
  - ◊ Hubs & Bottlenecks (yeast ppi & reg. net)
  - ◊ Tops of Heirarchies (yeast reg. net)
  - ◊ Identified by score (human miRNA-targ. net)
- Dynamics of Networks
  - ◊ Across environments (in prokaryote metab. pathways)
- Protein Networks & Variation (human ppi & miRNA-targ. net)

Example: yeast PPI network



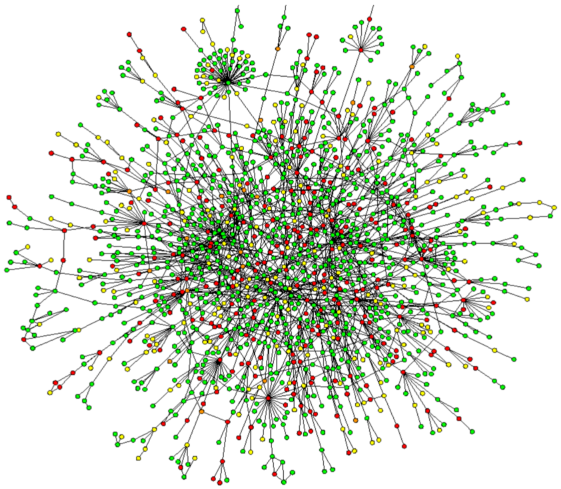
Actual size:

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

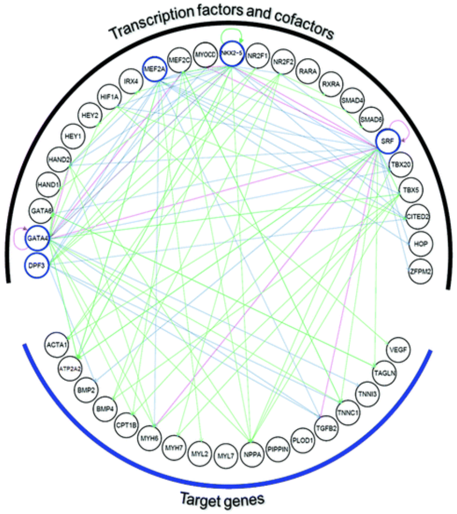
Known interactions:

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

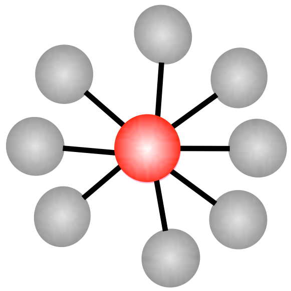
Different Types of Molecular Networks



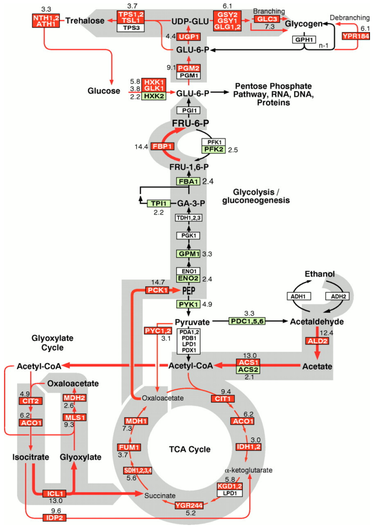
Protein-protein Interaction networks



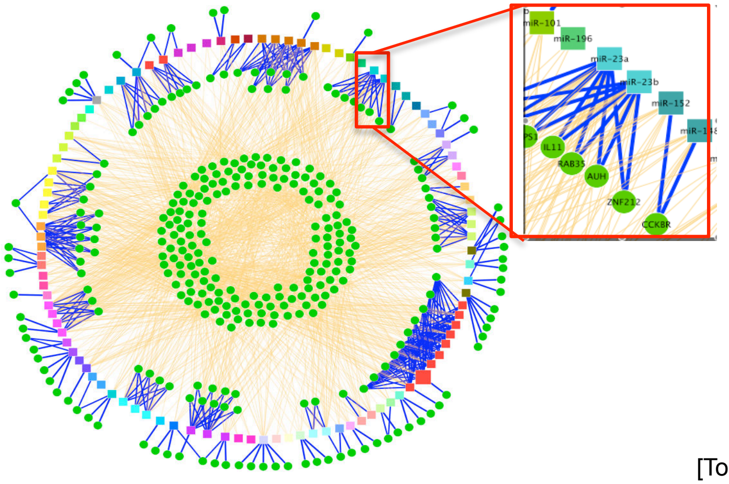
TF-target-gene Regulatory networks



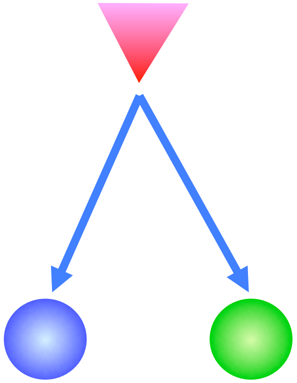
Undirected



Metabolic pathway networks



miRNA-target networks

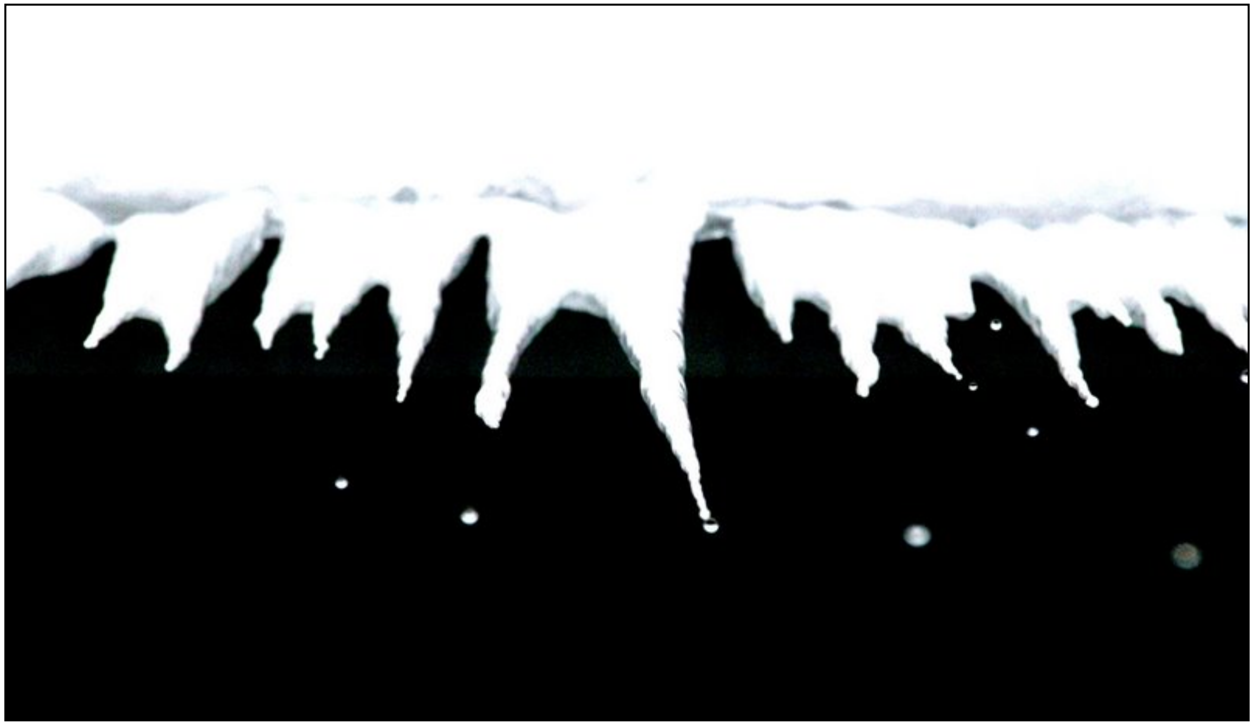


Directed

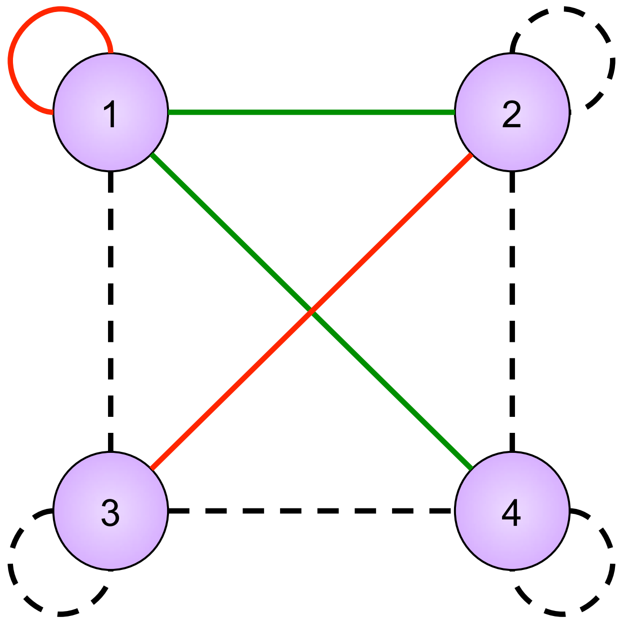
[Toenjes, *et al*, *Mol. BioSyst.* (2008); Jeong *et al*, *Nature* (2001); [Horak, et al, Genes & Development, 16:3017-3033; DeRisi, Iyer, and Brown, Science, 278:680-686]

Predicting Networks

How do we construct large molecular networks?
From extrapolating correlations between functional genomics data with fairly small sets of known interactions, making best use of the known training data.



Training sets



	1	2	3	4
1	0	1	?	1
2	1	?	0	?
3	?	0	?	?
4	1	?	?	?

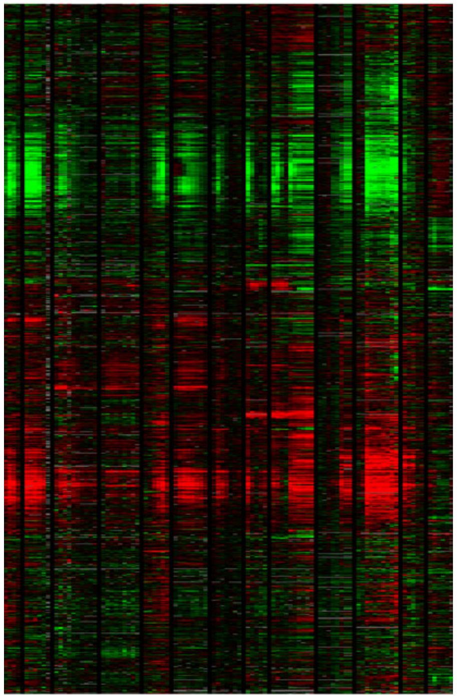
🟩 Known interactions

🟥 Known non-interactions

- - - Unknown

Network prediction: features

- Example 1: gene expression



Gasch et al., 2000

$x_1 = (0.2, 2.4, 1.5, \ldots)$

$x_2 = (0.8, 2.2, 1.5, \ldots)$

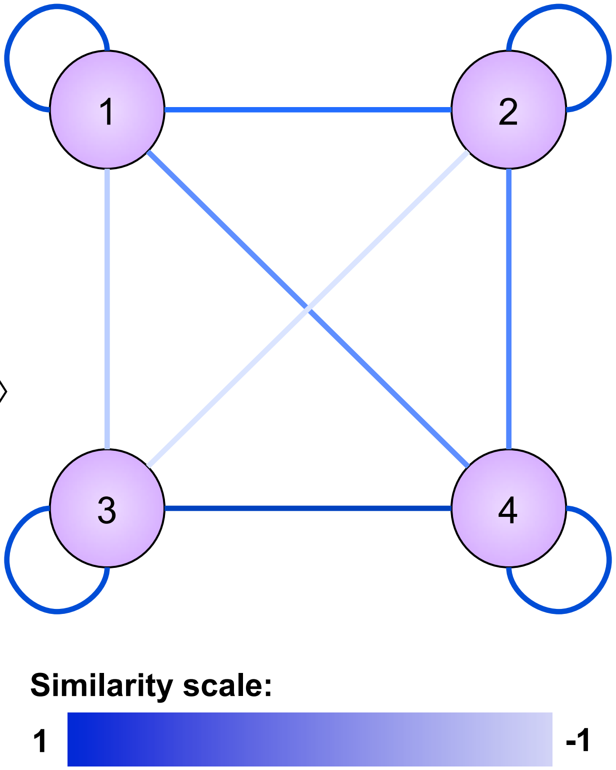
$x_3 = (4.3, 0.1, 7.5, \ldots)$

…

$\mathrm{sim}(x_1, x_2) = 0.62$

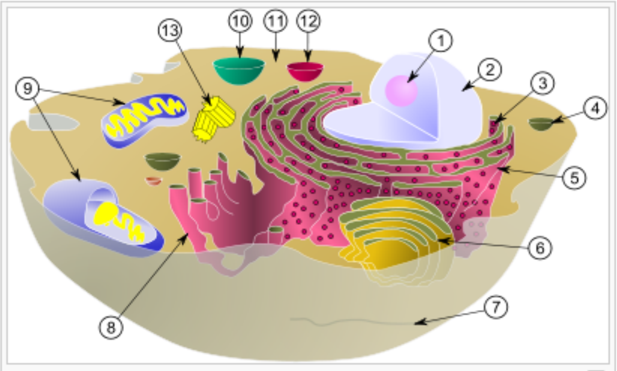
$\mathrm{sim}(x_1, x_3) = -0.58$

…



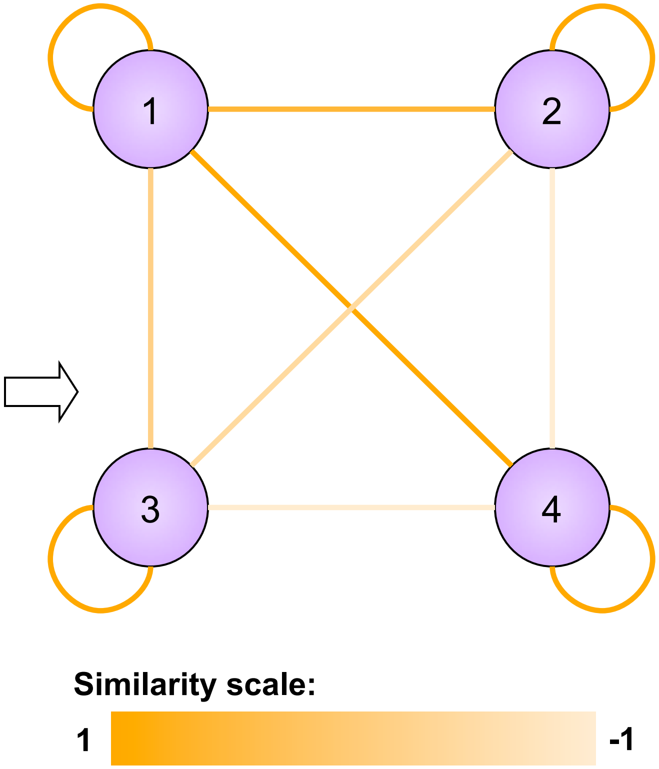
Network prediction: features

- Example 2: sub-cellular localization

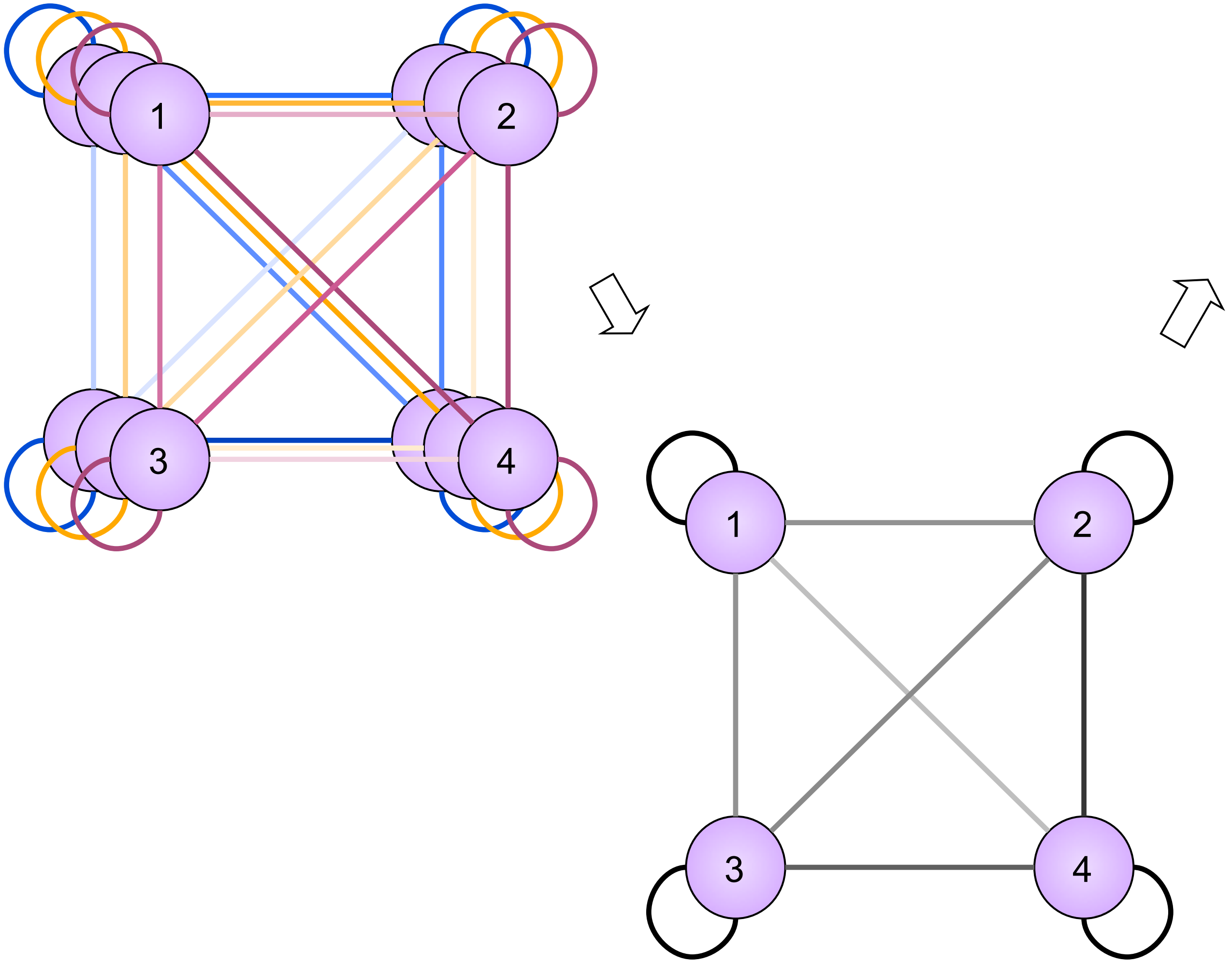


http://www.scq.ubc.ca/wp-content/yeasttwohybridtranscript.gif

$x_1 = (1, 1, 0, 0, \ldots)$
$x_2 = (1, 1, 1, 0, \ldots)$
$x_3 = (1, 0, 1, 0, \ldots)$
$\ldots$
$\mathrm{sim}(x_1, x_2) = 0.81$
$\mathrm{sim}(x_1, x_3) = 0.12$
$\ldots$



Data integration & Similarity Matrix



	1	2	3	4
1	1.00	0.57	0.55	0.40
2	0.57	1.00	0.66	0.89
3	0.55	0.66	1.00	0.79
4	0.40	0.89	0.79	1.00

Learning methods

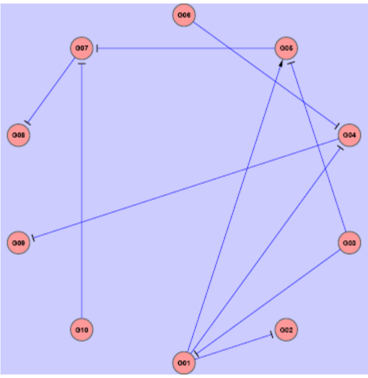
An endless list:

- Docking (e.g. Schoichet and Kuntz 1991)
- Evolutionary (e.g. Ramani and Marcotte, 2003)
- Topological (e.g. Yu et al., 2006)
- Bayesian (e.g. Jansen et al., 2003)
- **Kernel methods**
  - ◊ Global modeling:
    - em (Tsuda et al., 2003)
    - kCCA (Yamanishi et al., 2004)
    - kML (Vert and Yamanishi, 2005)
    - Pairwise kernel (Pkernel) (Ben-Hur and Noble, 2005)
  - ◊ Local modeling:
    - Local modeling (Bleakley et al., 2007)

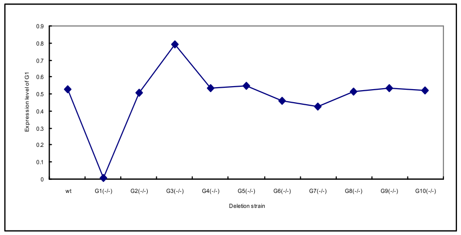
Let's compare in a public challenge!
(DREAM: Dialogue for Reverse Engineering Assessment and Methods)

DREAM3: *in silico* regulatory network reconstruction

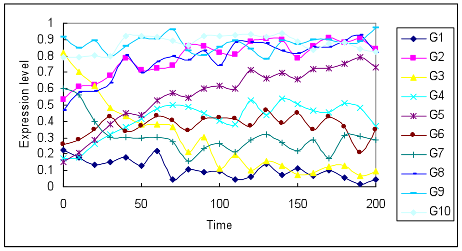
Actual network



Expression data

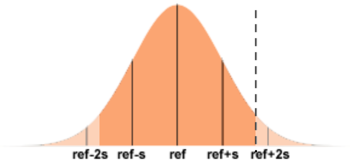


Deletion strains



Time series after initial perturbation

Modeling



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

Noise models

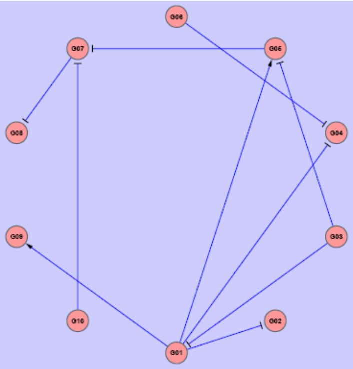
$$\frac{dy_j}{dt} = a_{j0} - a_{jj}y_j + \sum_{k \in S} a_{jk}y_k$$

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

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

Expression rate models

Predictions



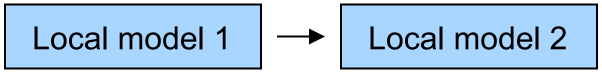
Accuracy (AUC)	E. Coli 1	E. Coli 2	Yeast 1	Yeast 2	Yeast 3
Size-10	0.928	0.912	0.949	0.747	0.714
Size-50	0.930	0.924	0.917	0.792	0.805
Size-100	0.948	0.960	0.915	0.856	0.783

[Yip et al., DREAM3]

Our work: efficiently propagating known information

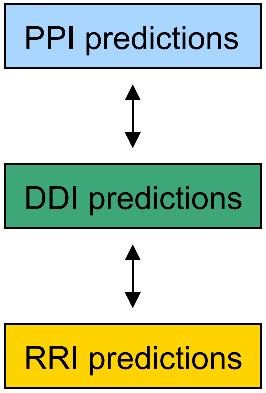
Training set expansion

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



Multi-level learning

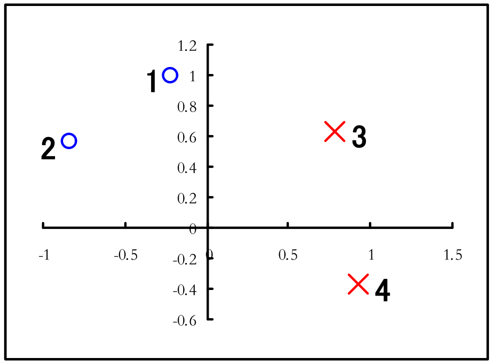
- Motivation: hierarchical nature of interaction
- Expand training sets vertically



DREAM3 *in silico* regulatory network reconstruction challenge

Kernels

Kernel: a similarity matrix that is positive semi-definite (p.s.d.)



Objects in an feature space

Compute inner products →

← p.s.d. implies

	1	2	3	4
1	1.00	0.72	0.45	-0.56
2	0.72	1.00	-0.30	-0.98
3	0.45	-0.30	1.00	0.49
4	-0.56	-0.98	0.49	1.00

Similarity matrix

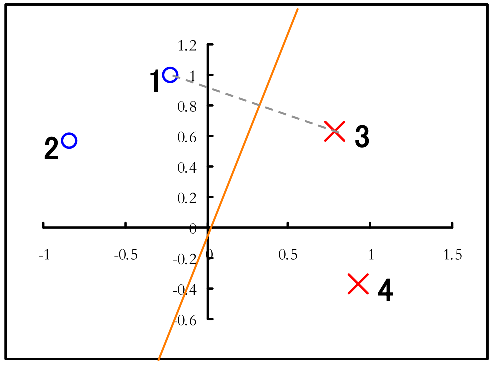
Good for integrating heterogeneous datasets (protein sequences, PSSM, gene expression, …)

– no need to explicitly place them in a common feature space

Kernel methods

Use the kernel as proxy to work in the feature space

Example: SVM (finding the best separating hyperplane)



← Equivalent to

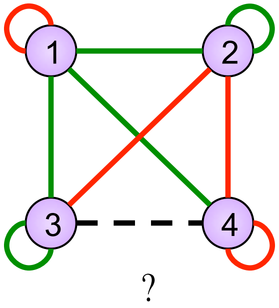
Maximize $$\sum_i \lambda_i - \frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j \boxed{\langle x_i, x_j \rangle}$$

Subject to $$\lambda \geq 0$$

$$\sum_i \lambda_i y_i = 0$$

The only thing that we need to know about the objects: their similarity values (inner products)

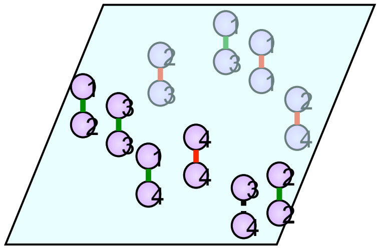
Kernel methods for predicting networks: local vs. global modeling



Global modeling: build one model for the whole network

Pairwise kernel: consider object pairs instead of individual objects

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



Direct methods: threshold the kernel to make predictions

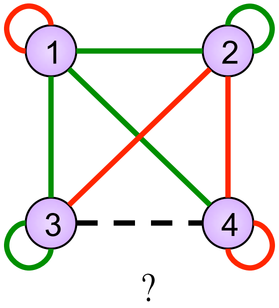
Problem: One single global model, may not be able to handle subclasses

	1	2	3	4
1	1.00	0.57	0.55	0.40
2	0.57	1.00	0.66	0.89
3	0.55	0.66	1.00	0.79
4	0.40	0.89	0.79	1.00

Threshold: 0.7 →

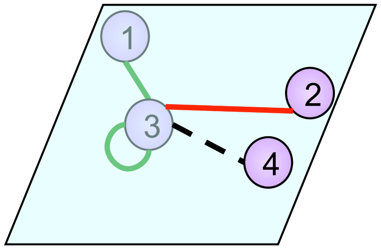
	1	2	3	4
1	1.00	0.57	0.55	0.40
2	0.57	1.00	0.66	0.89
3	0.55	0.66	1.00	0.79
4	0.40	0.89	0.79	1.00

Kernel methods for predicting networks: local vs. global modeling



Local modeling: build one model for each node

Model for node 3:



