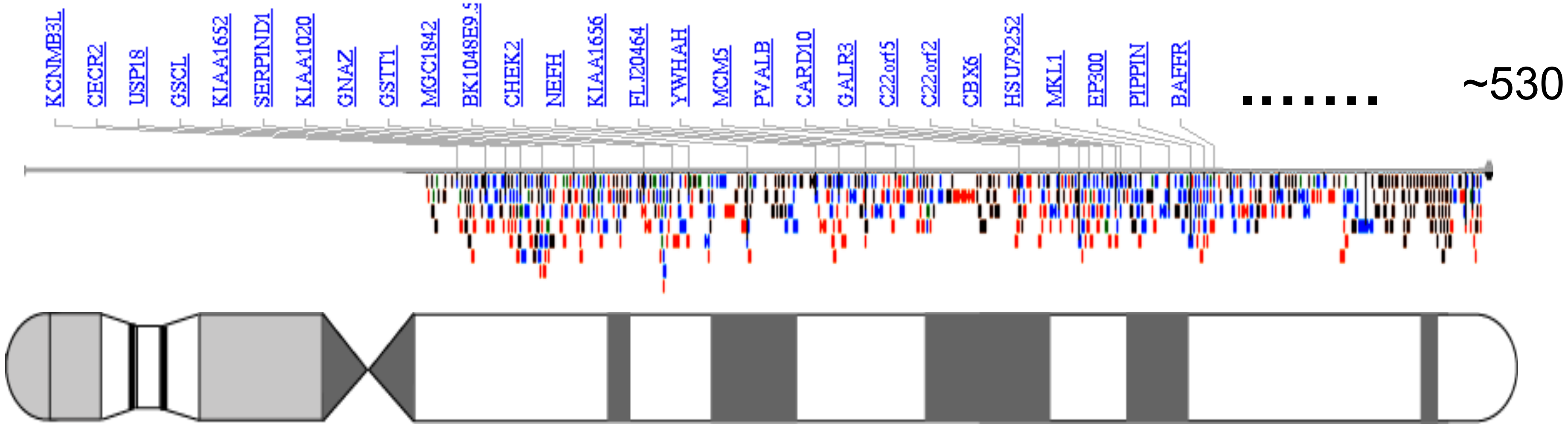
Understanding Protein Function on a Genome-scale through the Analysis of Molecular Networks

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)

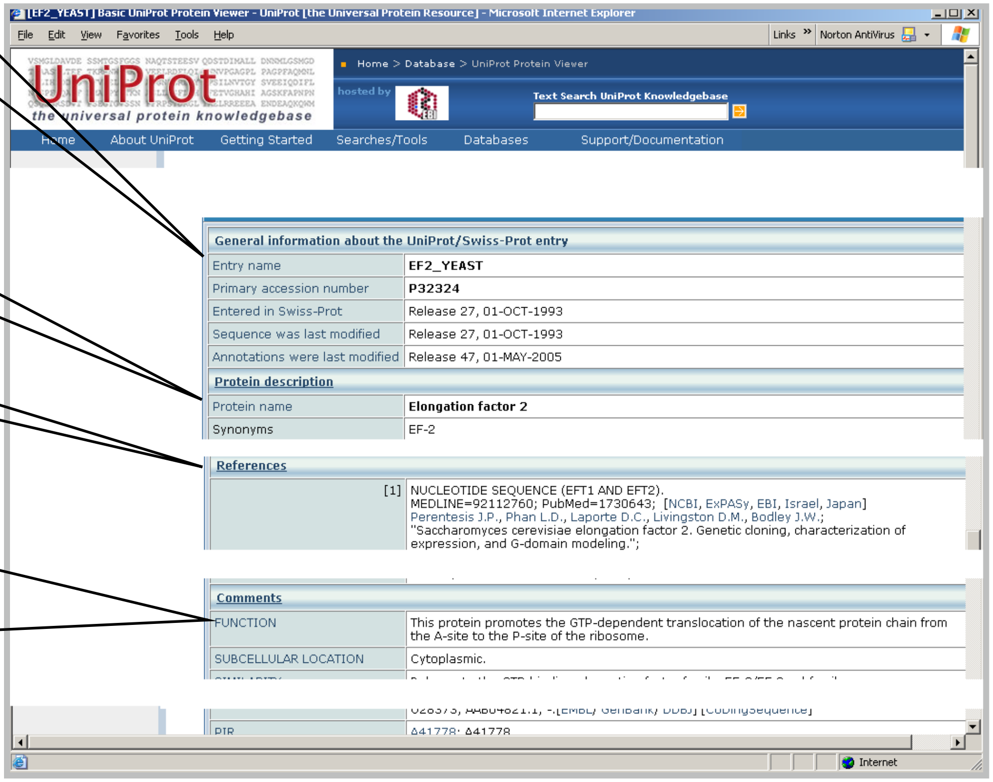
Traditional single molecule way to integrate evidence & describe function

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
EBI
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.

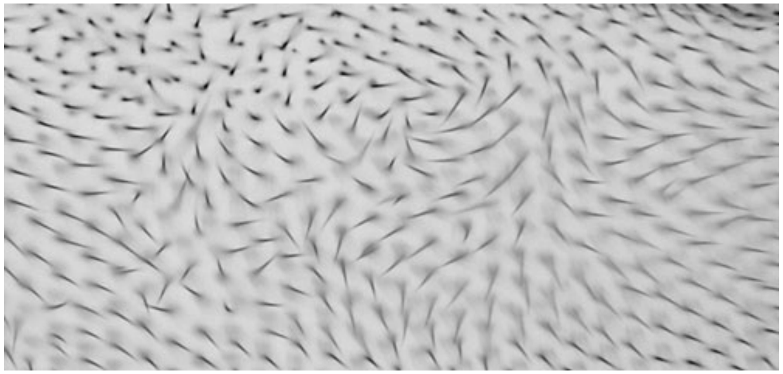
	U28373, AAB04821.1, -.[EMBL/ GenBank/ DDBJ] [CodingSequence]
PIR	A41778; A41778

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

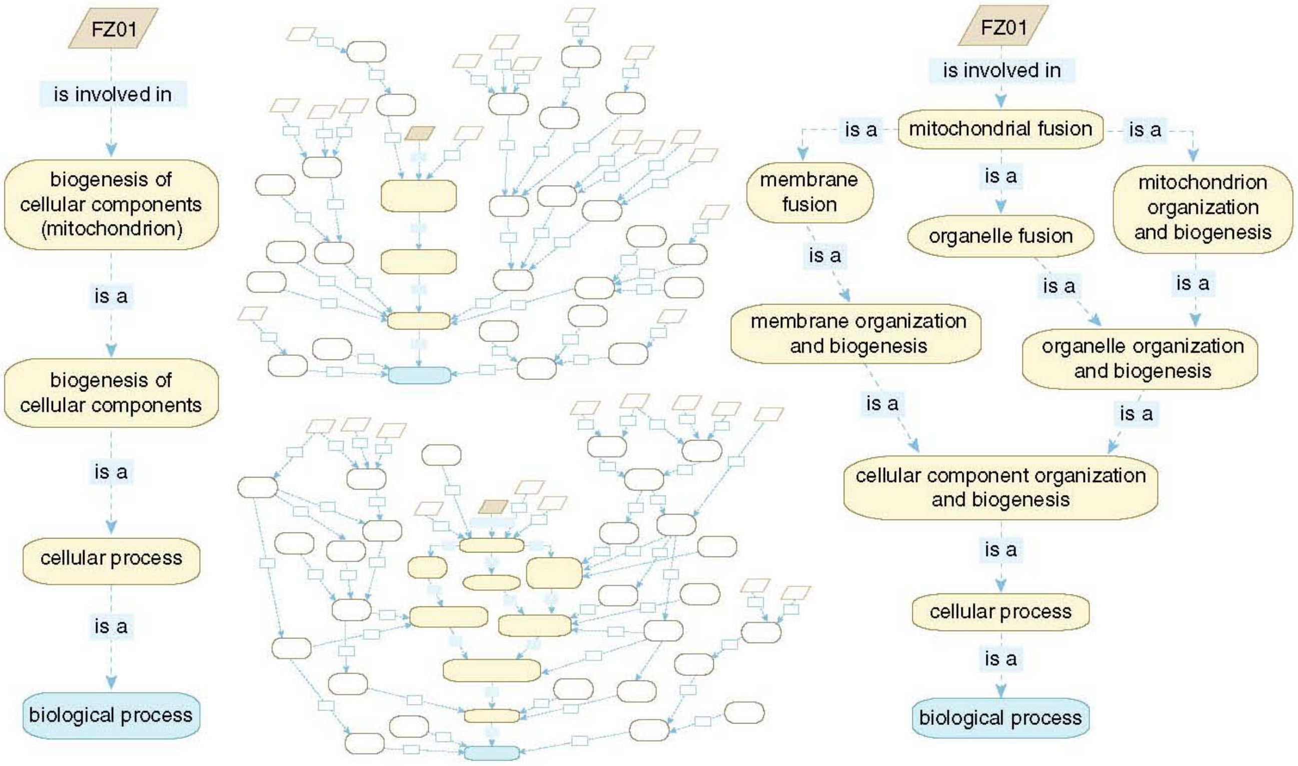
Some obvious issues in scaling single molecule definition to a genomic scale

- 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
is a
mitochondrial fusion
is a
membrane fusion
is a
organelle fusion
mitochondrion organization and biogenesis
is a
is a
is a
membrane organization and biogenesis
organelle organization and biogenesis
is a
is a
cellular component organization and biogenesis
is a
cellular process
is a
biological process

MIPS (Mewes et al.)

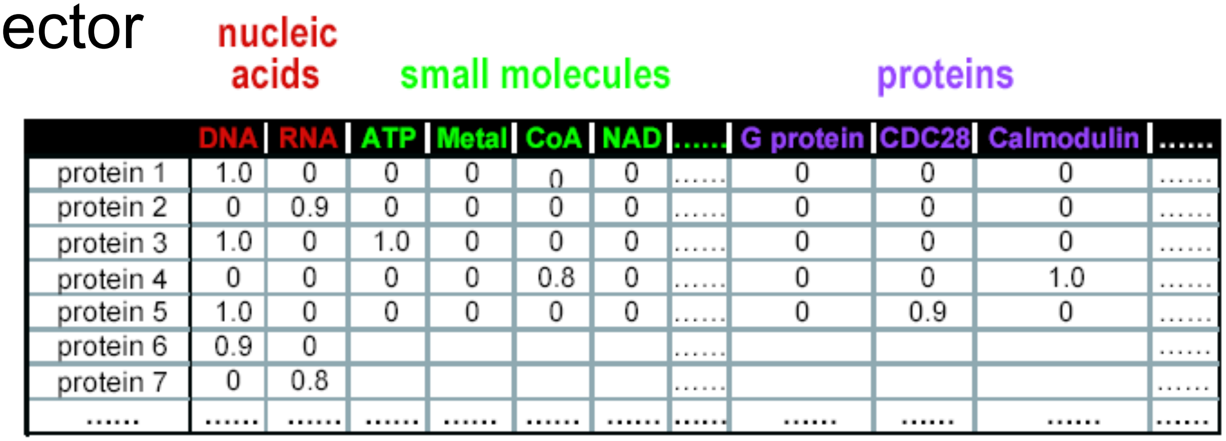
GO (Ashburner et al.)

[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)

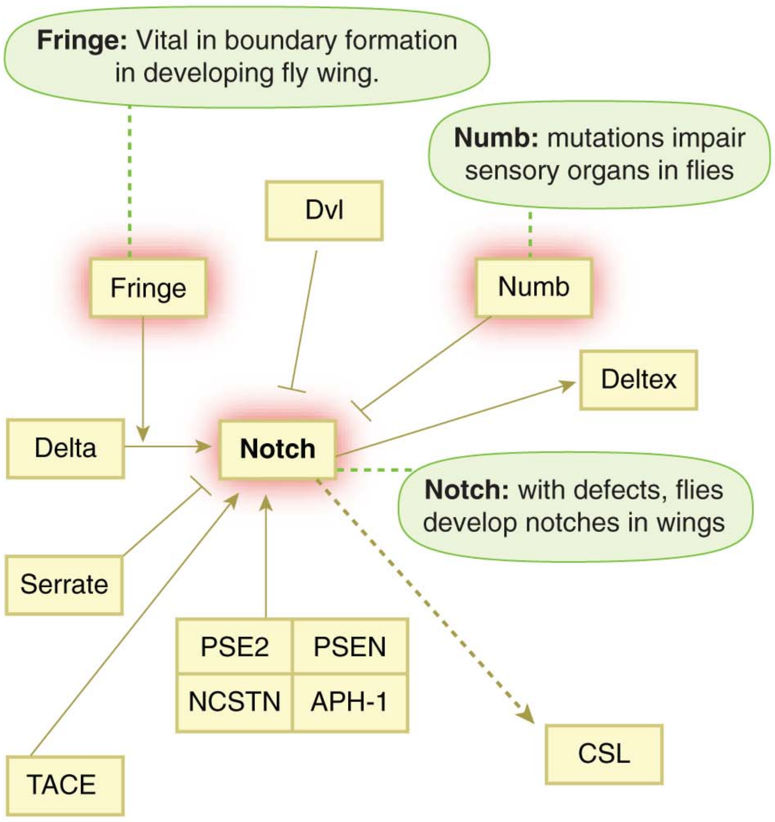
- Function as a vector



	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]

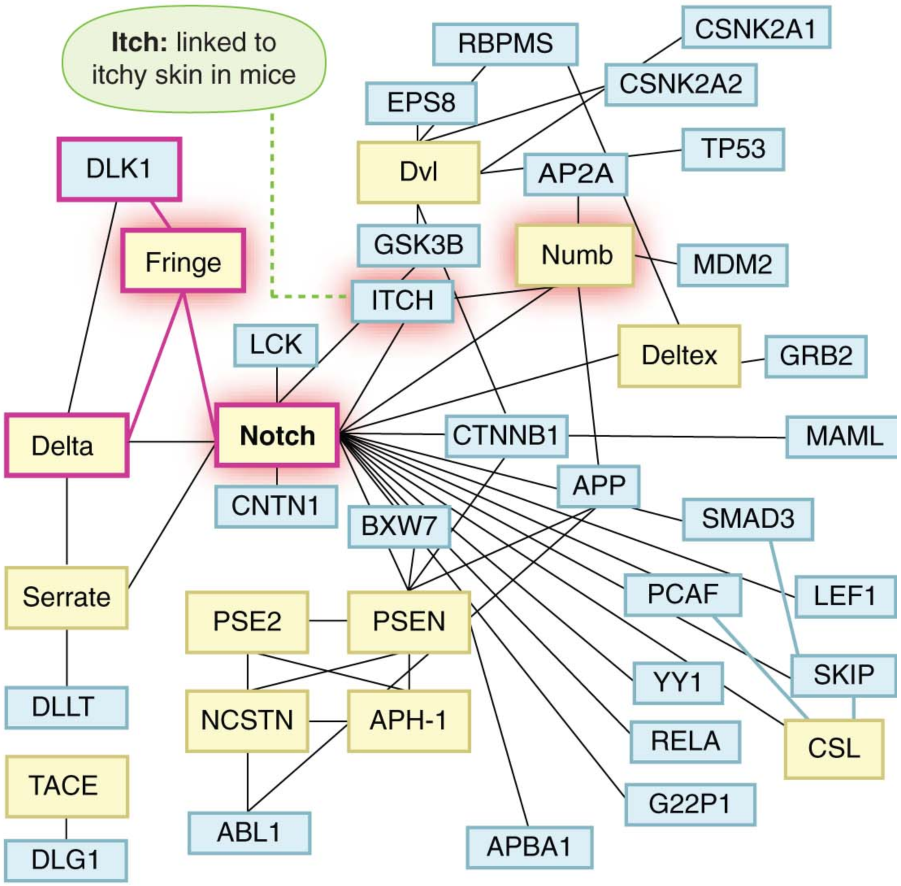
Networks (Old & New)


Fringe: Vital in boundary formation in developing fly wing.

Numb: mutations impair sensory organs in flies

Notch: with defects, flies develop notches in wings

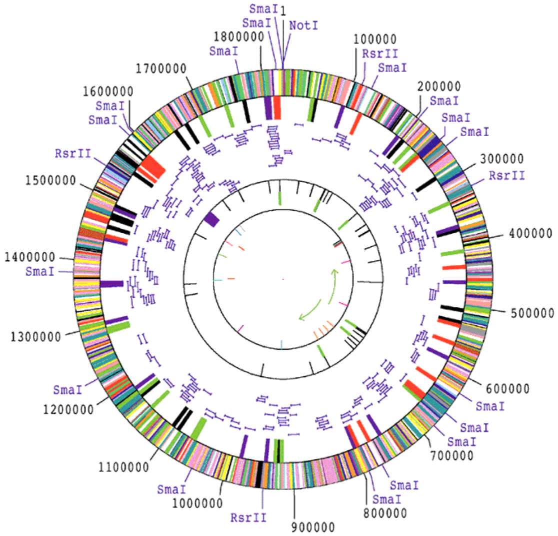
Classical KEGG pathway


Itch: linked to itchy skin in mice

Same Genes in High-throughput Network

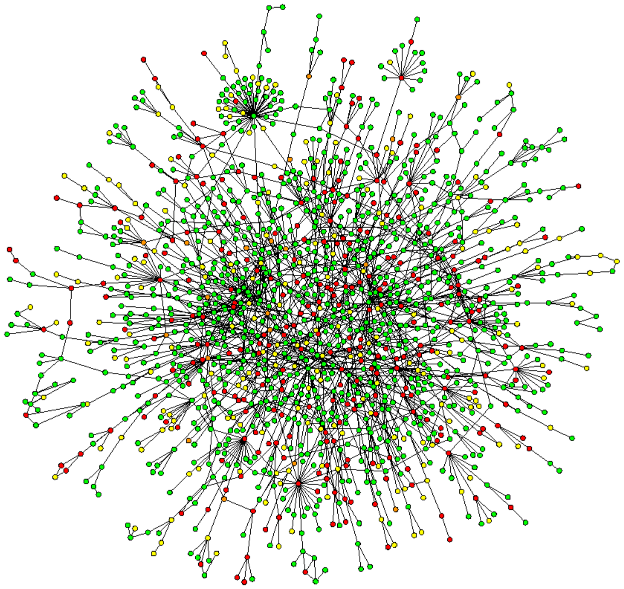
[Seringhaus & Gerstein, Am. Sci. '08]

Networks occupy a midway point in terms of level of understanding



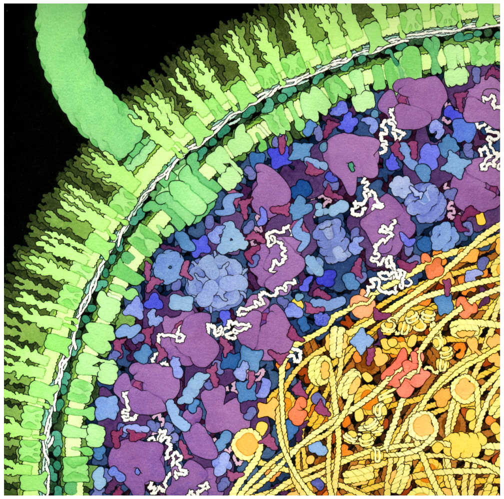
1D: Complete Genetic Partslist

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



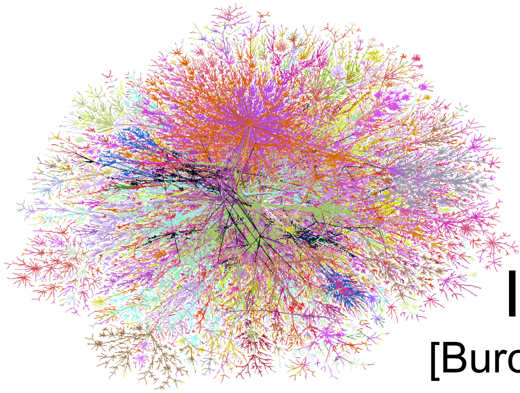
~2D: Bio-molecular Network Wiring Diagram

[Jeong et al. Nature, 41:411]

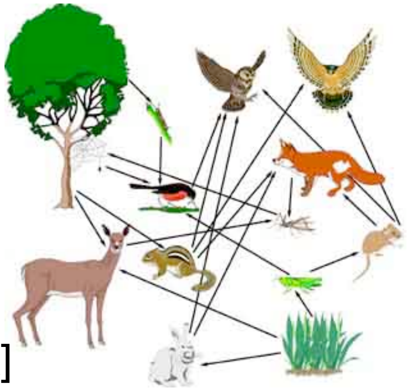


3D: Detailed structural understanding of cellular machinery

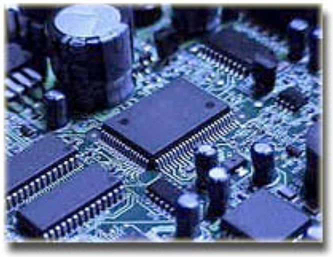
Networks as a universal language



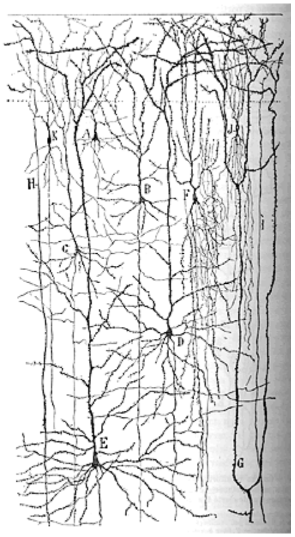
Internet
[Burch & Cheswick]



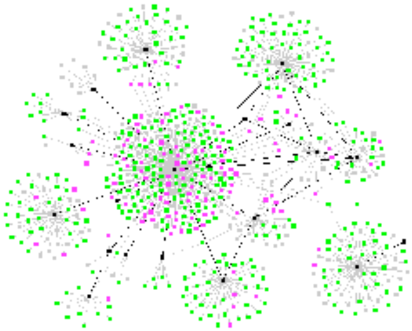
Food Web



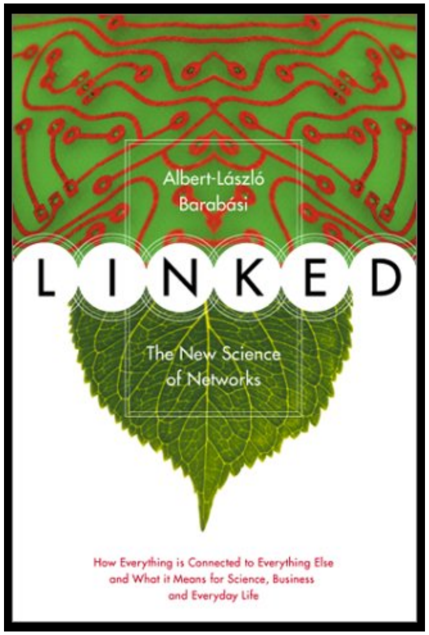
Electronic
Circuit



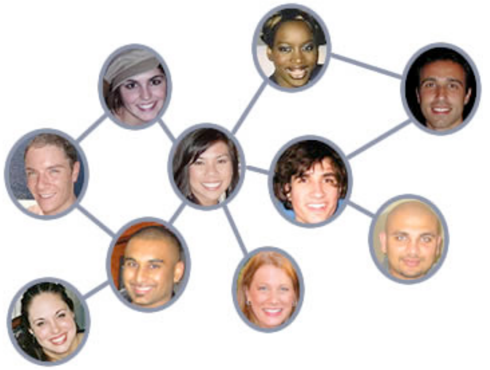
Neural Network
[Cajal]



Disease
Spread
[Krebs]

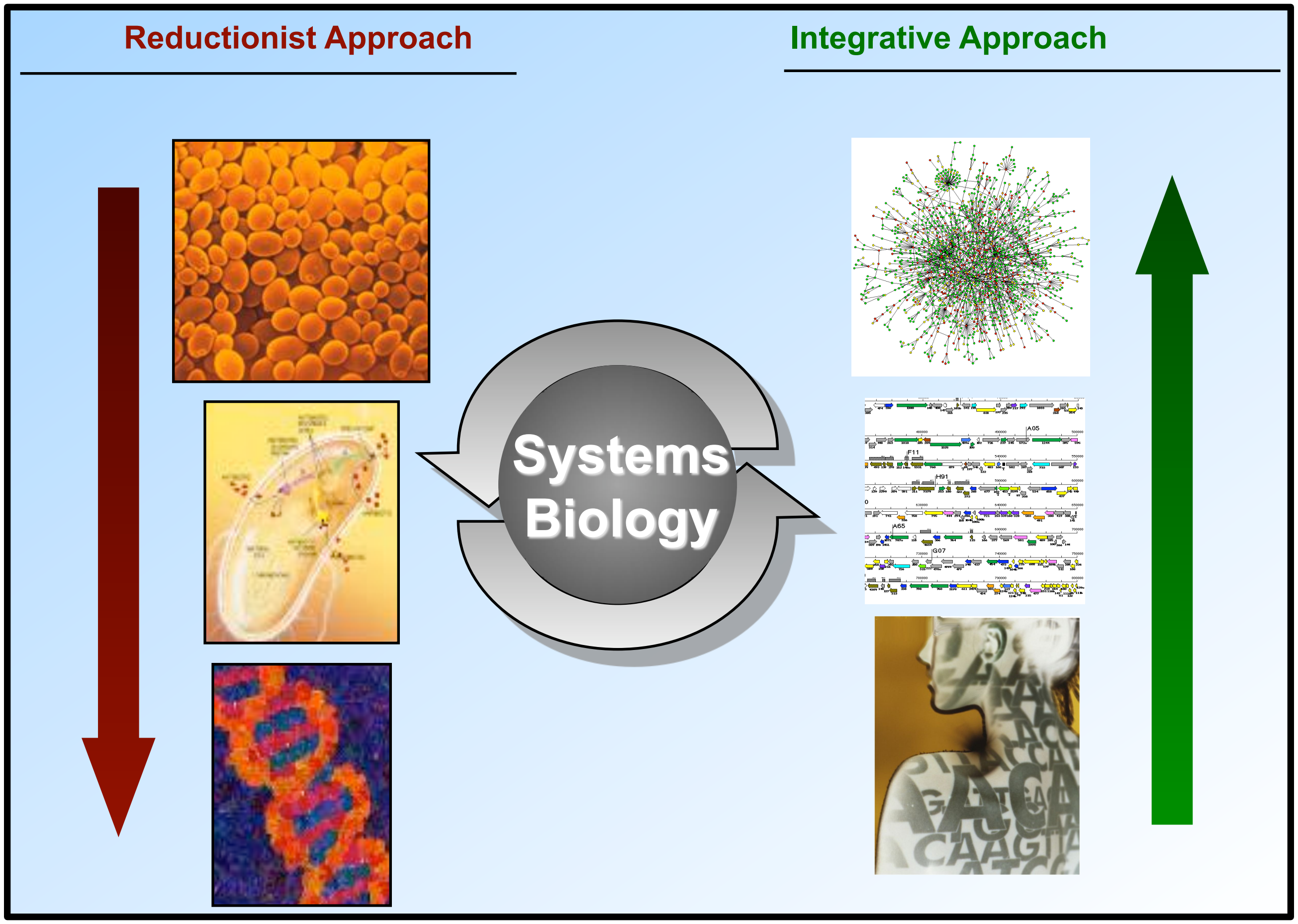

Albert-László
Barabási
LINKED
The New Science
of Networks
How Everything is Connected to Everything Else
and What it Means for Science, Business
and Everyday Life