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

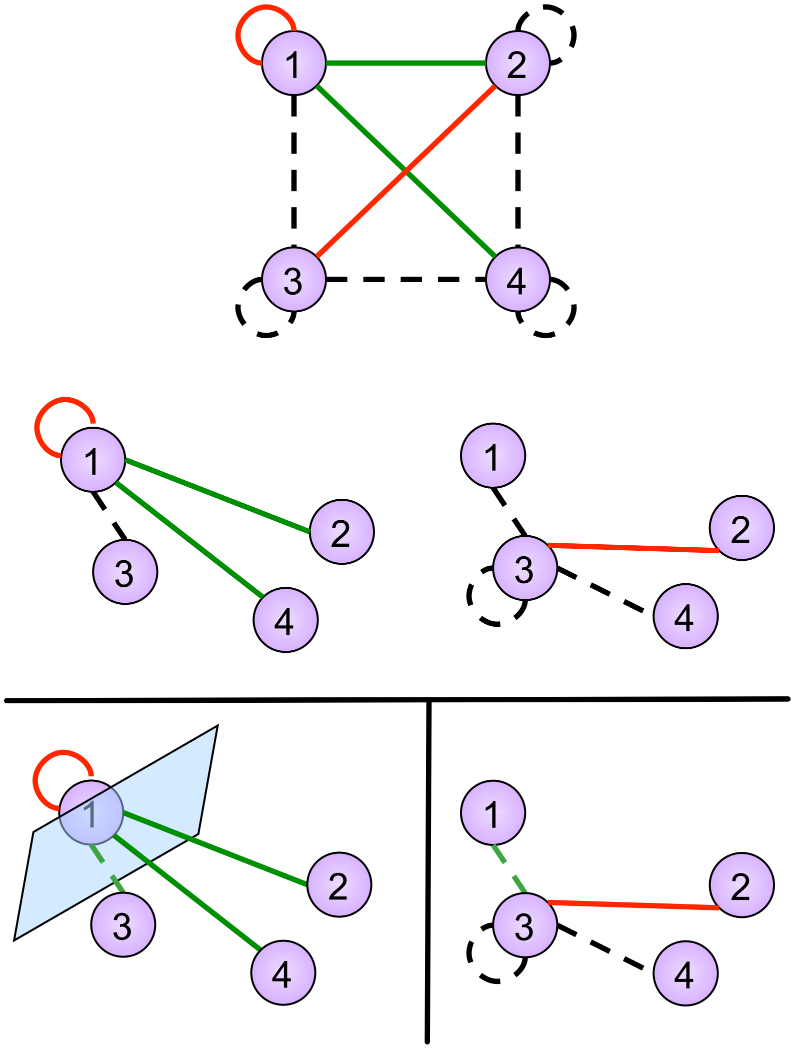
Our work: training set expansion

- Goal:
  - ◊ Utilize the flexibility of local modeling
  - ◊ Tackle the problem of insufficient training data
- Idea: generate auxiliary training data
  - ◊ Prediction propagation
  - ◊ Kernel initialization

[Yip and Gerstein, Bioinformatics ('09, in press)]

Prediction propagation

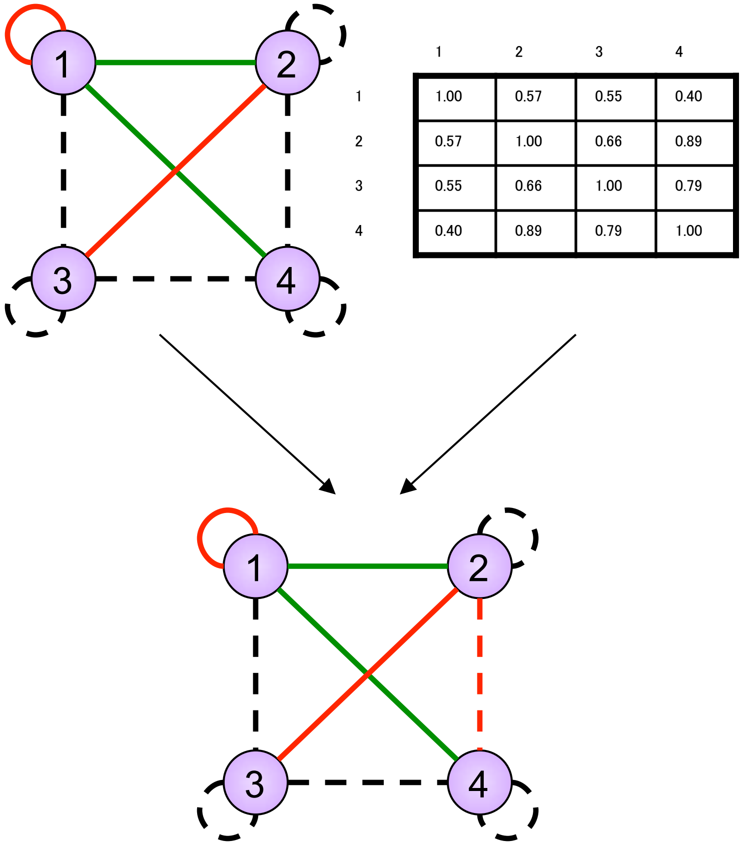
- Motivation: some objects have more examples than others
- Our approach:
  - ◊ Learn models for objects with more examples first
  - ◊ Propagate the most confident predictions as auxiliary examples of other objects



[Yip and Gerstein, Bioinformatics ('09, in press)]

Kernel initialization

- Motivation: what if most objects have very few examples?
- Our approach (inspired by the direct method):
  - ◊ Add the most similar pairs in the kernel as positive examples
  - ◊ Add the most dissimilar pairs in the kernel as negative examples



	1	2	3	4
1	1.00	0.57	0.55	0.40
2	0.57	1.00	0.66	0.89
3	0.55	0.66	1.00	0.79
4	0.40	0.89	0.79	1.00

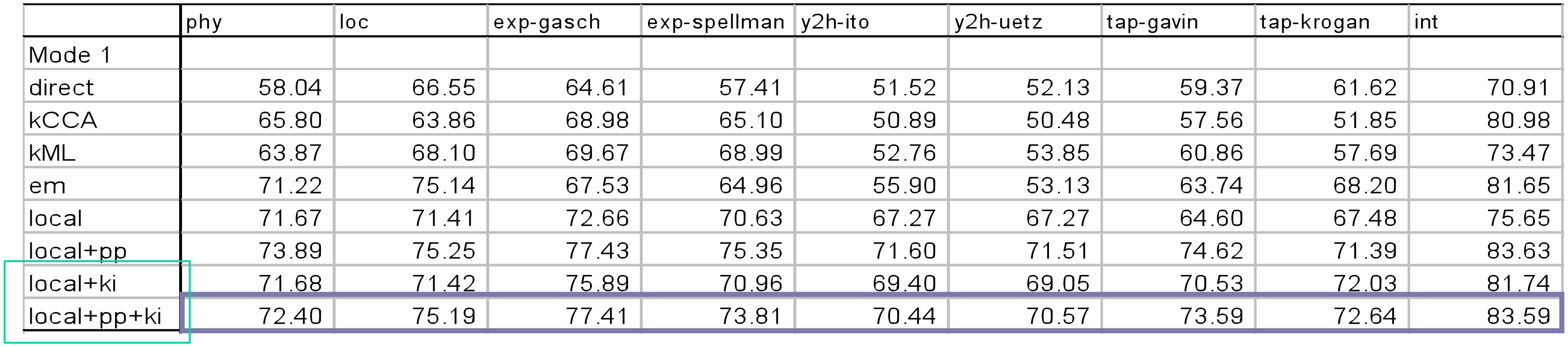
[Yip and Gerstein, Bioinformatics ('09, in press)]

Remarks

- Can be used in combination
- Prediction propagation theoretically related to co-training (Blum and Mitchell, 1998)
  - ◊ Semi-supervised
    - Similarity with PSI-BLAST
- Algorithm complexity O(nf(n)) of local modeling vs. O(f(n^2)) of global modeling

[Yip and Gerstein, Bioinformatics ('09, in press)]

Prediction accuracy (AUC)



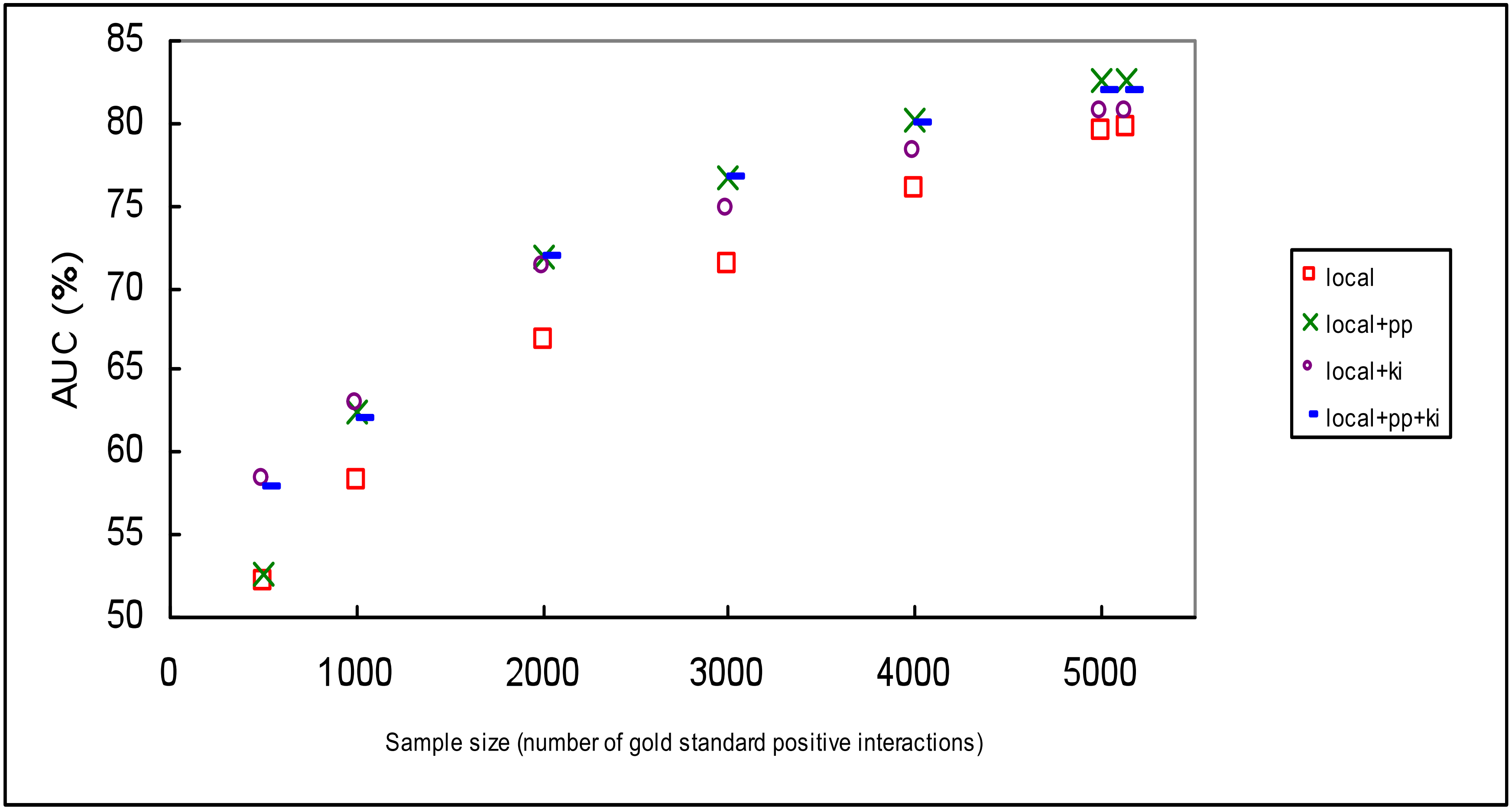
	phy	loc	exp-gasch	exp-spellman	y2h-ito	y2h-uetz	tap-gavin	tap-krogan	int
Mode 1									
direct	58.04	66.55	64.61	57.41	51.52	52.13	59.37	61.62	70.91
kCCA	65.80	63.86	68.98	65.10	50.89	50.48	57.56	51.85	80.98
kML	63.87	68.10	69.67	68.99	52.76	53.85	60.86	57.69	73.47
em	71.22	75.14	67.53	64.96	55.90	53.13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

Observations:

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

[Yip and Gerstein, Bioinformatics ('09)]

Complementarity of the two methods



[Yip and Gerstein, Bioinformatics ('09, in press)]

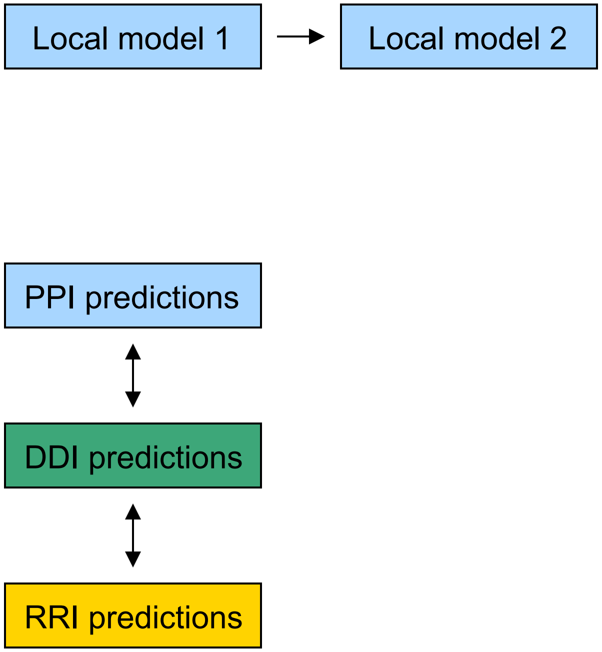
From horizontal to vertical

Training set expansion

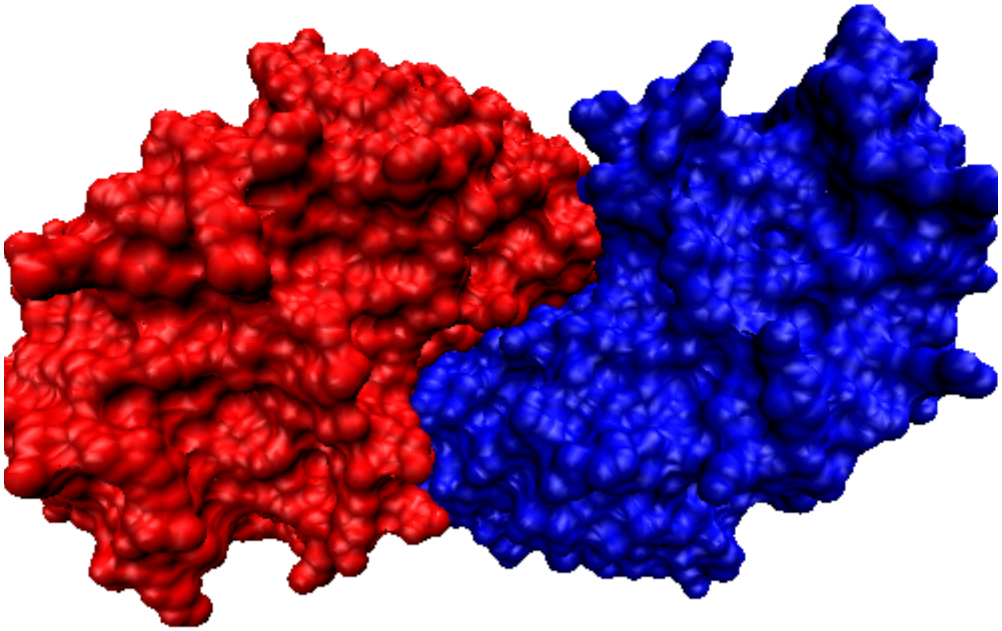
- Motivation: lack of training examples
- Expand training sets horizontally

Multi-level learning

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



Protein interaction

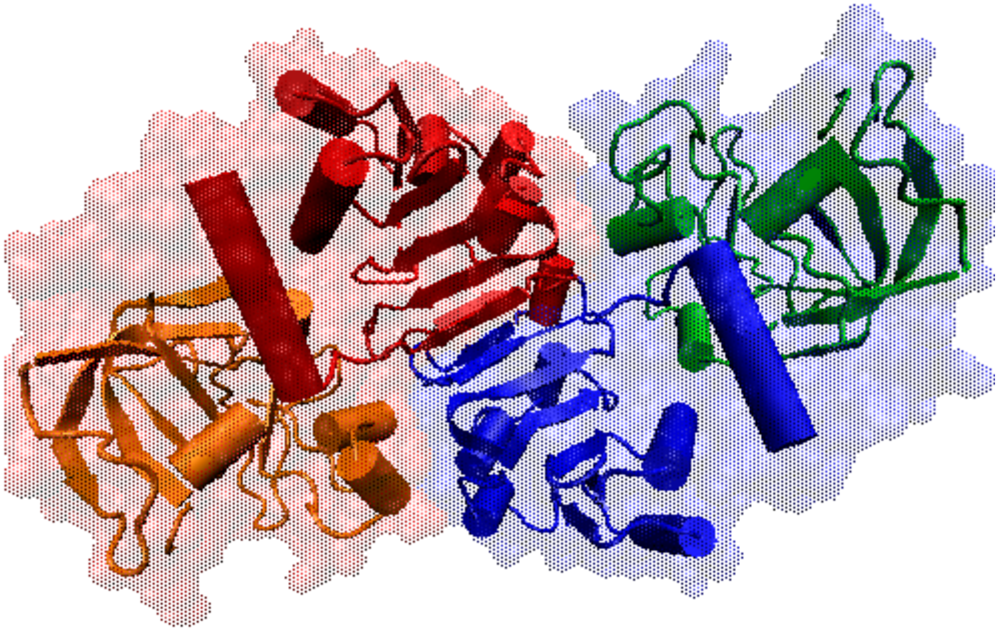


Yeast NADP-dependent alcohol dehydrogenase 6 (PDB: 1piw)

Protein-level features for interaction prediction: functional genomic information

[Yip and Gerstein, BMC Bioinfo. ('09, press)]

Domain interaction

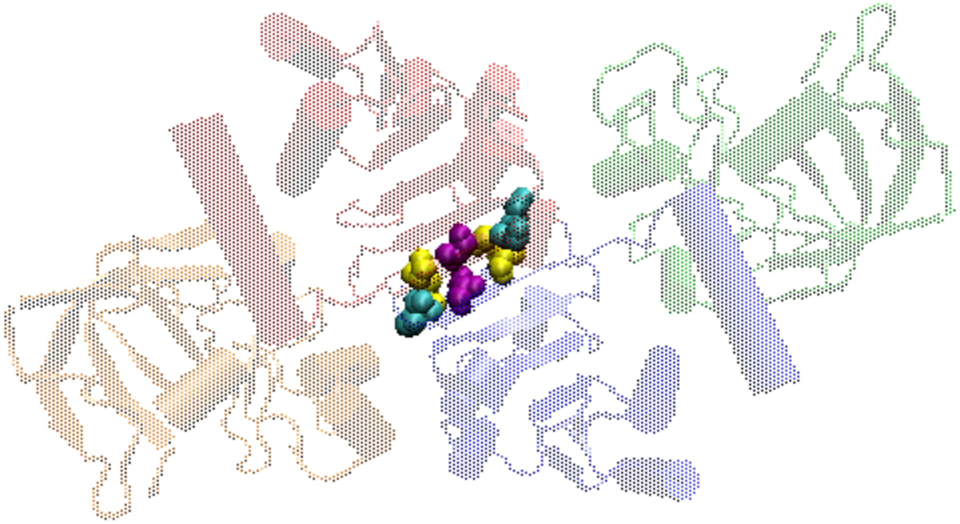


Pfam domains: PF00107 (inner) and PF08240 (outer)

Domain-level features for interaction prediction: evolutionary information

[Yip and Gerstein, BMC Bioinfo. ('09, press)]

Residue interaction

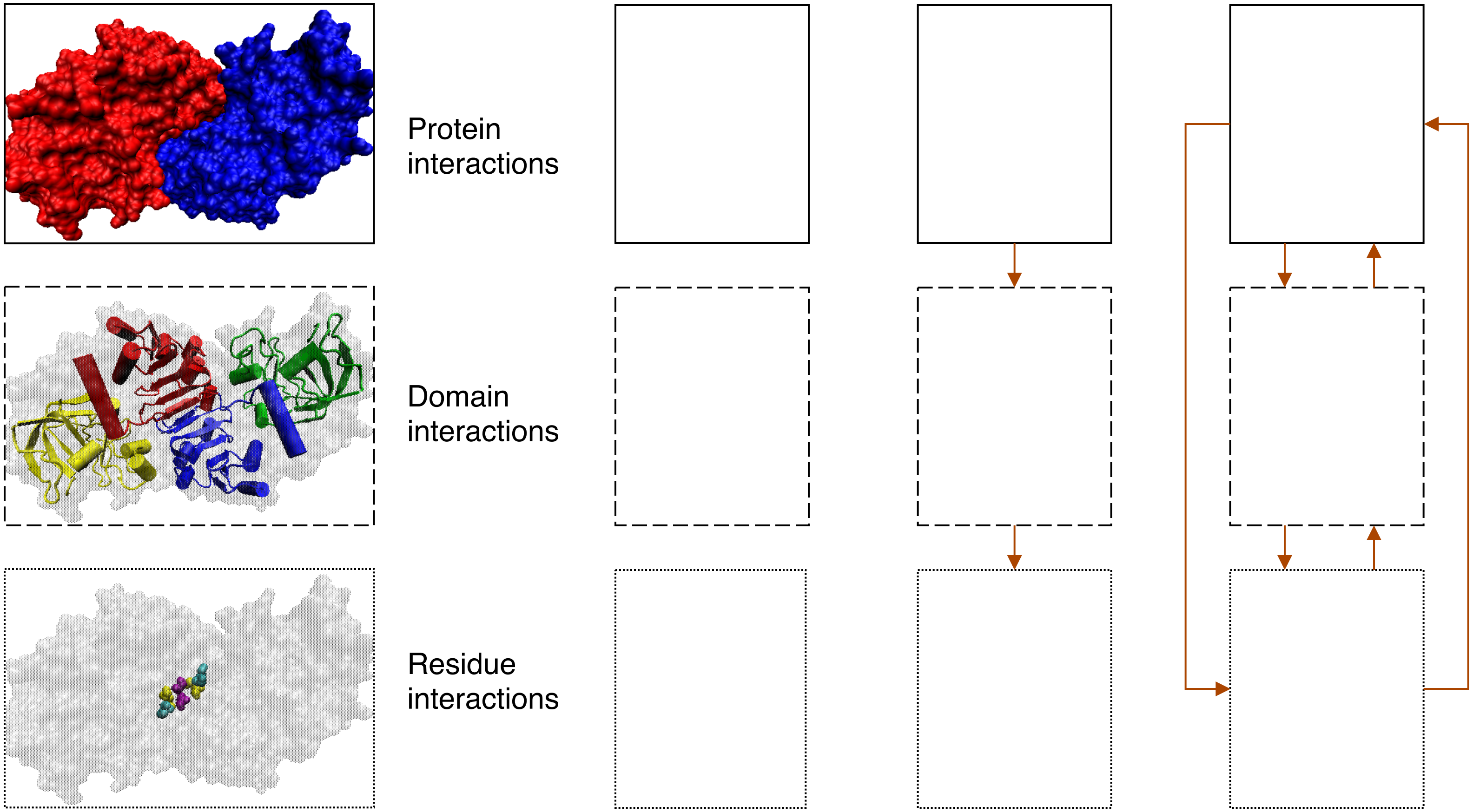


Interacting residues: 283 (yellow) with 287 (cyan), and 285 (purple) with 285

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

[Yip and Gerstein, BMC Bioinfo. ('09, press)]

Combining the three problems



Protein interactions

Domain interactions

Residue interactions

i. Independent levels

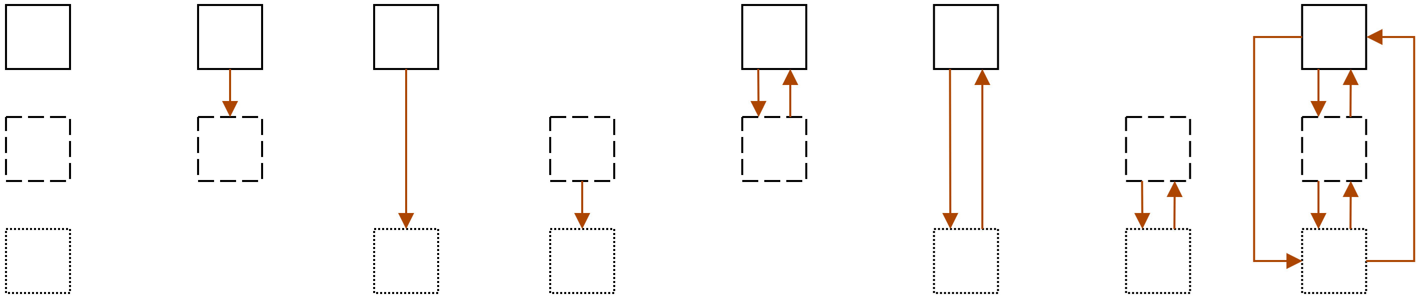
ii. Unidirectional flow

iii. Bidirectional flow

[Yip and Gerstein, BMC Bioinfo. ('09, press)]

Empirical results (AUCs)

	Ind. levels	Unidirectional flow			Bidirectional flow			
Level		PD	PR	DR	PD	PR	DR	PDR
Proteins	71.68				72.23	72.50		**72.82**
Domains	53.18	61.51			**71.71**		68.94	71.20
Residues	57.36		54.89	53.81		72.26	63.16	**77.86**



- Highest accuracy by bidirectional flow
- Additive effect: 2 vs. 3 levels

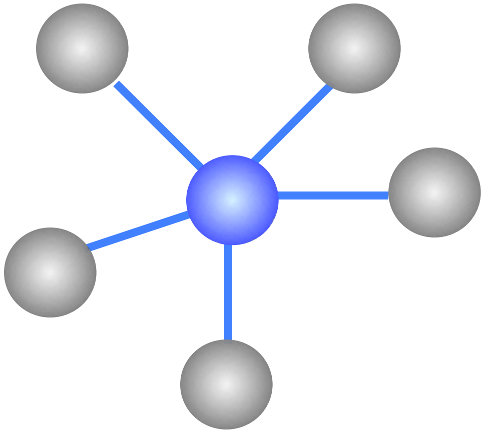
[Yip and Gerstein, BMC Bioinfo. ('09, press)]

Finding Central Points in Networks: Hubs & Bottlenecks

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

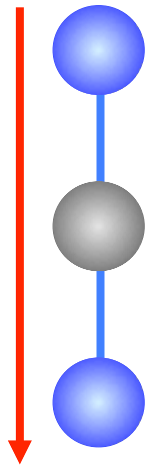
Global topological measures

Indicate the gross topological structure of the network



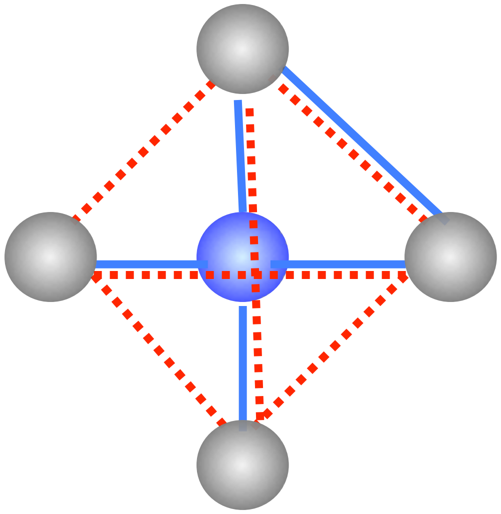
Degree (*K*)

5



Path length (*L*)

2



Clustering coefficient (*C*)

1/6

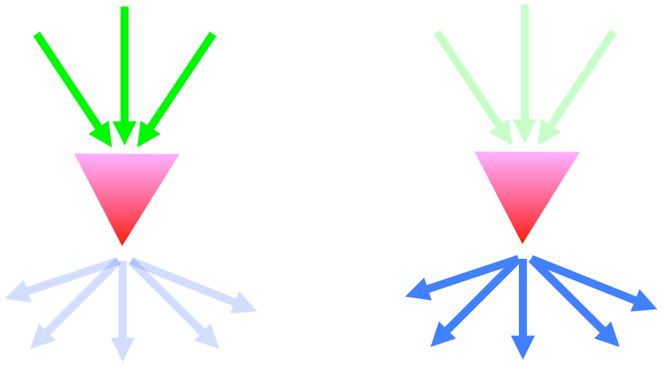
Interaction and expression networks are ***undirected***

[Barabasi]