Protein
Interactions
[Barabasi]



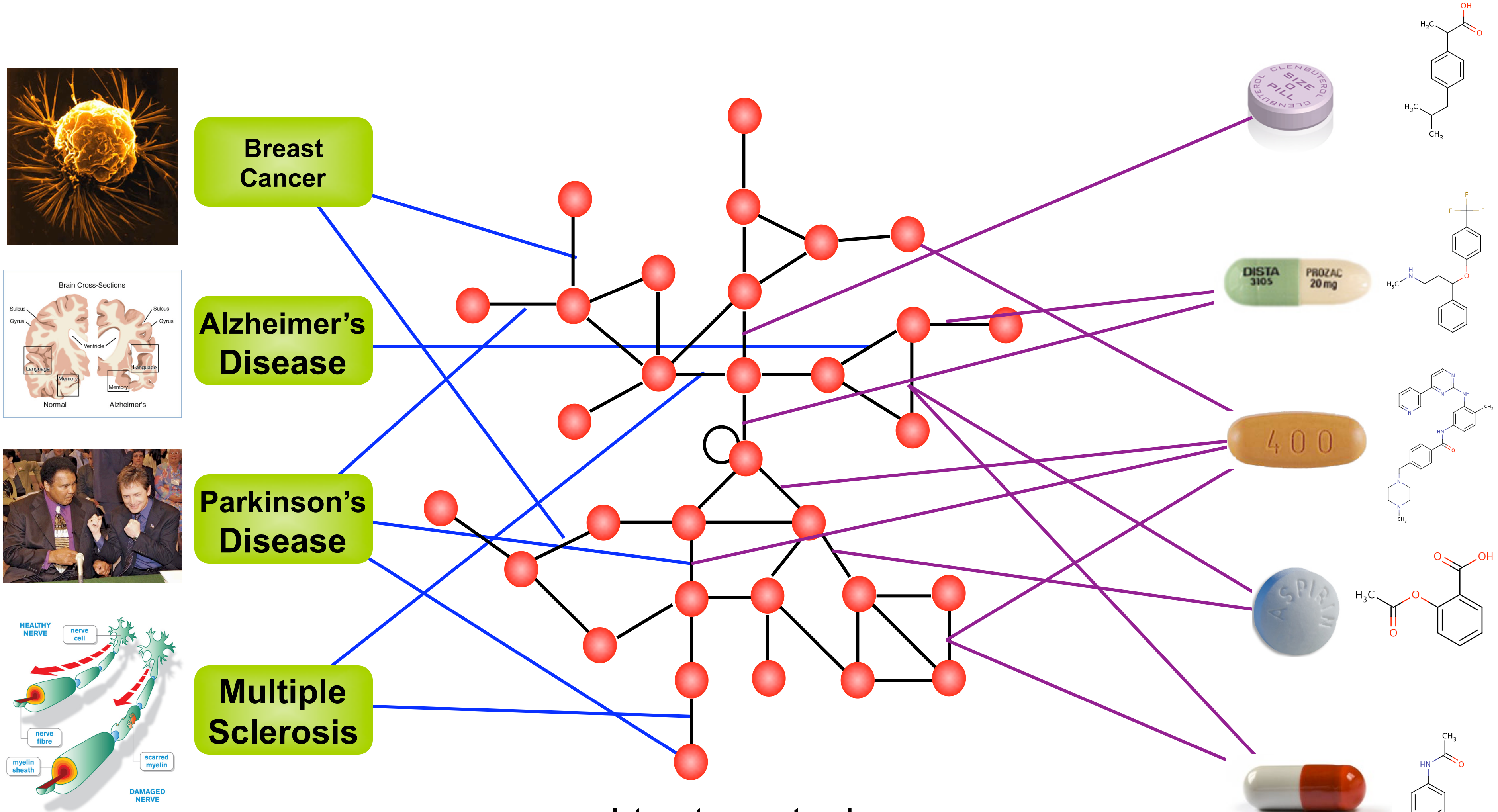
Social Network

Networks as a Central Theme in Systems Biology



[Adapted from H Yu]

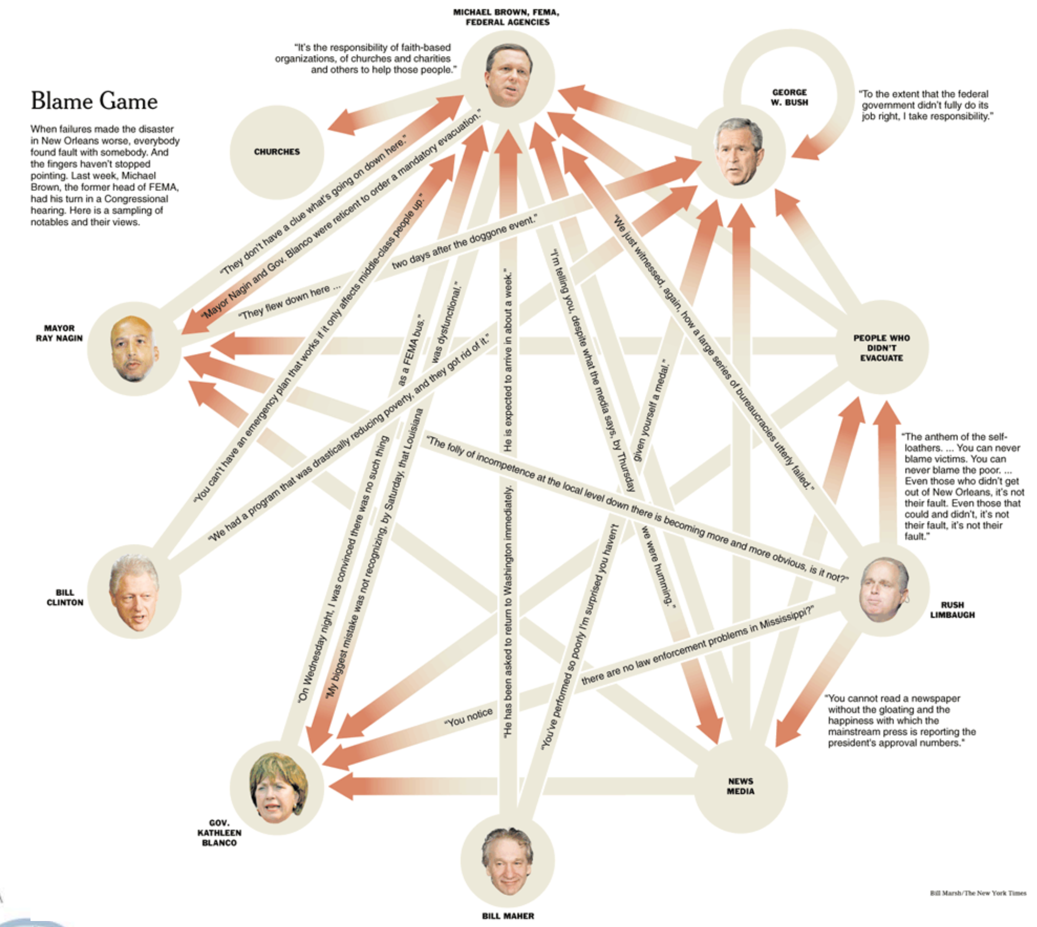
Network pathology & pharmacology

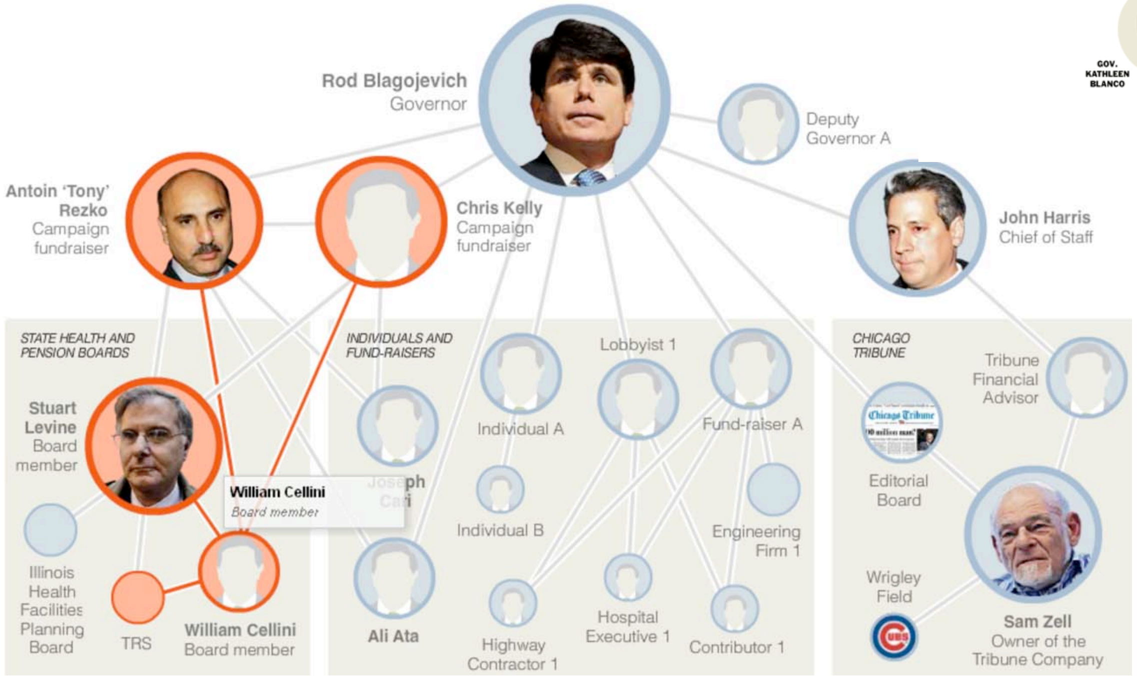


Interactome networks

[Adapted from H Yu]

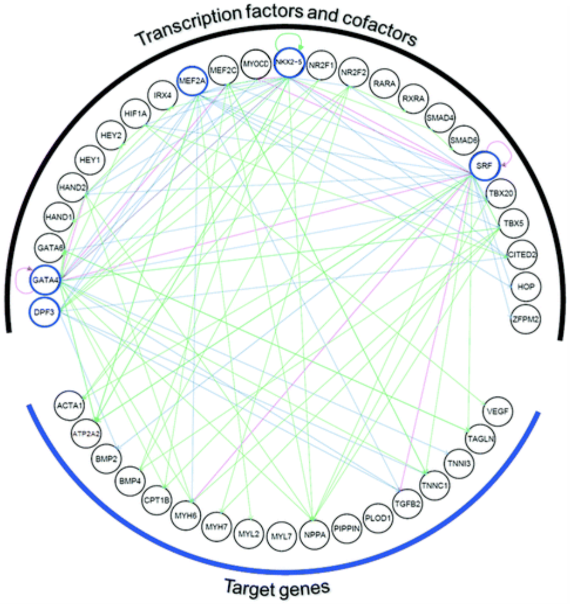
Using the position in networks to describe function


Blame Game
When failures made the disaster in New Orleans worse, everybody found fault with somebody. And the fingers haven't stopped pointing. Last week, Michael Brown, the former head of FEMA, had his turn in a Congressional hearing. Here is a sampling of notables and their views.
MICHAEL BROWN, FEMA, FEDERAL AGENCIES
"It's the responsibility of faith-based organizations, of churches and charities and others to help those people."
CHURCHES
GEORGE W. BUSH
"To the extent that the federal government didn't fully do its job right, I take responsibility."
MAYOR RAY NAGIN
PEOPLE WHO DIDN'T EVACUATE
"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."
BILL CLINTON
RUSH LIMBAUGH
"You cannot read a newspaper without the gloating and the happiness with which the mainstream press is reporting the president's approval numbers."
GOV. KATHLEEN BLANCO
NEWS MEDIA
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?"
Bill Marsh/The New York Times

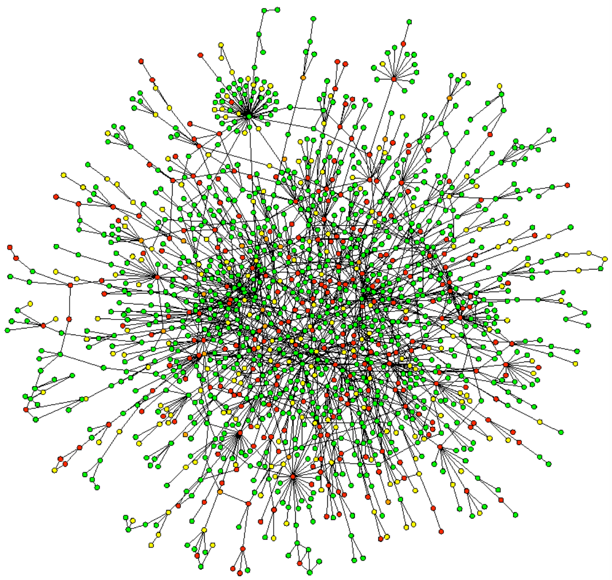


[NY Times, 2-Oct-05, 9-Dec-08]

Types of Networks

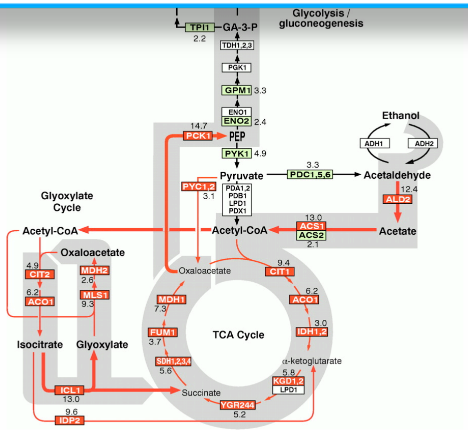


Regulatory networks



Interaction networks

Nodes: proteins or genes
Edges: interactions



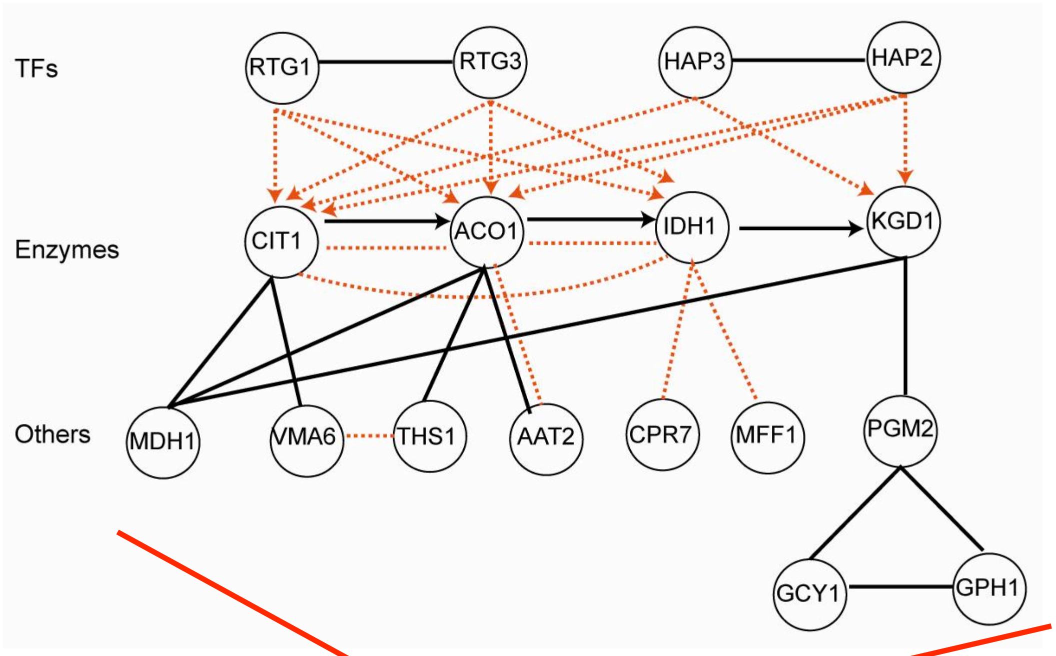
Metabolic networks

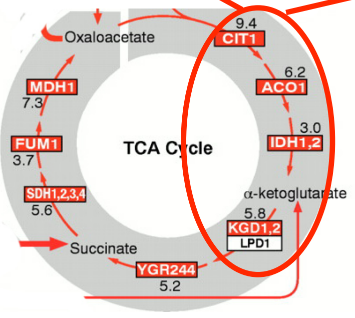
[Horak, et al, Genes & Development, 16:3017-3033]

[DeRisi, Iyer, and Brown, Science, 278:680-686]

[Jeong et al, Nature, 41:411]

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

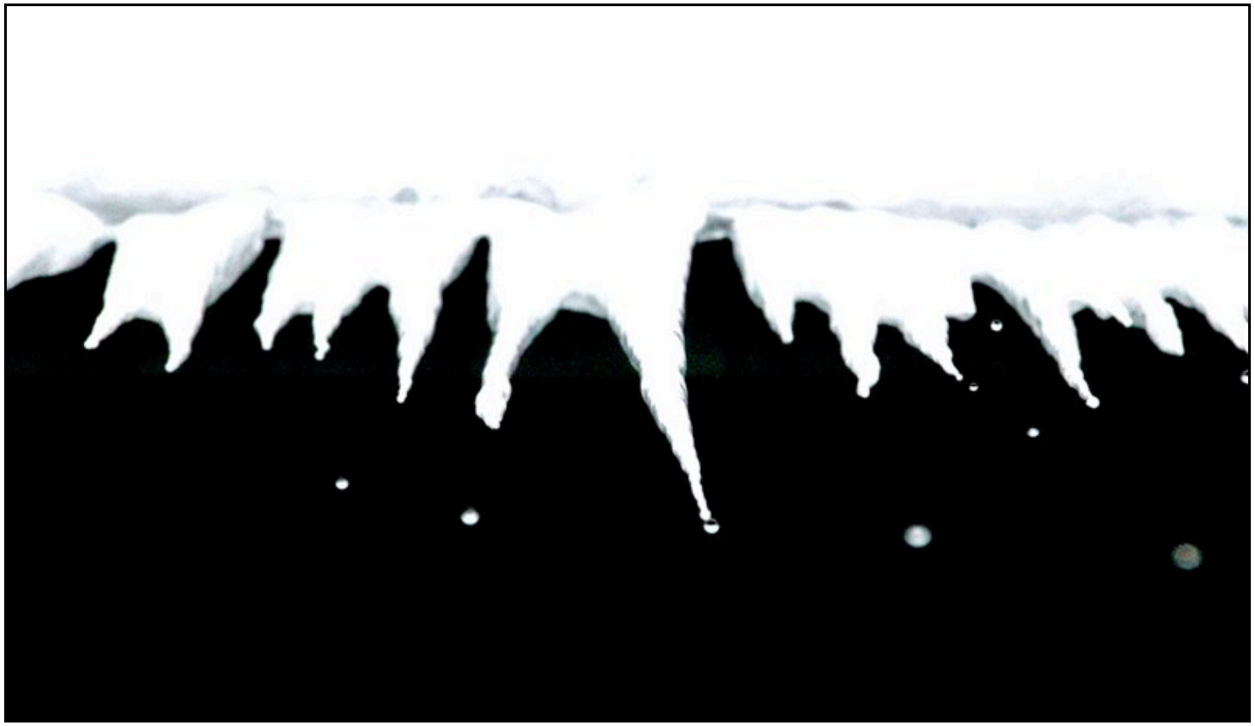
Genetic interaction (synthetic lethal)
Signaling pathways

Outline

- Predicting Networks
  - ◊ Training set expansion
- Properties of Protein Networks
  - ◊ Hubs
- Dynamics of Networks
  - ◊ Dynamics across cellular states
  - ◊ Dynamics across environments
- Protein Networks and Human Variation

Predicting Networks

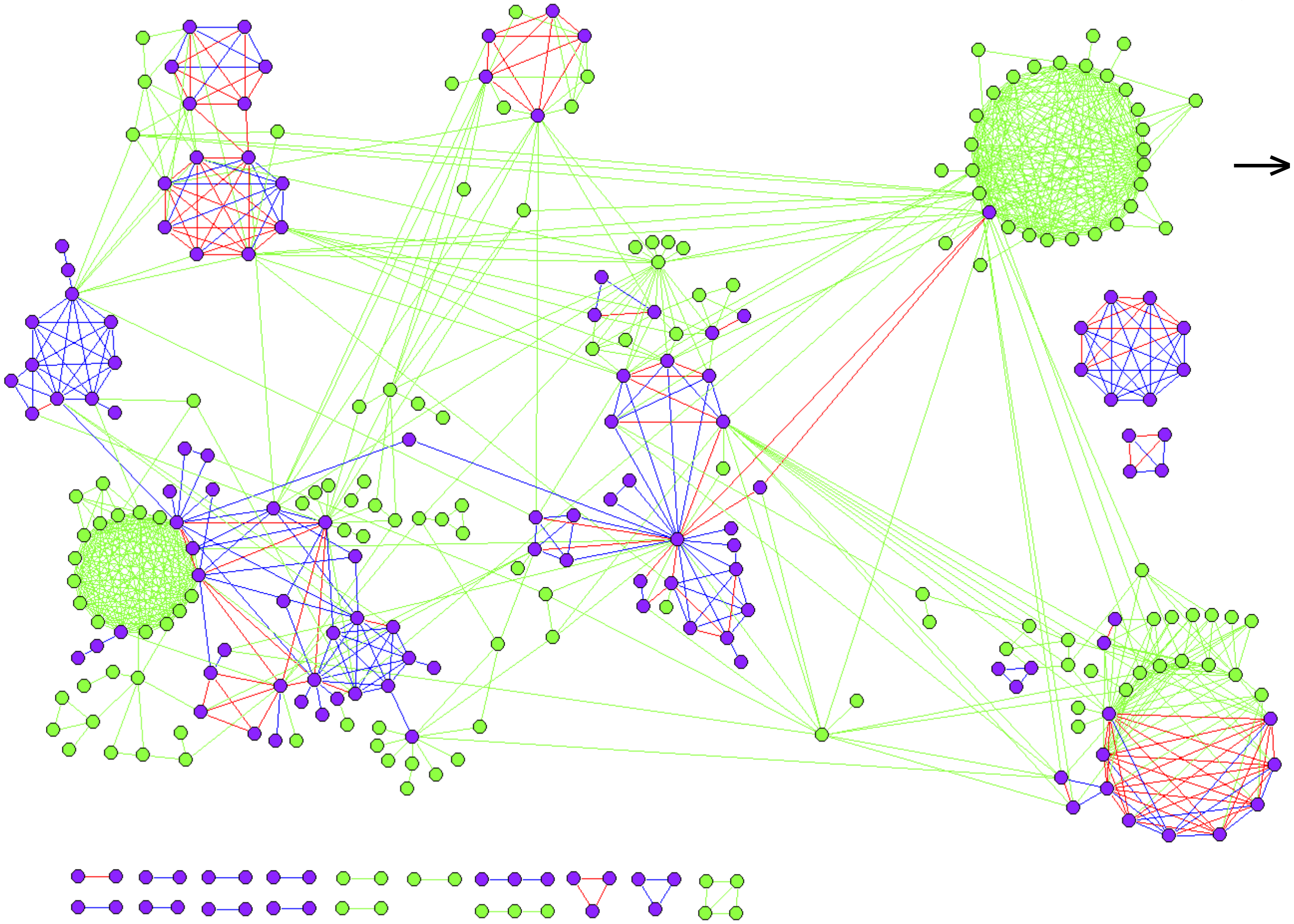
How do we construct large molecular networks?
From extrapolating correlations between functional genomics data with fairly small sets of known interactions, making best use of the known training data.