Global topological measures for directed networks

▼ TFs

● Targets



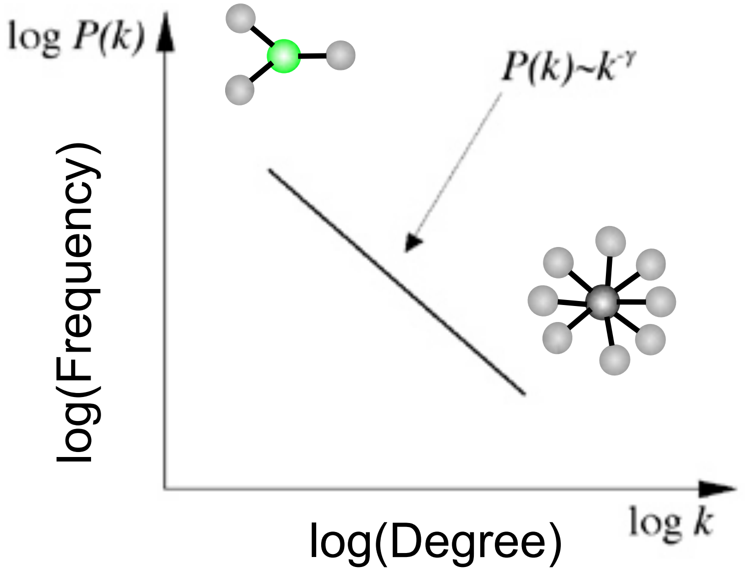
In-degree
3

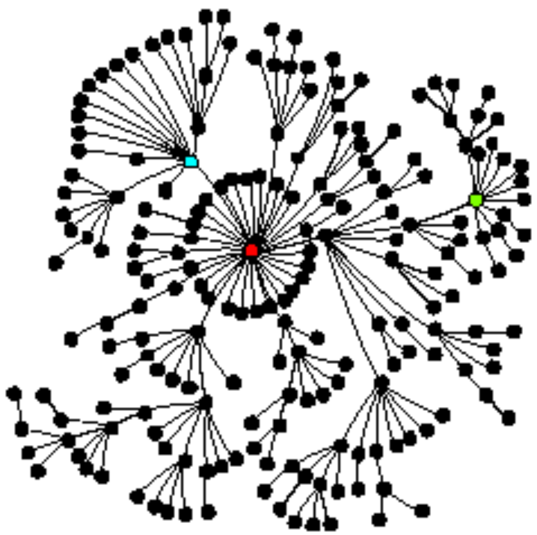
Out-degree
5

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

Scale-free networks

Power-law distribution





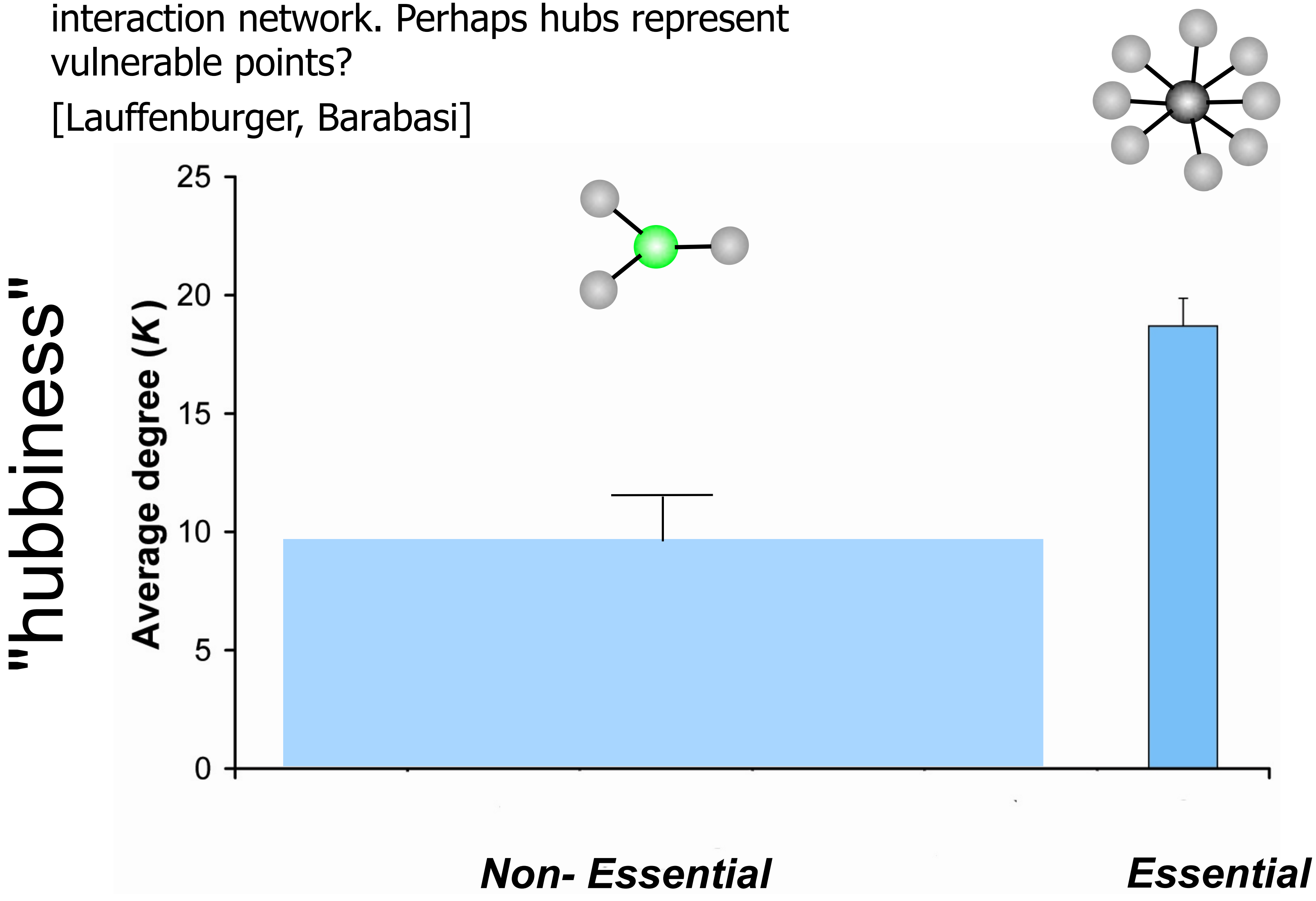
Hubs dictate the structure of the network

[Barabasi]

Hubs tend to be Essential

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

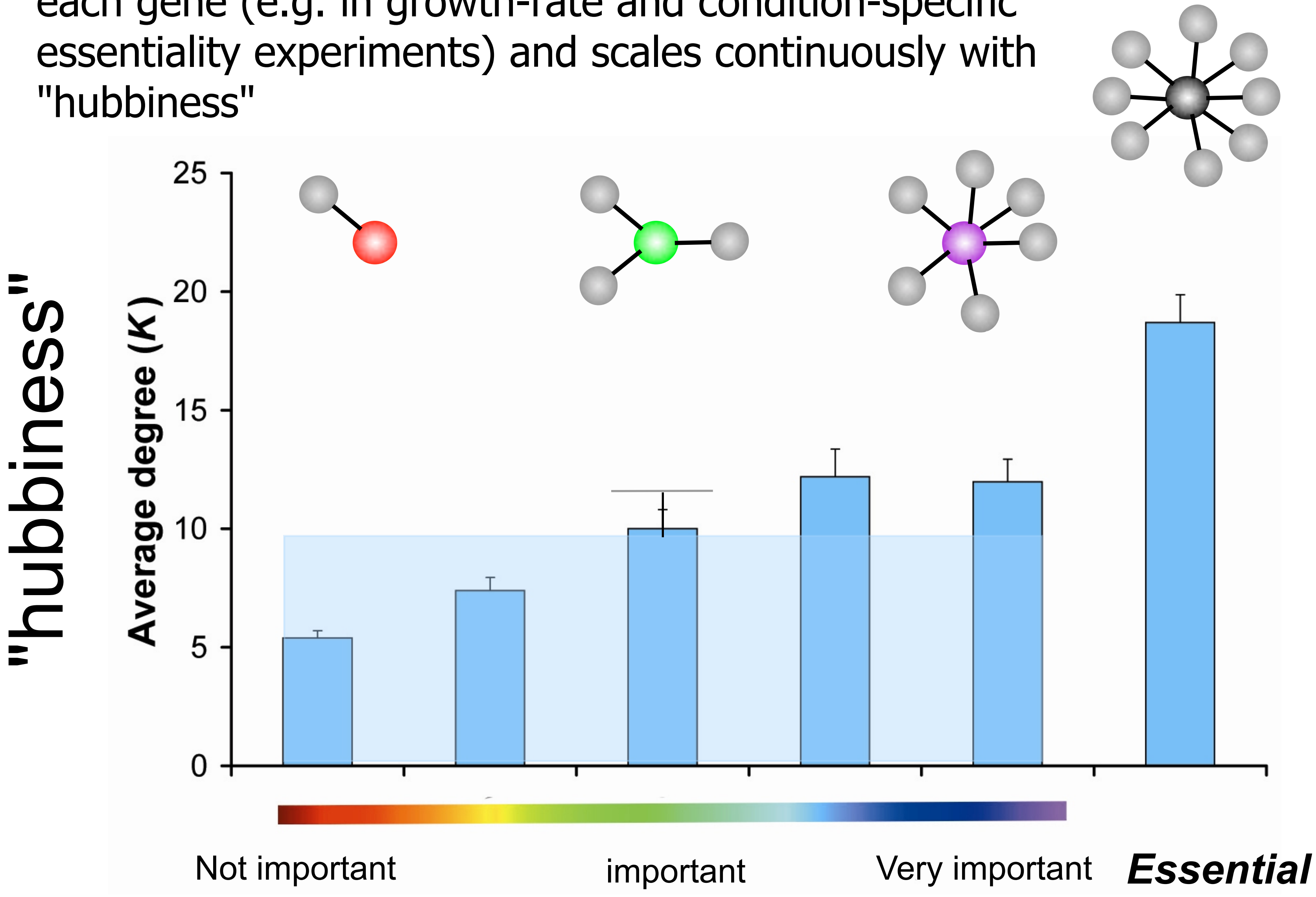
[Lauffenburger, Barabasi]



[Yu et al., 2003, TIG]

Relationships extends to "Marginal Essentiality"

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



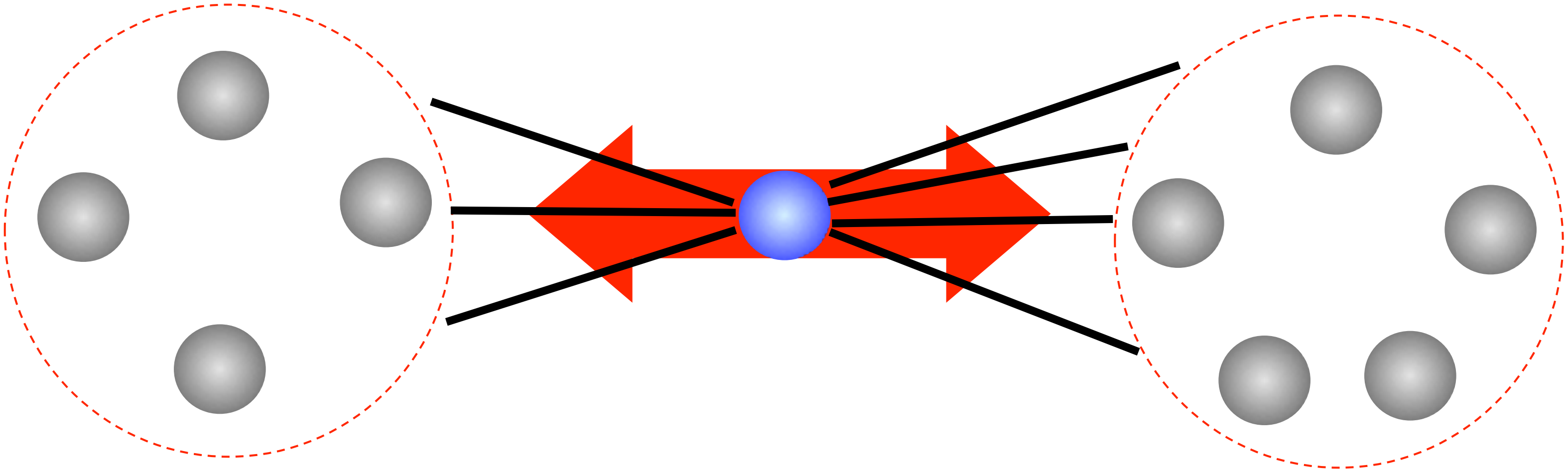
[Yu et al., 2003, TIG]

Another measure of Centrality: Betweenness centrality

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

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

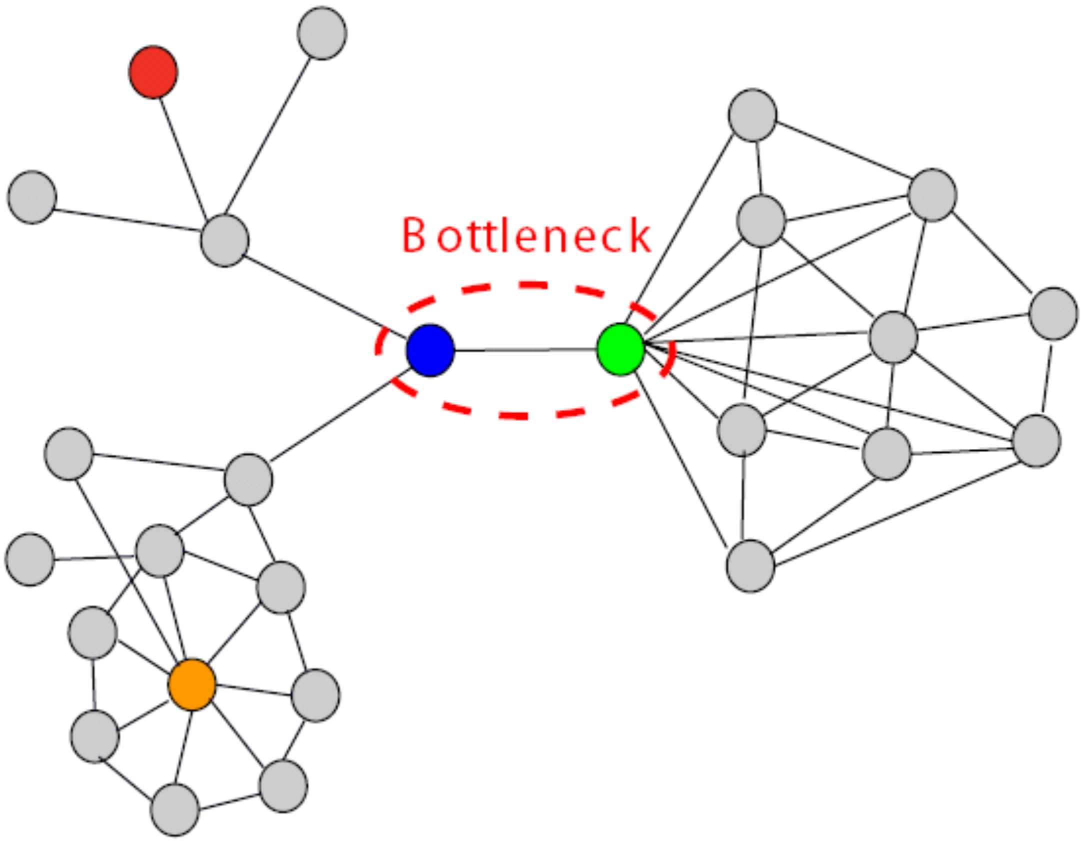
Girvan & Newman (2002) PNAS 99: 7821.



Betweenness centrality -- Bottlenecks

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

Bottlenecks & Hubs



🟢 Hub-bottleneck **node**

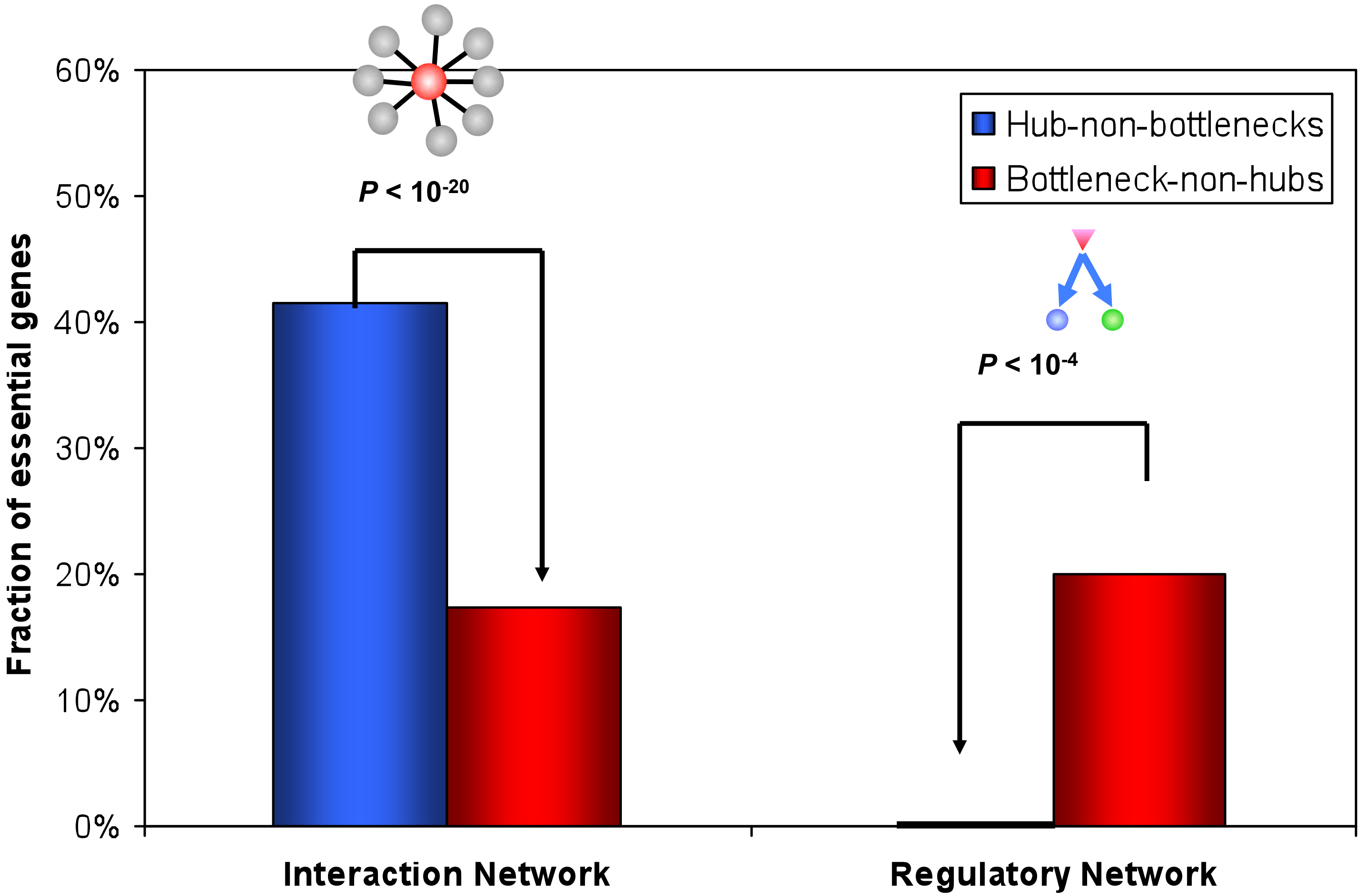
🔵 Non-hub-bottleneck **node**

🟠 Hub-non-bottleneck **node**

🔴 Non-hub-non-bottleneck **node**

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

Bottlenecks are what matters in regulatory networks

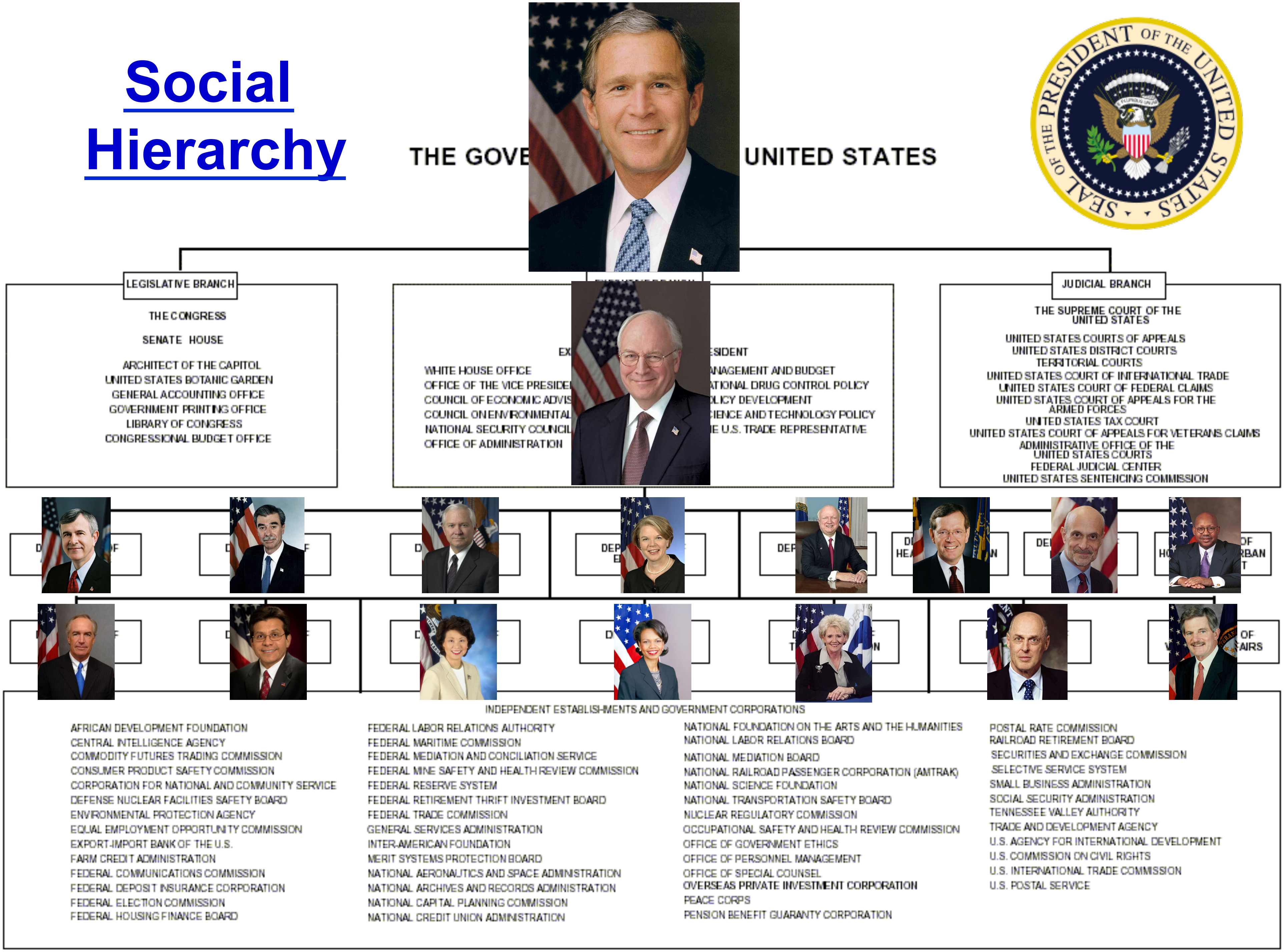


[Yu *et al.*, *PLoS Comput Biol* (2007)]

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

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

Social Hierarchy



THE GOVE
UNITED STATES
LEGISLATIVE BRANCH
THE CONGRESS
SENATE HOUSE
ARCHITECT OF THE CAPITOL
UNITED STATES BOTANIC GARDEN
GENERAL ACCOUNTING OFFICE
GOVERNMENT PRINTING OFFICE
LIBRARY OF CONGRESS
CONGRESSIONAL BUDGET OFFICE
EX
ESIDENT
WHITE HOUSE OFFICE
OFFICE OF THE VICE PRESIDE
COUNCIL OF ECONOMIC ADVIS
COUNCIL ON ENVIRONMENTAL
NATIONAL SECURITY COUNCIL
OFFICE OF ADMINISTRATION
ANAGEMENT AND BUDGET
ATIONAL DRUG CONTROL POLICY
OLICY DEVELOPMENT
CIENCE AND TECHNOLOGY POLICY
E U.S. TRADE REPRESENTATIVE
JUDICIAL BRANCH
THE SUPREME COURT OF THE UNITED STATES
UNITED STATES COURTS OF APPEALS
UNITED STATES DISTRICT COURTS
TERRITORIAL COURTS
UNITED STATES COURT OF INTERNATIONAL TRADE
UNITED STATES COURT OF FEDERAL CLAIMS
UNITED STATES COURT OF APPEALS FOR THE ARMED FORCES
UNITED STATES TAX COURT
UNITED STATES COURT OF APPEALS FOR VETERANS CLAIMS
ADMINISTRATIVE OFFICE OF THE UNITED STATES COURTS
FEDERAL JUDICIAL CENTER
UNITED STATES SENTENCING COMMISSION
DEP
HEA
DE
HO
OF
RBAN
OF
AIRS

INDEPENDENT ESTABLISHMENTS AND GOVERNMENT CORPORATIONS

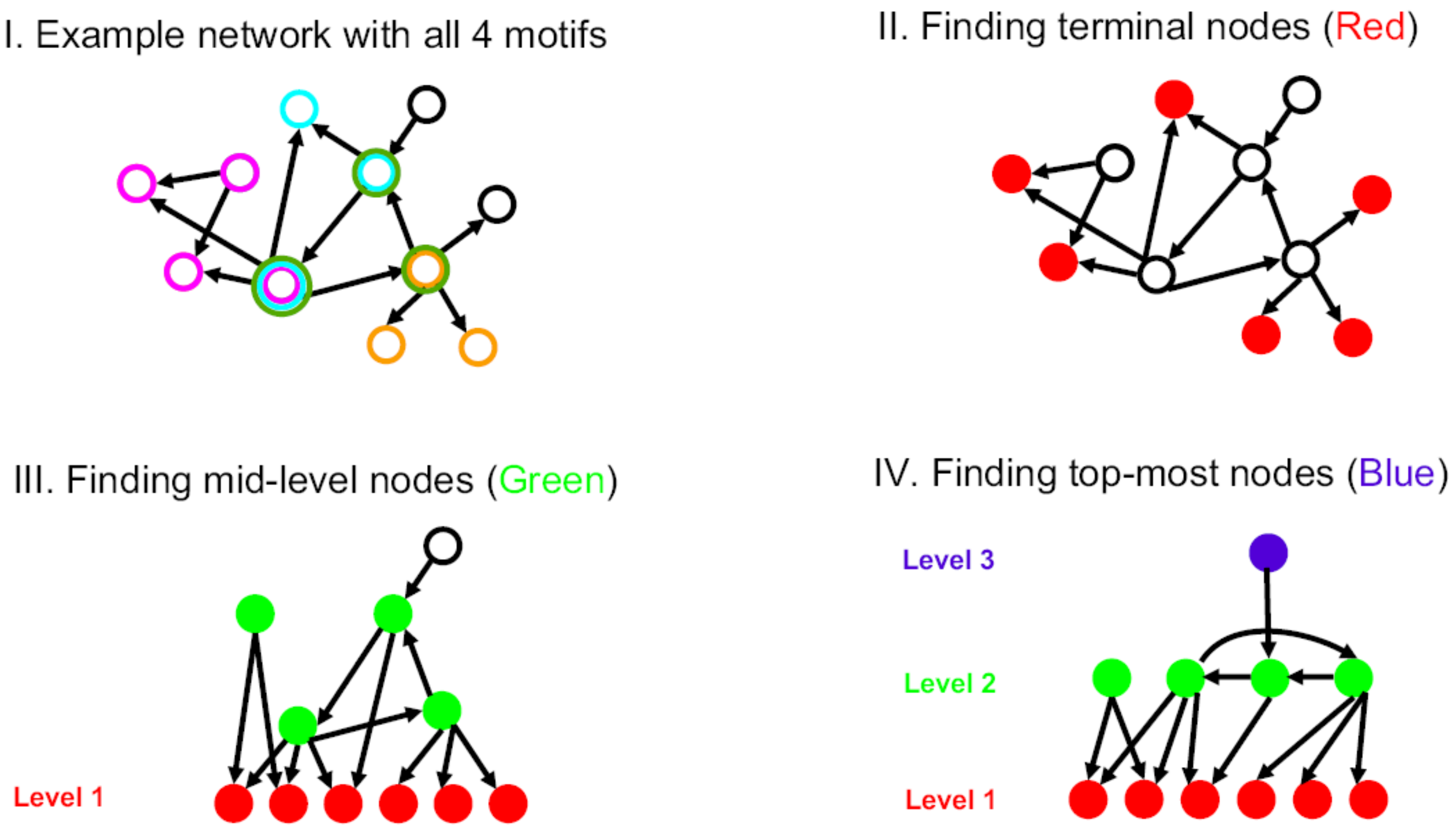
AFRICAN DEVELOPMENT FOUNDATION
CENTRAL INTELLIGENCE AGENCY
COMMODITY FUTURES TRADING COMMISSION
CONSUMER PRODUCT SAFETY COMMISSION
CORPORATION FOR NATIONAL AND COMMUNITY SERVICE
DEFENSE NUCLEAR FACILITIES SAFETY BOARD
ENVIRONMENTAL PROTECTION AGENCY
EQUAL EMPLOYMENT OPPORTUNITY COMMISSION
EXPORT-IMPORT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

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

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

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

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



I. Example network with all 4 motifs

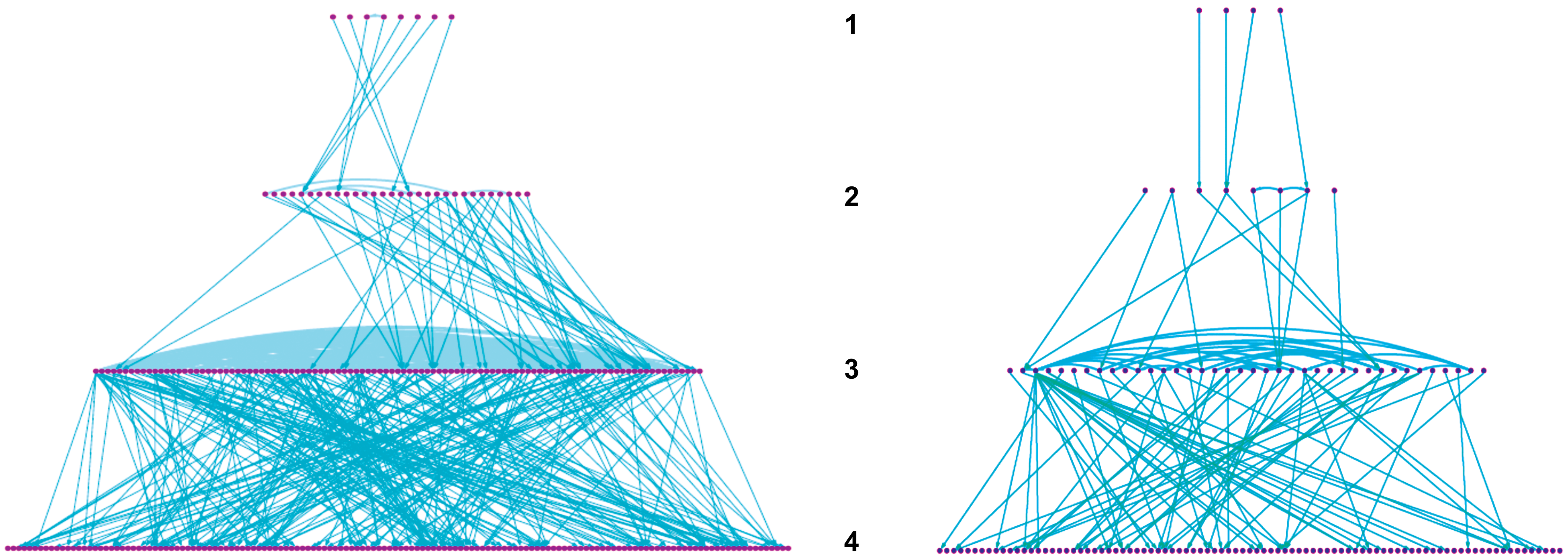
II. Finding terminal nodes (Red)

III. Finding mid-level nodes (Green)

IV. Finding top-most nodes (Blue)

[Yu et al., PNAS (2006)]

Regulatory Networks have similar hierarchical structures

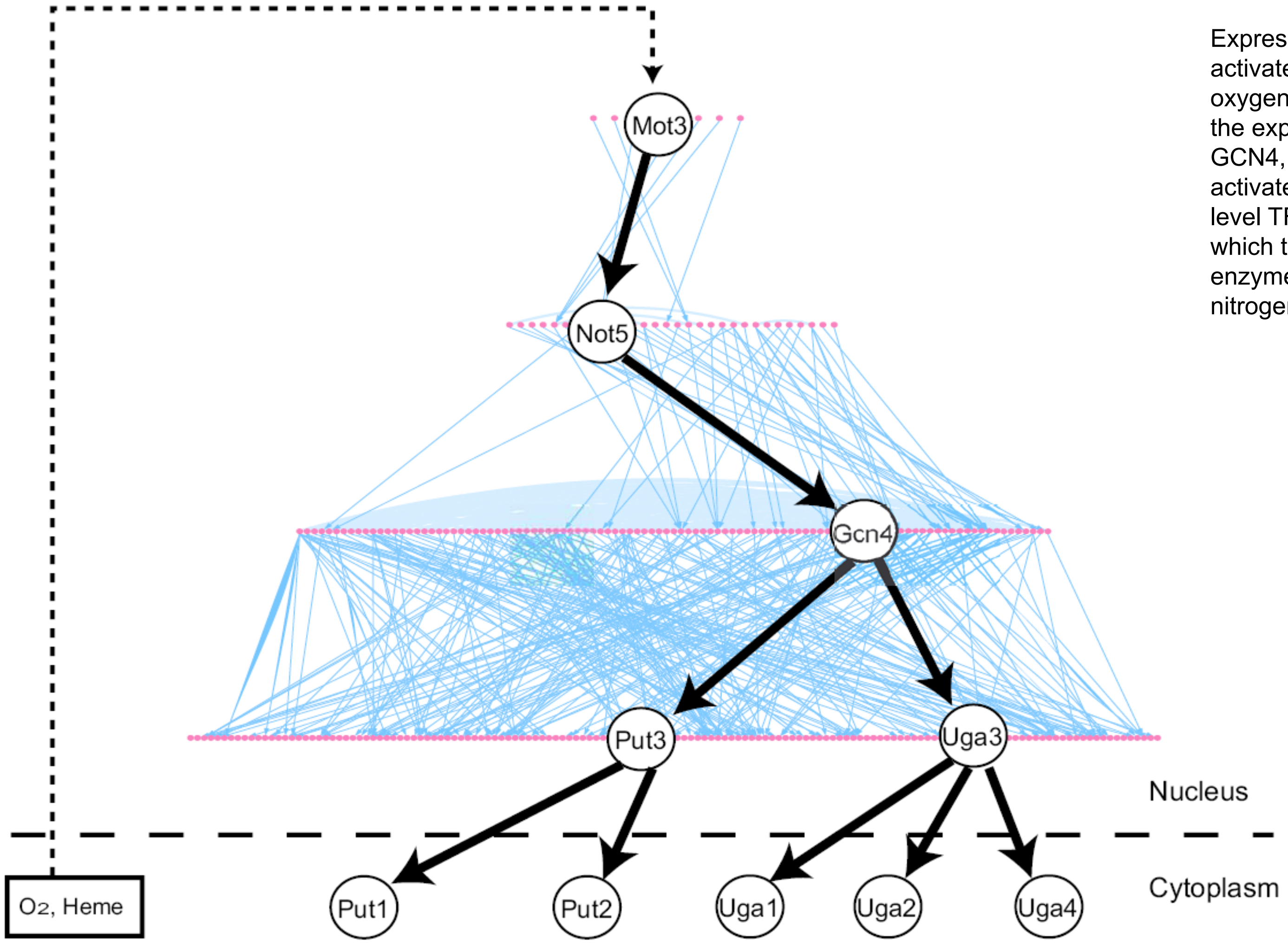


S. cerevisiae

E. coli

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

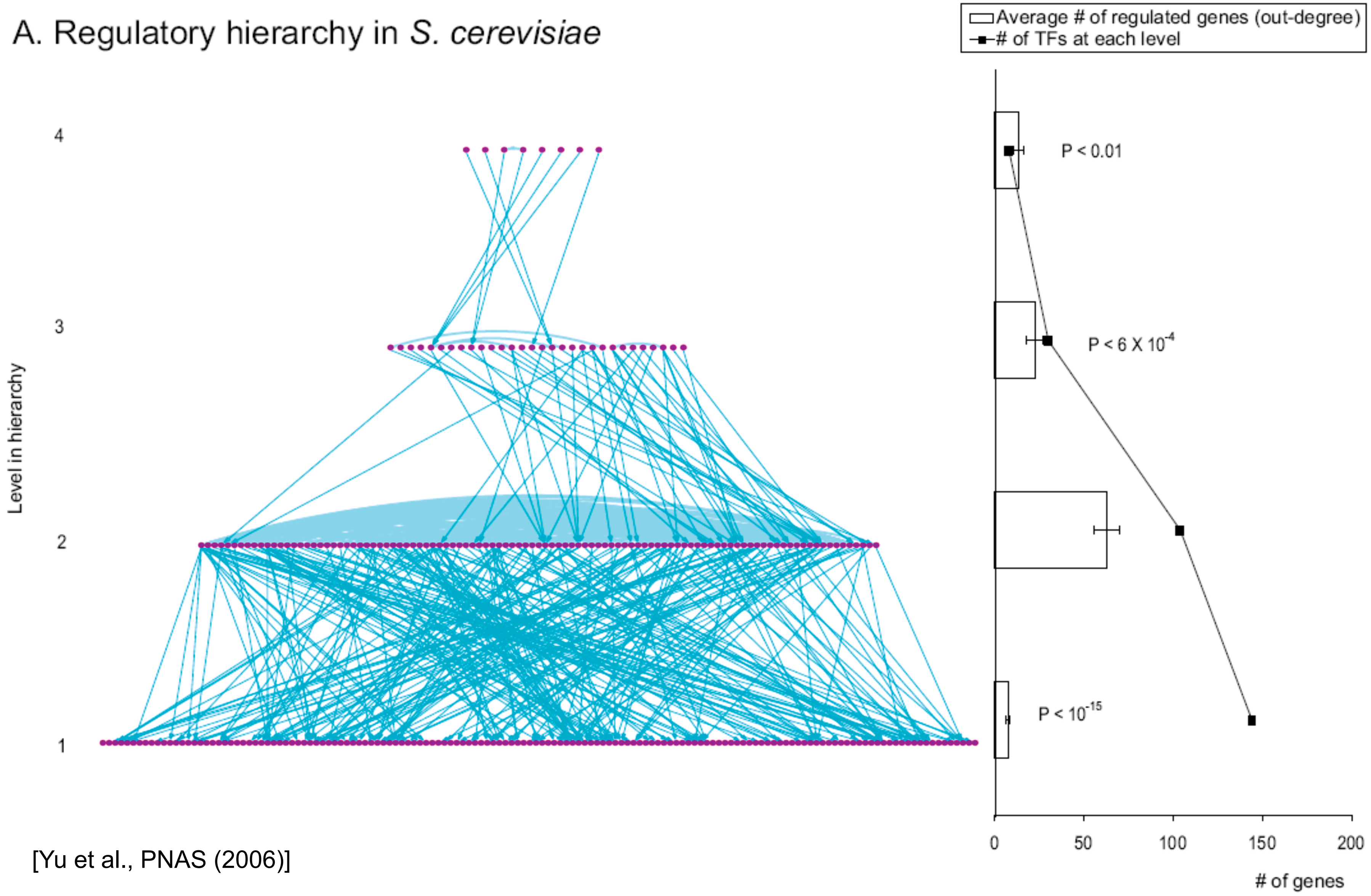
Example of Path Through Regulatory Network



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

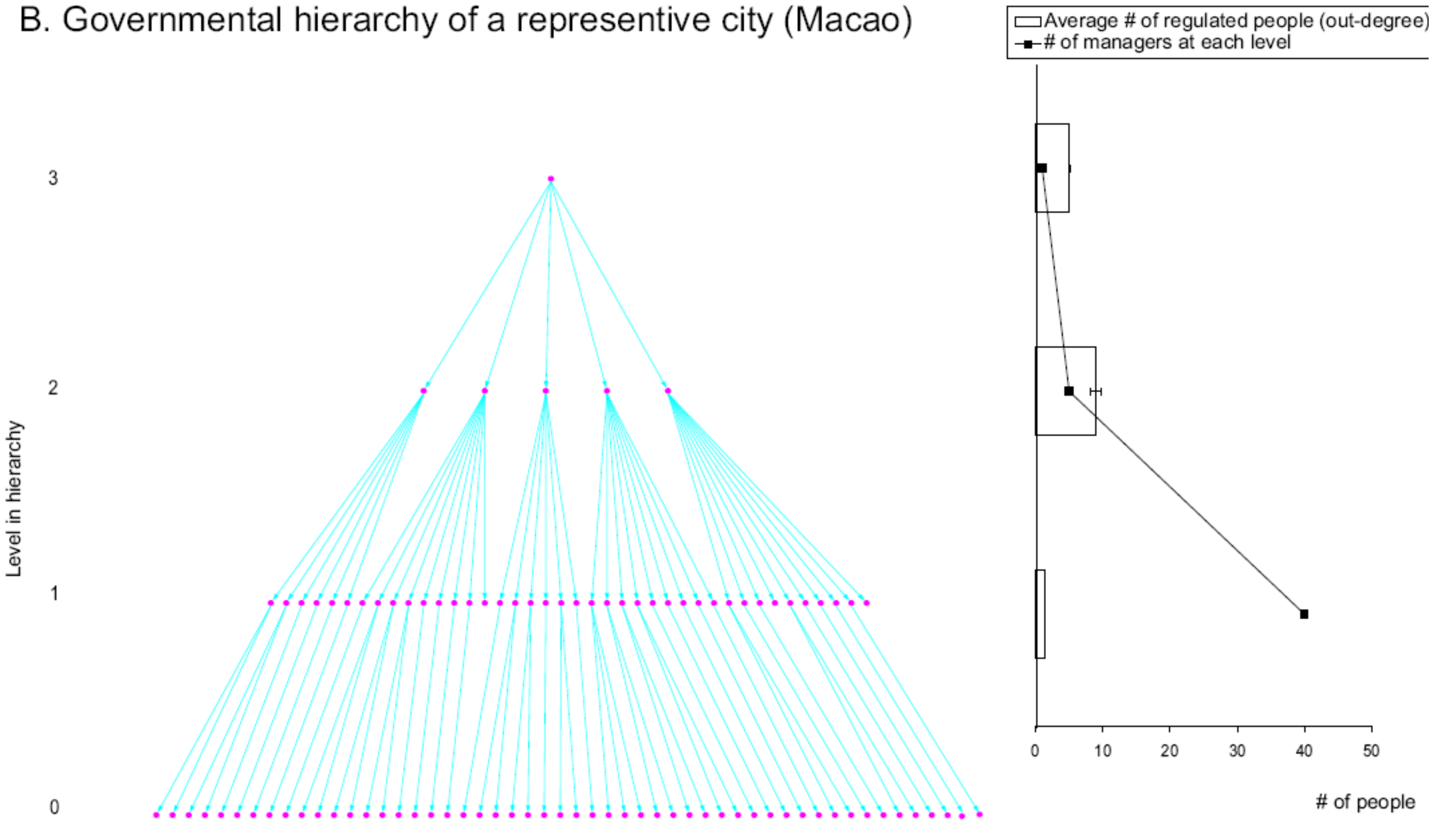
[Yu et al., PNAS (2006)]

Yeast Regulatory Hierarchy: the Middle-managers Rule

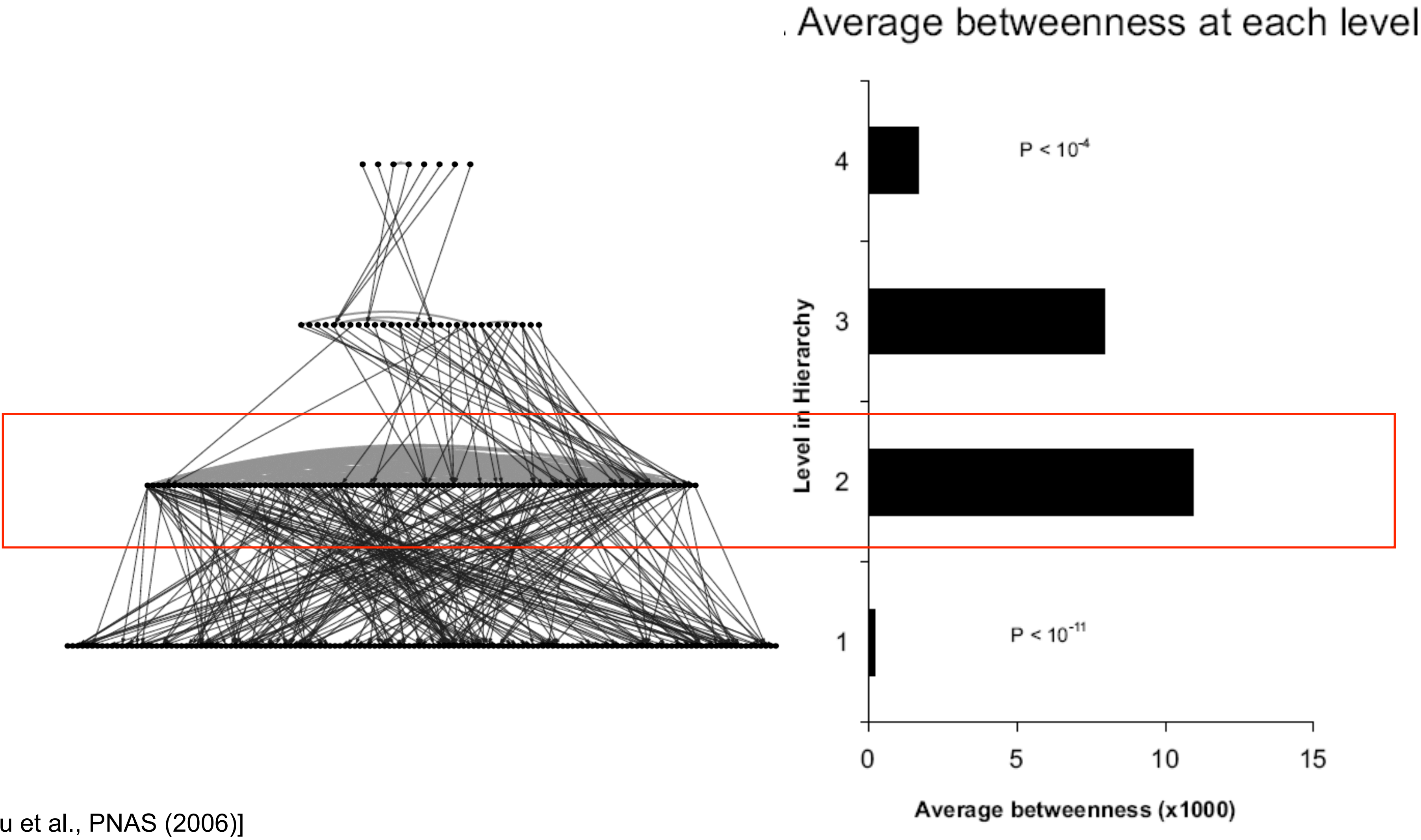


[Yu et al., PNAS (2006)]

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

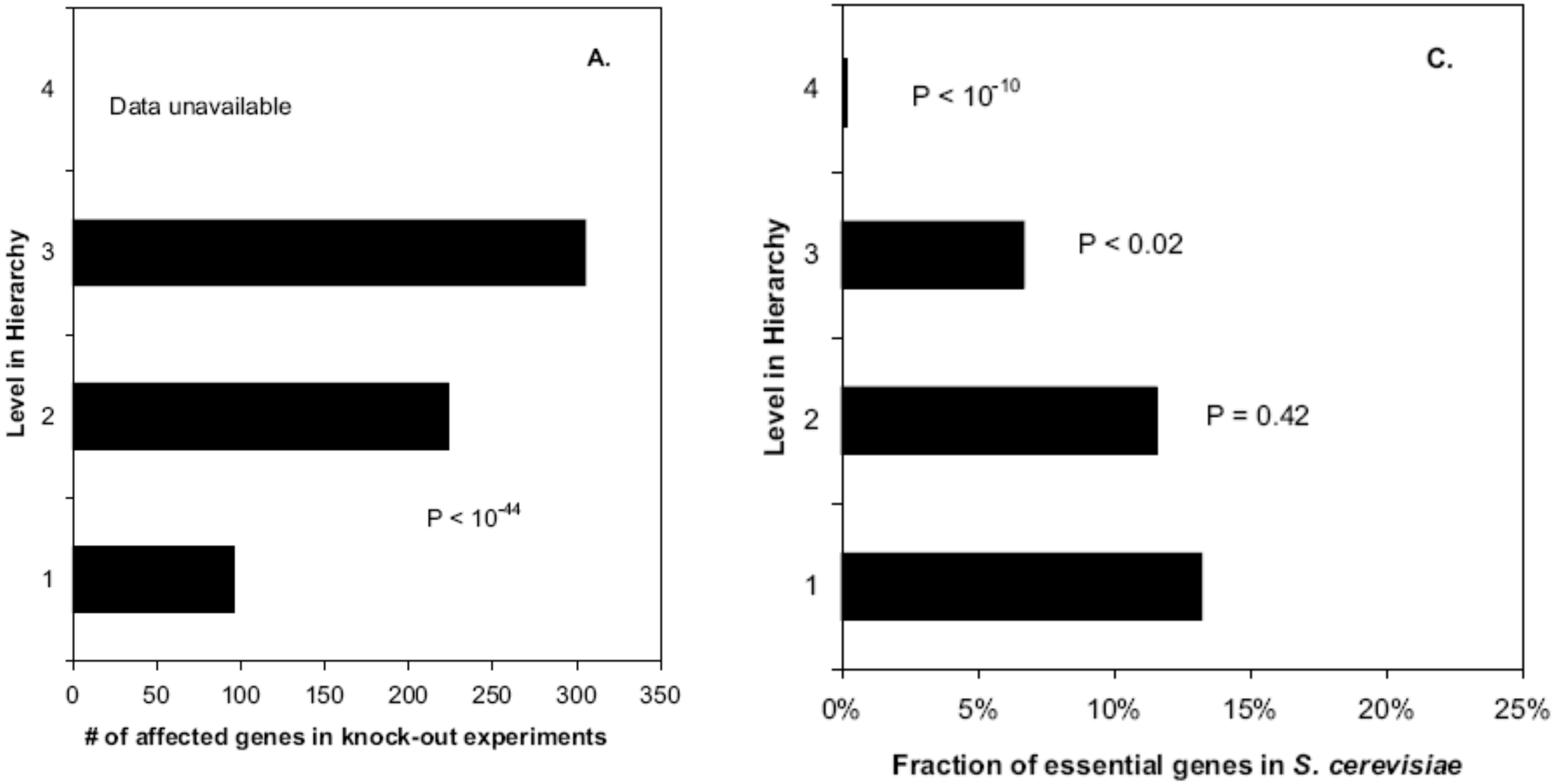


Characteristics of Regulatory Hierarchy: Middle Managers are Information Flow Bottlenecks



[Yu et al., PNAS (2006)]

Characteristics of Regulatory Hierarchy: The Paradox of Influence and Essentiality

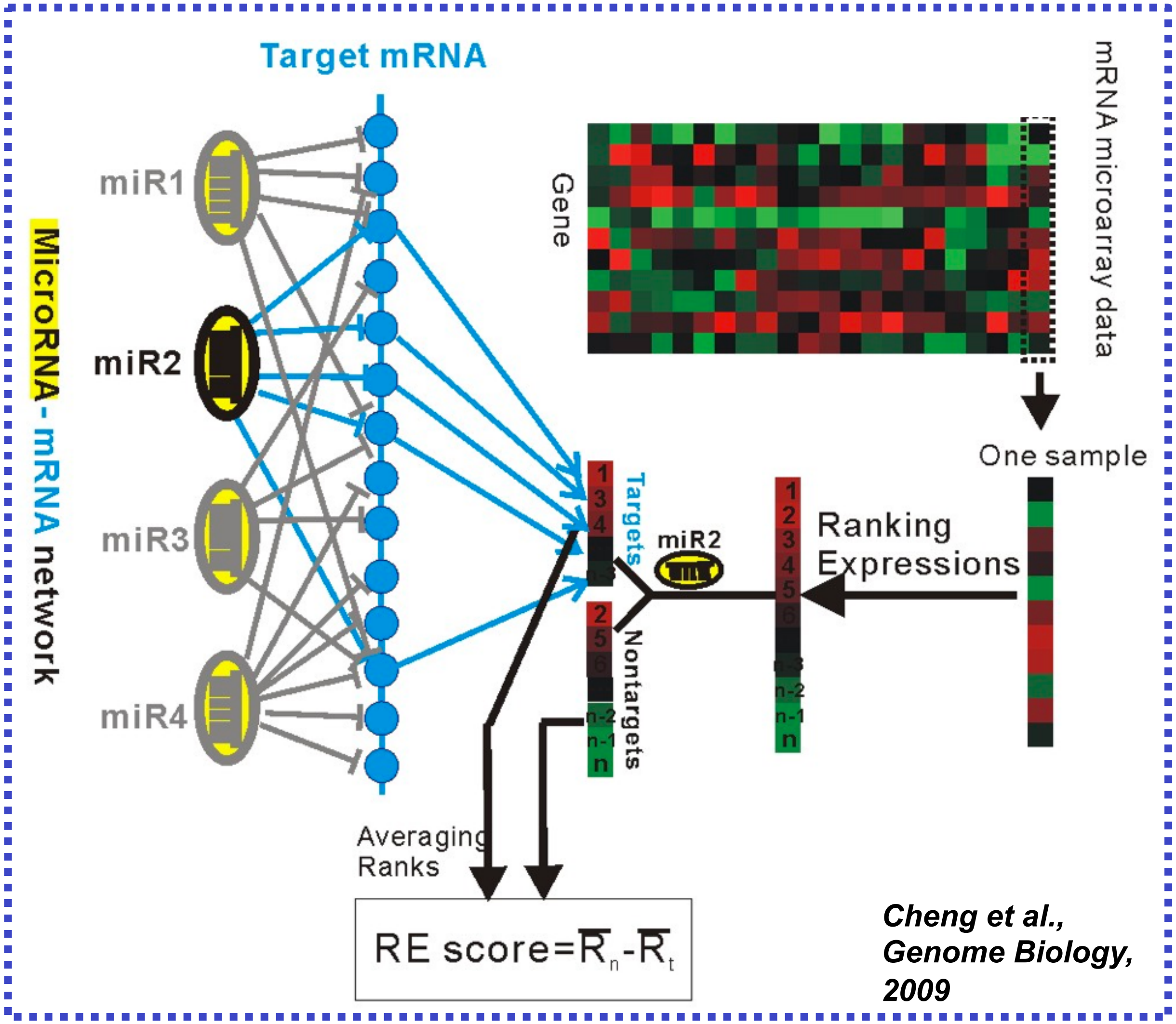


[Yu et al., PNAS (2006)]