Network Prediction

- Only small portions are already known
- Many other kinds of data available

→ Use them to learn models for predicting the unknown portions



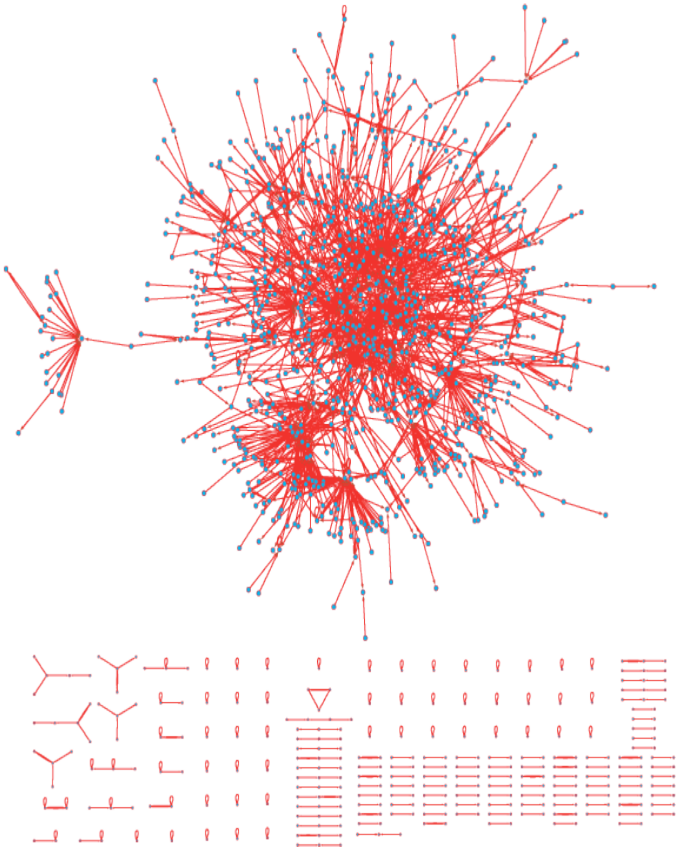
Known

New

Ex. of Predicted Membrane Protein Interactome in Xia et al. JMB (2006)

Figure 6: A map of known and a subset of predicted interactions among helical membrane proteins. Nodes represent helical

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:

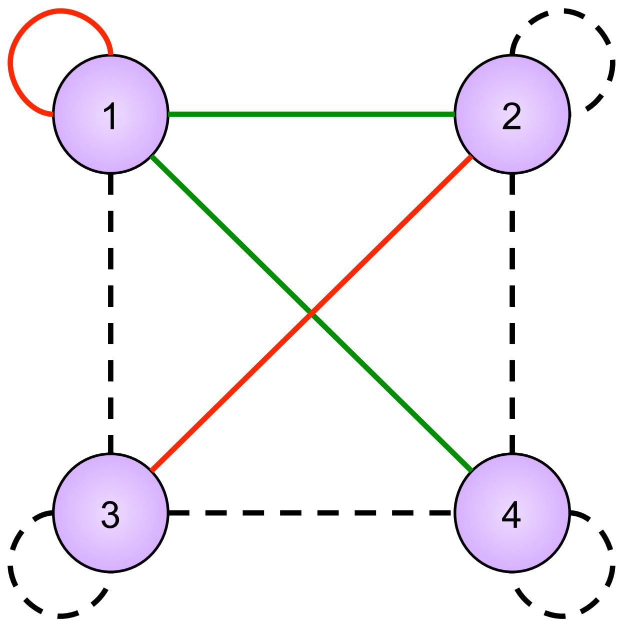
- ◊ Small-scale experiments: accurate but few
  → Overfitting: ~5,000 in BioGRID, involving ~2,300 proteins
- ◊ Large-scale experiments: abundant but noisy
  → Noise: false +ve/-ve for yeast two-hybrid data up to 45% and 90% (Huang et al., 2007)

Learning

Concepts in machine learning:

- Training sets:
  - ◊ Positive set: known interactions
  - ◊ Negative set: known non-interactions
- Features:
  - ◊ Data describing the objects
- Model:
  - ◊ A function that takes two objects as input and predicts whether they interact

Training sets



━━━ Known interactions

━━━ Known non-interactions

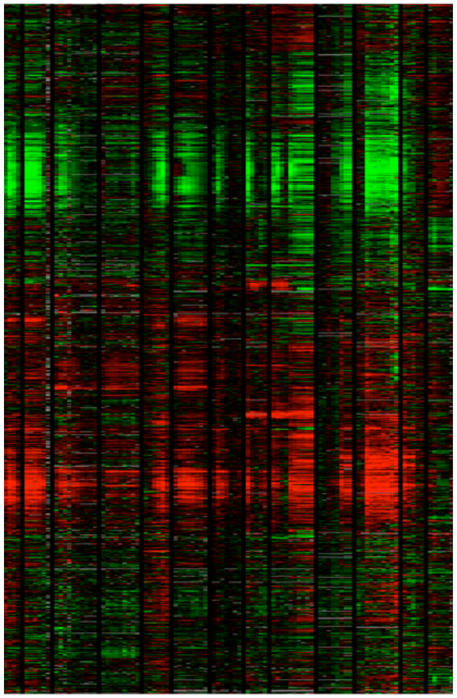
- - - - Unknown

⇨

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

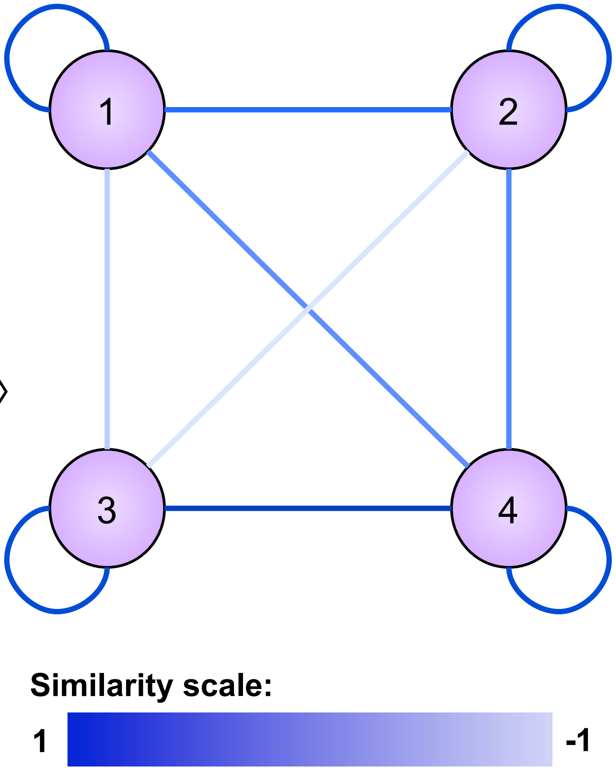
Features

- Example 1: gene expression



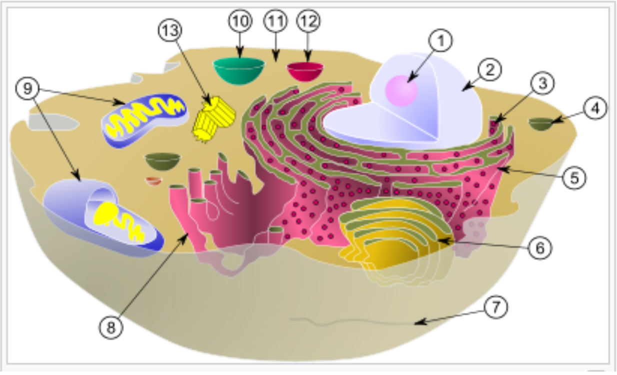
Gasch et al., 2000

$x_1 = (0.2, 2.4, 1.5, \ldots)$
$x_2 = (0.8, 2.2, 1.5, \ldots)$
$x_3 = (4.3, 0.1, 7.5, \ldots)$
$\ldots$
$\mathrm{sim}(x_1, x_2) = 0.62$
$\mathrm{sim}(x_1, x_3) = -0.58$
$\ldots$



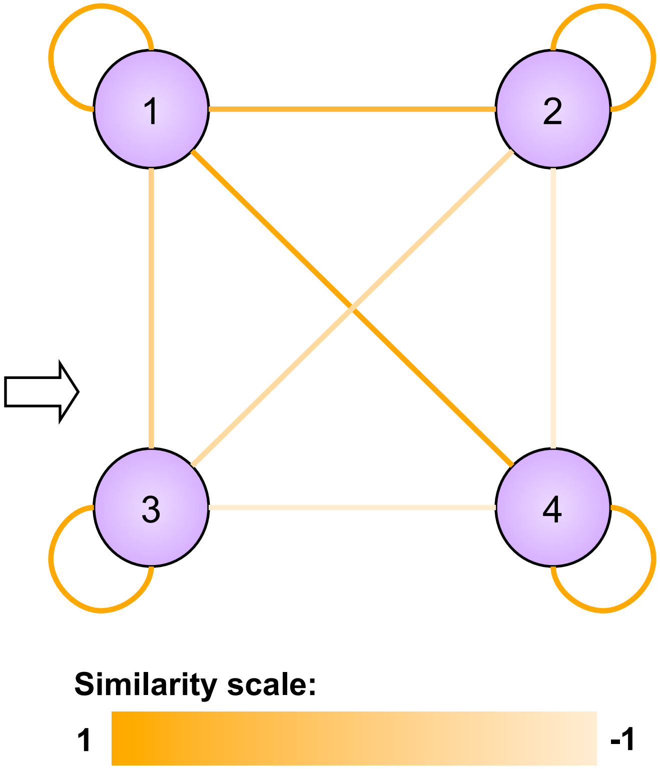
Features

- Example 2: sub-cellular localization

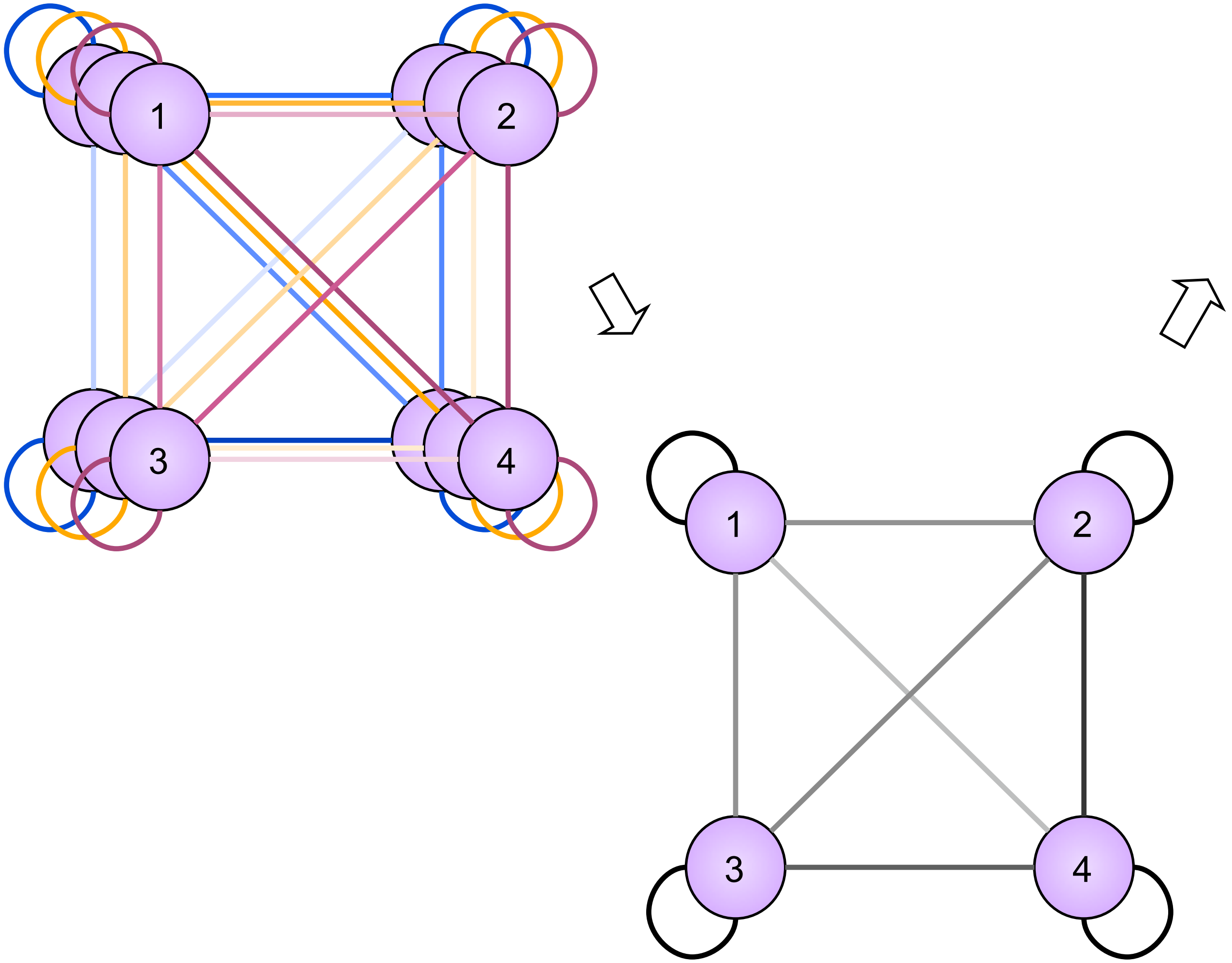


http://www.scq.ubc.ca/wp-content/yeasttwohybridtranscript.gif

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



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

Evaluation

- Computational:
  - Cross-validation
  - Indirect evidence (e.g. same GO category)
- Experimental:
  - Validation of de novo predictions

Learning methods

An endless list:

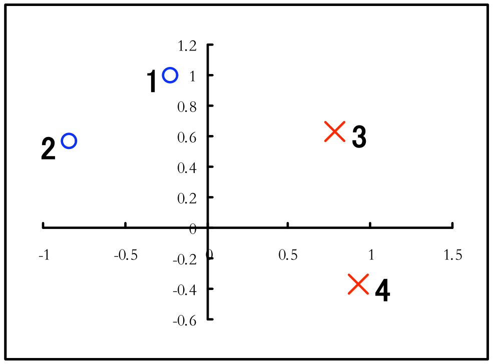
- Docking (e.g. Schoichet and Kuntz 1991)
- Evolutionary (e.g. Ramani and Marcotte, 2003)
- Topological (e.g. Yu et al., 2006)
- Bayesian (e.g. Jansen et al., 2003)
- Kernel methods
  - ◊ Global modeling:
    - em (Tsuda et al., 2003)
    - kCCA (Yamanishi et al., 2004)
    - kML (Vert and Yamanishi,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

…

Let's compare fairly in a public challenge! (DREAM)

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

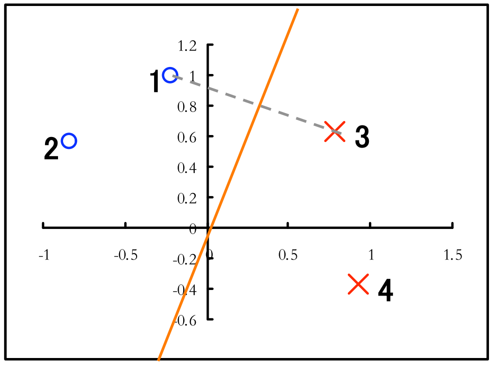
Good for integrating heterogeneous datasets (protein sequences, PSSM, gene expression, …)

– 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

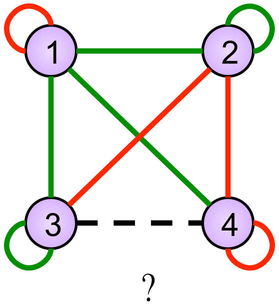
Maximize $\sum_i \lambda_i - \frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j \boxed{\langle x_i , x_j \rangle}$

Subject to $\lambda \geq 0$

$$\sum_i \lambda_i y_i = 0$$

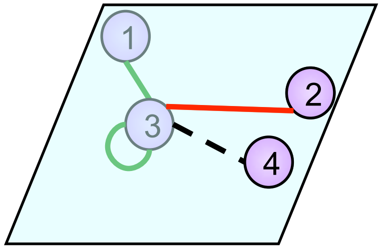
The only thing that we need to know about the objects: their similarity values (inner products)

Kernel methods for predicting networks: local vs. global modeling



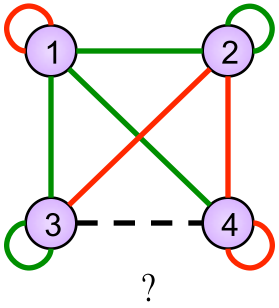
Local modeling: build one model for each node

Model for node 3:



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

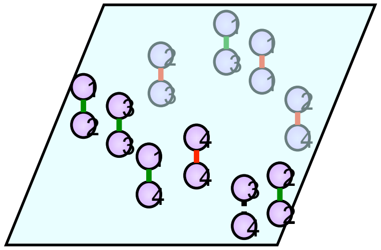
Kernel methods for predicting networks: local vs. global modeling



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

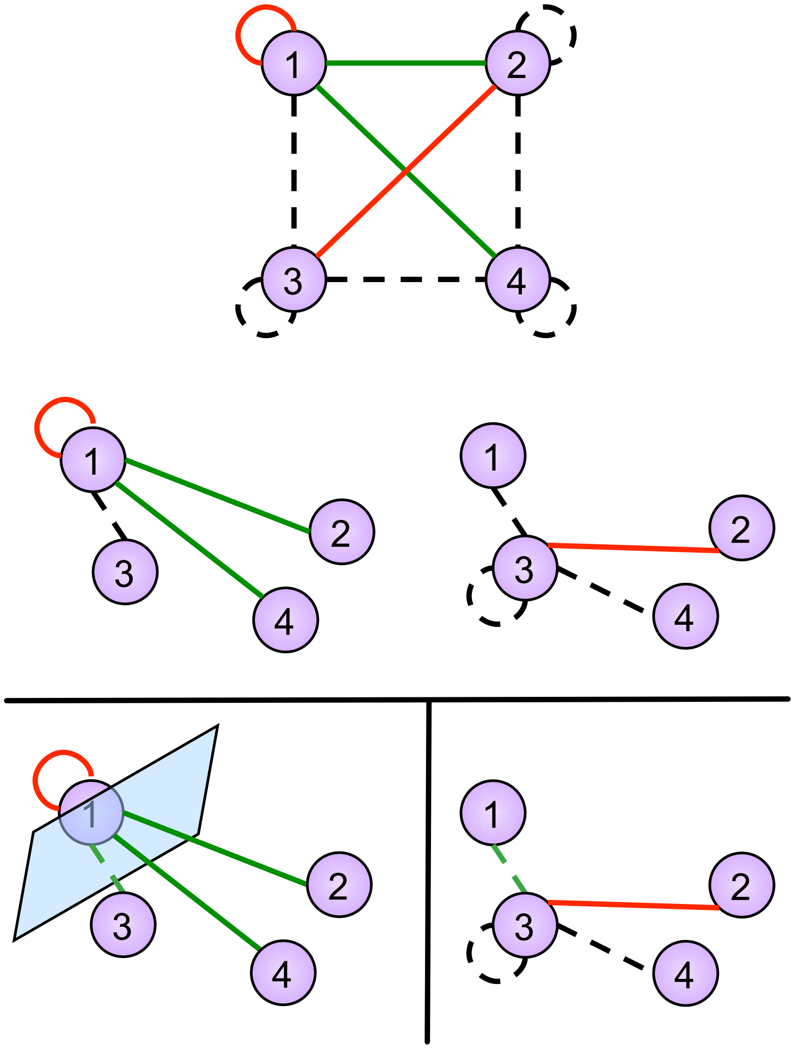
Our work: training set expansion

- 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

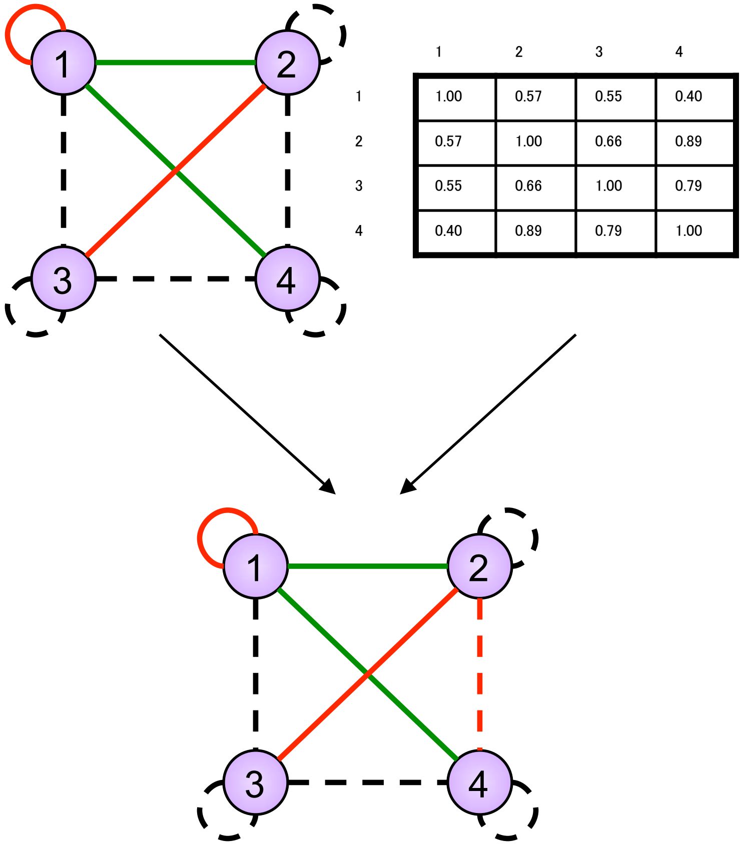
- Motivation: some objects have more examples than others
- Our approach:
  - ◊ Learn models for objects with more examples first
  - ◊ Propagate the most confident predictions as auxiliary examples of other objects



[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)]

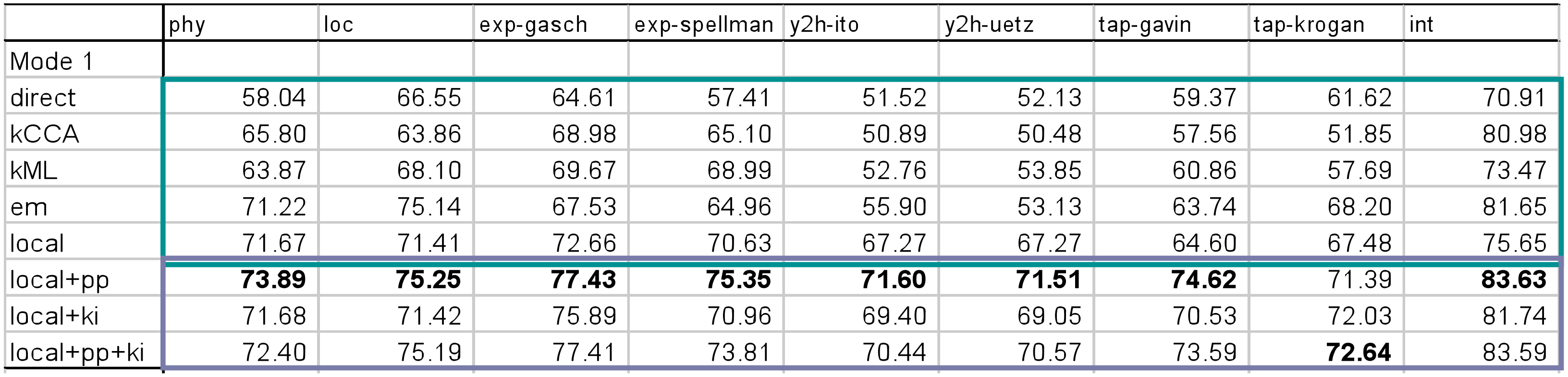
Experiments

- Gold-standard interactions: BioGRID, from studies that report less than 10 interactions
- Features:

Code	Data type	Source	Kernel
phy	Phylogenetic profiles	COG v7 (Tatusov et al., 1997)	RBF (σ=3,8)
loc	Sub-cellular localization	(Huh et al., 2003)	Linear
exp-gasch	Gene expression (environmental response)	(Gasch et al., 2000)	RBF (σ=3,8)
exp-spellman	Gene expression (cell-cycle)	(Spellman et al., 1998)	RBF (σ=3,8)
y2h-ito	Yeast two-hybrid	(Ito et al., 2000)	Diffusion (β=0.01)
y2h-uetz	Yeast two-hybrid	(Uetz et al., 2000)	Diffusion (β=0.01)
tap-gavin	Tandem affinity purification	(Gavin et al., 2006)	Diffusion (β=0.01)
tap-krogan	Tandem affinity purification	(Krogan et al., 2006)	Diffusion (β=0.01)
int	Integration		Summation

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

Prediction accuracy



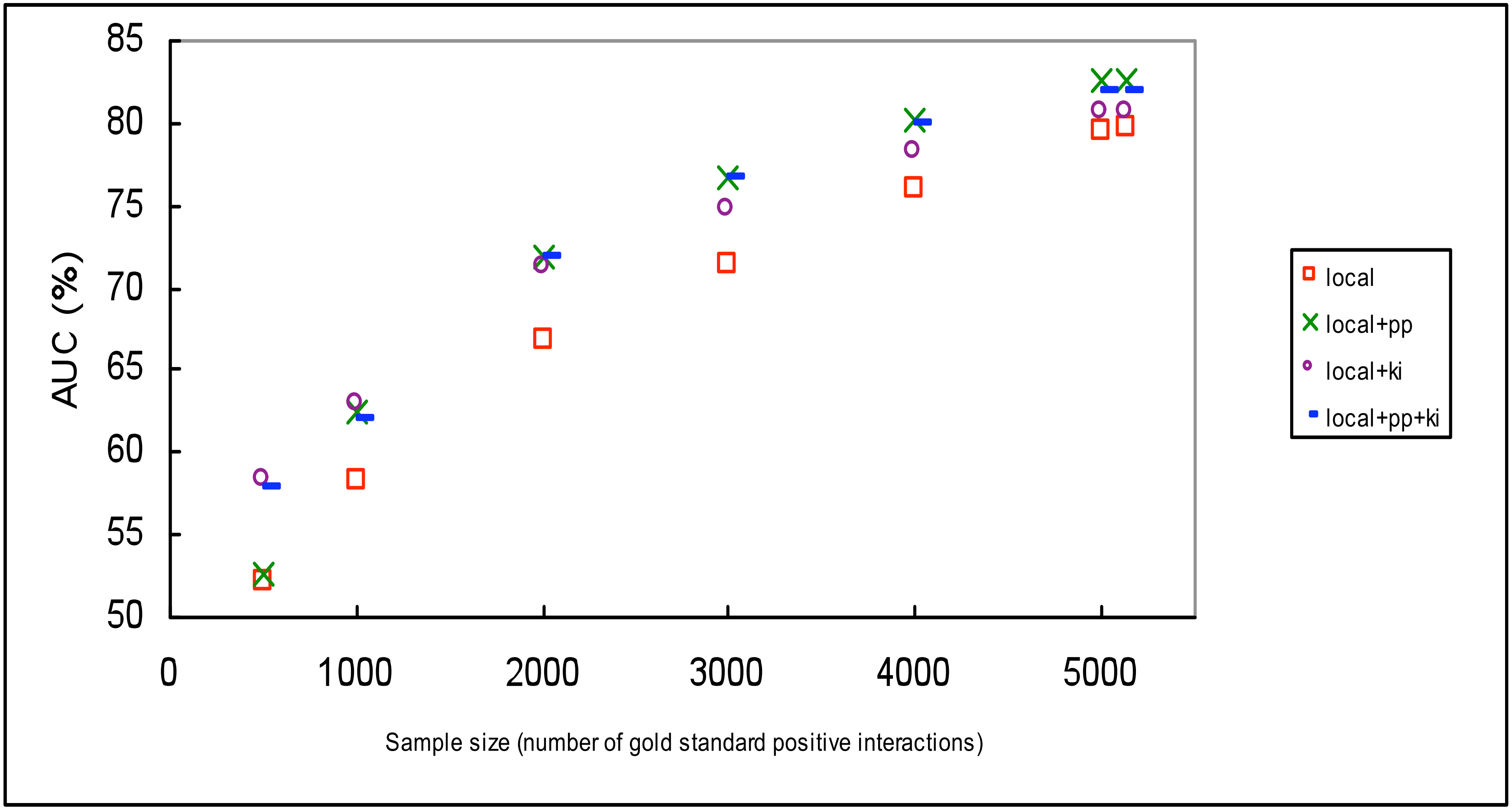
	phy	loc	exp-gasch	exp-spellman	y2h-ito	y2h-uetz	tap-gavin	tap-krogan	int
Mode 1									
direct	58.04	66.55	64.61	57.41	51.52	52.13	59.37	61.62	70.91
kCCA	65.80	63.86	68.98	65.10	50.89	50.48	57.56	51.85	80.98
kML	63.87	68.10	69.67	68.99	52.76	53.85	60.86	57.69	73.47
em	71.22	75.14	67.53	64.96	55.90	53.13	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
- Overfitting of local modeling without training set expansion
- Comparing prediction propagation and kernel initialization

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

Complementarity of the two methods



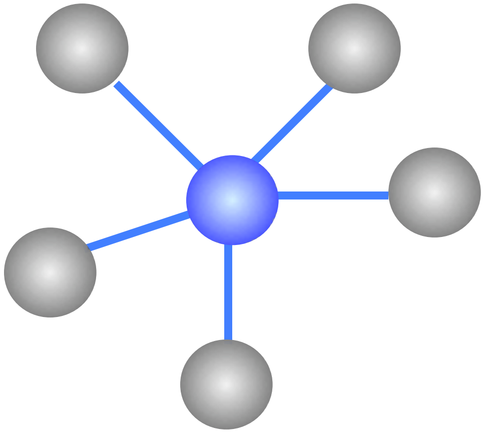
[Yip and Gerstein, Bioinformatics ('09, in press)]

Network Dynamics #1: Cellular States

How do networks change across different cellular states?
How can this be used to assign function to a protein?

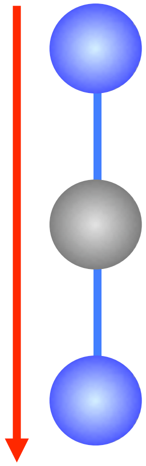
Global topological measures

Indicate the gross topological structure of the network



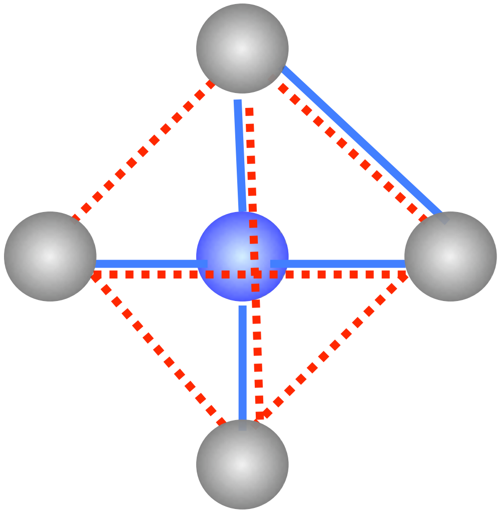
Degree (*K*)

5



Path length (*L*)

2



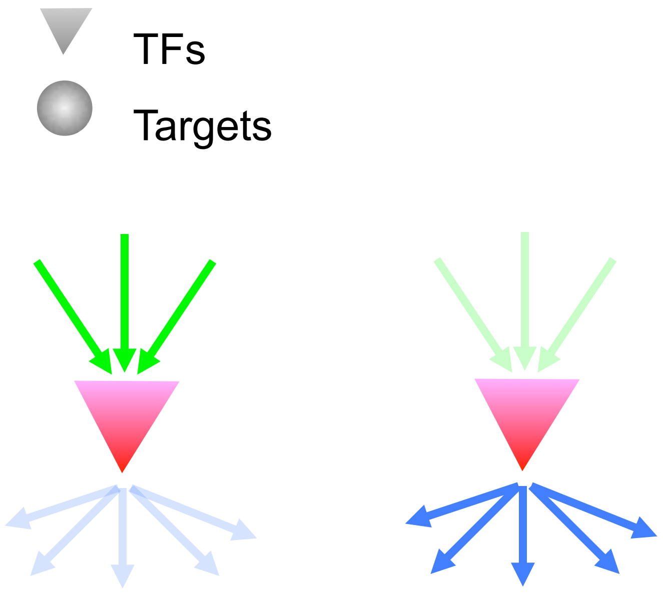
Clustering coefficient (*C*)

1/6

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

[Barabasi]

Global topological measures for directed networks



▼ TFs

● Targets

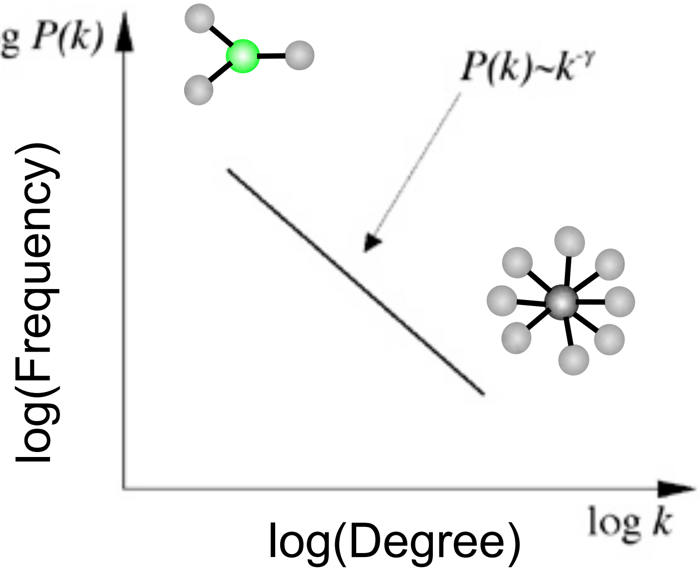
In-degree
3

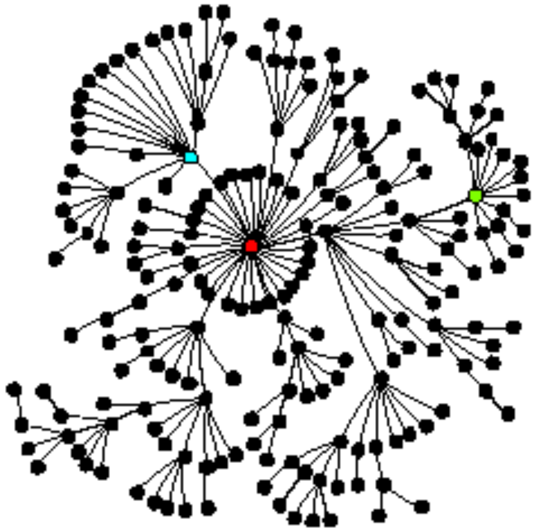
Out-degree
5

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

Scale-free networks

Power-law distribution





Hubs dictate the structure of the network

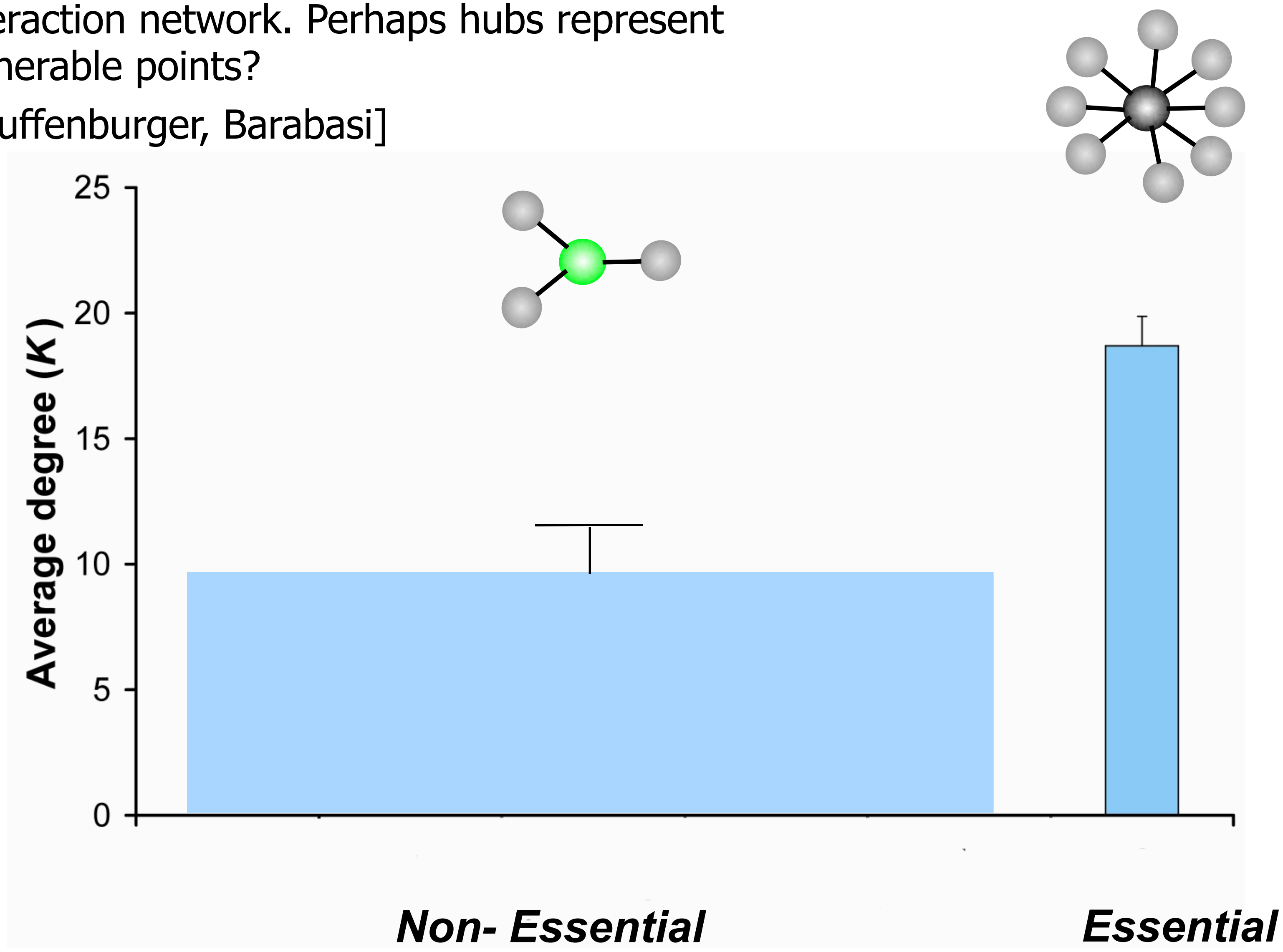
[Barabasi]

Hubs tend to be Essential

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

[Lauffenburger, Barabasi]

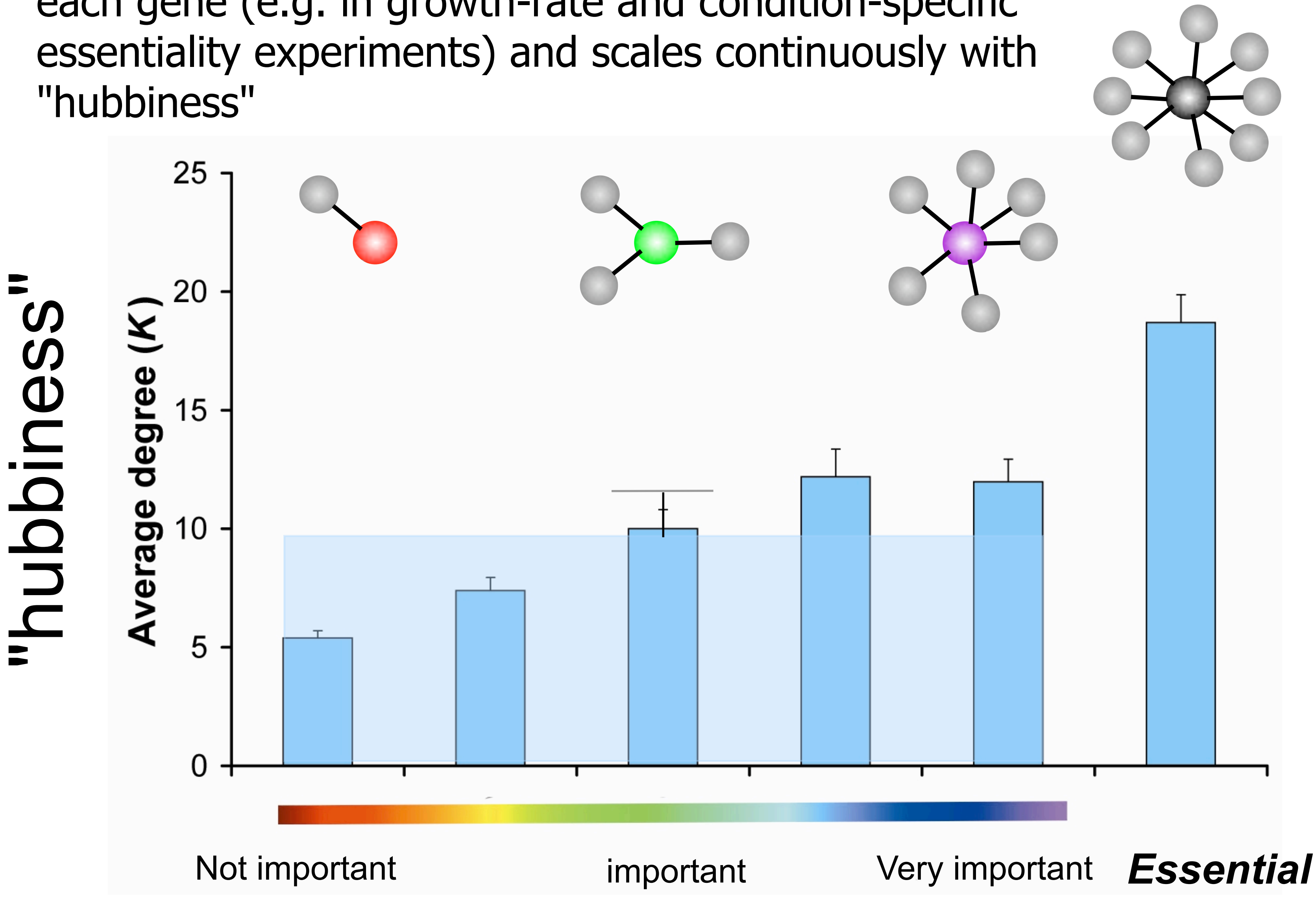
"hubbiness"



[Yu et al., 2003, TIG]

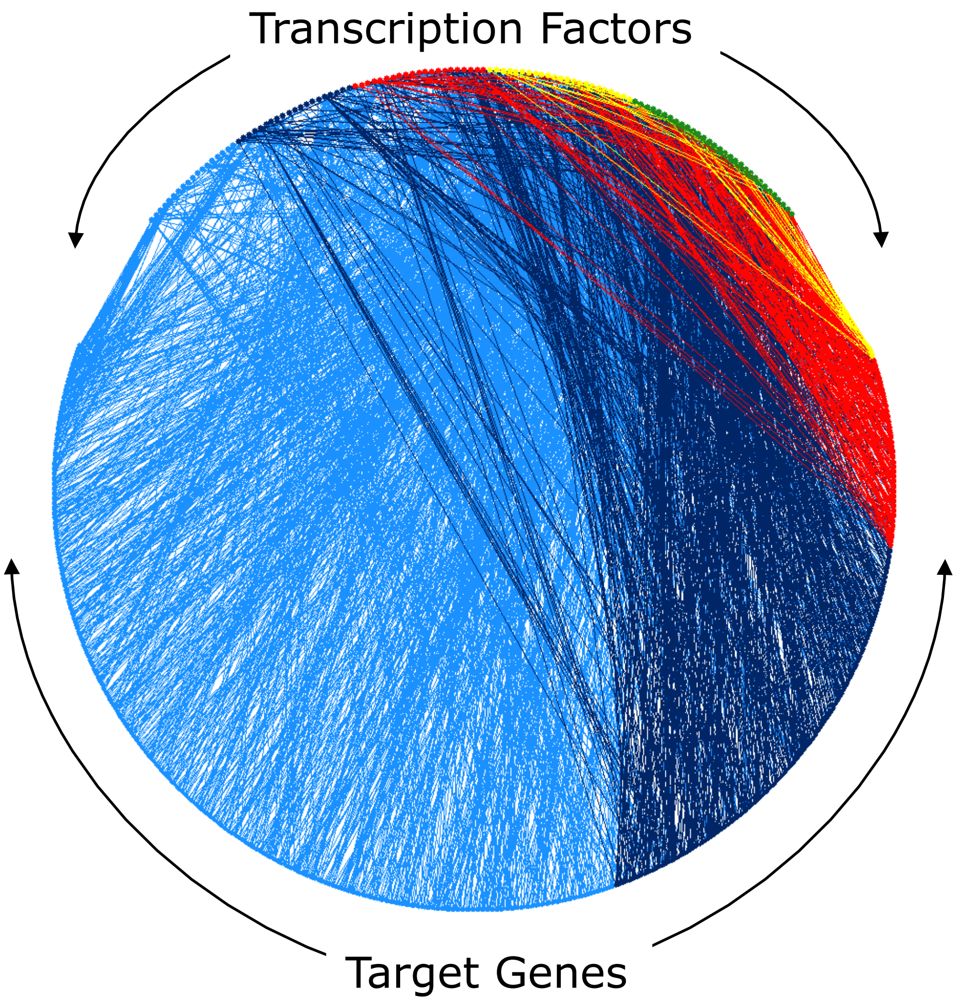
Relationships extends to "Marginal Essentiality"

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]

Dynamic Yeast TF network



- Analyzed network as a static entity
- But network is *dynamic*
  - ◊ Different sections of the network are active under different cellular conditions
- Integrate gene expression data

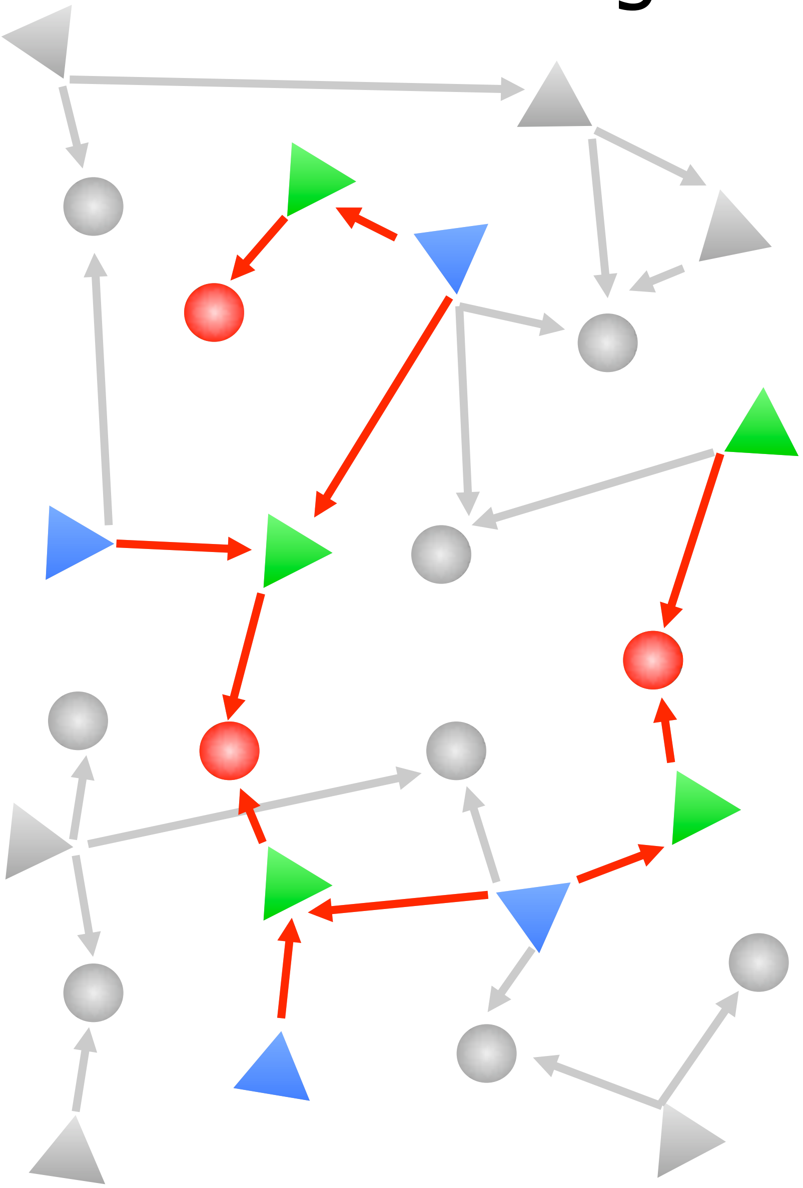
Luscombe et al. Nature 431: 308

Gene expression data for five cellular conditions in yeast

	Cellular condition
Multi-stage	Cell cycle
	Sporulation
Binary	Diauxic shift
	DNA damage
	Stress response

[Brown, Botstein, Davis….]

Backtracking to find active sub-network

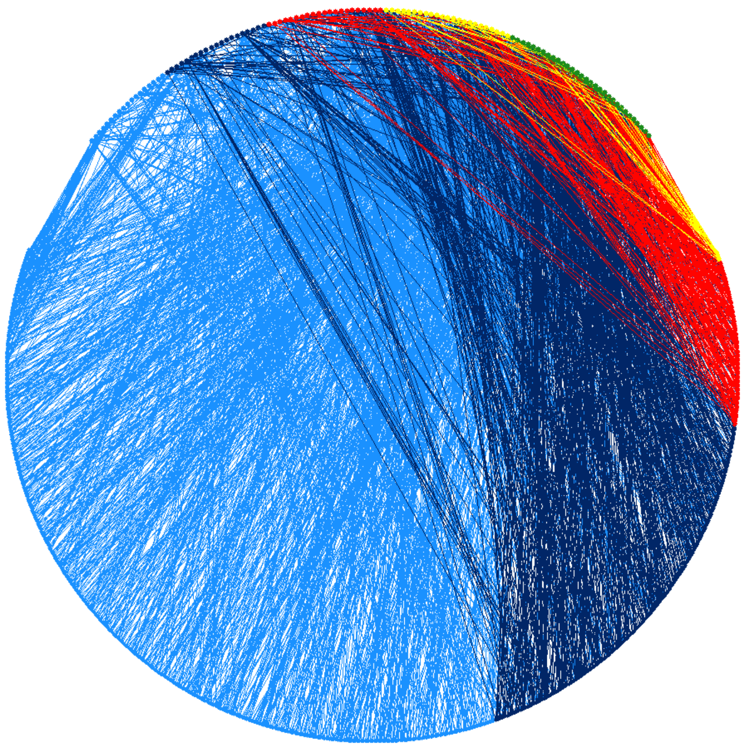


- Define differentially expressed genes
- Identify TFs that regulate these genes
- Identify further TFs that regulate these TFs

Active regulatory sub-network

Network usage under different conditions

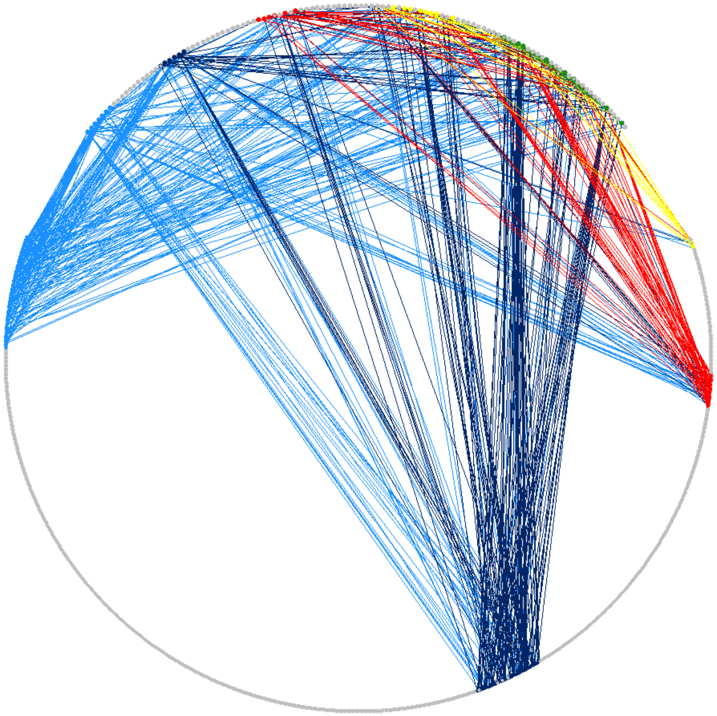
static



Luscombe et al. Nature 431: 308

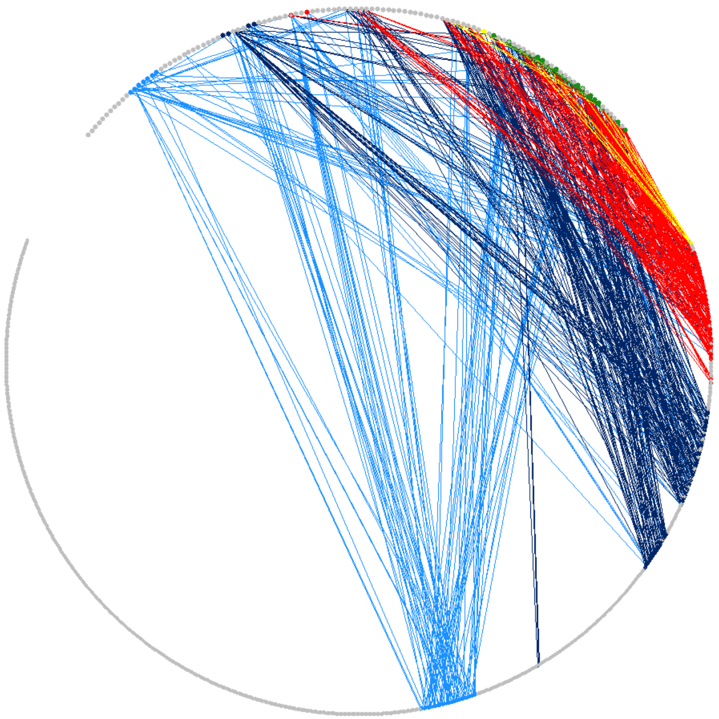
Network usage under different conditions

cell cycle



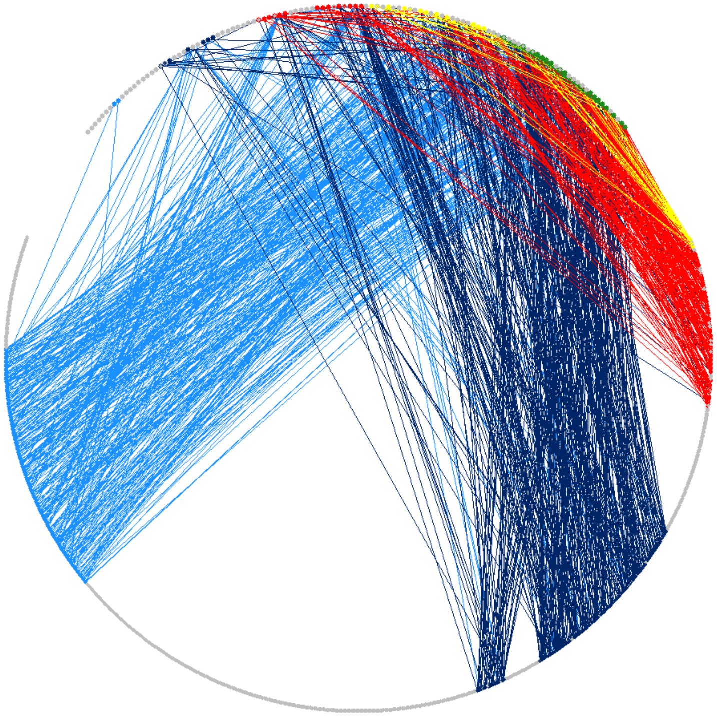
Network usage under different conditions

sporulation



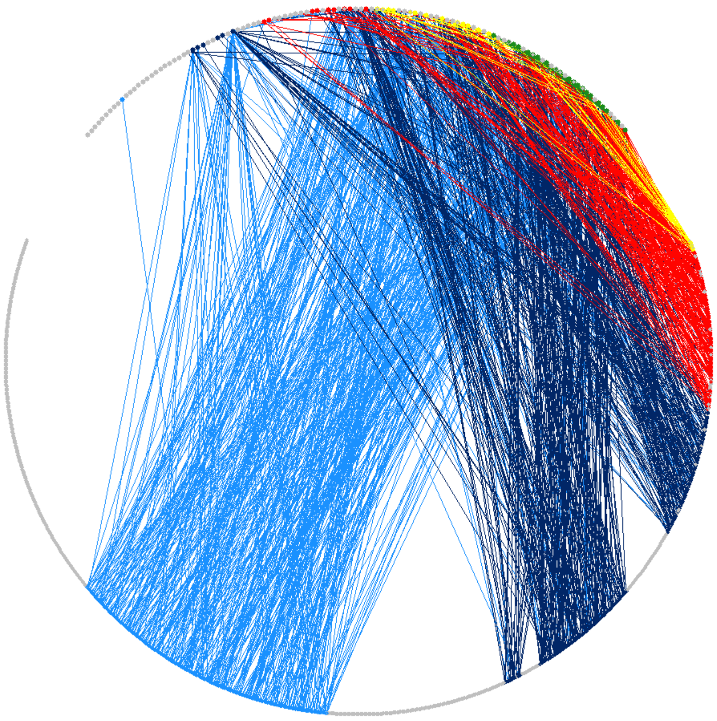
Network usage under different conditions

diauxic shift



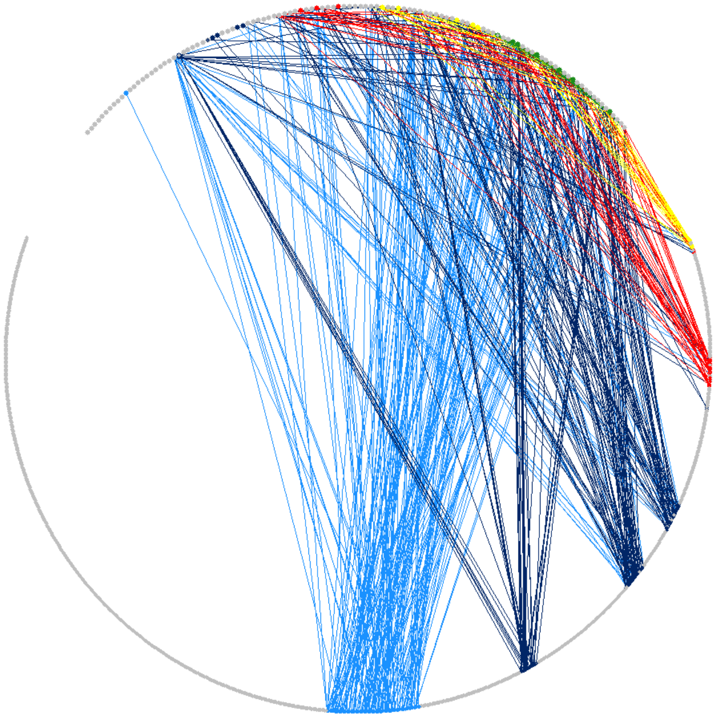
Network usage under different conditions