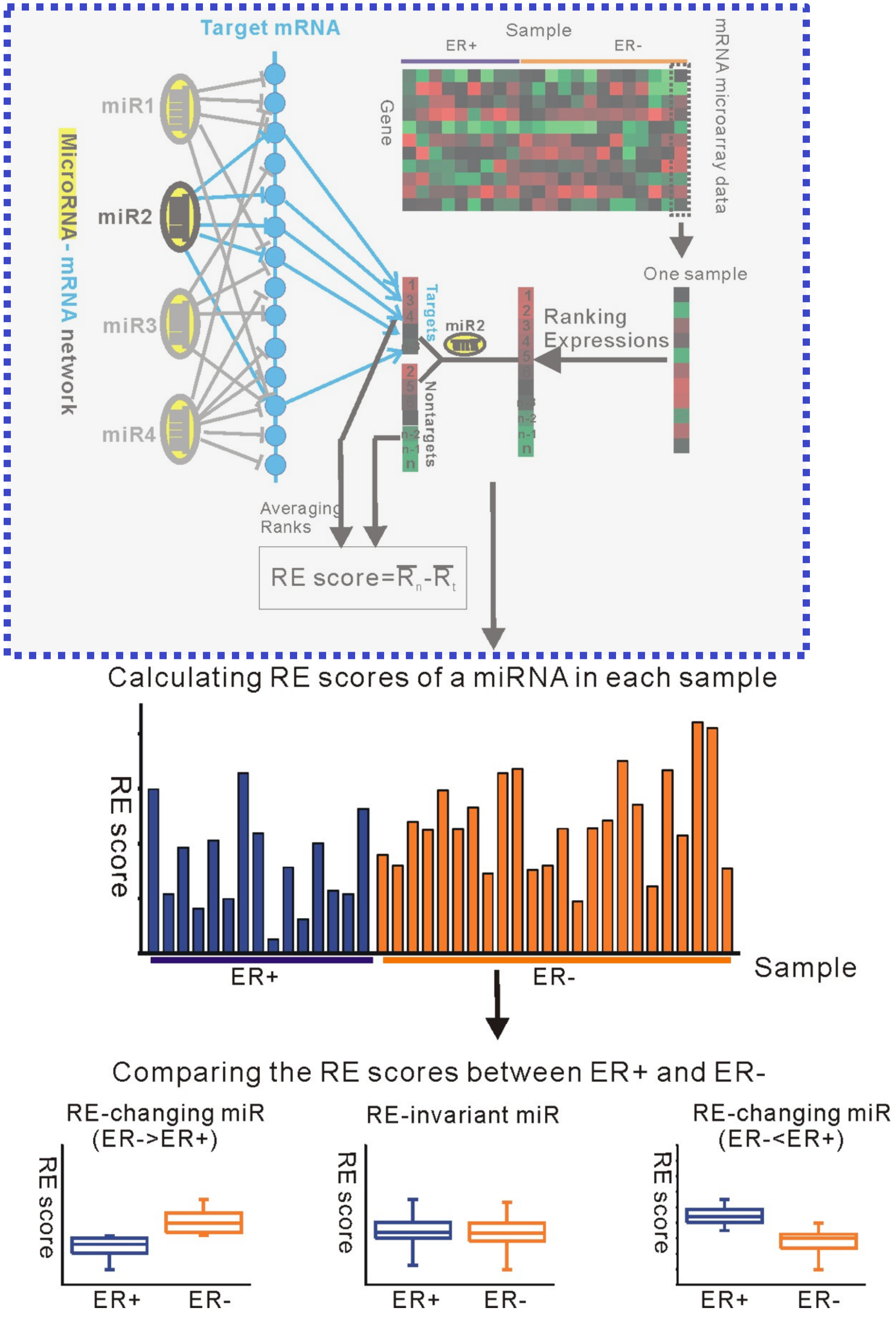
Finding Central Points in Networks #3: Points of Maximal Regulatory Effect

RE-score: Another way to measure "importance" in networks

- How much does a regulator influence its targets?
- For micro-RNA-target networks easy to calculate, as all influence is down-regulation
  - ◊ target prediction methods: 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



Cheng et al., Genome Biology, 2009

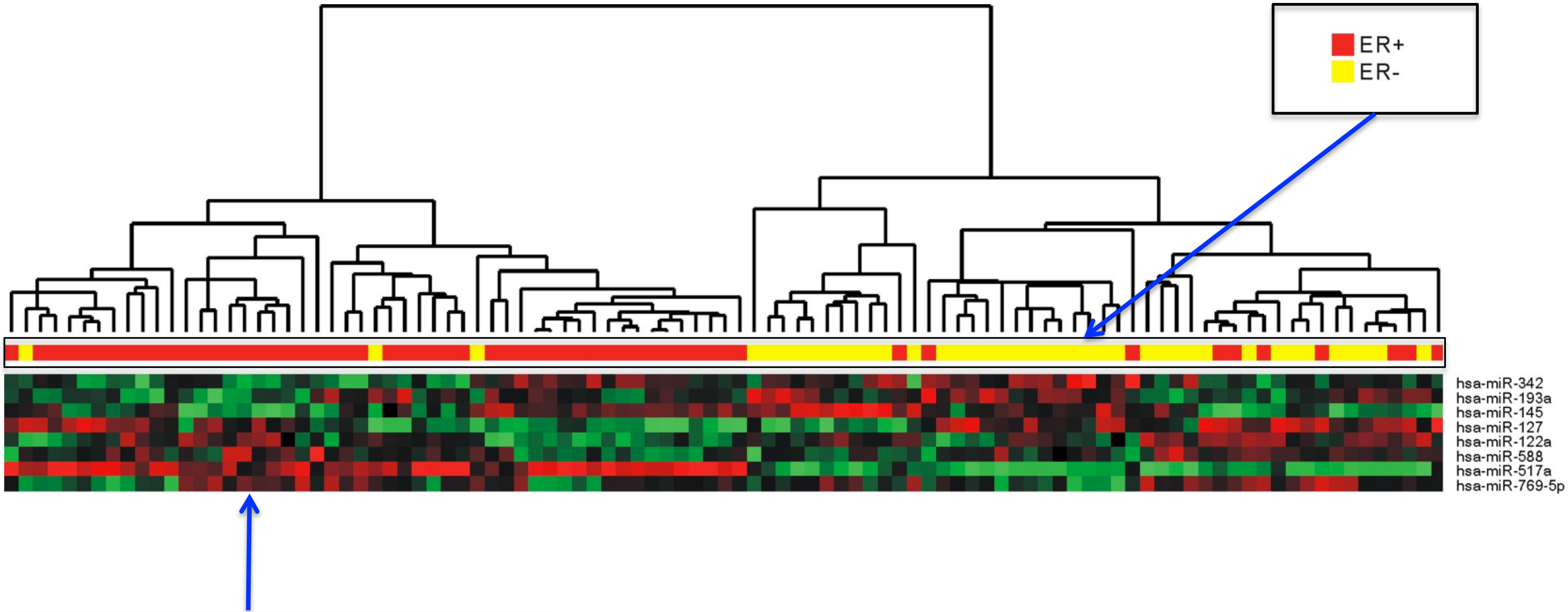


Application of RE-score to measure changing miRNA effect in different conditions

(ER- and ER+ breast cancer)

Cheng et al., Genome Biology, 2009

miRNA RE-scores can be used to classify cancers

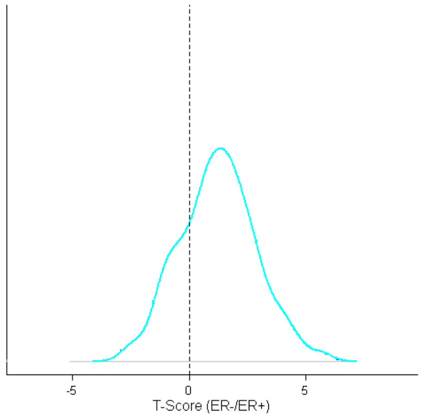


RE-score profiles for 8 miRNAs

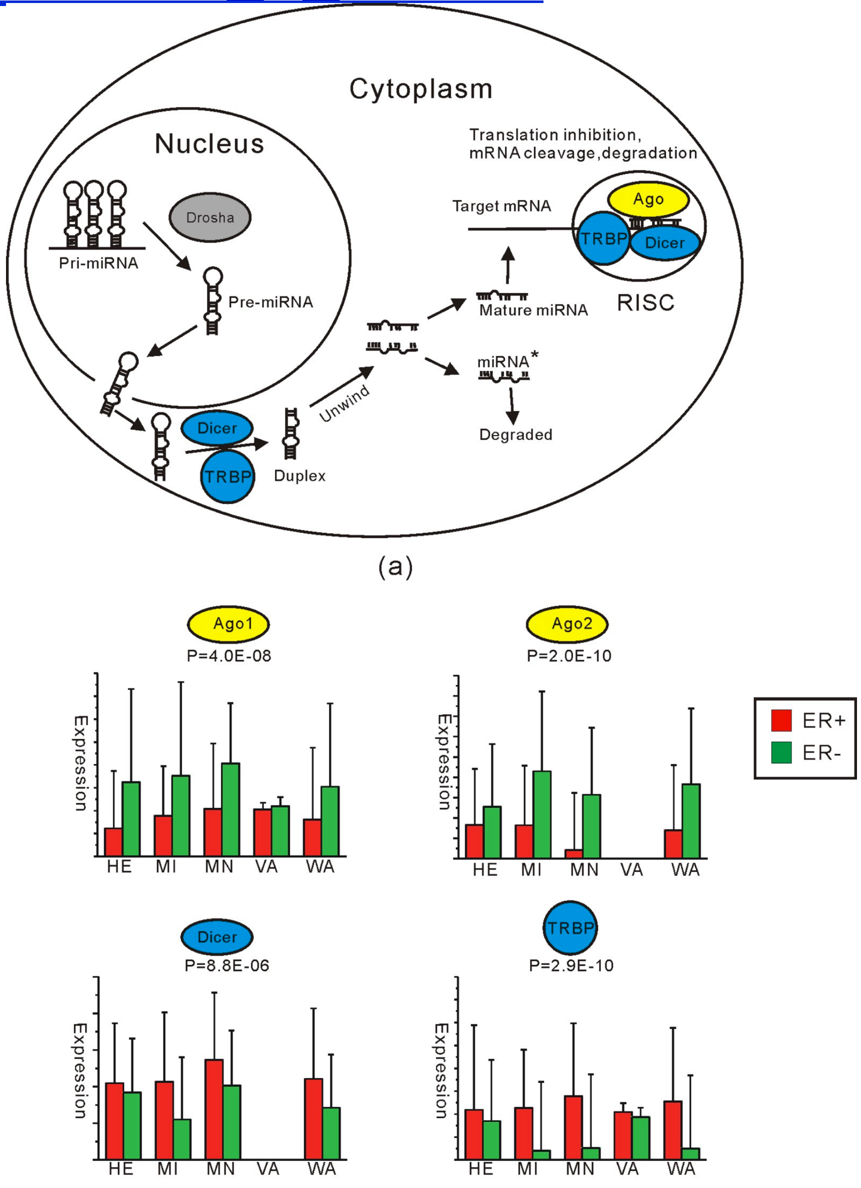
Cheng et al., Genome Biology, 2009

Differential expression of miRNA processing genes

Distribution of ER-/ER+ T-scores for all miRNAs



The majority of miRNAs have higher RE-score in ER- than in ER+

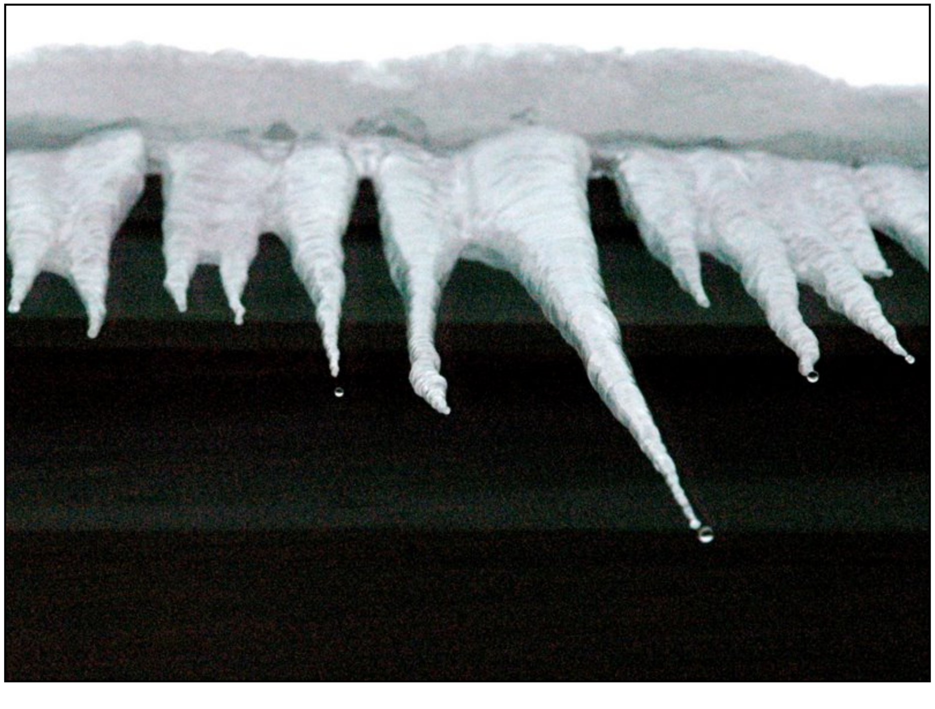


(a)

(b)

Network Dynamics #2: Environments

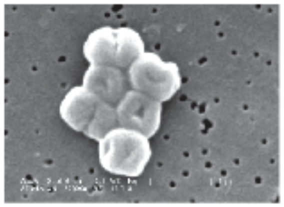
How do molecular networks change across environments?
What pathways are used more ?
Used as a biosensor ?



What is metagenomics?

Genomics Approach

Culture Microbes



→

Extract DNA

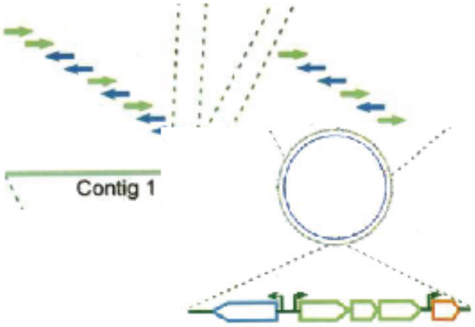
→

Sequence

ATCGTATA
CGCGAAG
ACGTCTGA
AGTGCTGCT

→

Assemble and Annotate



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

Metagenomics Approach

Collect Sample

→

Extract DNA

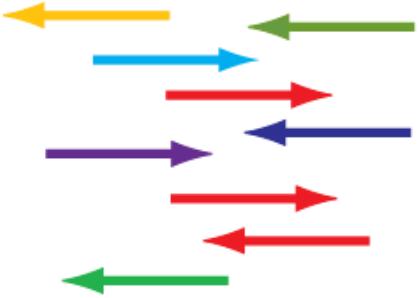
→

Sequence

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

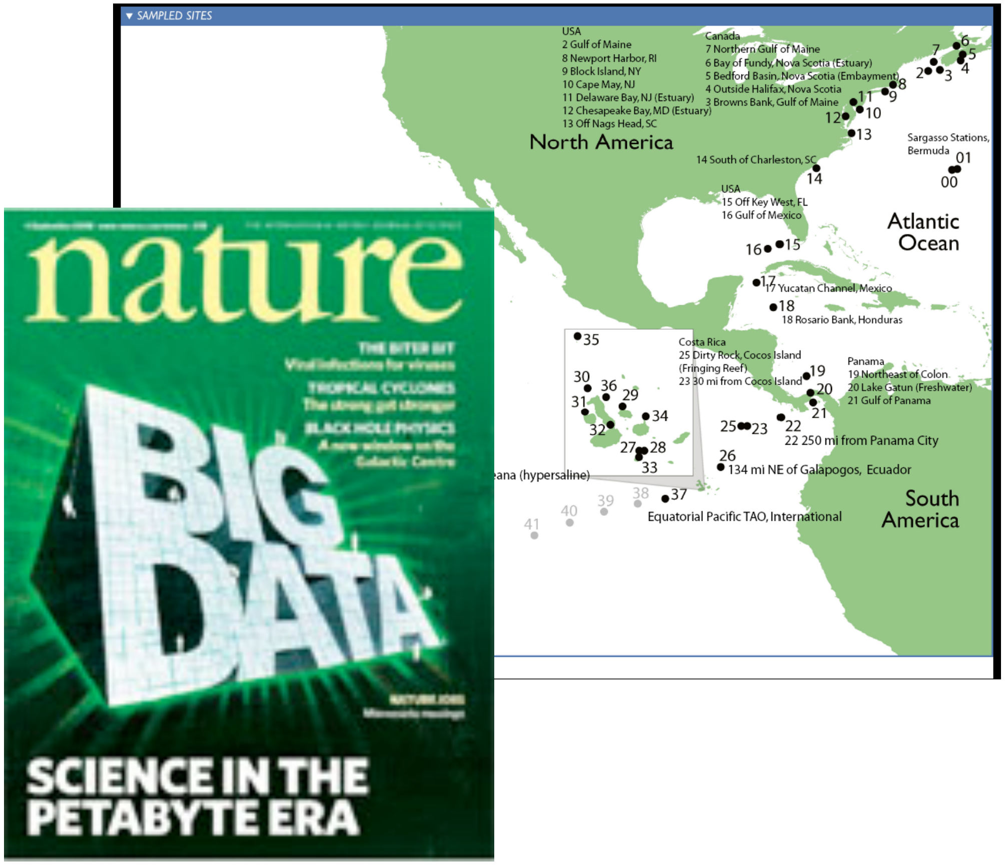
→

Partially Assemble and Annotate



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

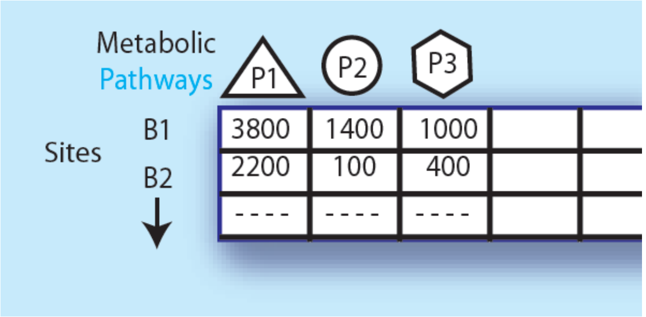
Global Ocean Survey Statistics (GOS)


SAMPLED SITES
USA
2 Gulf of Maine
8 Newport Harbor, RI
9 Block Island, NY
10 Cape May, NJ
11 Delaware Bay, NJ (Estuary)
12 Chesapeake Bay, MD (Estuary)
13 Off Nags Head, SC
Canada
7 Northern Gulf of Maine
6 Bay of Fundy, Nova Scotia (Estuary)
5 Bedford Basin, Nova Scotia (Embayment)
4 Outside Halifax, Nova Scotia
3 Browns Bank, Gulf of Maine
North America
14 South of Charleston, SC
Sargasso Stations, Bermuda
01
00
USA
15 Off Key West, FL
16 Gulf of Mexico
Atlantic Ocean
17 Yucatan Channel, Mexico
18 Rosario Bank, Honduras
Costa Rica
25 Dirty Rock, Cocos Island (Fringing Reef)
23 30 mi from Cocos Island
Panama
19 Northeast of Colon
20 Lake Gatun (Freshwater)
21 Gulf of Panama
22 250 mi from Panama City
26 134 mi NE of Galapogos, Ecuador
ana (hypersaline)
Equatorial Pacific TAO, International
South America
35 30 36 29 31 34 32 27 28 33 37 38 39 40 41
nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
SCIENCE IN THE PETABYTE ERA

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

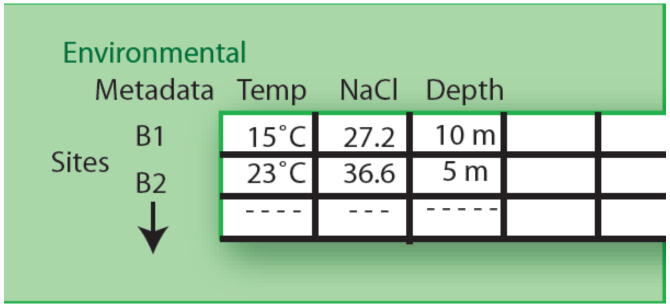
Rusch, et al., PLOS Biology 2007

Pathway Sequences
(Community Function)

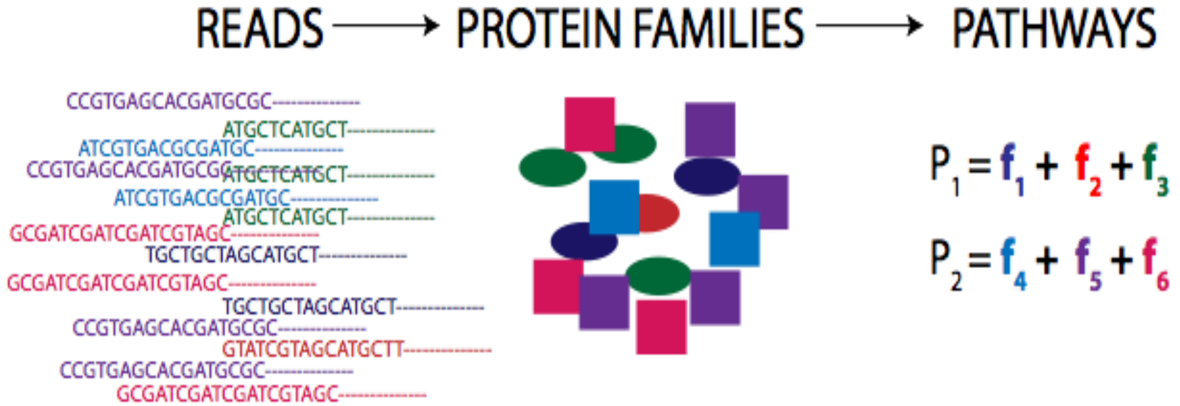


Metabolic Pathways		P1	P2	P3		
Sites	B1	3800	1400	1000		
	B2	2200	100	400		
	↓	----	----	----		

Environmental
Features

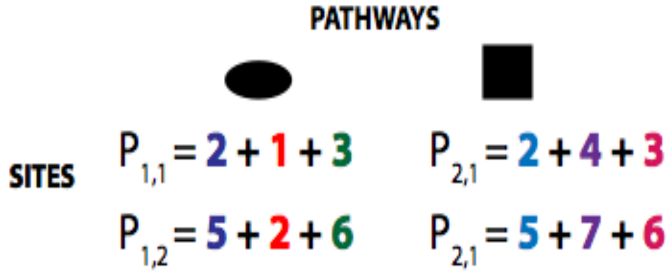


Environmental Metadata		Temp	NaCl	Depth		
Sites	B1	15˚C	27.2	10 m		
	B2	23˚C	36.6	5 m		
	↓	----	---	-----		



$$P_1 = f_1 + f_2 + f_3$$

$$P_2 = f_4 + f_5 + f_6$$



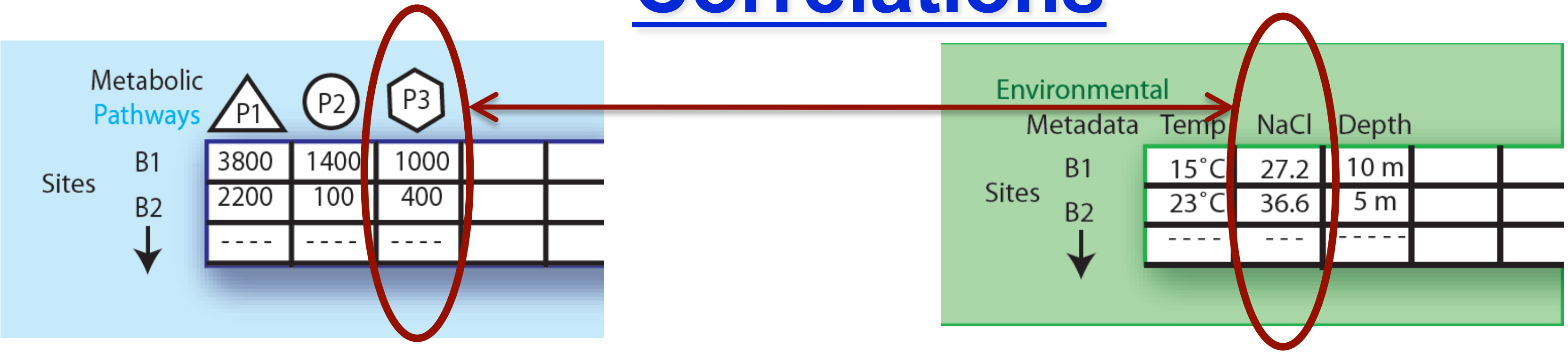
$$P_{1,1} = 2 + 1 + 3 \qquad P_{2,1} = 2 + 4 + 3$$

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

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

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

Simple Relationships: Pairwise Correlations

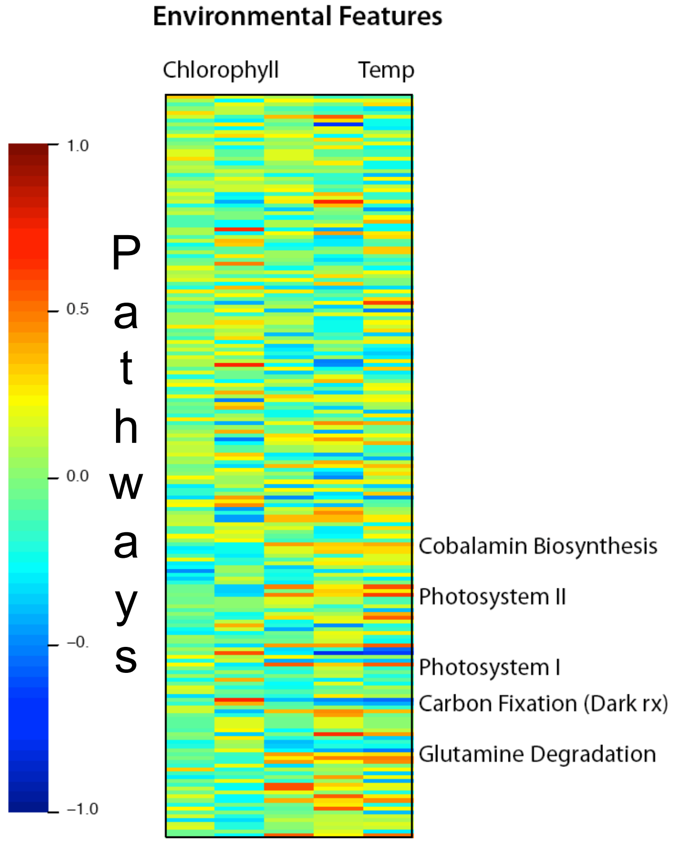


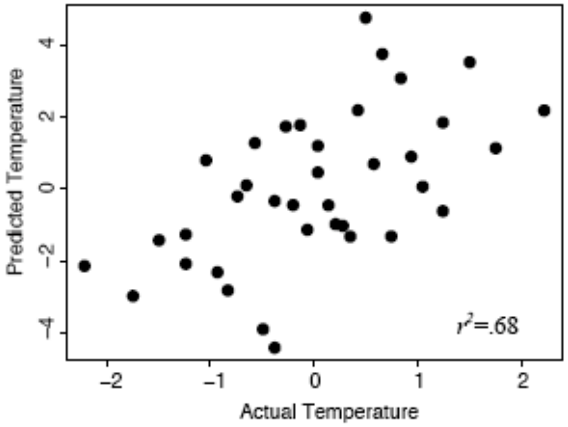
Metabolic Pathways

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

Environmental Metadata

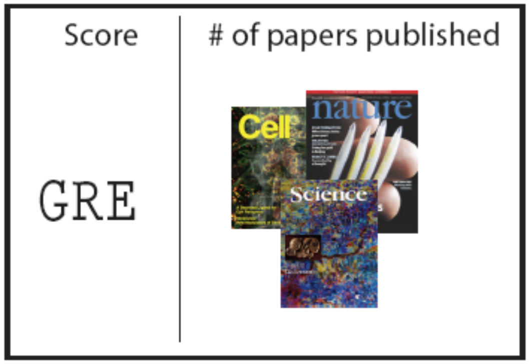
Sites		Temp	NaCl	Depth
	B1	15˚C	27.2	10 m
	B2	23˚C	36.6	5 m
		----	---	-----

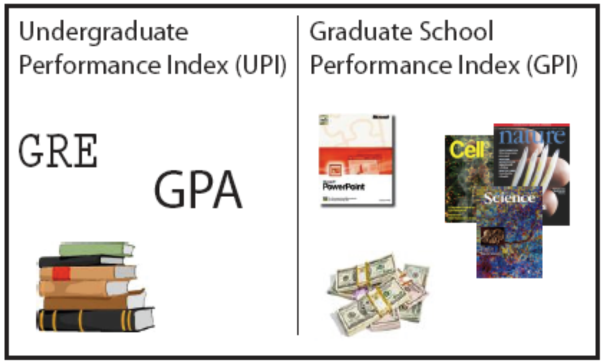




[Gianoulis et al., PNAS (in press, 2009)]

Canonical Correlation Analysis: Simultaneous weighting



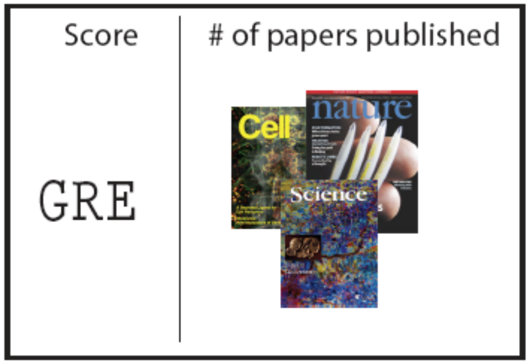

Undergraduate Performance Index (UPI)
Graduate School Performance Index (GPI)
GRE
GPA

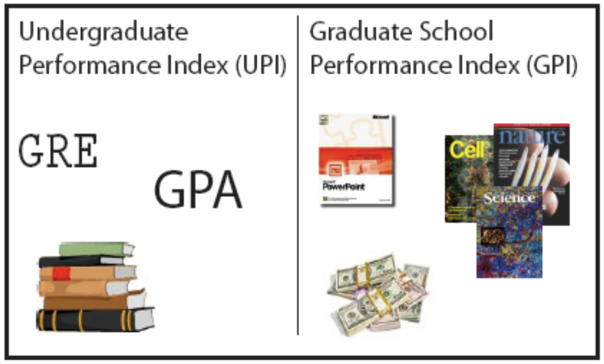
UPI = a GRE + b + c GPA

GPI = a’ + b’ + c’

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

Canonical Correlation Analysis: Simultaneous weighting

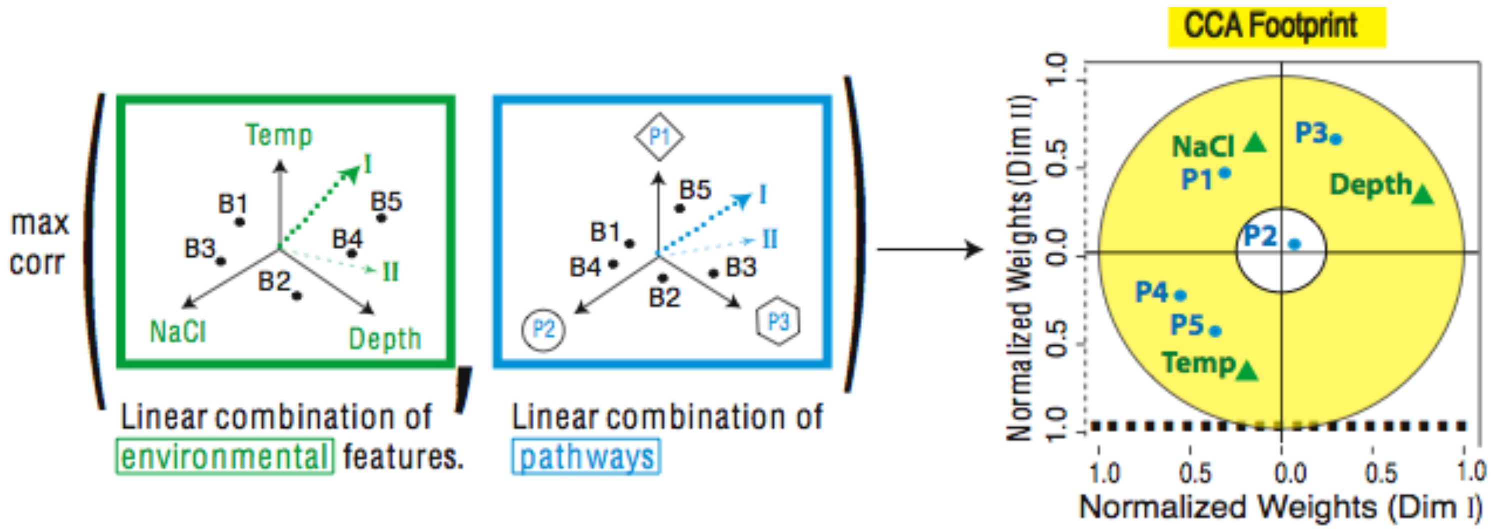



Undergraduate Performance Index (UPI)
Graduate School Performance Index (GPI)
GRE
GPA

Environmental Features	Metabolic Pathways
Temp etc	Photosynthesis etc
Chlorophyll	Lipid Metabolism

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

Environmental-Metabolic Space

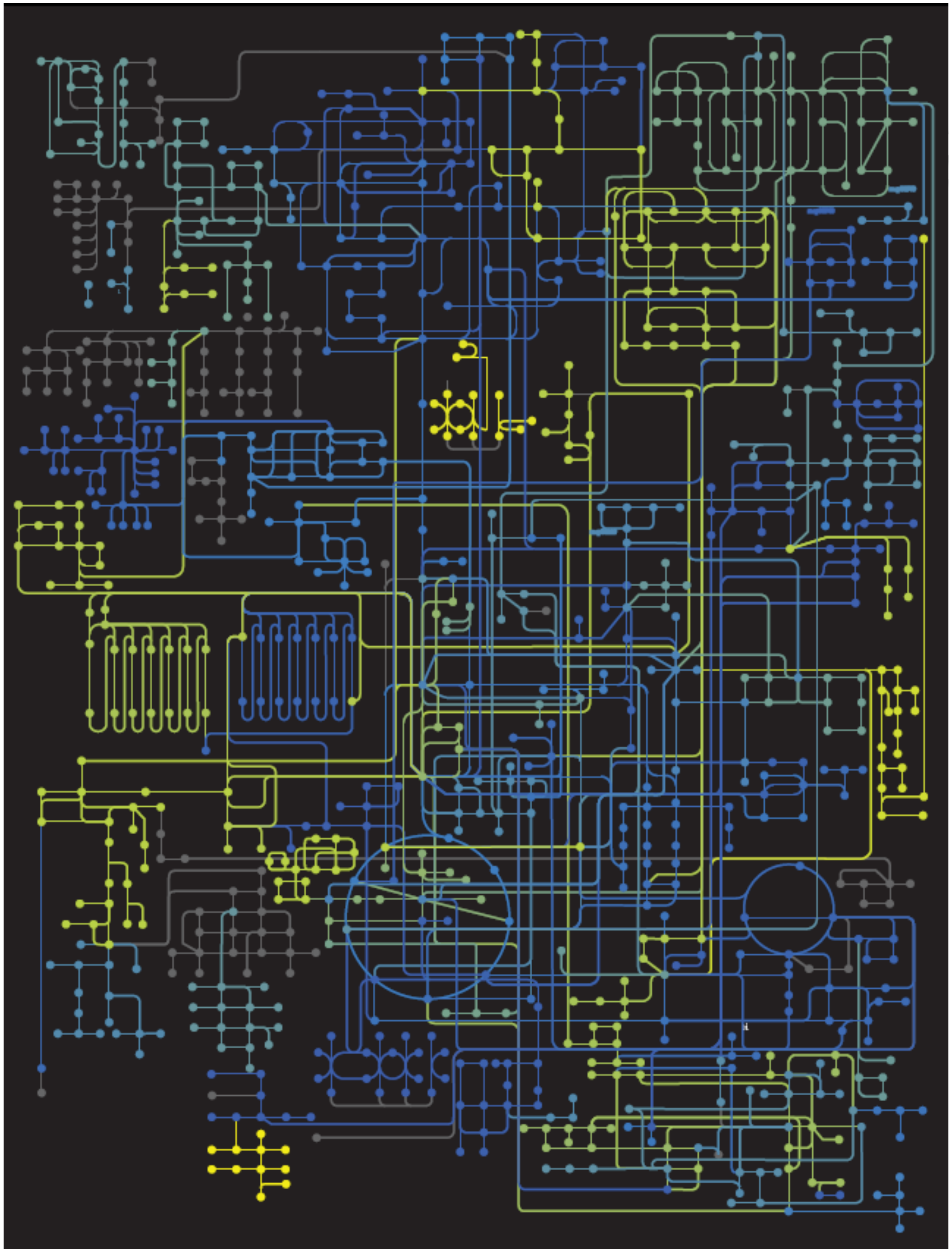

max
corr
Temp
I
II
B1
B5
B3
B4
B2
NaCl
Depth
Linear combination of environmental features.
P1
B5
I
II
B1
B4
B3
B2
P2
P3
Linear combination of pathways
CCA Footprint
NaCl
P1
P3
Depth
P2
P4
P5
Temp
Normalized Weights (Dim II)
1.0
0.5
0.0
0.5
1.0
1.0
0.5
0.0
0.5
1.0
Normalized Weights (Dim I)

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

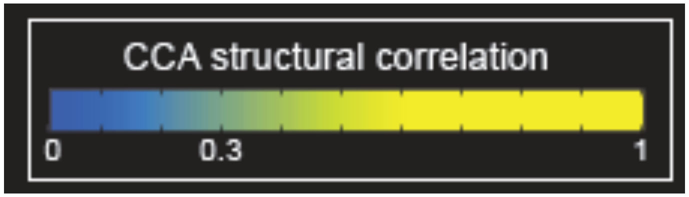
$$\text{Given } X = \{x_1, x_2, \ldots, x_n\} \text{ and } Y = \{y_1, y_2, \ldots, y_m\}$$

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

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

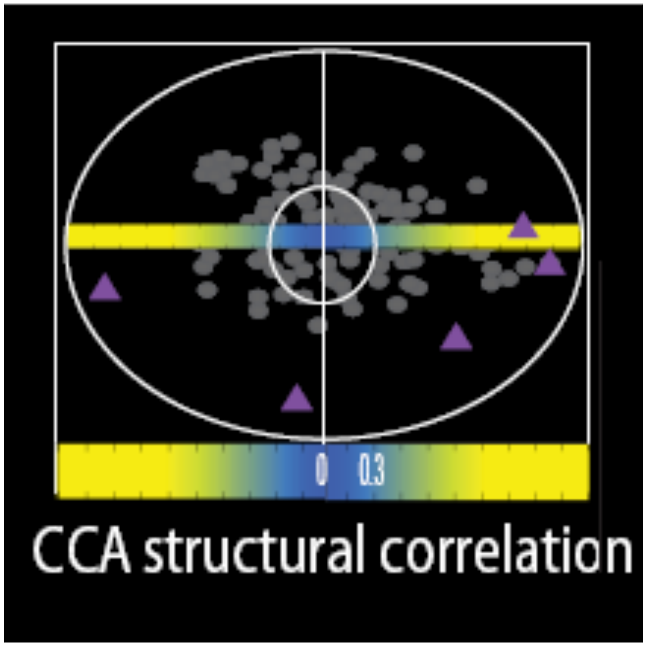


Strength of Pathway co-variation with environment



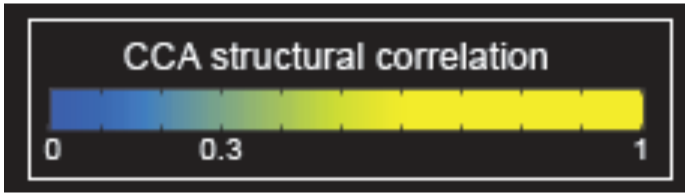
Environmentally invariant

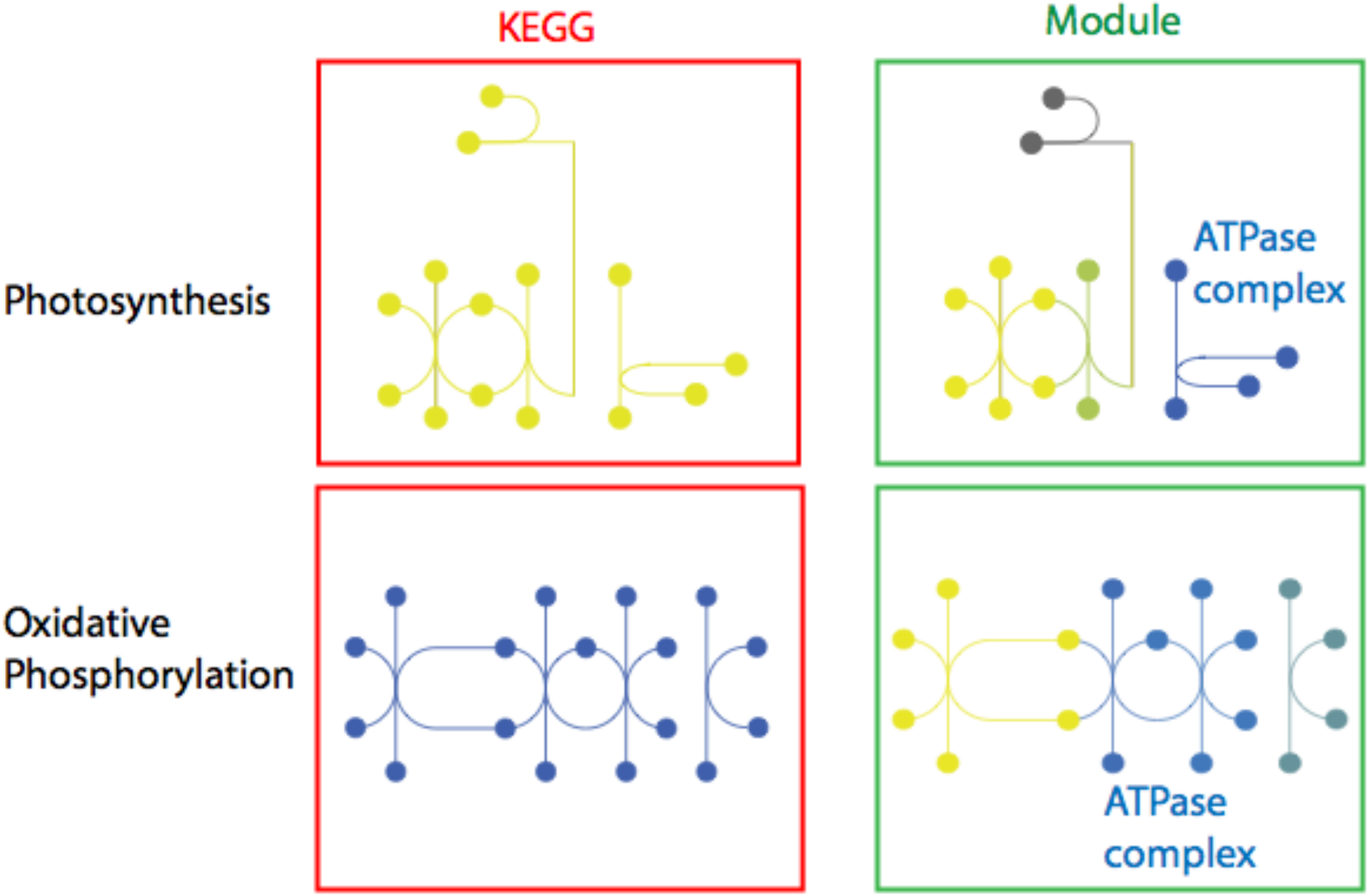
Environmentally variant



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

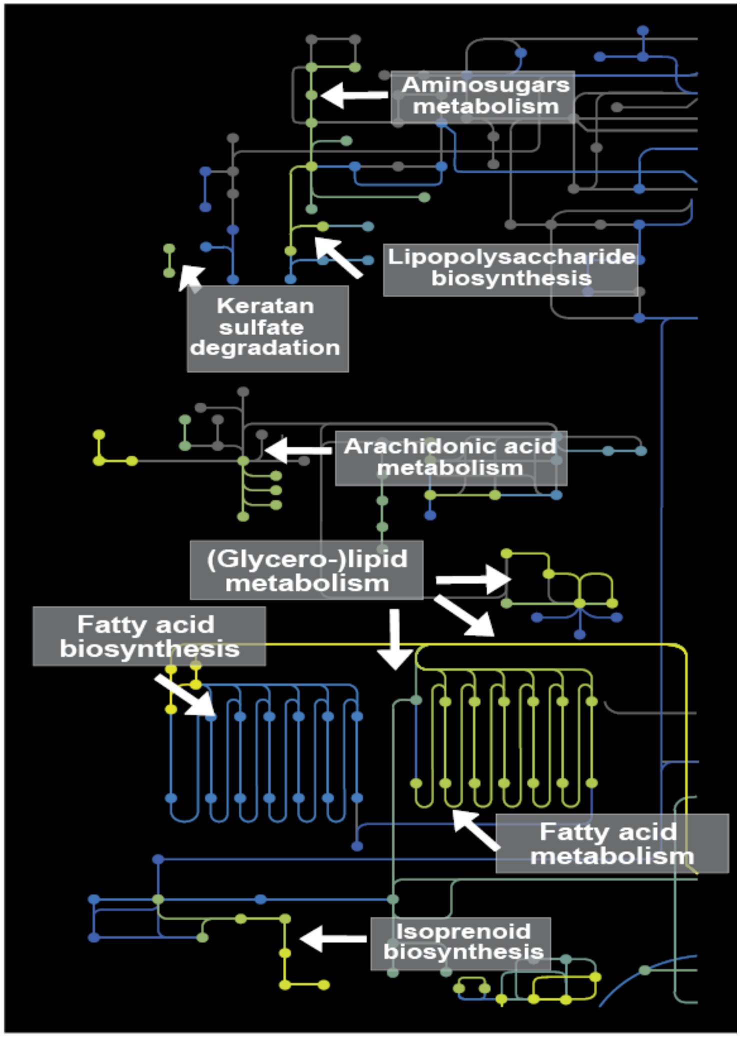
Conclusion #1: energy conversion strategy, temp and depth

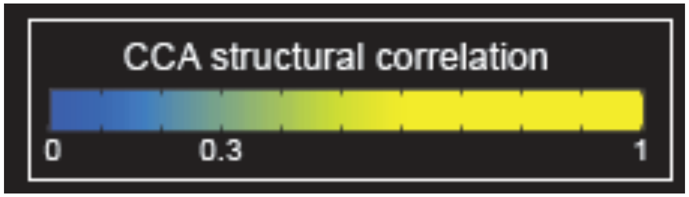




[Gianoulis et al., PNAS (in press, 2009)]

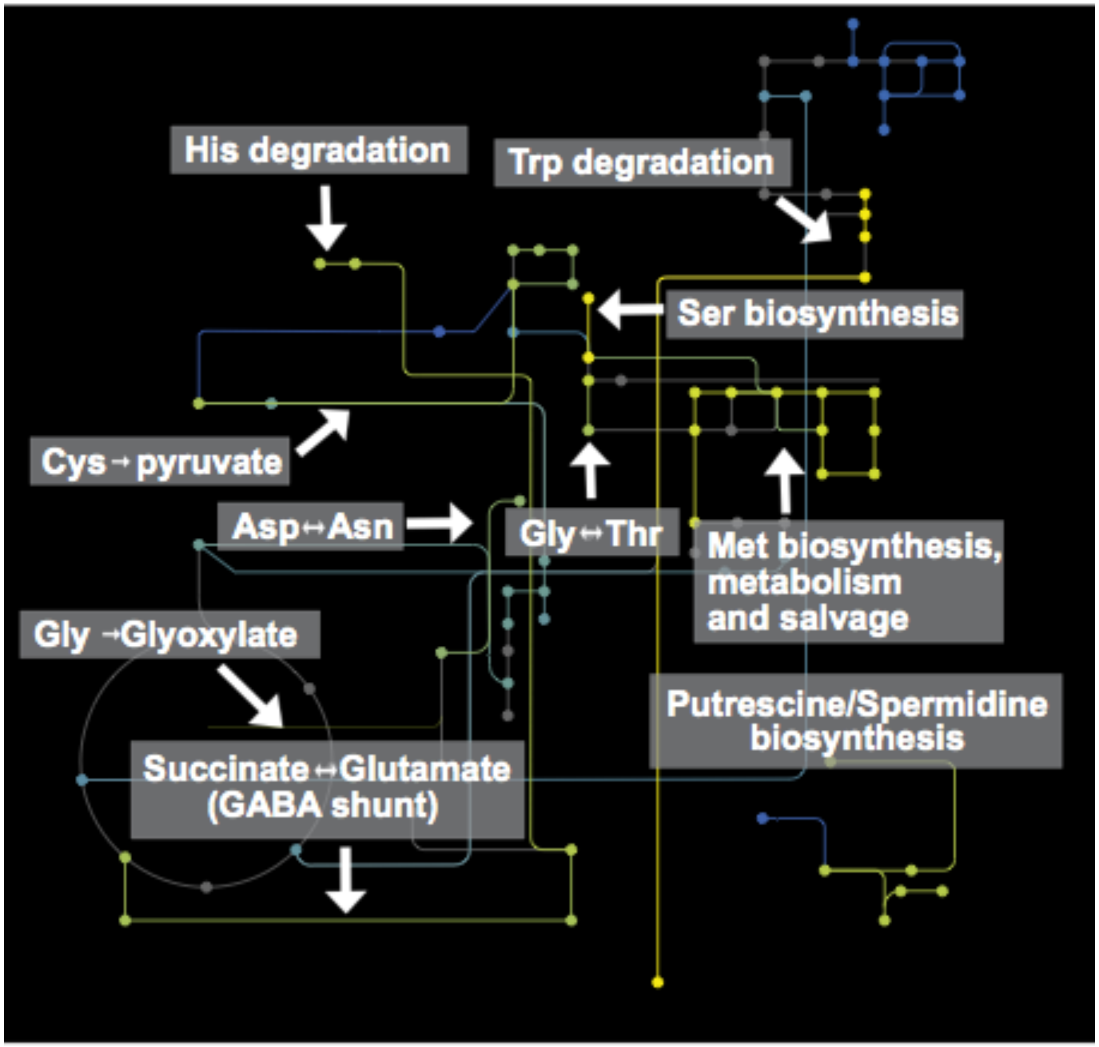
Conclusion #2: Outer Membrane components vary the environment


Aminosugars metabolism
Lipopolysaccharide biosynthesis
Keratan sulfate degradation
Arachidonic acid metabolism
(Glycero-)lipid metabolism
Fatty acid biosynthesis
Fatty acid metabolism
Isoprenoid biosynthesis



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

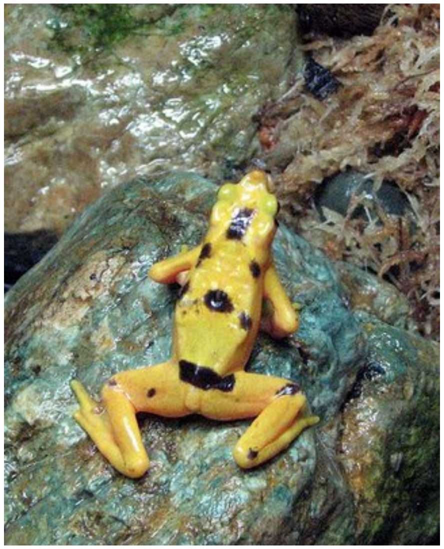
Conclusion #3: Covariation of AA biosynthesis and Import


His degradation
Trp degradation
Ser biosynthesis
Cys→pyruvate
Asp↔Asn
Gly↔Thr
Met biosynthesis, metabolism and salvage
Gly →Glyoxylate
Putrescine/Spermidine biosynthesis
Succinate↔Glutamate (GABA shunt)

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



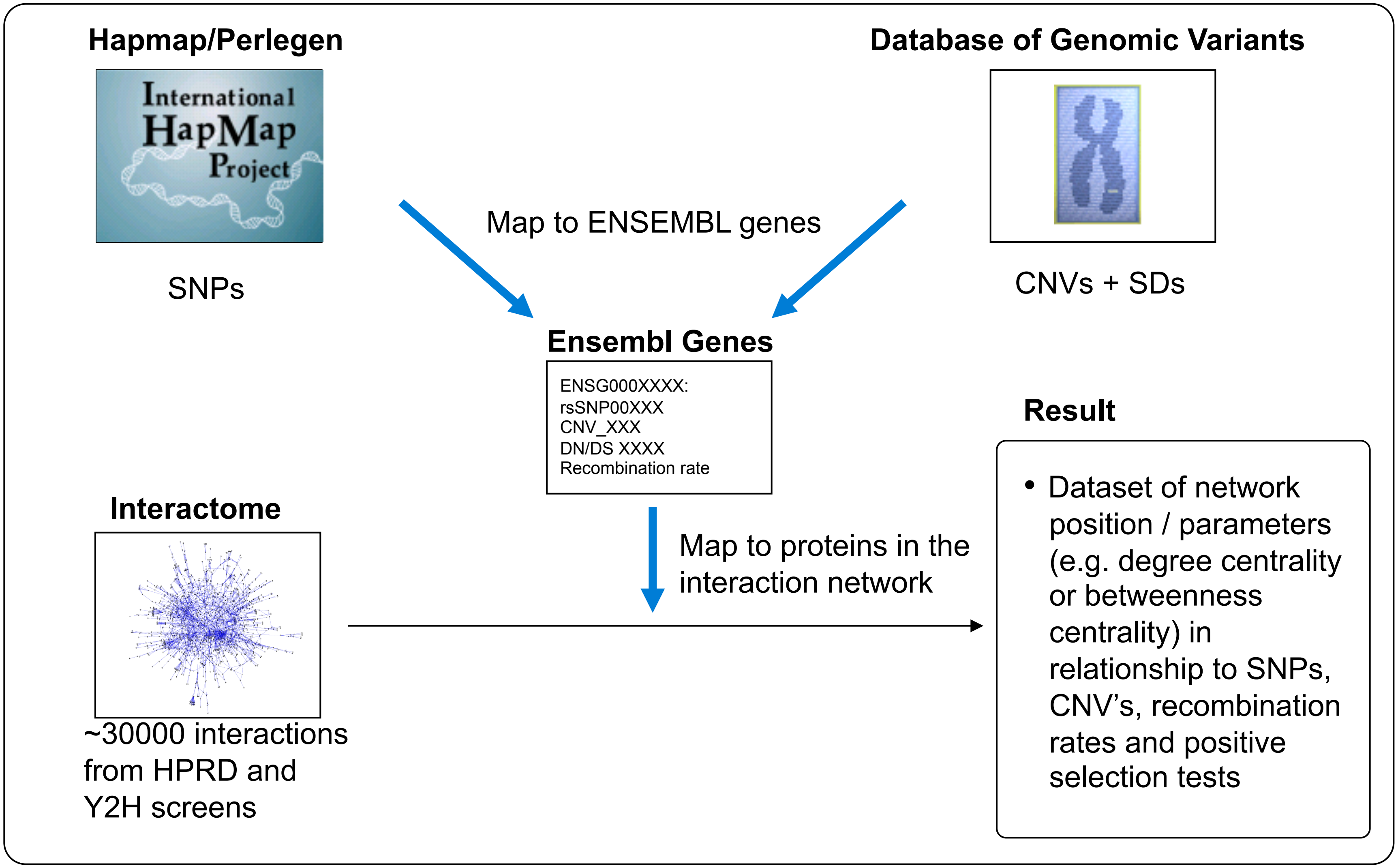
[Gianoulis et al., PNAS (in press, 2009)]