DNA damage

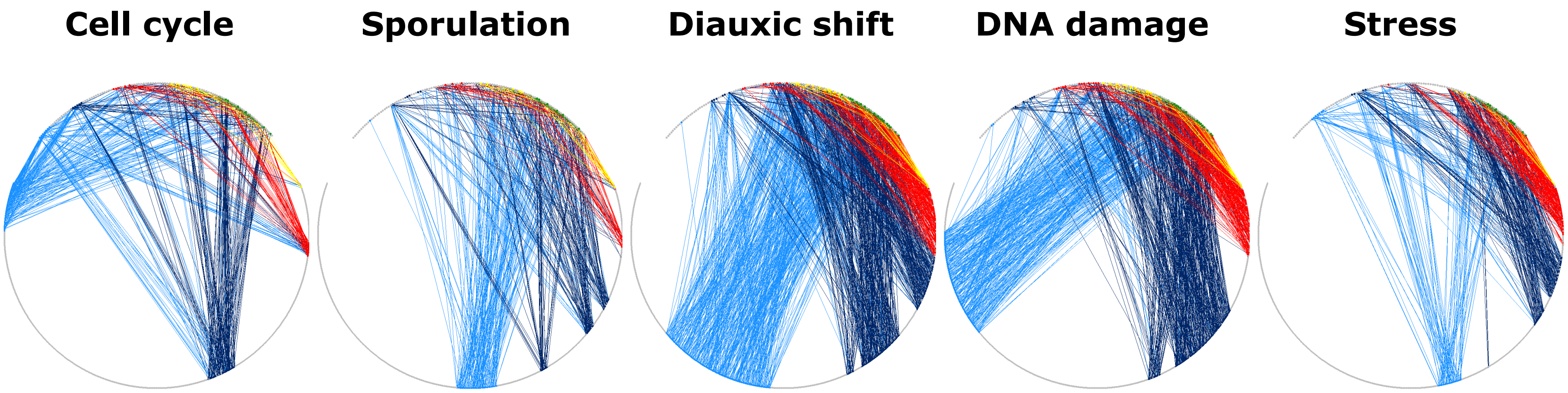


Network usage under different conditions

stress response



Network usage under different conditions



Cell cycle

Sporulation

Diauxic shift

DNA damage

Stress

SANDY:

1. Standard graph-theoretic statistics:

- Global topological measures
- Local network motifs

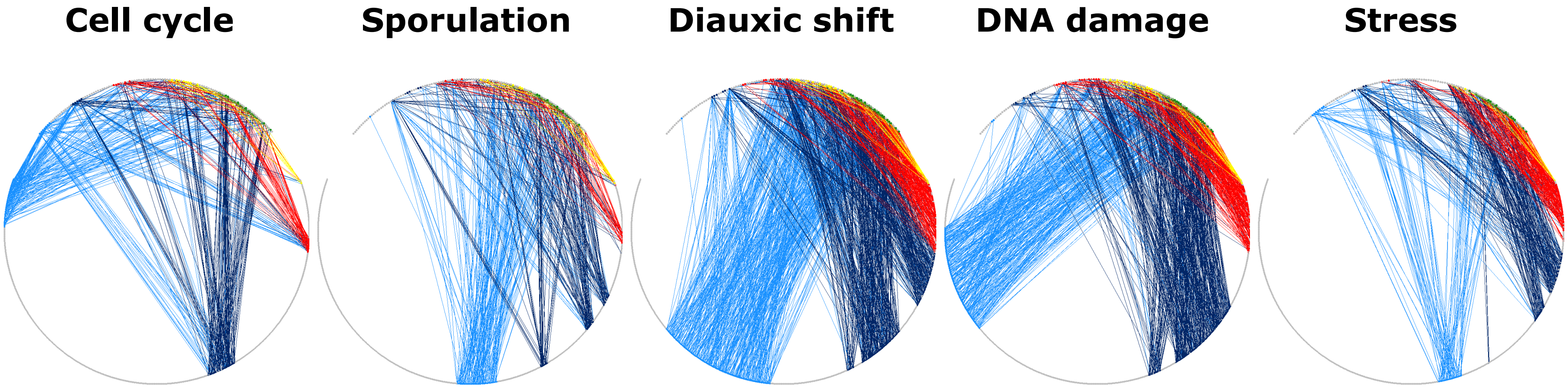
2. Newly derived follow-on statistics:

- Hub usage
- Interaction rewiring

3. Statistical validation of results

Luscombe et al. Nature 431: 308

Network usage under different conditions



Cell cycle Sporulation Diauxic shift DNA damage Stress

SANDY:

1. Standard graph-theoretic statistics:

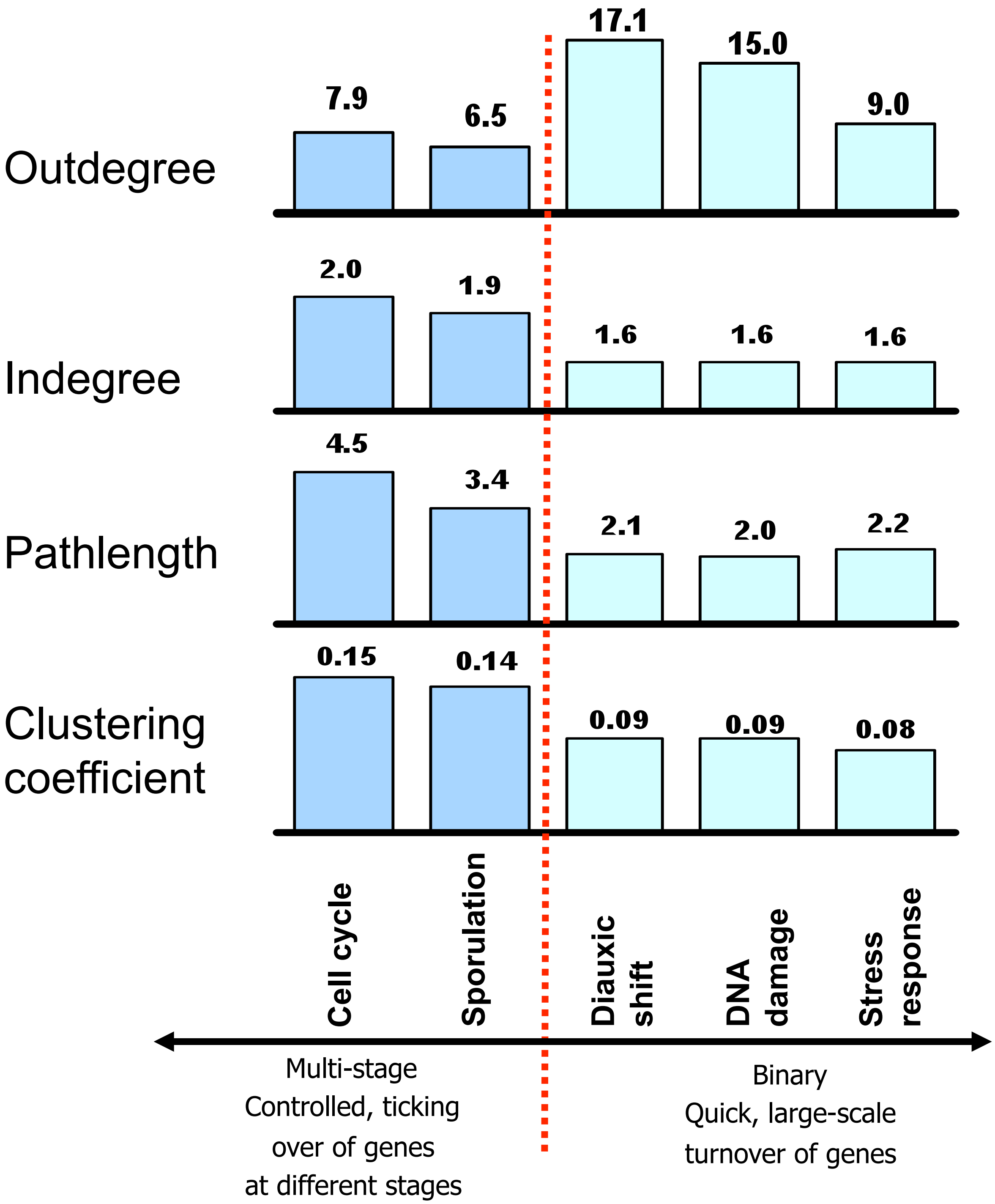
- Global topological measures
- Local network motifs

2. Newly derived follow-on statistics:

- Hub usage
- Interaction rewiring

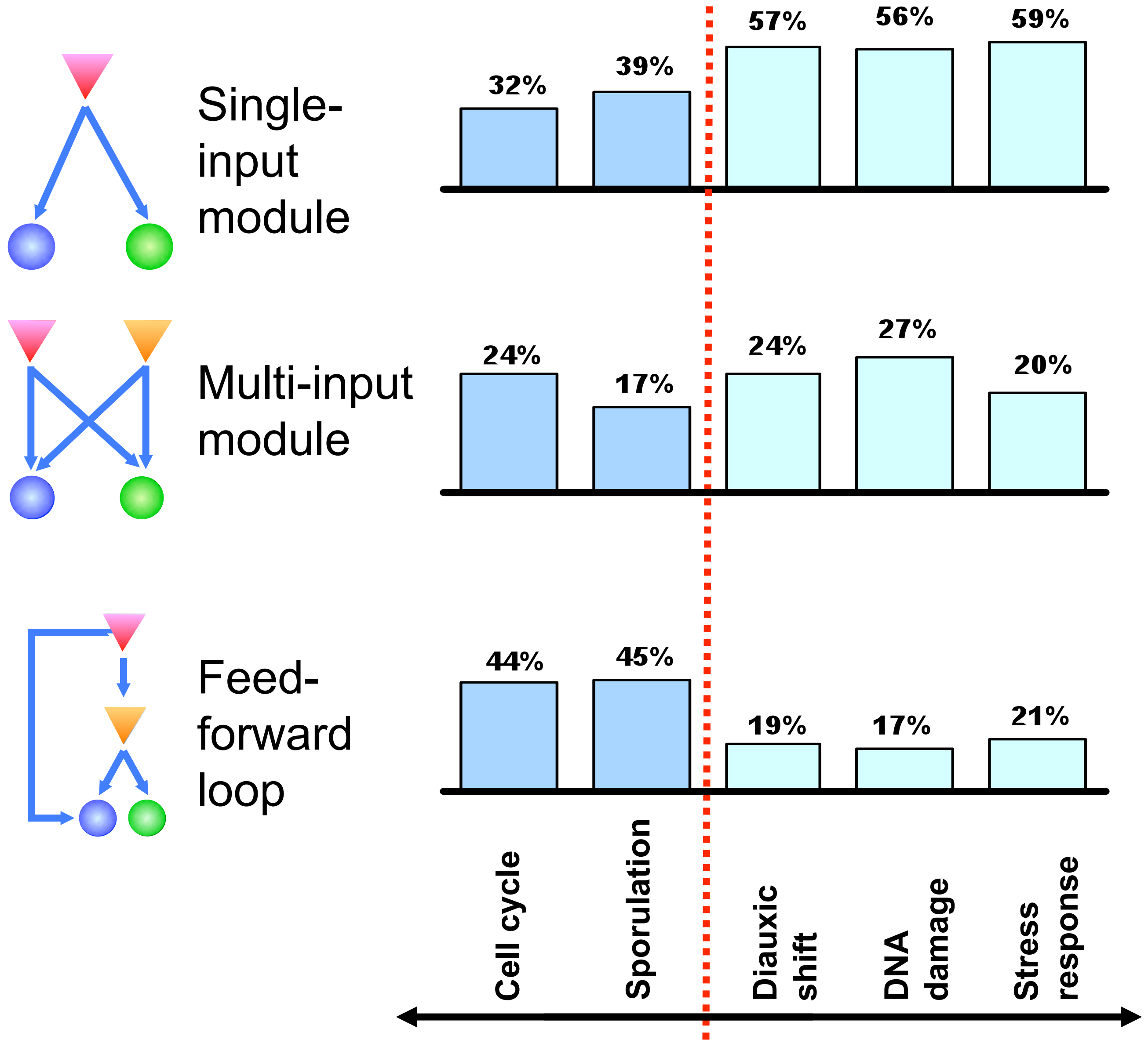
3. Statistical validation of results

Analysis of condition-specific subnetworks in terms of global topological statistics



Luscombe et al. Nature 431: 308

Analysis of condition-specific subnetworks in terms of occurrence of local motifs

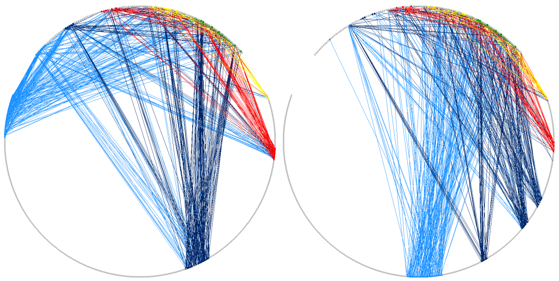


Multi-stage
Controlled, ticking
over of genes
at different stages

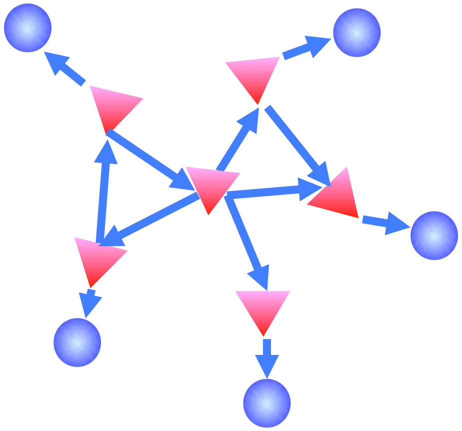
Binary
Quick, large-scale
turnover of genes

Luscombe et al. Nature 431: 308

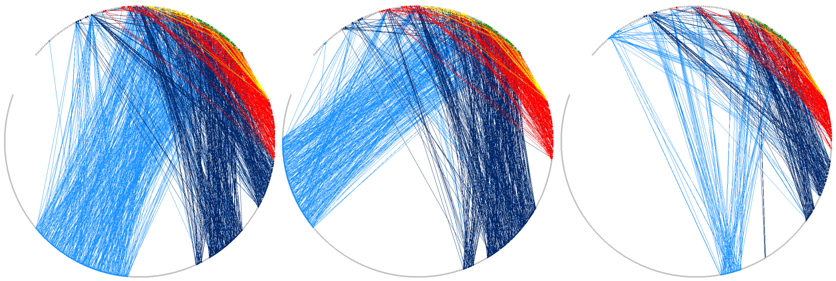
Cell cycle **Sporulation**



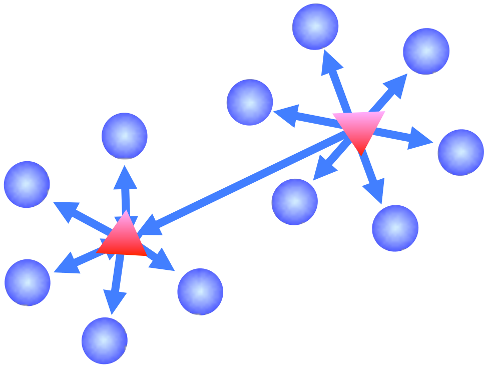
multi-stage conditions



Diauxic shift **DNA damage** **Stress**



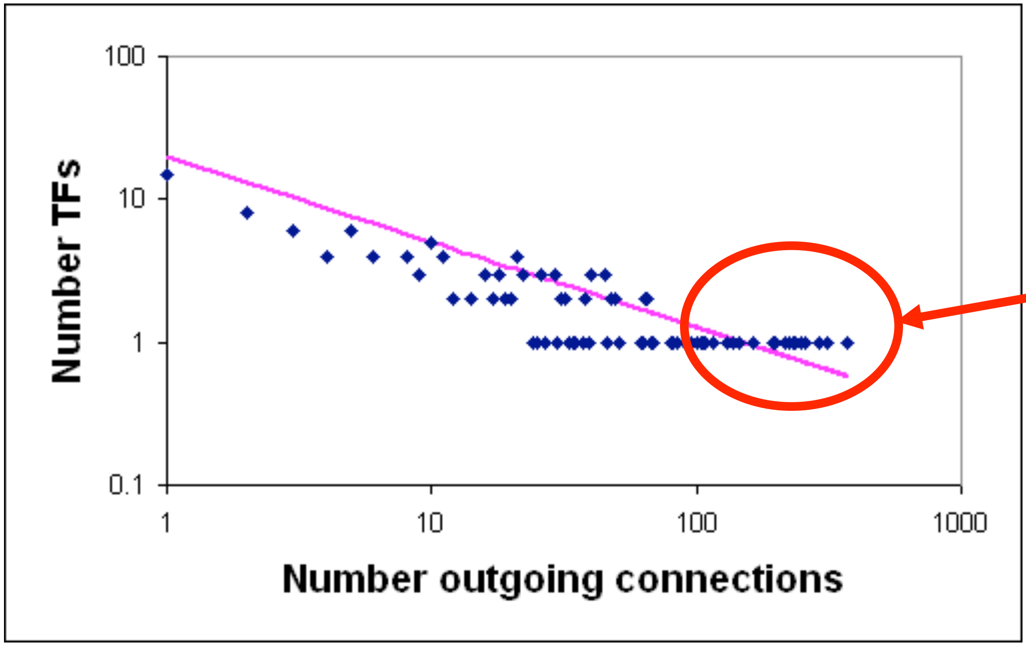
binary conditions

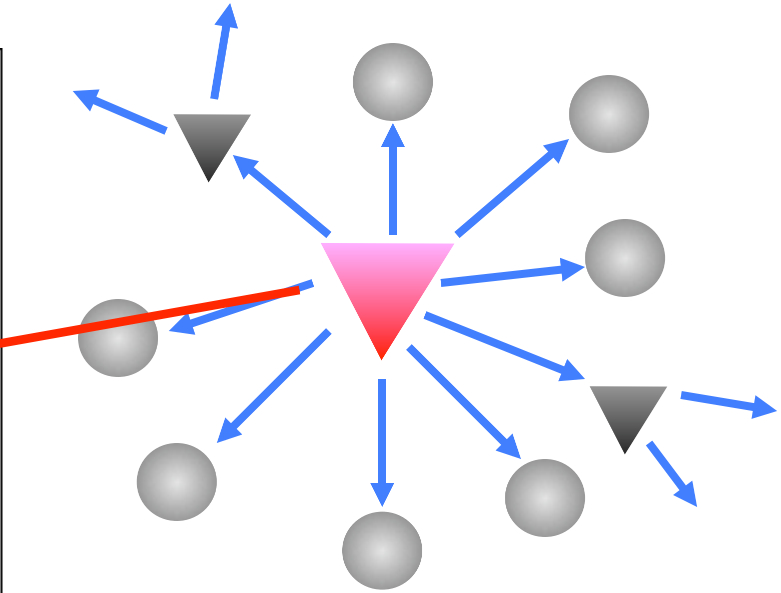


Summary

less pronounced	Hubs	more pronounced
longer	Path Lengths	shorter
more	TF inter-regulation	less
complex loops (FFLs)	Motifs	simpler (SIMs)

Transient Hubs

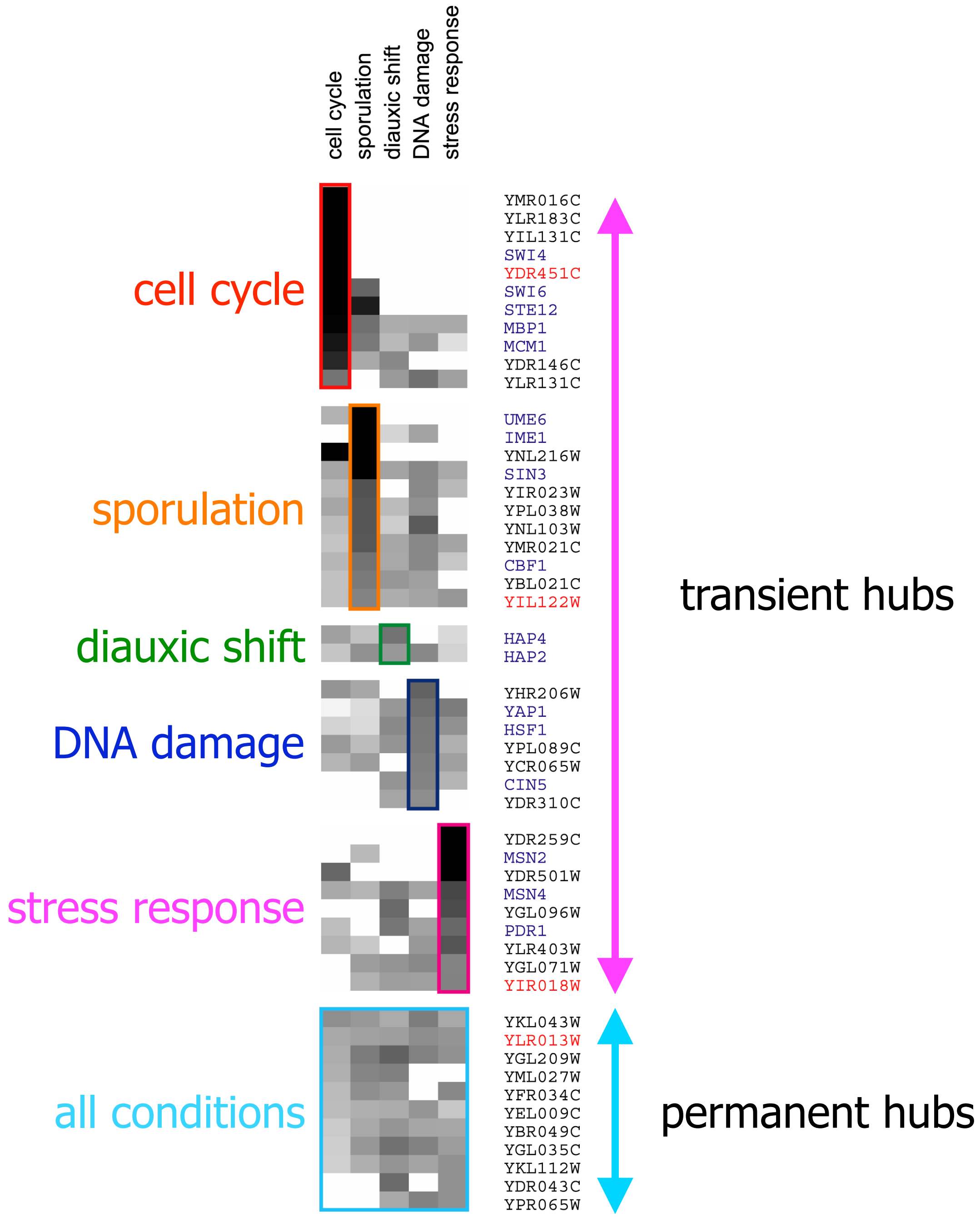




Regulatory hubs

- Questions:
  - Do hubs stay the same or do they change over between conditions?
  - Do different TFs become important?
- Our Expectations
  - Literature:
    - Hubs are permanent features of the network regardless of condition
  - Random networks (sampled from complete regulatory network)
    - Random networks converge on same TFs
    - 76-97% overlap in TFs classified as hubs (*ie* hubs are permanent)

Luscombe et al. Nature 431: 308

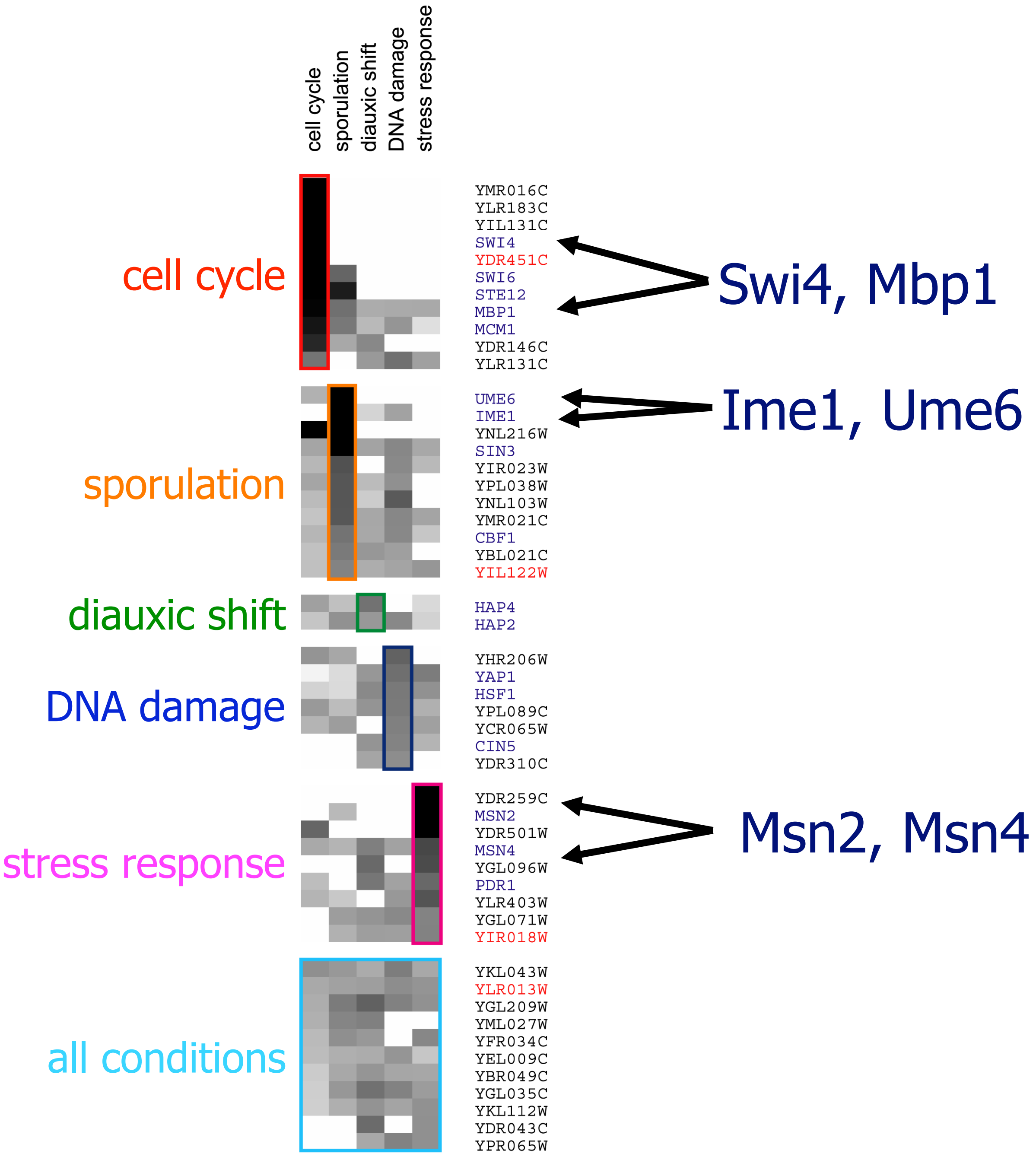


- Some permanent hubs
  - ◊ house-keeping functions

- Most are transient hubs
  - ◊ Different TFs become key regulators in the network

- Implications for condition-dependent vulnerability of network

Luscombe et al. Nature 431: 308

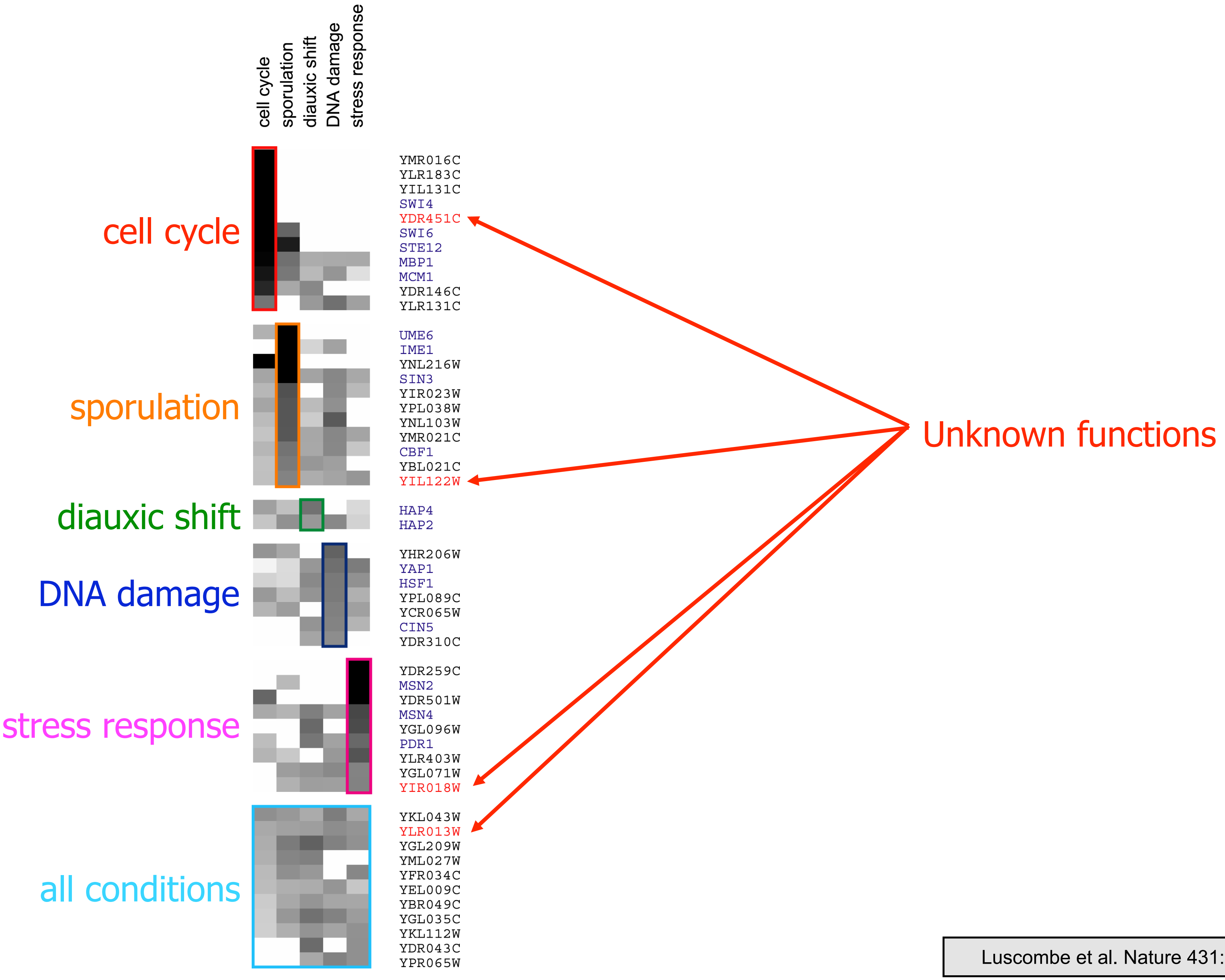


Swi4, Mbp1

Ime1, Ume6

Msn2, Msn4

Luscombe et al. Nature 431: 308



cell cycle
sporulation
diauxic shift
DNA damage
stress response

cell cycle

YMR016C
YLR183C
YIL131C
SWI4
YDR451C
SWI6
STE12
MBP1
MCM1
YDR146C
YLR131C

sporulation

UME6
IME1
YNL216W
SIN3
YIR023W
YPL038W
YNL103W
YMR021C
CBF1
YBL021C
YIL122W

diauxic shift

HAP4
HAP2

DNA damage

YHR206W
YAP1
HSF1
YPL089C
YCR065W
CIN5
YDR310C

stress response

YDR259C
MSN2
YDR501W
MSN4
YGL096W
PDR1
YLR403W
YGL071W
YIR018W

all conditions

YKL043W
YLR013W
YGL209W
YML027W
YFR034C
YEL009C
YBR049C
YGL035C
YKL112W
YDR043C
YPR065W

Unknown functions

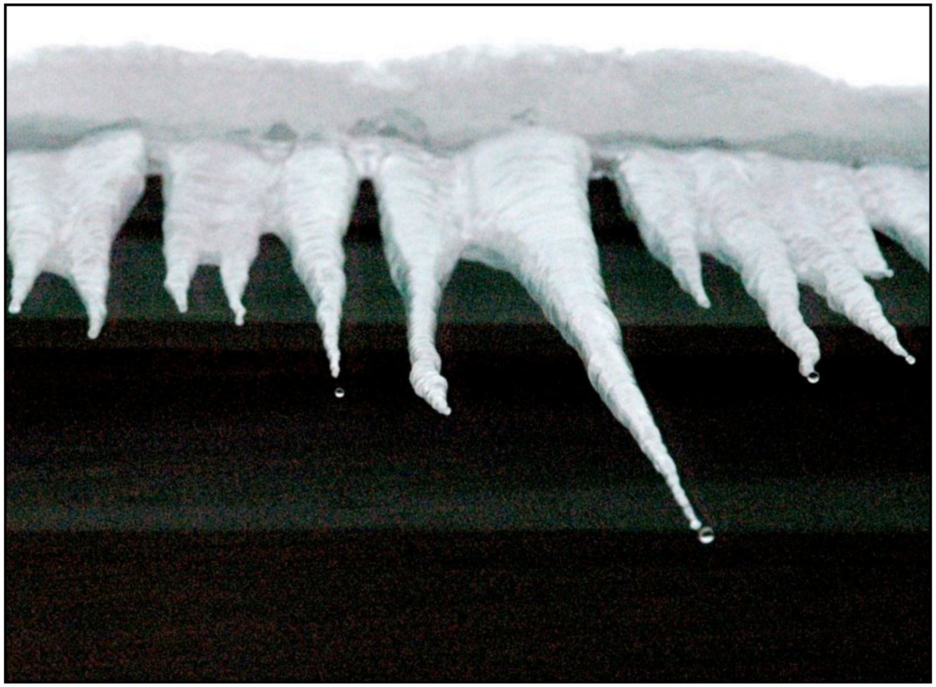
Luscombe et al. Nature 431: 308

Network Dynamics #2: Environments

How do molecular networks change across environments?

What pathways are used more ?

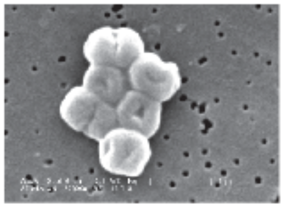
Used as a biosensor ?



What is metagenomics?

Genomics Approach

Culture Microbes

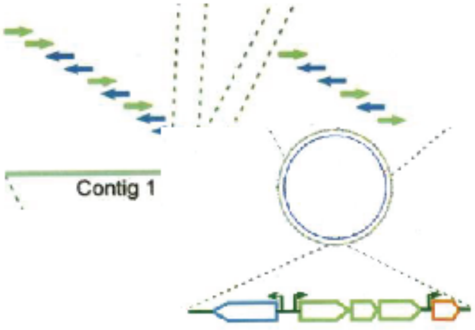


Extract DNA

Sequence

ATCGTATA
CGCGAAG
ACGTCTGA
AGTGCTGCT

Assemble and Annotate



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

Metagenomics Approach

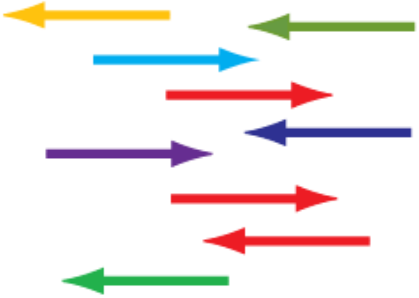
Collect Sample

Extract DNA

Sequence

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

Partially Assemble and Annotate



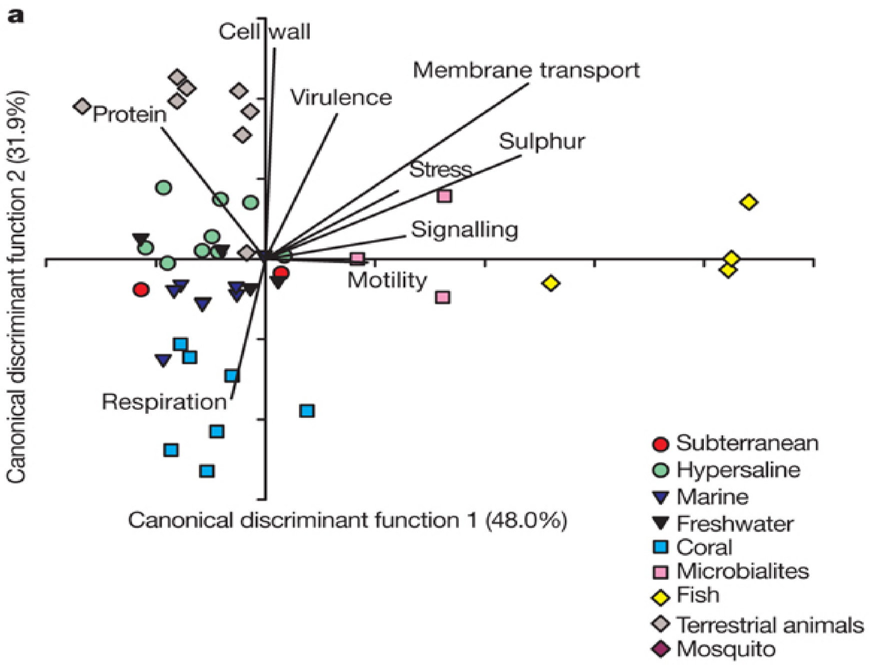
PROBLEM: Lose information about which gene belongs to which microbe.

Comparative Metagenomics

Water

Soil

Do the proportions of pathways represented in these two samples differ?



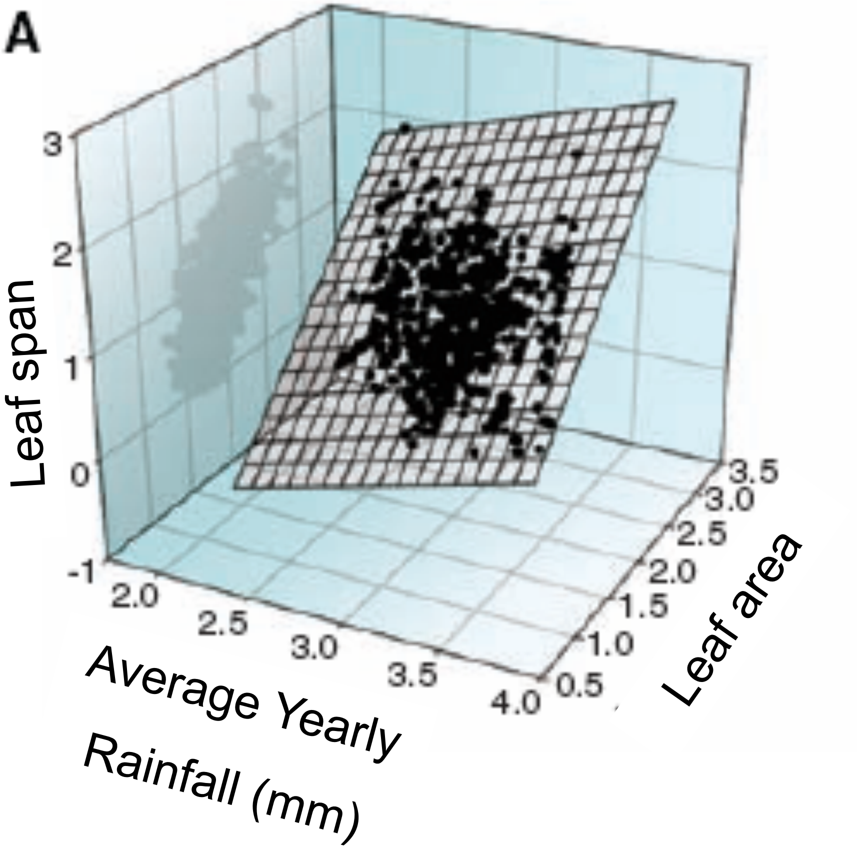
Dinsdale et. al., Nature 2008

Trait-based Biogeography

Charles River,
MA

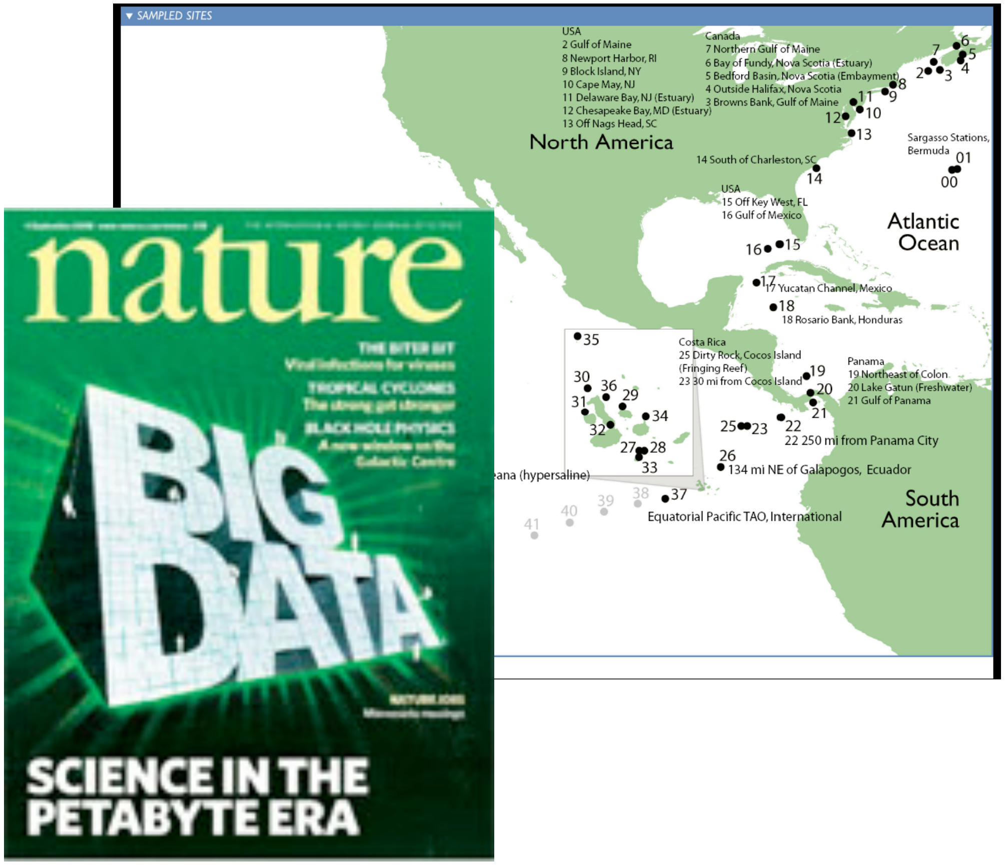
Long Island
Sound, CT

Do the proportions of pathways represented in these two samples CHANGE as a function of their environments?



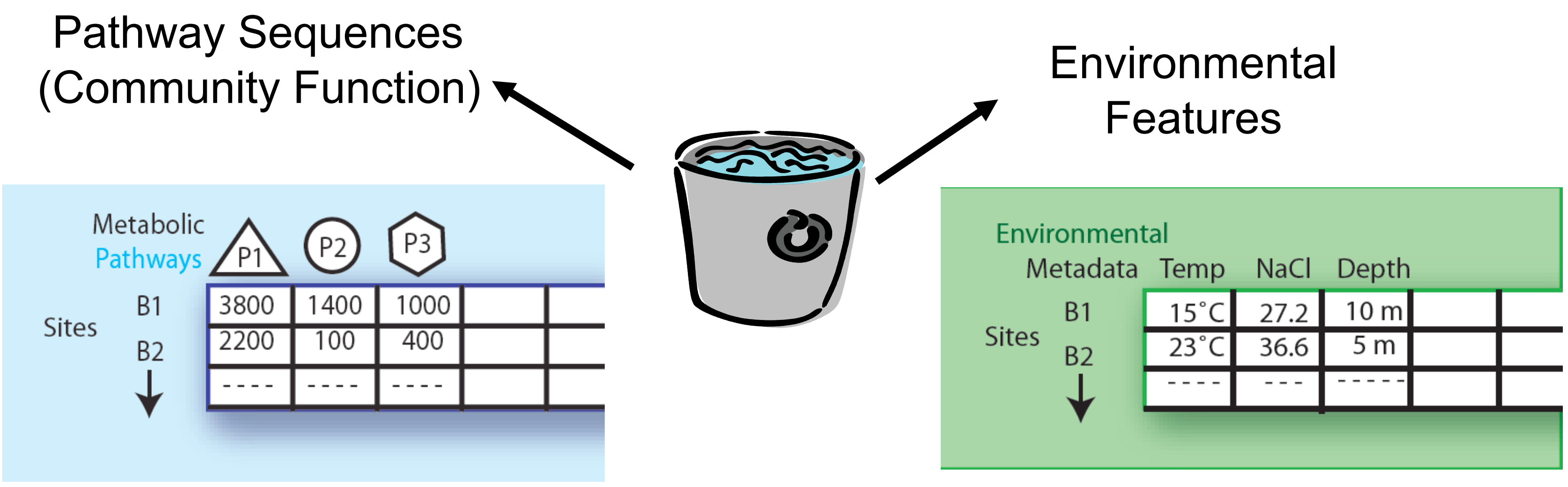
Green et. al., Science 2008

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

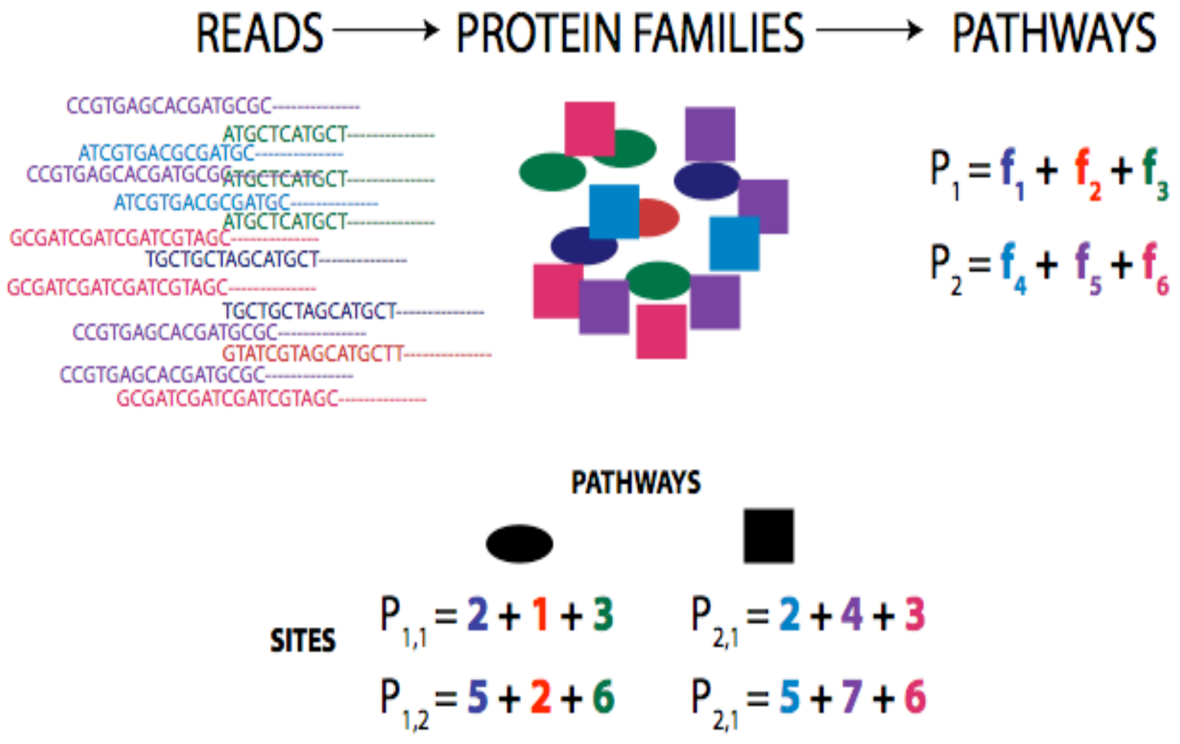
6.25 GB of data
7.7M Reads
1 million CPU hours
to process

Rusch, et al., PLOS Biology 2007



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

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



$P_1 = f_1 + f_2 + f_3$

$P_2 = f_4 + f_5 + f_6$

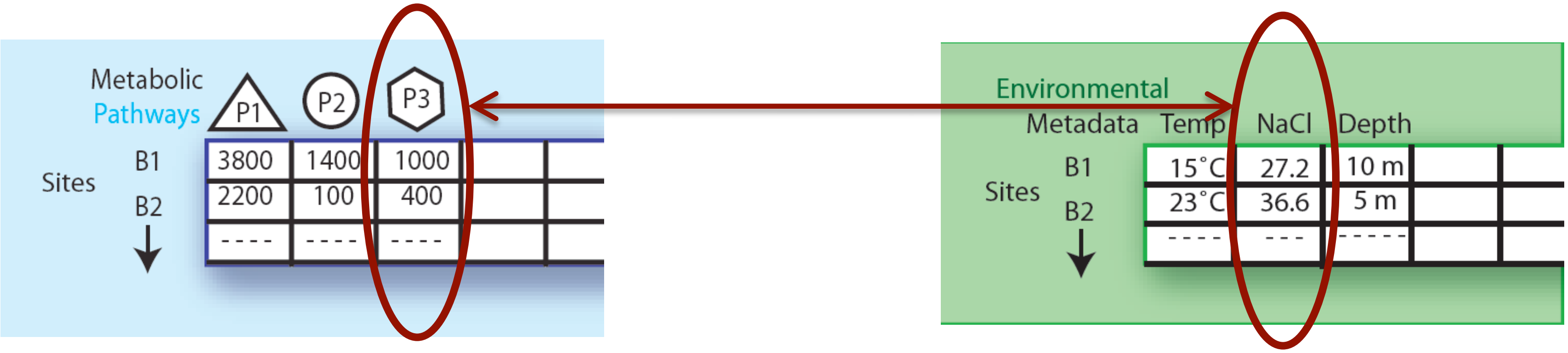
$P_{1,1} = 2 + 1 + 3$ $P_{2,1} = 2 + 4 + 3$