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

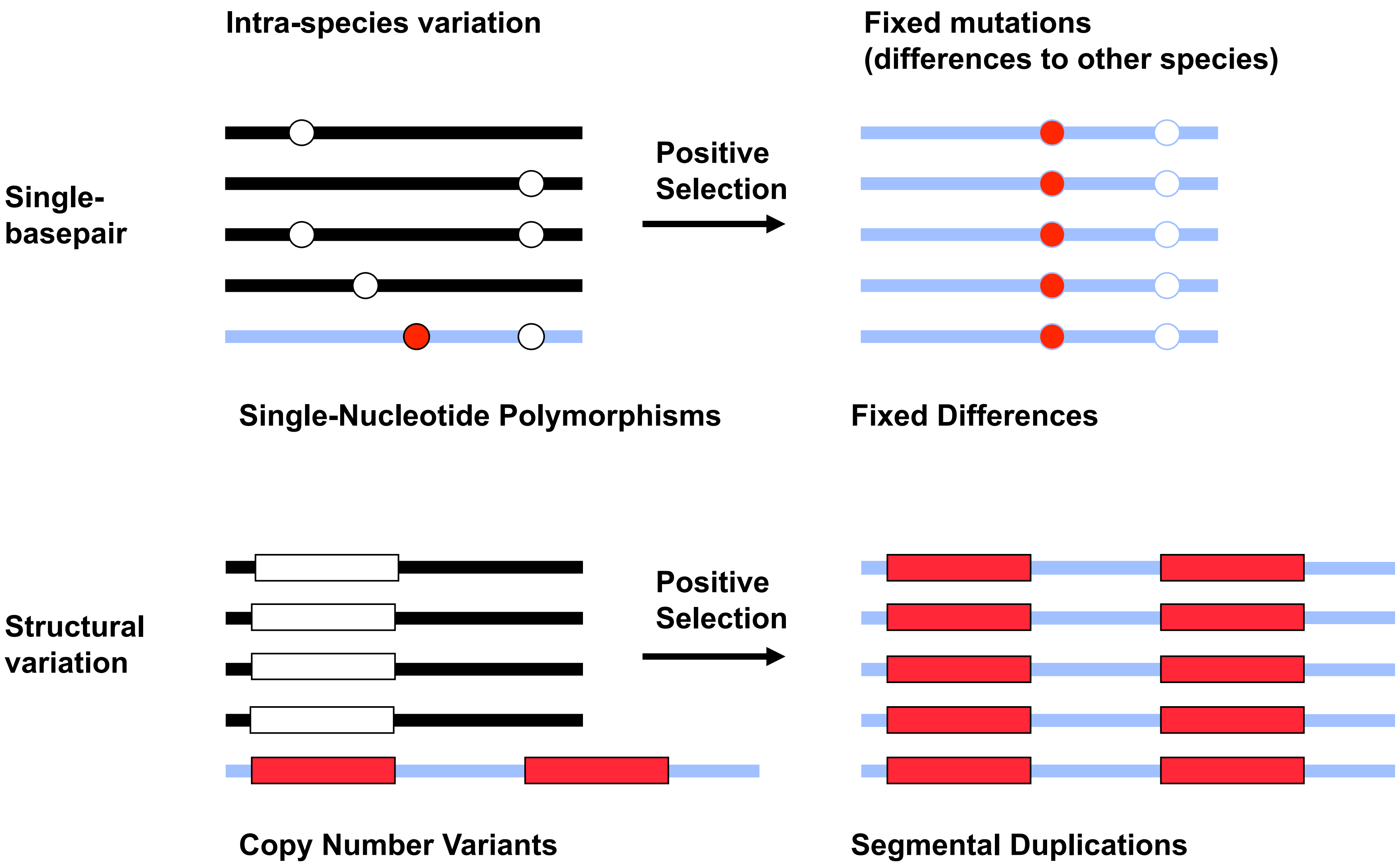
Result

- 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

ADAPTIVE EVOLUTION CAN BE SEEN ON TWO DIFFERENT LEVELS



Intra-species variation

Fixed mutations
(differences to other species)

Single-basepair

Positive Selection

Single-Nucleotide Polymorphisms

Fixed Differences

Structural variation

Positive Selection

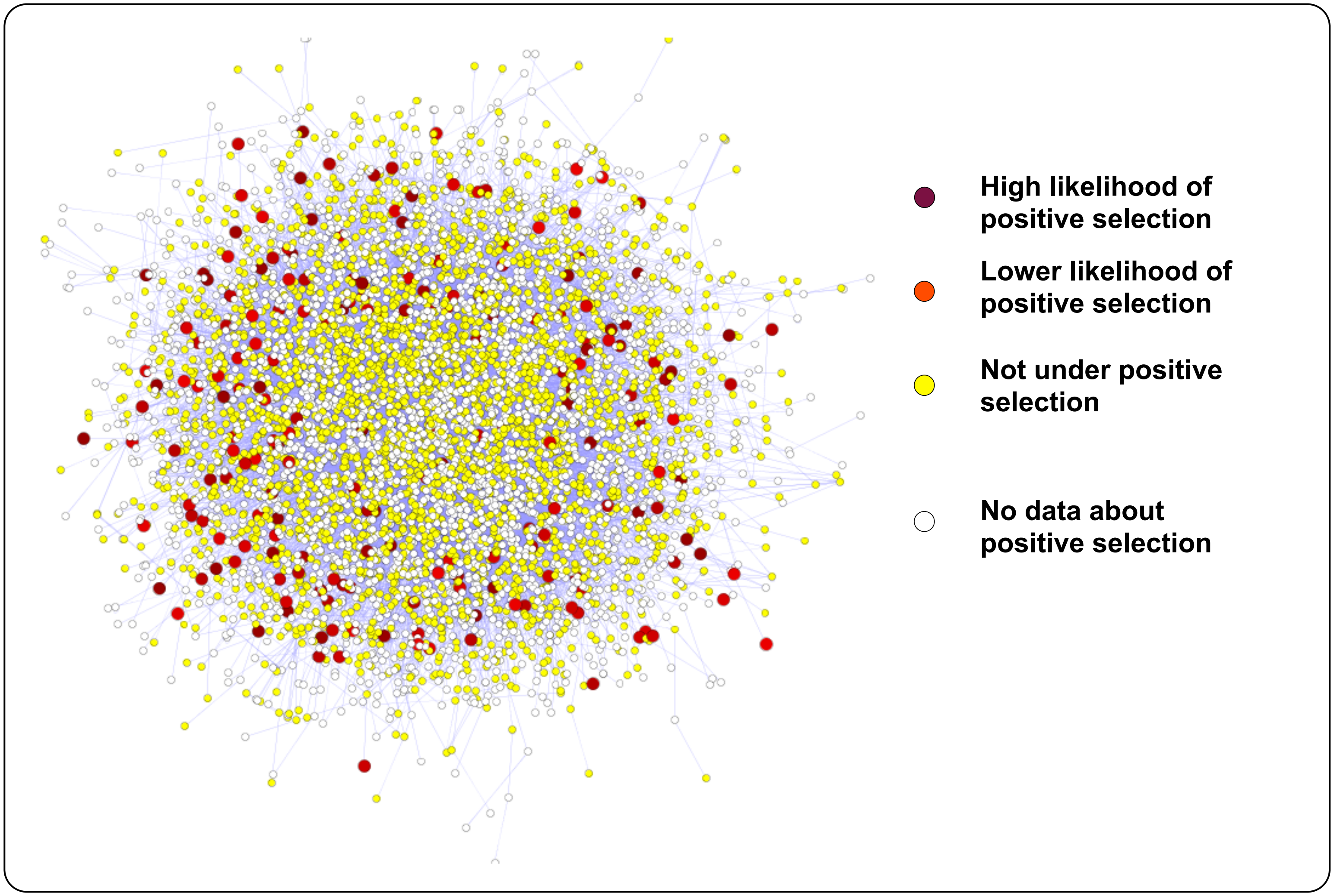
Copy Number Variants

Segmental Duplications

Source: PMK

POSITIVE SELECTION LARGELY TAKES PLACE AT THE NETWORK PERIPHERY

Positive selection in the human interactome

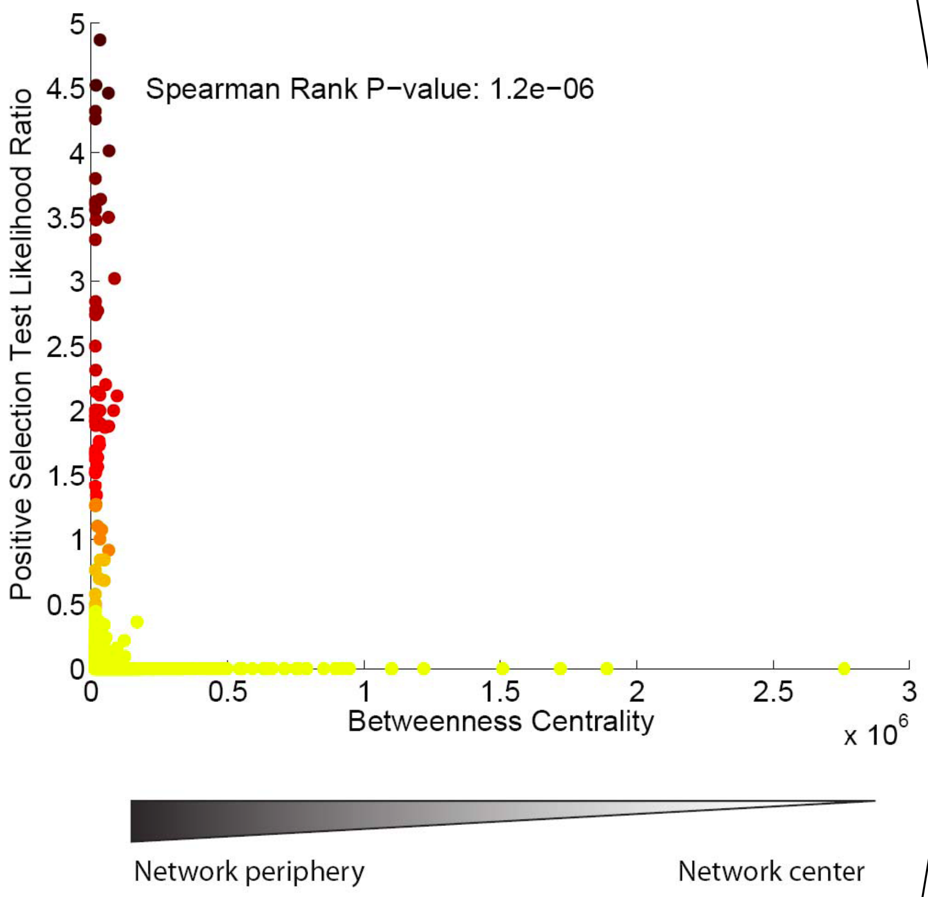

● High likelihood of positive selection
● Lower likelihood of positive selection
● Not under positive selection
○ No data about positive selection

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

CENTRAL PROTEINS ARE LESS LIKELY TO BE UNDER POSITIVE SELECTION

□ Hubs

Degree vs. Positive Selection



Reasoning

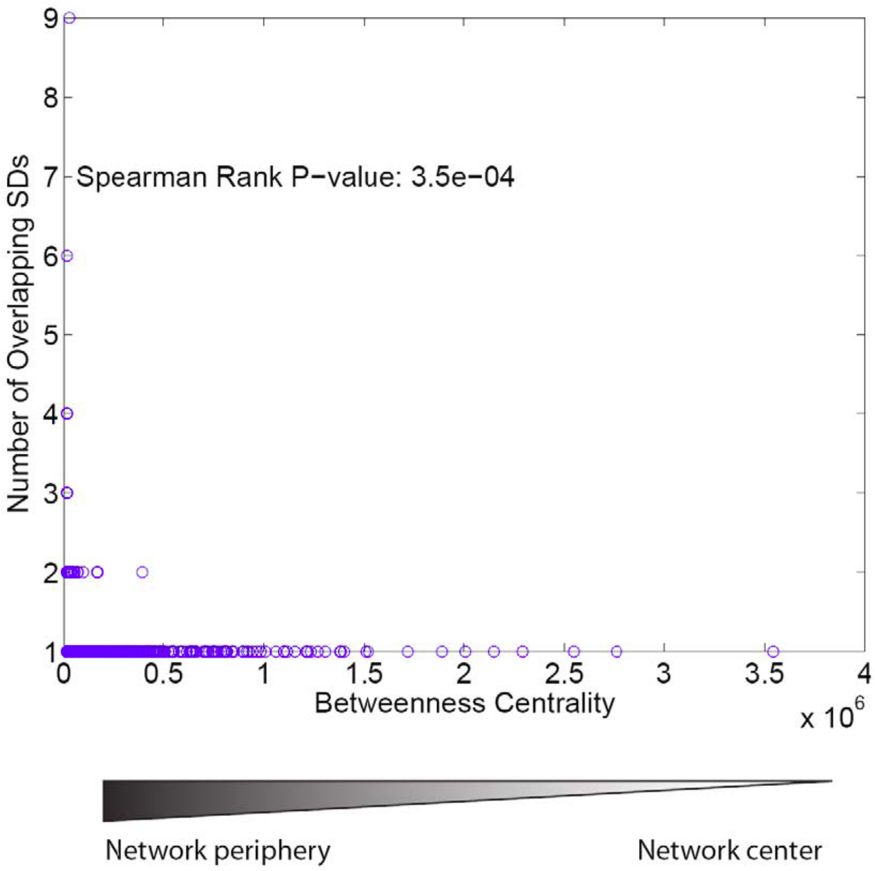
- Peripheral genes are likely to under positive selection, whereas hubs aren’t
- This is likely due to the following reasons:
  - Hubs have stronger structural constraints, the network periphery doesn’t
  - Most recently evolved functions (e.g. “environmental interaction genes” such as sensory perception genes etc.) would probably lie in the network periphery
- Effect is independent of any bias due to gene expression differences

* With a probability of over 80% to be positively selected as determined by Ka/Ks. Other tests of positive selection (McDonald Kreitmann and LDD) corroborate this result.

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

CENTRAL NODES ARE LESS LIKELY TO LIE INSIDE OF SDs

Centrality vs. SD occurrence



Reasoning

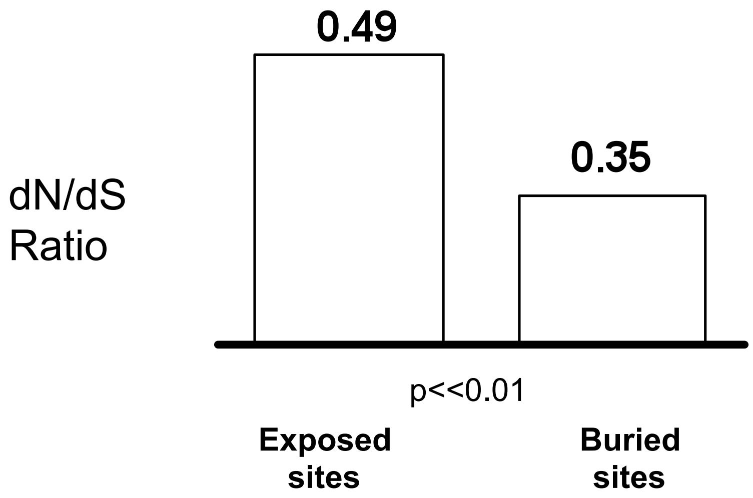
- This result also confirms our initial hypothesis – peripheral nodes tend to lie in regions rich in SDs.
- Since segmental duplications are a different mechanism of ongoing evolution, the less constrained peripheral proteins are enriched in them.
- Note that despite the small size of our dataset for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

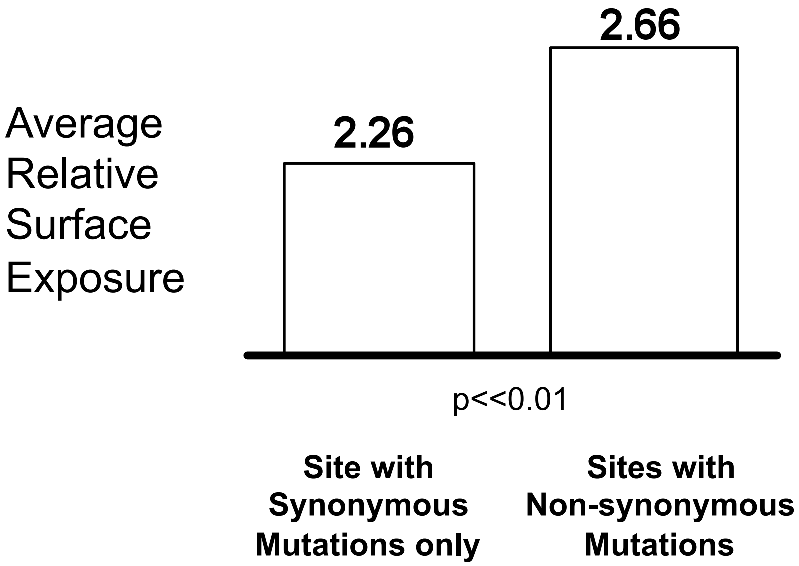
* Specifically, a number of the SDs are likely not fixed, but rather common CNVs in the reference genome

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

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

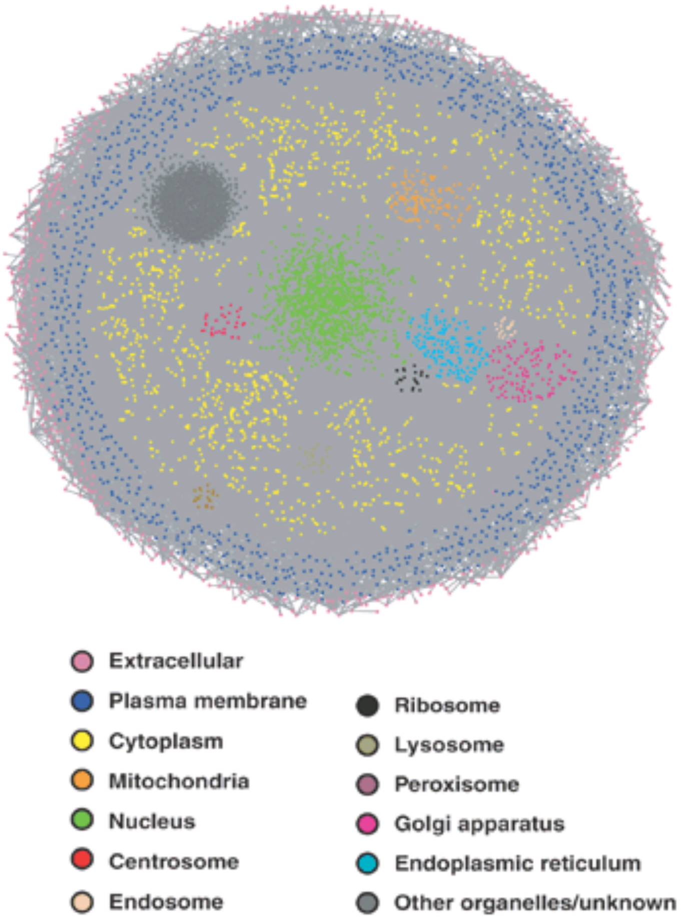
BURIED SITES ARE CONSERVED AND MUCH LESS LIKELY TO HARBOR NON-SYNONYMOUS MUTATIONS

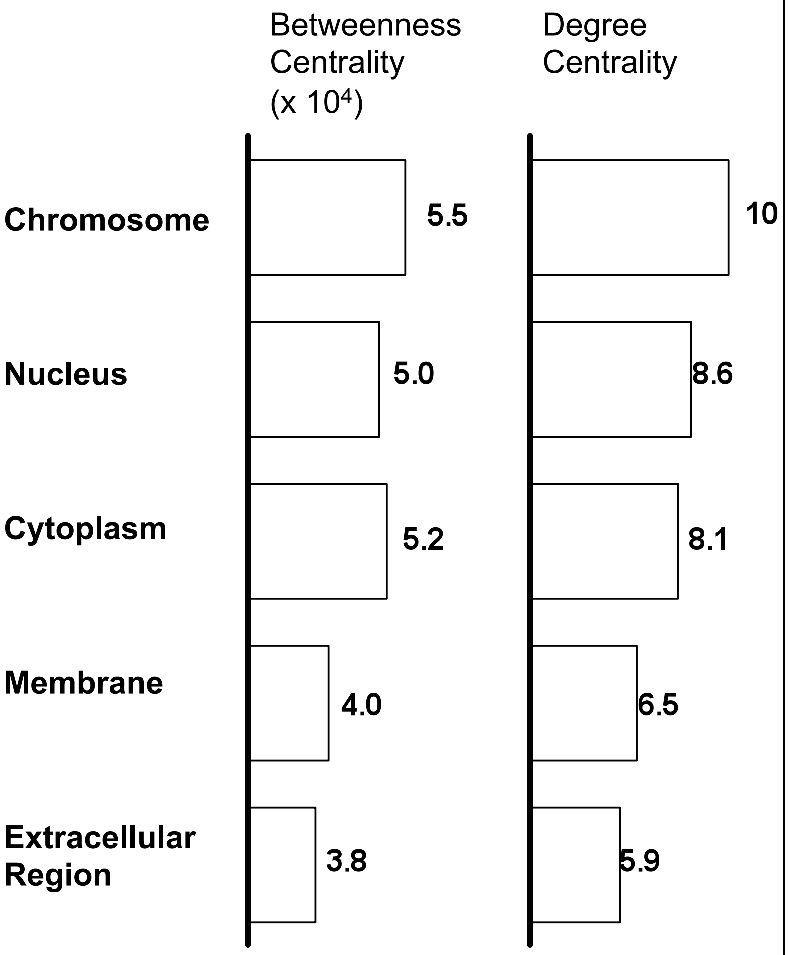




Source: Kim et al. PNAS (2007)

Another explanation: THE NETWORK PERIPHERY CORRESPONDS TO THE CELLULAR PERIPHERY





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

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

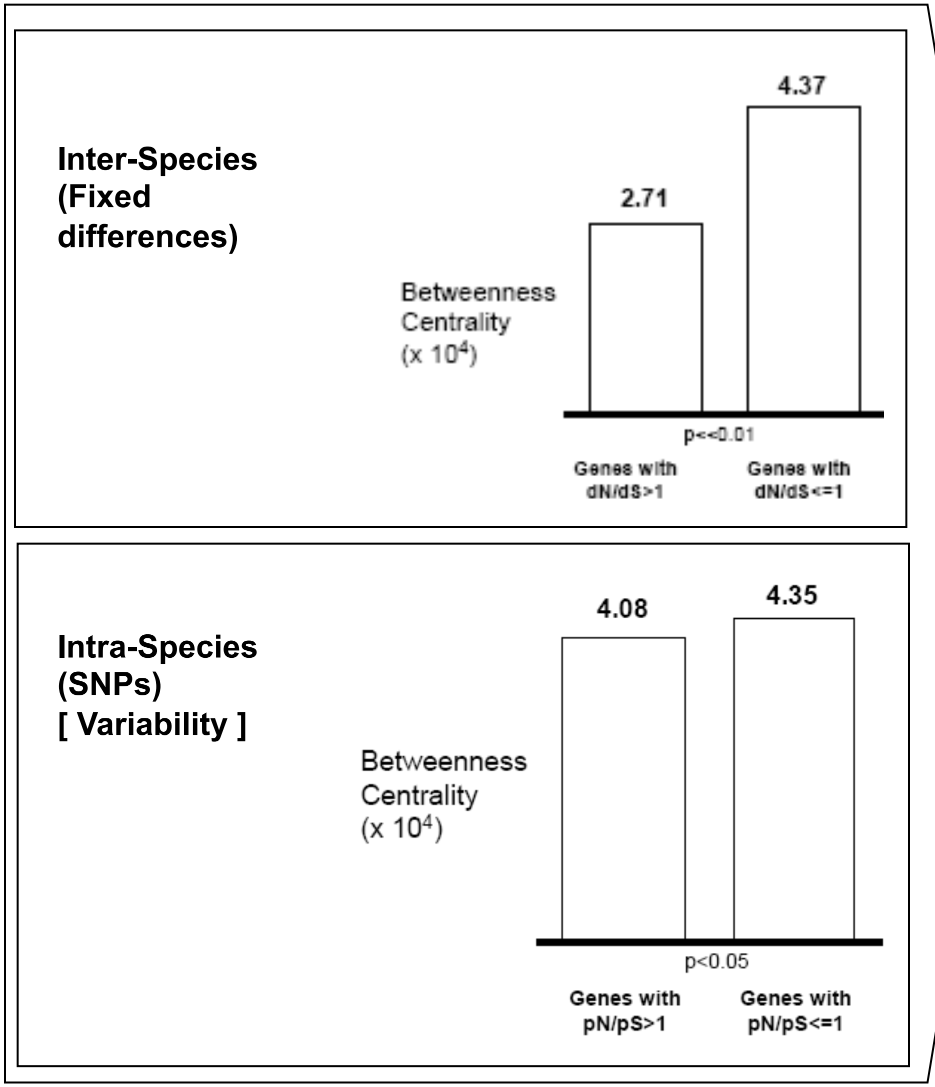
ILLUSTRATIVE

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

Source: Kim et al. PNAS (2007)

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

Inter vs. Intra-Species Variation in Networks



Reasoning

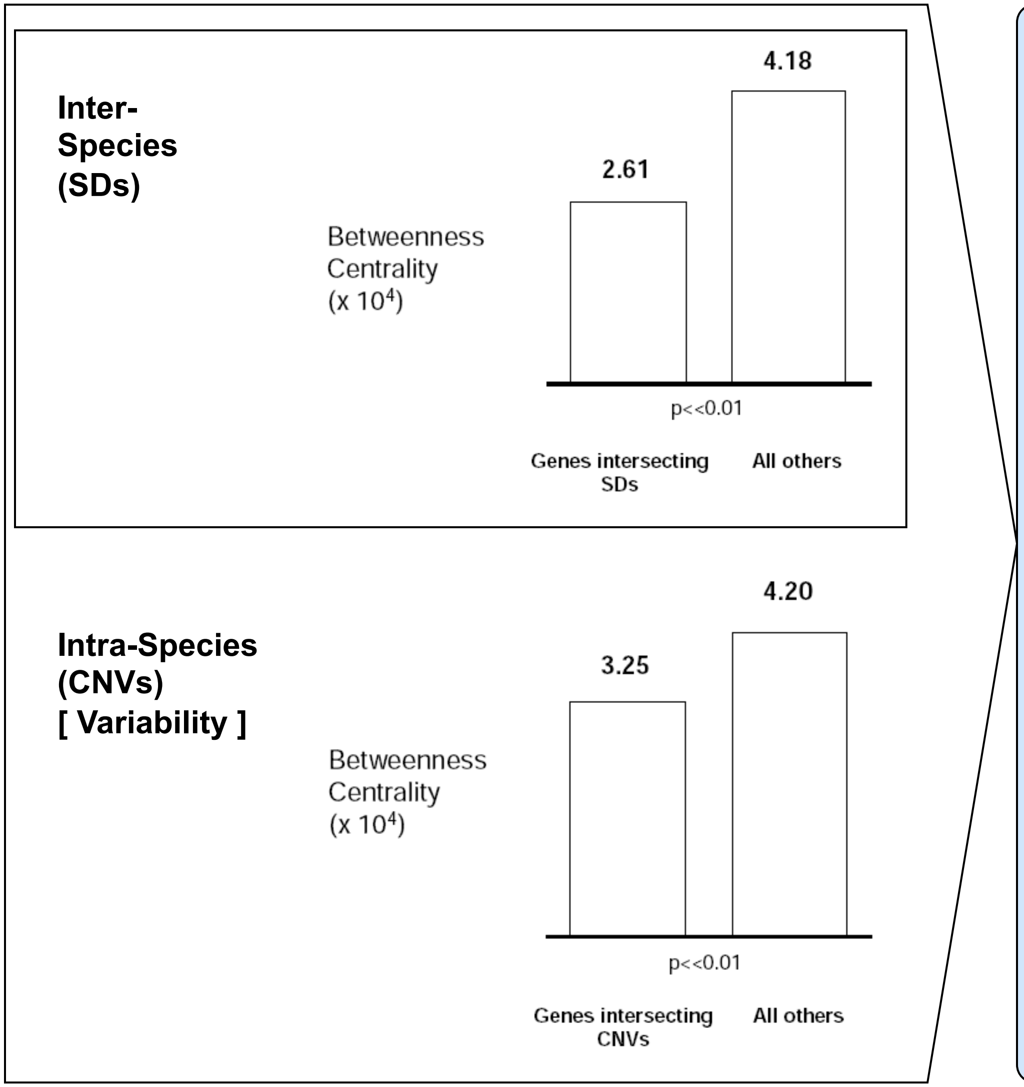
- There is a difference in **variability** (in terms of SNPs) between the network periphery and the center
- However, this difference is much smaller than the difference in **selection**
- This most likely means, that part of the effect we're seeing is due to relaxed constraint (and higher variability)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