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

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

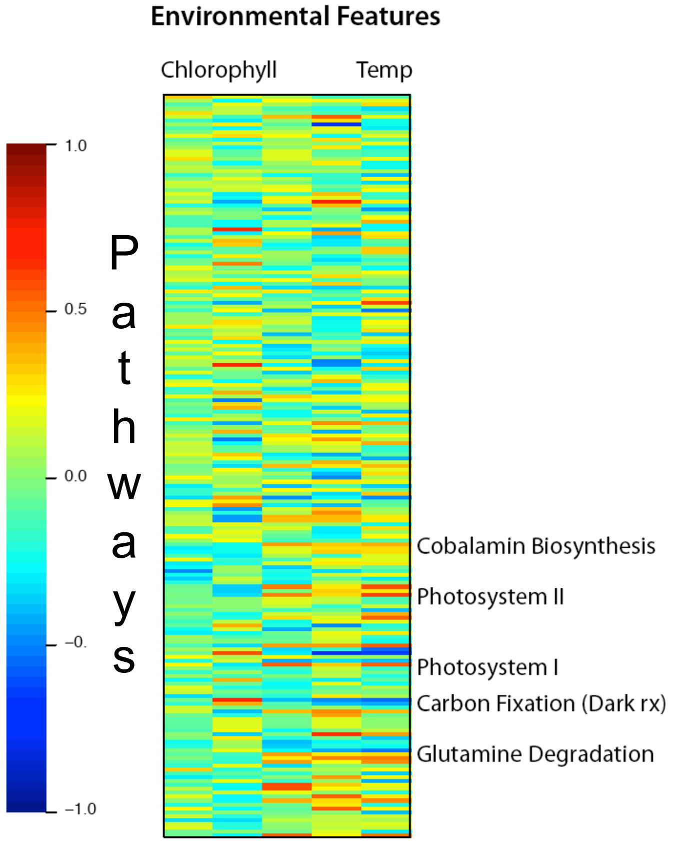
[Rusch et. al., (2007) PLOS Biology;
Gianoulis et al., PNAS (in press, 2009)]

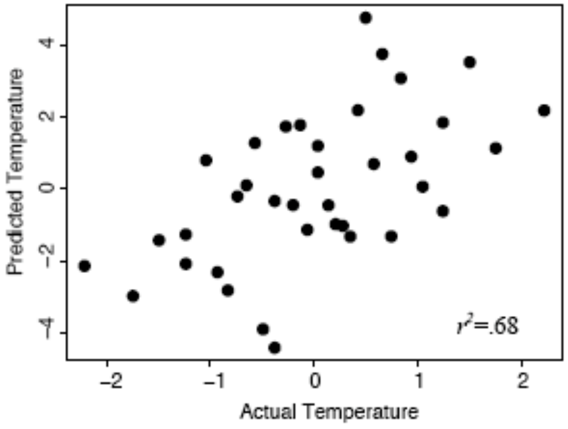
Simple Relationships: Pairwise Correlations



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

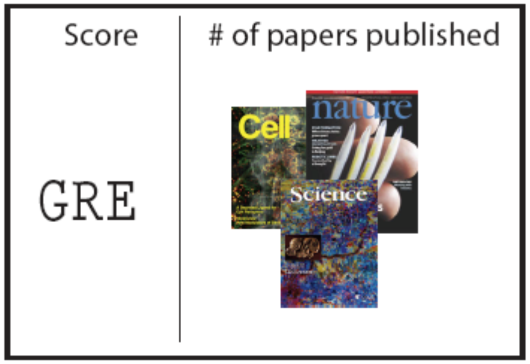
	Environmental Metadata	Temp	NaCl	Depth		
Sites	B1	15˚C	27.2	10 m		
	B2	23˚C	36.6	5 m		
		----	---	-----		

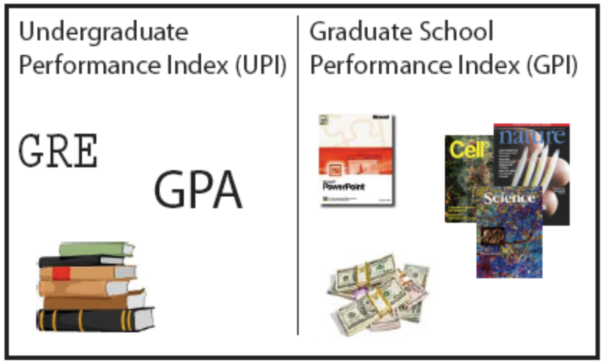




[Gianoulis et al., PNAS (in press, 2009)]

Canonical Correlation Analysis: Simultaneous weighting



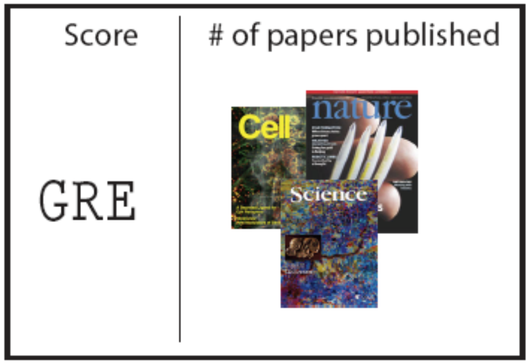

Undergraduate Performance Index (UPI)
Graduate School Performance Index (GPI)
GRE
GPA

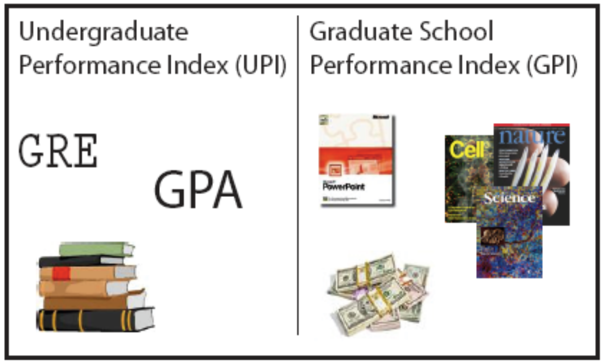
UPI = a GRE + b + c GPA

GPI = a' + b' + c'

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

Canonical Correlation Analysis: Simultaneous weighting

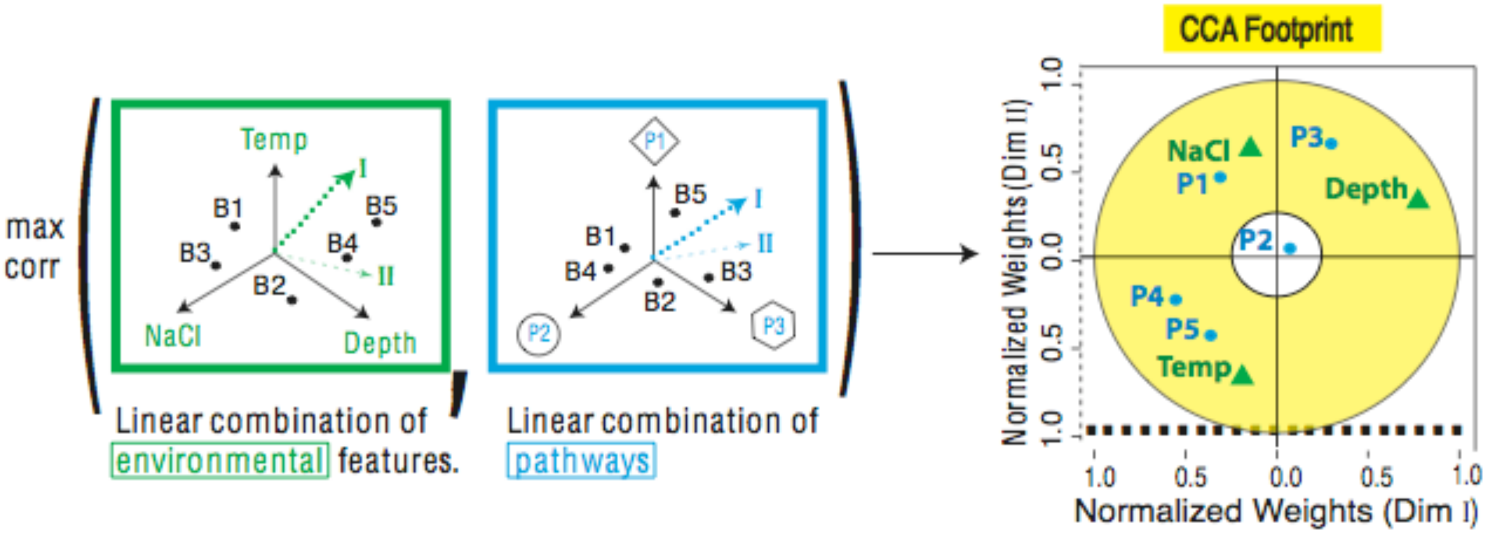



Undergraduate Performance Index (UPI)
Graduate School Performance Index (GPI)
GRE
GPA

Environmental Features	**Metabolic Pathways**
Temp	Photosynthesis
etc	etc
Chlorophyll	Lipid Metabolism

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

Environmental-Metabolic Space


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

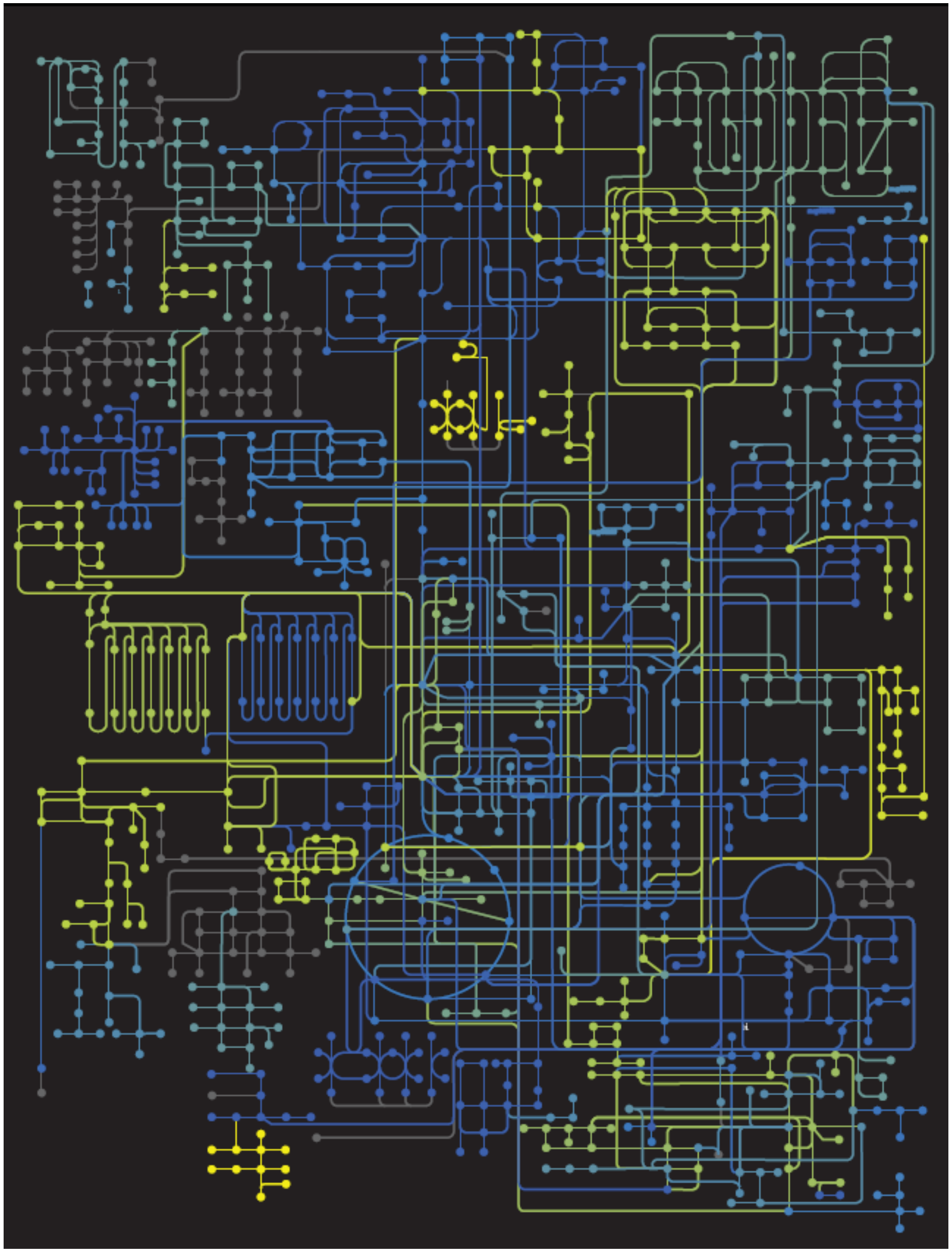
The goal of this technique is to interpret cross-variance matrices
We do this by defining a change of basis.

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

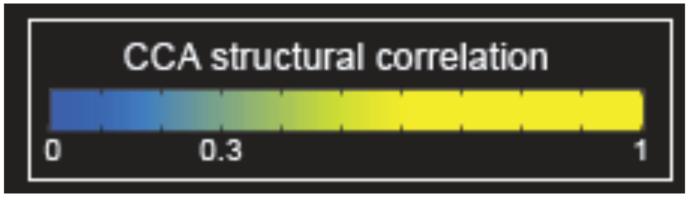
$$C = \begin{matrix} \Sigma_X & \Sigma_{X,Y} \\ \Sigma_Y & \Sigma_{Y,X} \end{matrix}$$

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

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

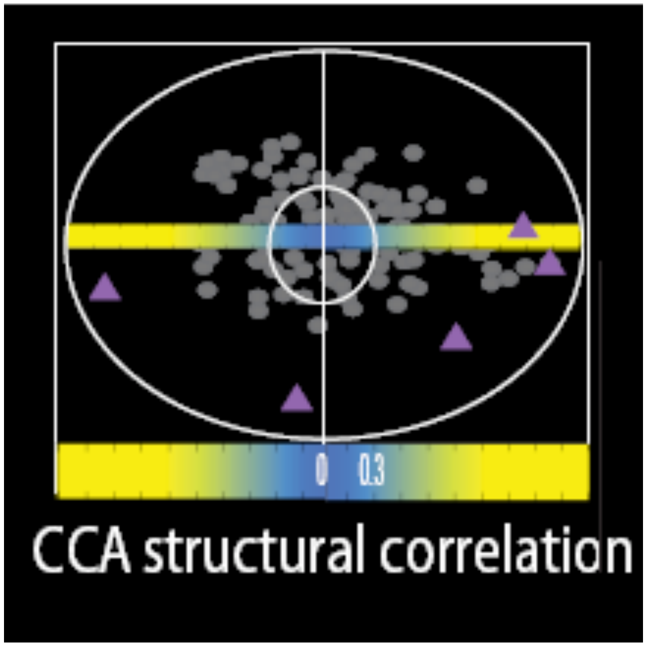


Strength of Pathway co-variation with environment



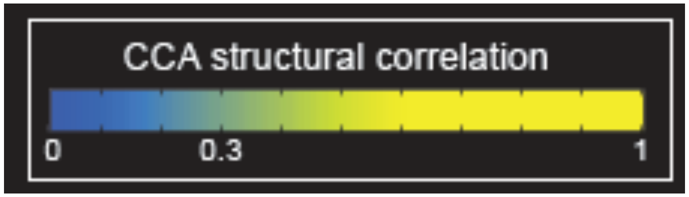
Environmentally invariant

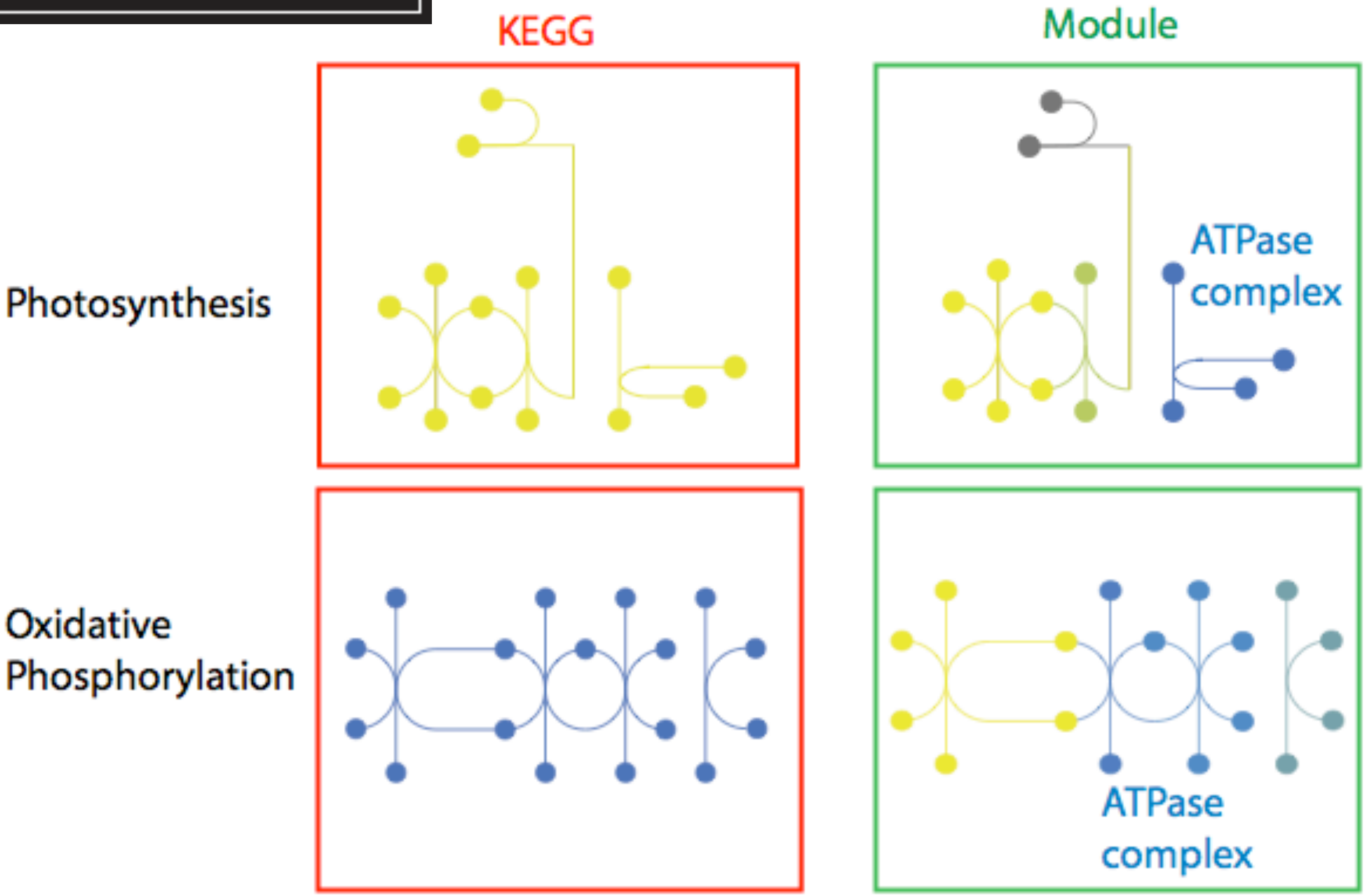
Environmentally variant



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

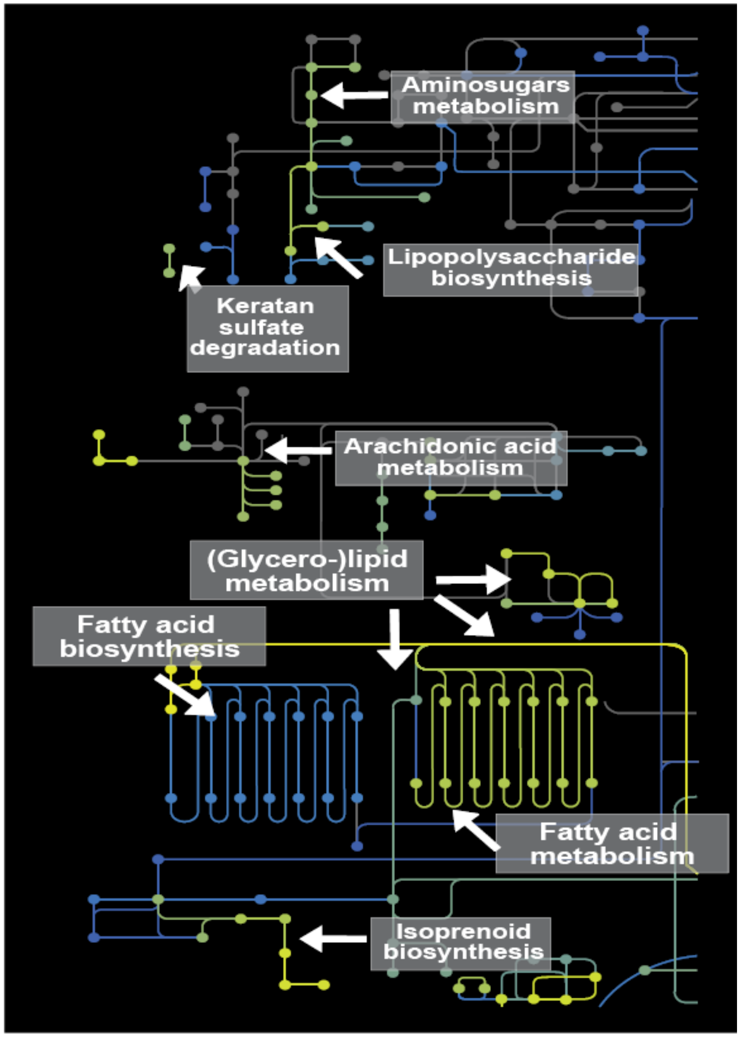
Conclusion #1: energy conversion strategy, temp and depth

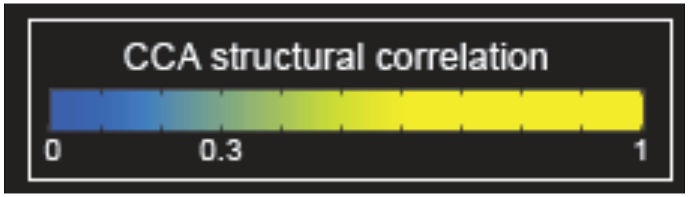




[Gianoulis et al., PNAS (in press, 2009)]

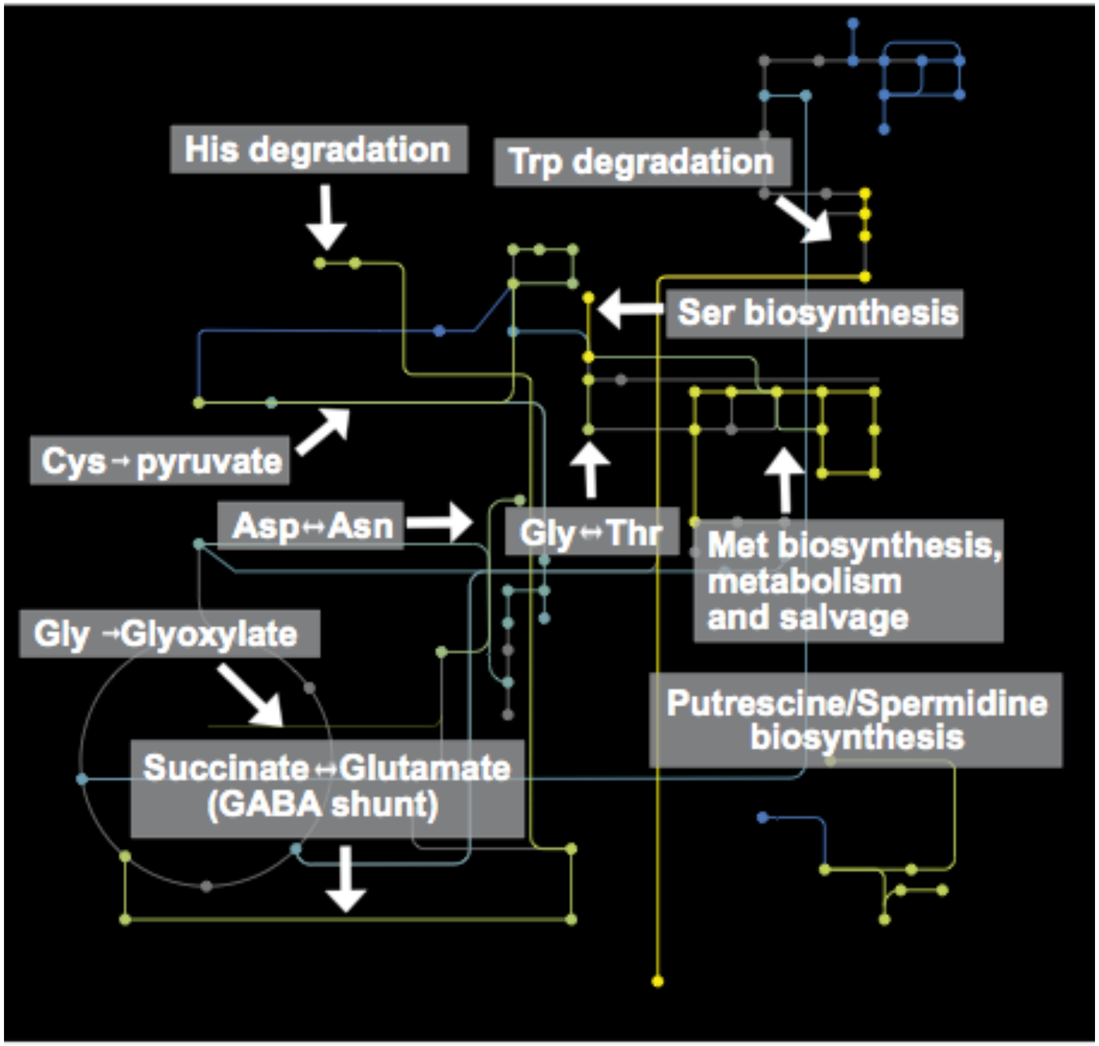
Conclusion #2: Outer Membrane components vary the environment


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)]

Conclusion #4: Cofactor (Metal) Optimization

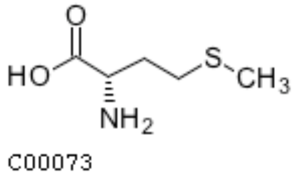
Methionine salvage, synthesis, and uptake, transport