* But it's hard to quantify

Source: Kim et al. (2007) PNAS

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

Inter vs. Intra-Species Variation in Networks



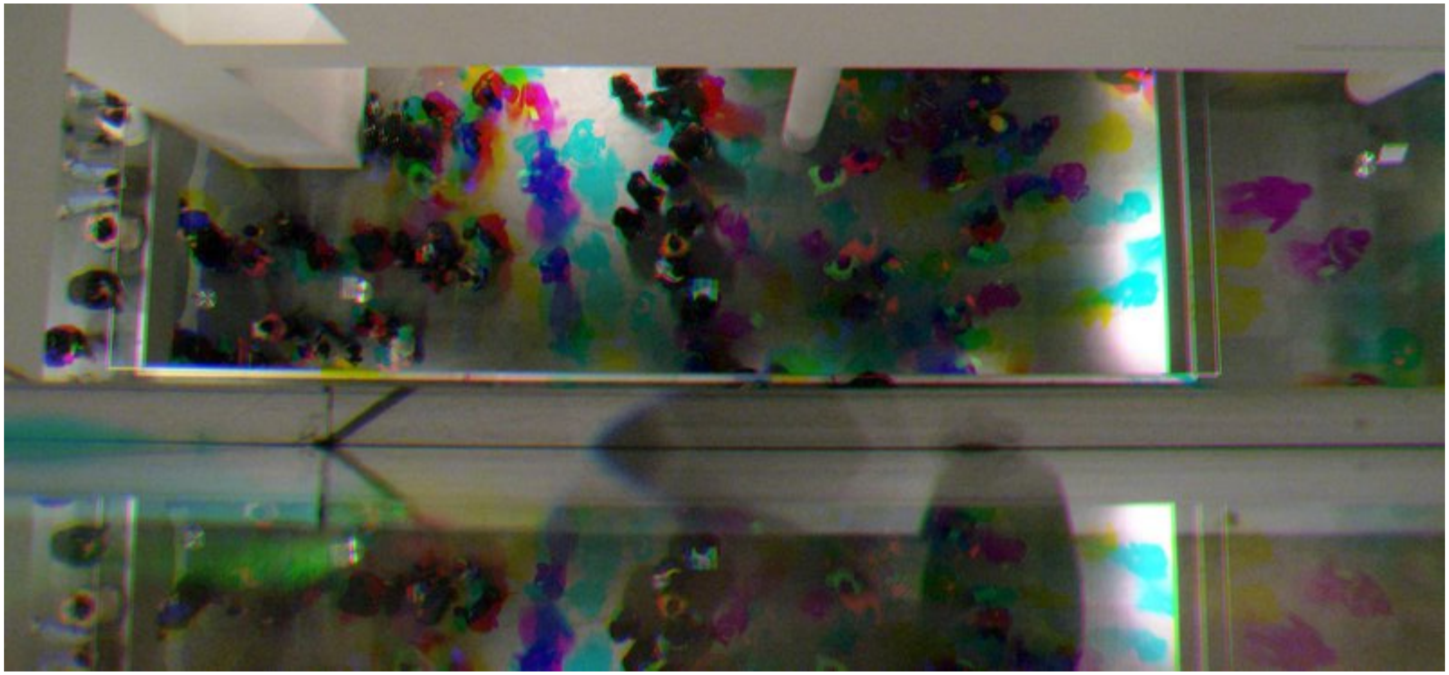
Reasoning

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

Source: Kim et al. (2007) PNAS

Networks & Variation 2

Variation in the miRNA network

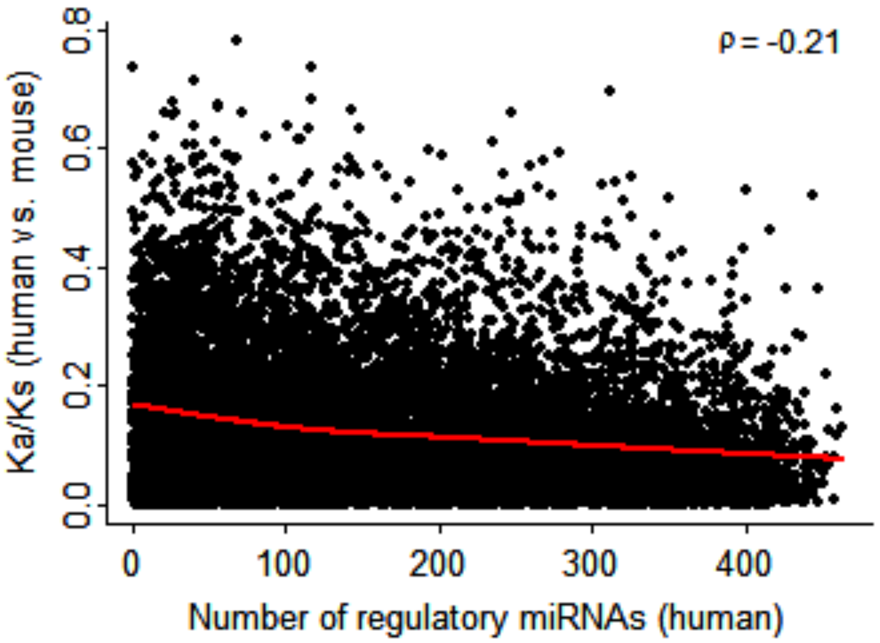


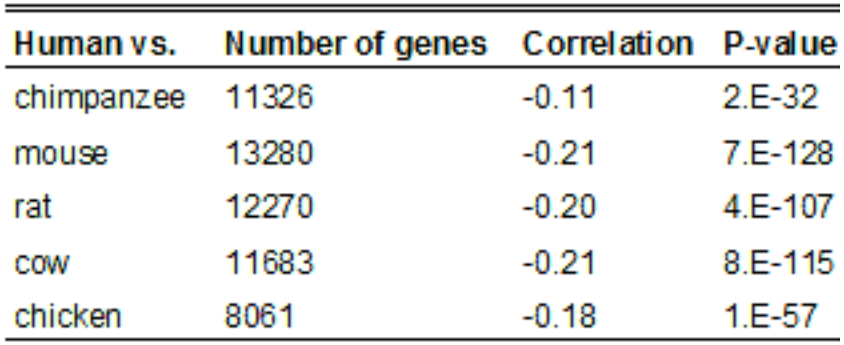
Analyze Regulation in microRNA-target Network

- Relationship between target in degree (number of micro-RNAs that regulate gene) & evolutionary rate of gene?
  - ◊ In deg. related 3' UTR size

- Expectation: more regulation, more constraint

Relationship between microRNA regulation and protein evolution





Human vs.	Number of genes	Correlation	P-value
chimpanzee	11326	-0.11	2.E-32
mouse	13280	-0.21	7.E-128
rat	12270	-0.20	4.E-107
cow	11683	-0.21	8.E-115
chicken	8061	-0.18	1.E-57

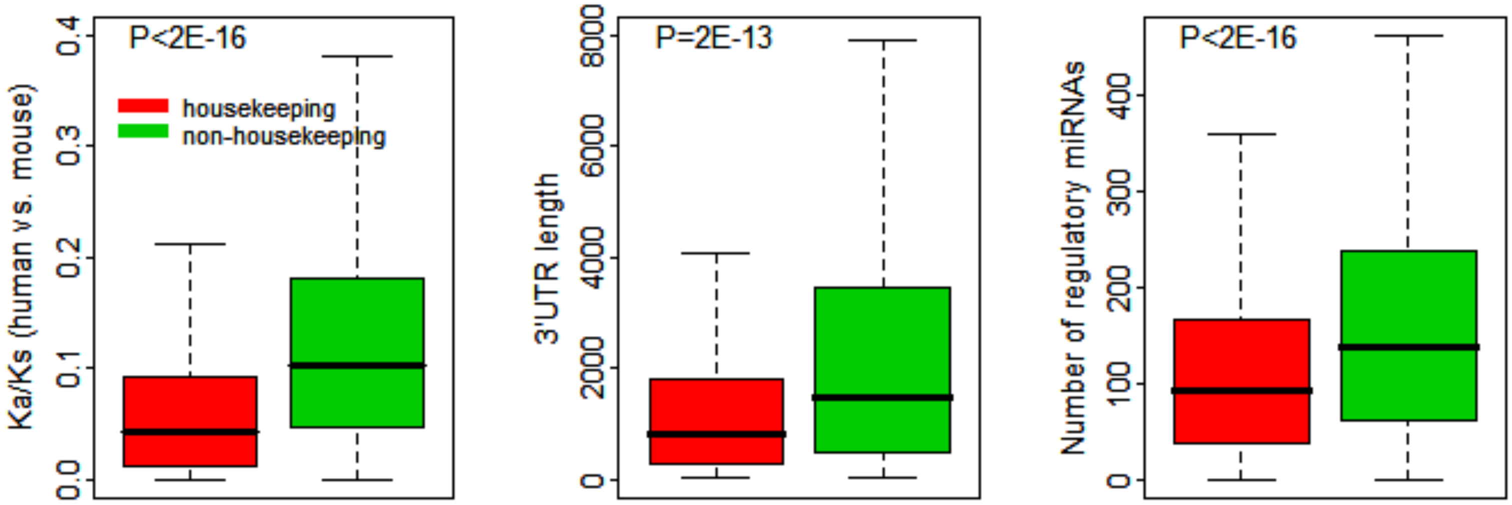
Important genes are regulated more intensively regulated by the microRNAs

[Cheng et al., BMC Genomics, 2009 (in press)]

MicroRNA regulation: a two-way strategy

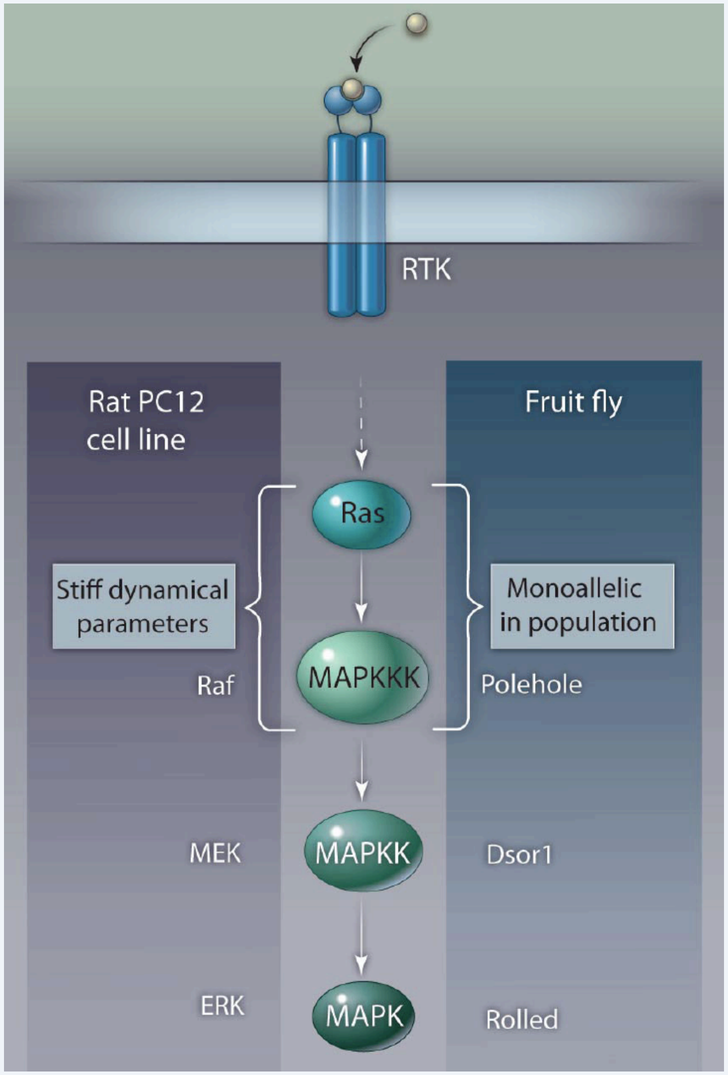
For non-housekeeping genes, functionally critical genes are intensively regulated by miRNAs and prefer long 3'UTR.

housekeeping genes, however conserved, are selected to have shorter 3'UTRs to avoid miRNA regulation.



[Cheng et al., BMC Genomics, 2009 (in press)]

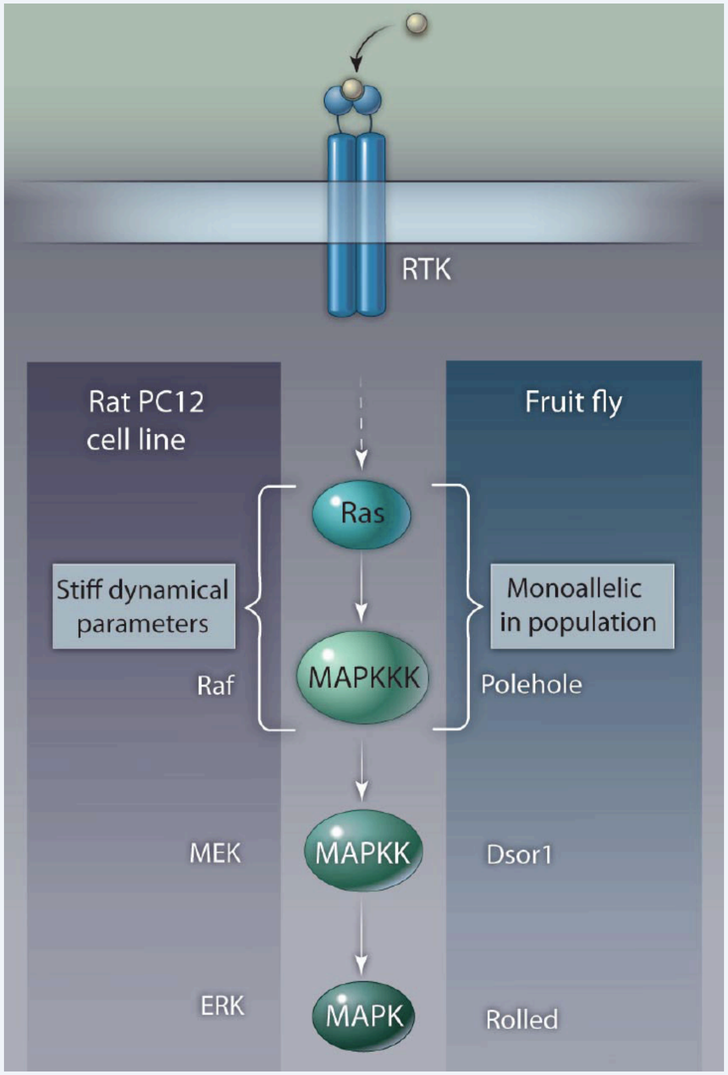
Network dynamics constrain evolution



Hypothesis: Nodes in a molecular network with the strongest impact on dynamic behavior should be under strong purifying selection and thus exhibit the least genetic variation.

Alexander et al. *Sci. Signal.* (2009) 2: pe44

Network dynamics constrain evolution



Hypothesis: Nodes in a molecular network with the strongest impact on dynamic behavior should be under strong purifying selection and thus exhibit the least genetic variation.

Algorithm:

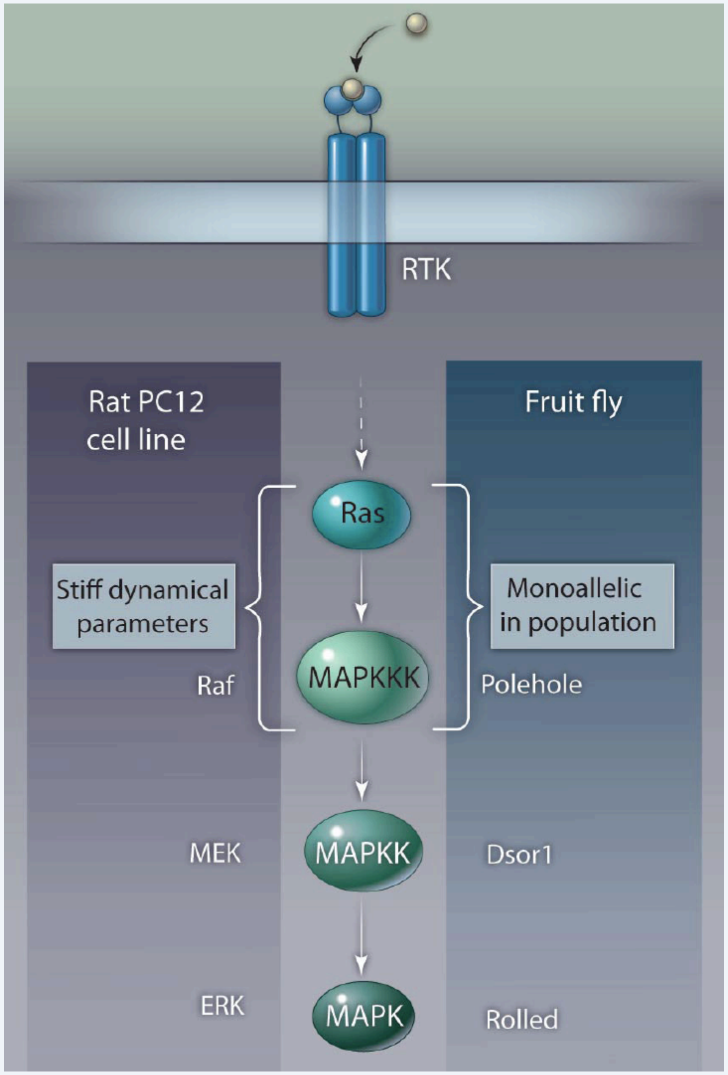
1) Reconstruct families of molecular networks from genomic data.

2) Map some kind of genetic variation onto the networks.

3) Analyze sensitivity of dynamical model of the generic network.

Alexander et al. *Sci. Signal.* (2009) 2: pe44

Speculation: Why more tightly regulated gene might have less variation



Alexander et al. *Sci. Signal.* (2009) 2: pe44

Example: MAP Kinase singaling pathway

Dynamic model:

- **ODE model with Michaelis-Menten kinetics**
- **parameters fit**
  to time series data of protein activities
  in response to EGF and NGF
  from rat PC12 cell line

In sensitivity analysis,
stiff parameters cluster around Ras and Raf.

Population study in fruit flies:

- **allele variation based on**
  PCR of pathway genes

Ras and Raf have less allele variation
than other proteins in the network.

Brown et al. *Phys. Biol.* (2004) 1: 184
Riley et al. *Molec. Ecol.* (2003) 12: 1315

Analogies show it reasonable for more variable part of network to be periphery

- Computer Networks
  - Servers in center have much depending on them; thus, can't be frequently updated & patched
  - Servers on periphery often attacked and so need frequent patches
- Social Networks
  - Individuals at center under more constraint (to conform), whereas those at periphery have more freedom to experiment

Outline: Molecular Networks

- Why Networks?
- Predicting Networks (yeast ppi)
  - ◊ Propagating known information
- Central Points in Networks
  - ◊ Hubs & Bottlenecks (yeast ppi & reg. net)
  - ◊ Tops of Heirarchies (yeast reg. net)
  - ◊ Identified by score (human miRNA-targ. net)
- Dynamics of Networks
  - ◊ Across environments (in prokaryote metab. pathways)
- Protein Networks & Variation (human ppi & miRNA-targ. net)

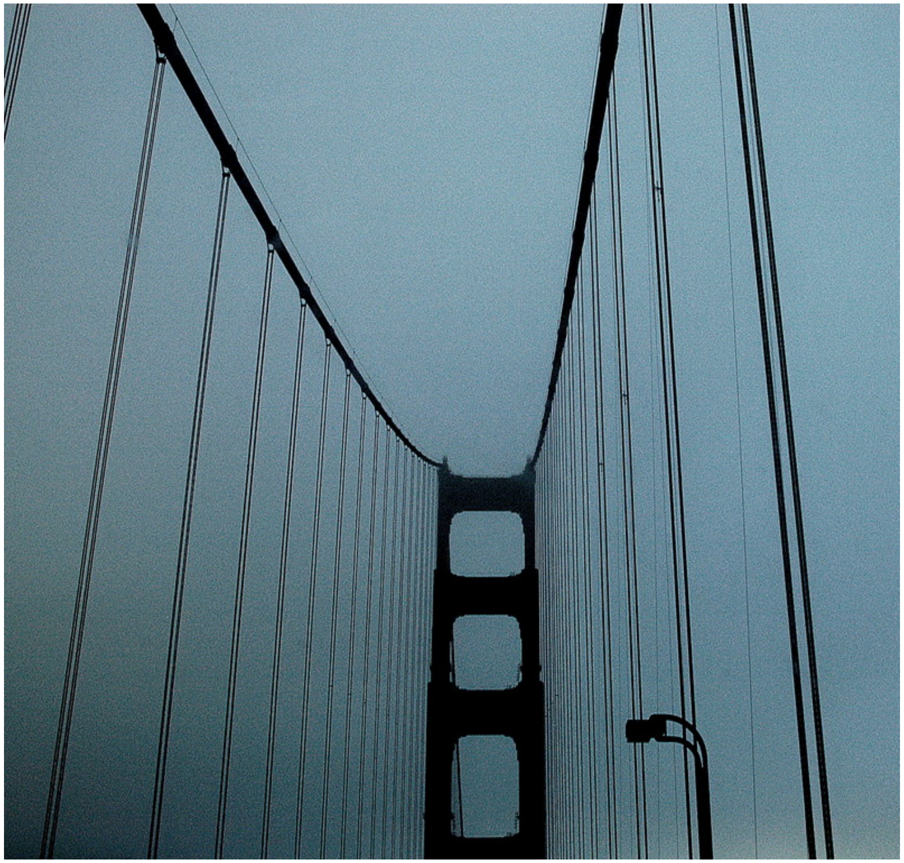
Conclusions on Networks: Predictions

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

Conclusions: Analysis of Network Structure

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

Conclusions: 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.

Conclusions: Connecting Networks & Variation

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

Conclusions: Connecting Networks & Variation 2

- More highly regulated genes are under more constraint in miRNA-target networks
- Exception for housekeeping genes

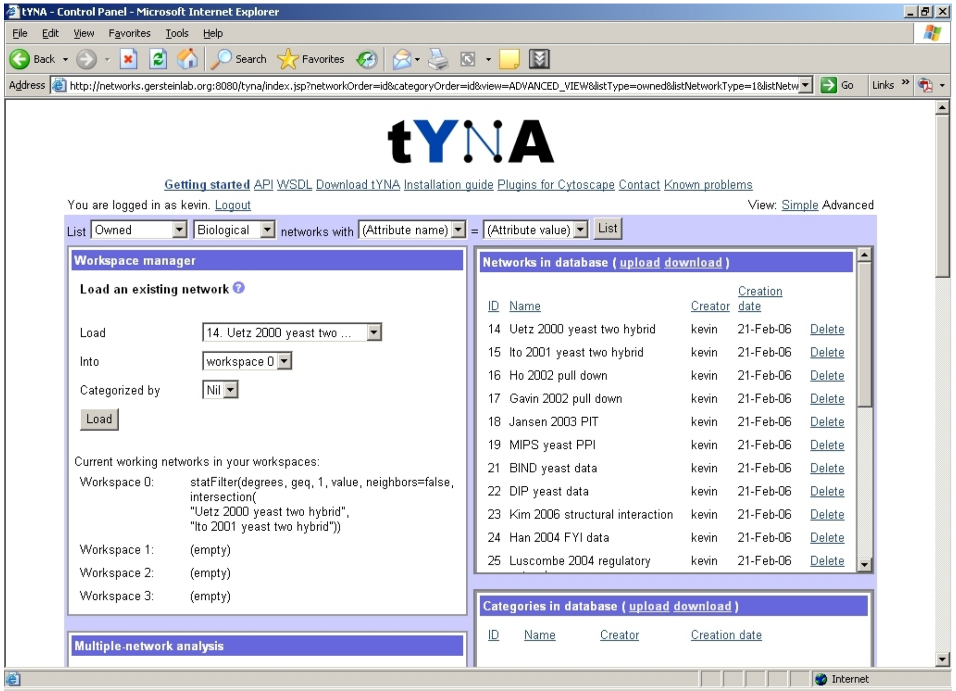
- Speculation as to why variation at periphery is quite reasonable

TopNet
Topology of Networks

- an automated web tool

tYNA

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



tYNA - Control Panel - Microsoft Internet Explorer

Address http://networks.gersteinlab.org:8080/tyna/index.jsp?networkOrder=id&categoryOrder=id&view=ADVANCED_VIEW&listType=owned&listNetworkType=1&listNetw

tYNA

Getting started API WSDL Download tYNA Installation guide Plugins for Cytoscape Contact Known problems

You are logged in as kevin. Logout

View: Simple Advanced

List Owned Biological networks with (Attribute name) = (Attribute value) List

Workspace manager

Load an existing network

Load: 14. Uetz 2000 yeast two ...

Into: workspace 0

Categorized by: Nil

Load

Current working networks in your workspaces:

Workspace 0:	statFilter(degrees, geq, 1, value, neighbors=false, intersection("Uetz 2000 yeast two hybrid", "Ito 2001 yeast two hybrid"))
Workspace 1:	(empty)
Workspace 2:	(empty)
Workspace 3:	(empty)

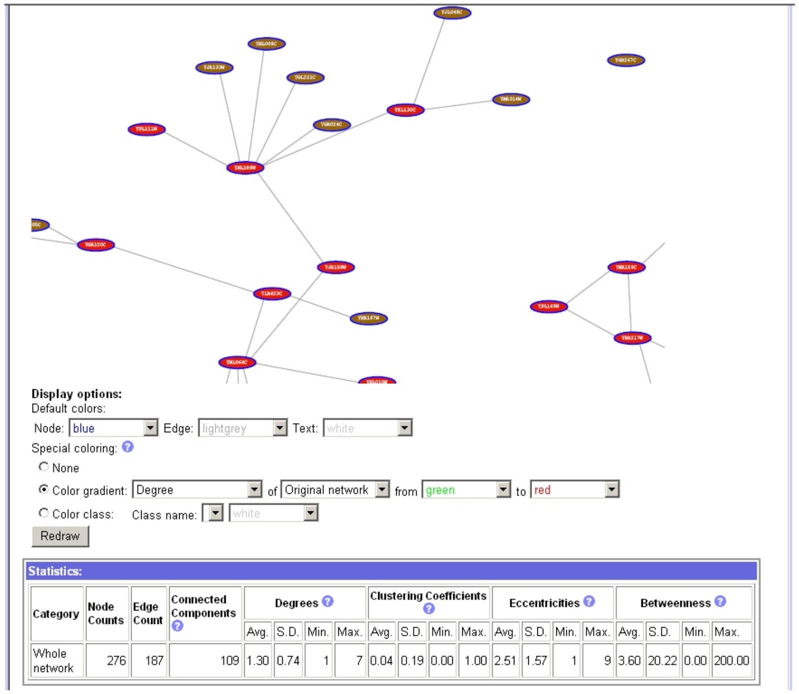
Multiple-network analysis

Networks in database (upload download)

ID	Name	Creator	Creation date	
14	Uetz 2000 yeast two hybrid	kevin	21-Feb-06	Delete
15	Ito 2001 yeast two hybrid	kevin	21-Feb-06	Delete
16	Ho 2002 pull down	kevin	21-Feb-06	Delete
17	Gavin 2002 pull down	kevin	21-Feb-06	Delete
18	Jansen 2003 PIT	kevin	21-Feb-06	Delete
19	MIPS yeast PPI	kevin	21-Feb-06	Delete
21	BIND yeast data	kevin	21-Feb-06	Delete
22	DIP yeast data	kevin	21-Feb-06	Delete
23	Kim 2006 structural interaction	kevin	21-Feb-06	Delete
24	Han 2004 FYI data	kevin	21-Feb-06	Delete
25	Luscombe 2004 regulatory	kevin	21-Feb-06	Delete

Categories in database (upload download)

ID	Name	Creator	Creation date



Display options:

Default colors:

Node: blue Edge: lightgrey Text: white

Special coloring:

○ None

● Color gradient: Degree of Original network from green to red

○ Color class: Class name: white

Redraw

Statistics:

Category	Node Counts	Edge Count	Connected Components	Degrees				Clustering Coefficients				Eccentricities				Betweenness			
				Avg.	S.D.	Min.	Max.	Avg.	S.D.	Min.	Max.	Avg.	S.D.	Min.	Max.	Avg.	S.D.	Min.	Max.
Whole network	276	187	109	1.30	0.74	1	7	0.04	0.19	0.00	1.00	2.51	1.57	1	9	3.60	20.22	0.00	200.00

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]

H Yu
P Kim
K Yip
T Gianoulis
C Cheng

Acknowledgements

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

SUBJECT: Networks

DESCRIPTION:

Network Biology: Understanding metabolic and protein interactions, Mathematical Biosciences Institute, Columbus, OH; 2009.09.14, 13:30-14:30; [I:**MBINETS**] (Long networks talk, adding in for the first time: **rescore***, **mirnatargevolrate*** & **netdynamicsrev***. Fits easily into 55′ w. 5′ questions. 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)

PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2008. Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html . Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to 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** .