IS DEPENDENT-ON

Methionine synthesis

Cobalamin biosynthesis

Cobalt transporters



Methionine

IS NEEDED FOR

Methionine degradation

S-adenosyl Methionine Biosynthesis (synthesize SAM one of the most important methyl donors)

Polyamine biosynthesis

RELIES ON

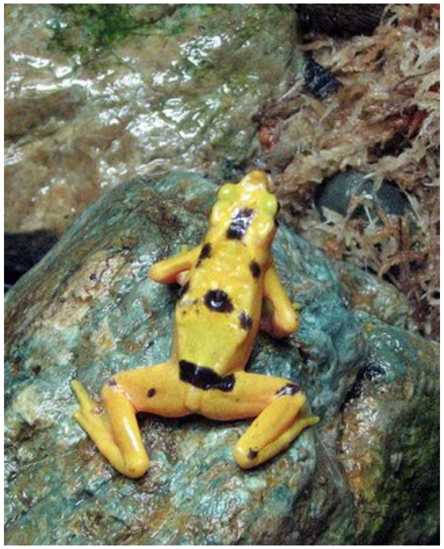
Methionine Salvage

Spermidine/Putrescine transporters

Arg/His/Ornithine transporters

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

Biosensors: Beyond Canaries in a Coal Mine



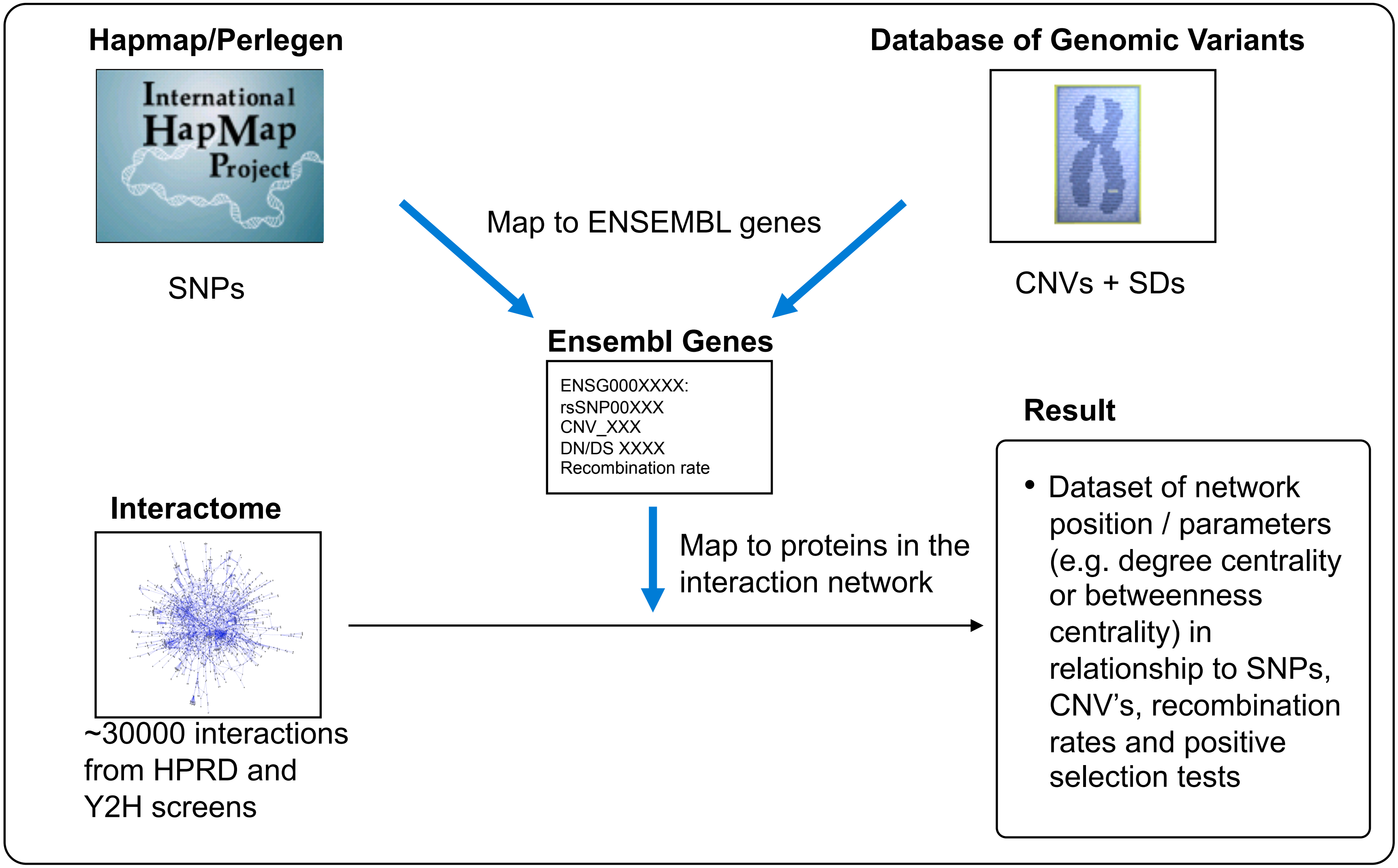
[Gianoulis et al., PNAS (in press, 2009)]

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

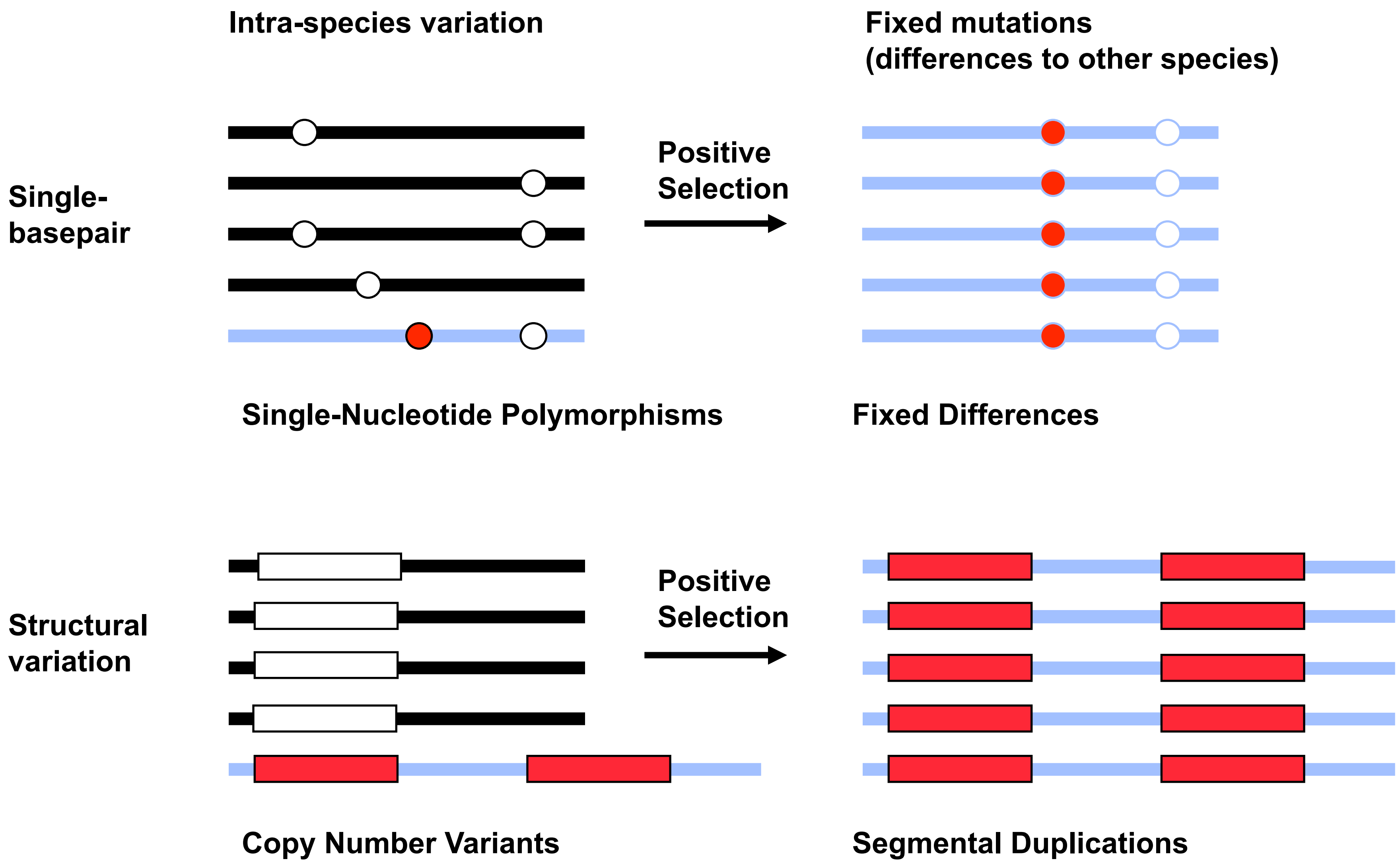
Result

- 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

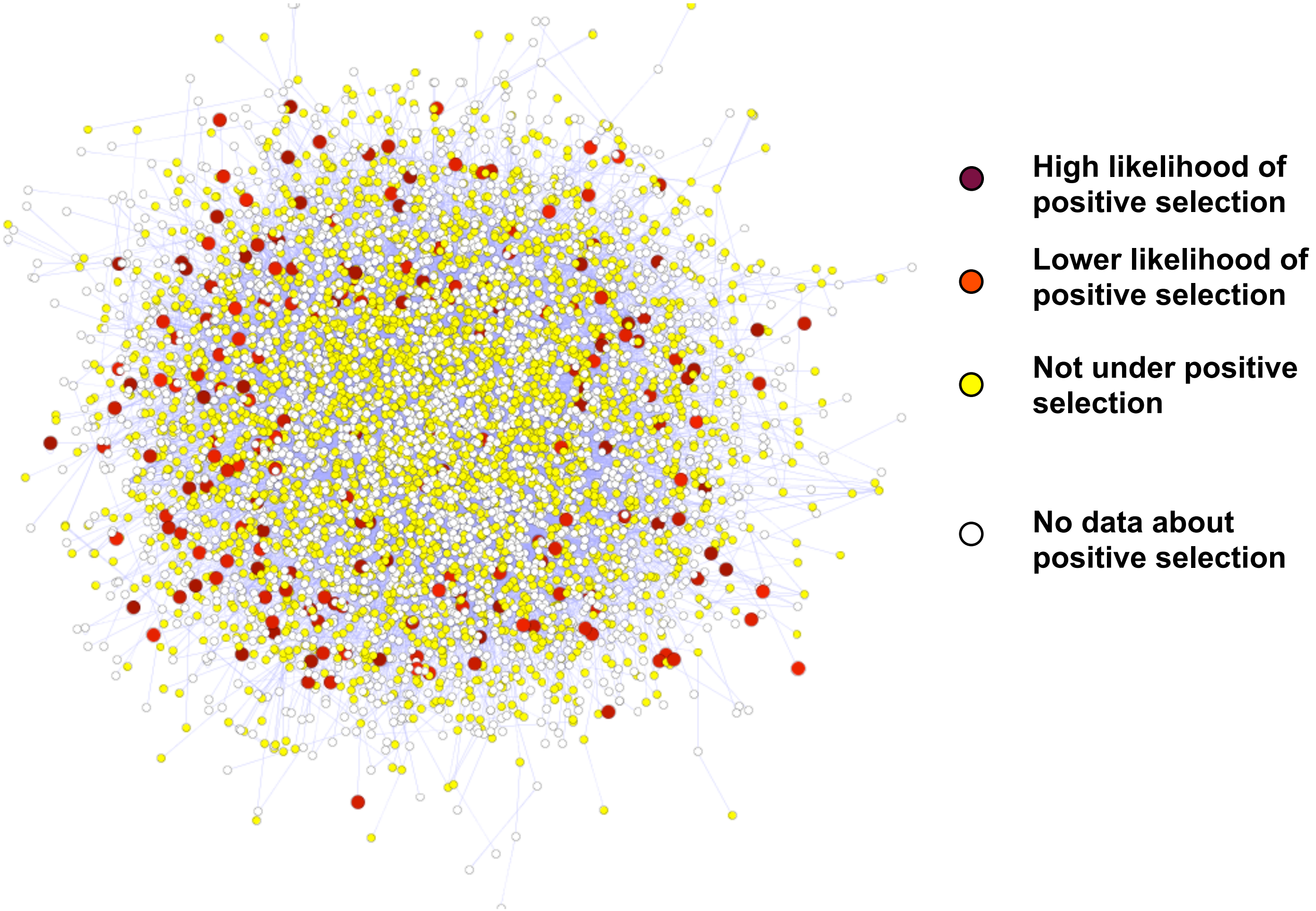
Copy Number Variants

Segmental Duplications

Source: PMK

POSITIVE SELECTION LARGELY TAKES PLACE AT THE NETWORK PERIPHERY

Positive selection in the human interactome

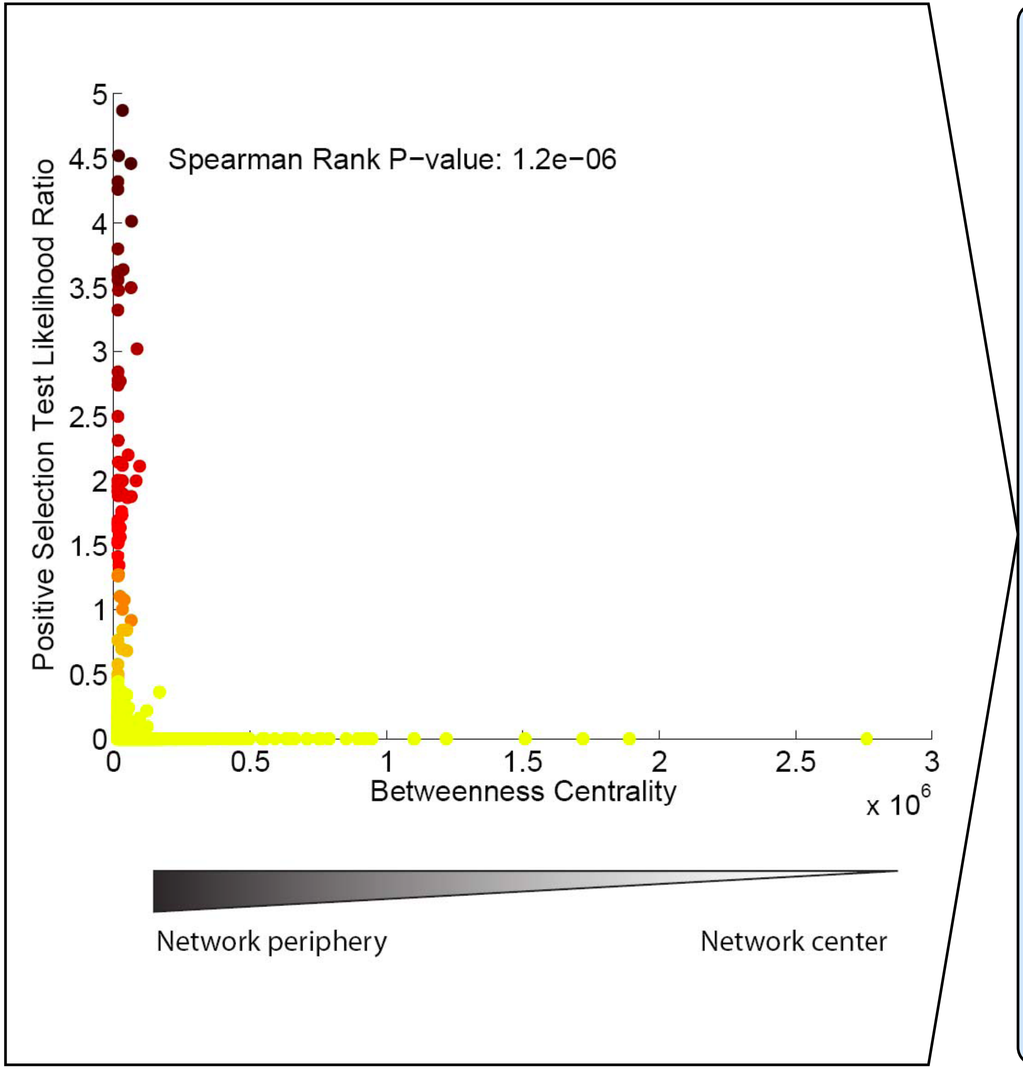

● High likelihood of positive selection
● Lower likelihood of positive selection
● Not under positive selection
○ No data about positive selection

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

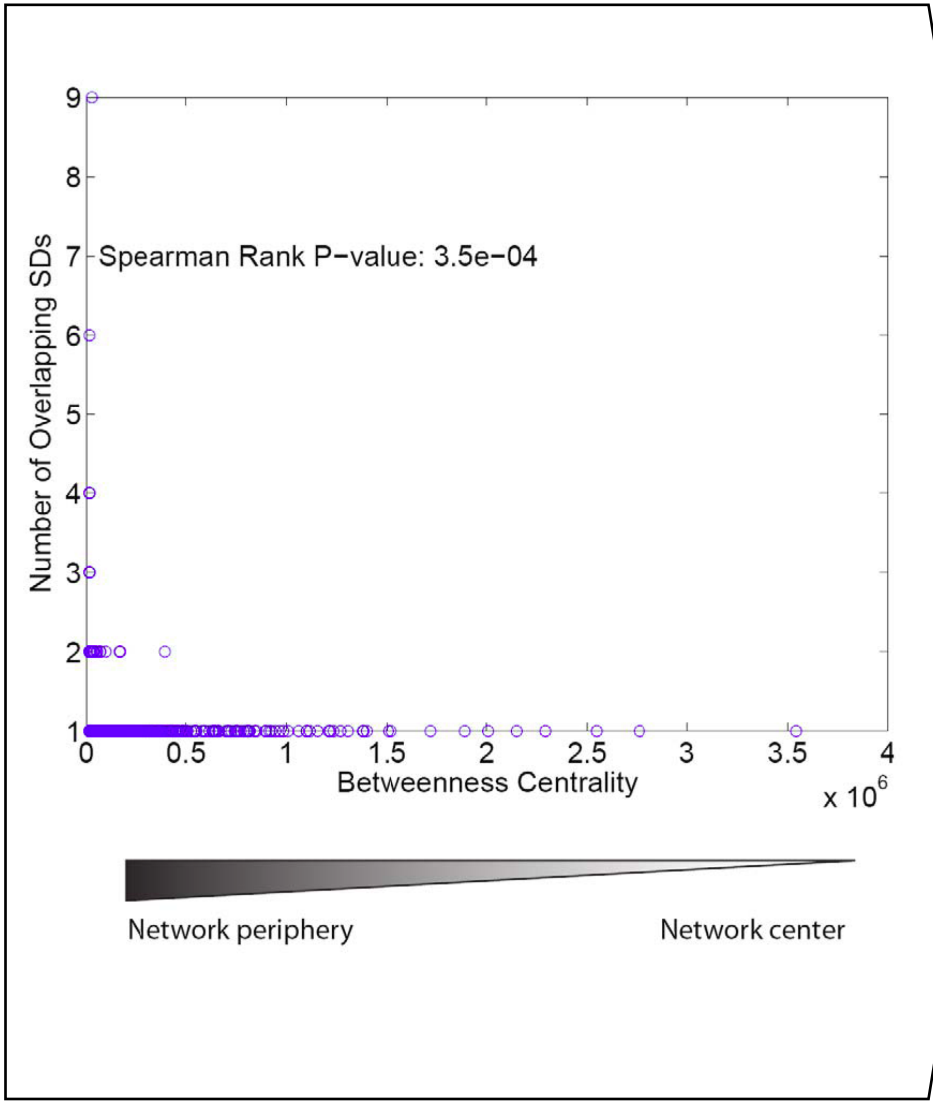
- Peripheral genes are likely to under positive selection, whereas hubs aren’t
- This is likely due to the following reasons:
  - Hubs have stronger structural constraints, the network periphery doesn’t
  - Most recently evolved functions (e.g. “environmental interaction genes” such as sensory perception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

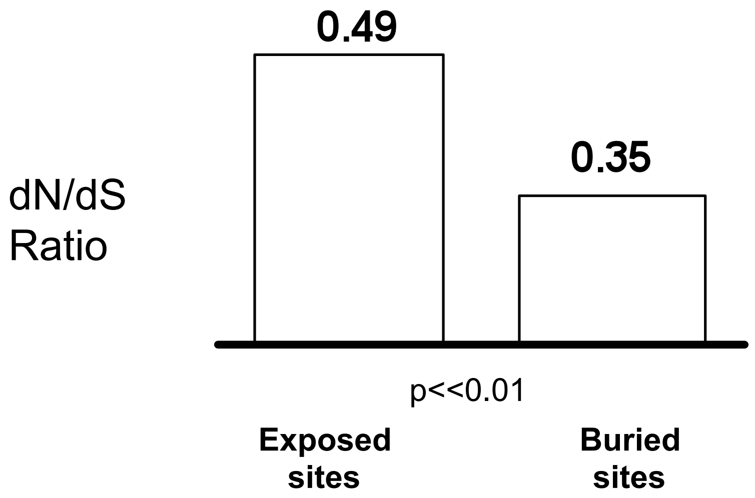
- This result also confirms our initial hypothesis – peripheral nodes tend to lie in regions rich in SDs.
- Since segmental duplications are a different mechanism of ongoing evolution, the less constrained peripheral proteins are enriched in them.
- Note that despite the small size of our dataset for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

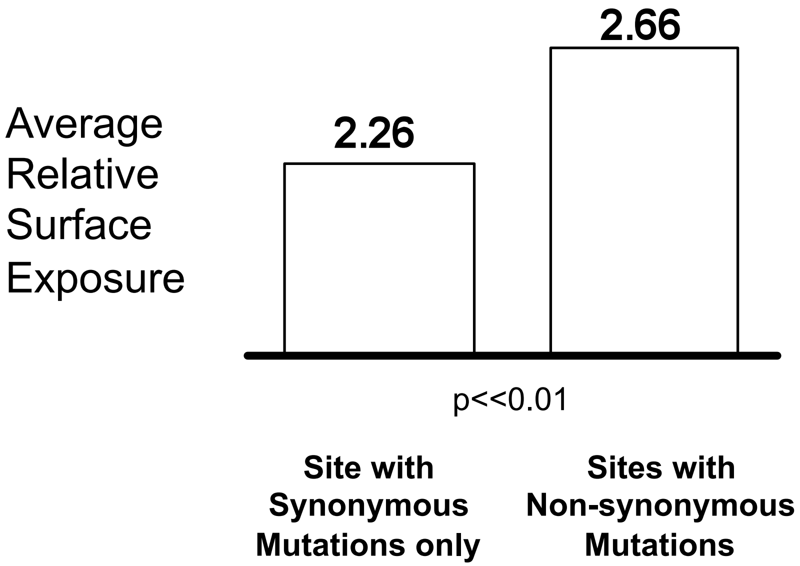
*Specifically, a number of the SDs are likely not fixed, but rather common CNVs in the reference genome

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.

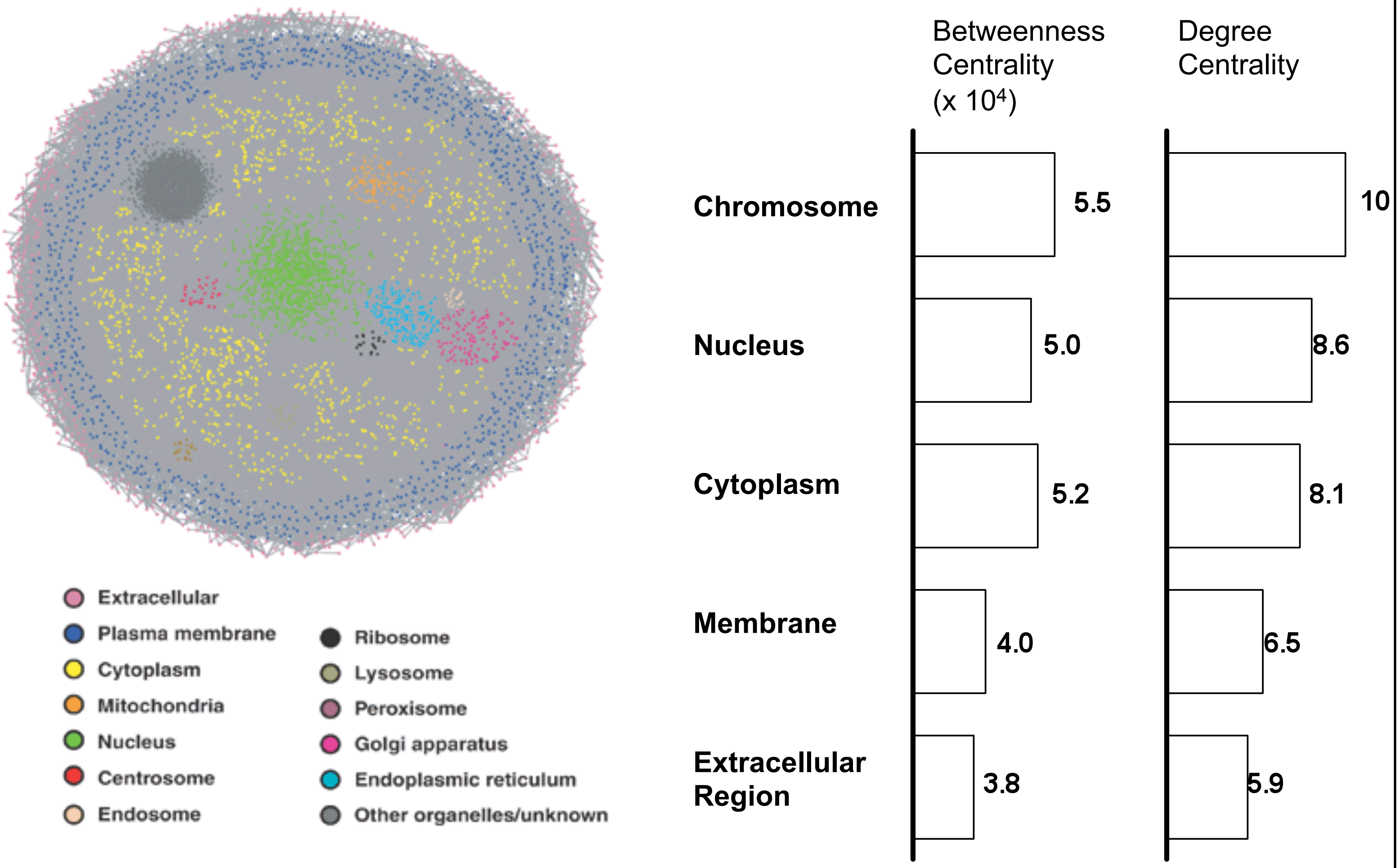
BURIED SITES ARE CONSERVED AND MUCH LESS LIKELY TO HARBOR NON-SYNONYMOUS MUTATIONS





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?

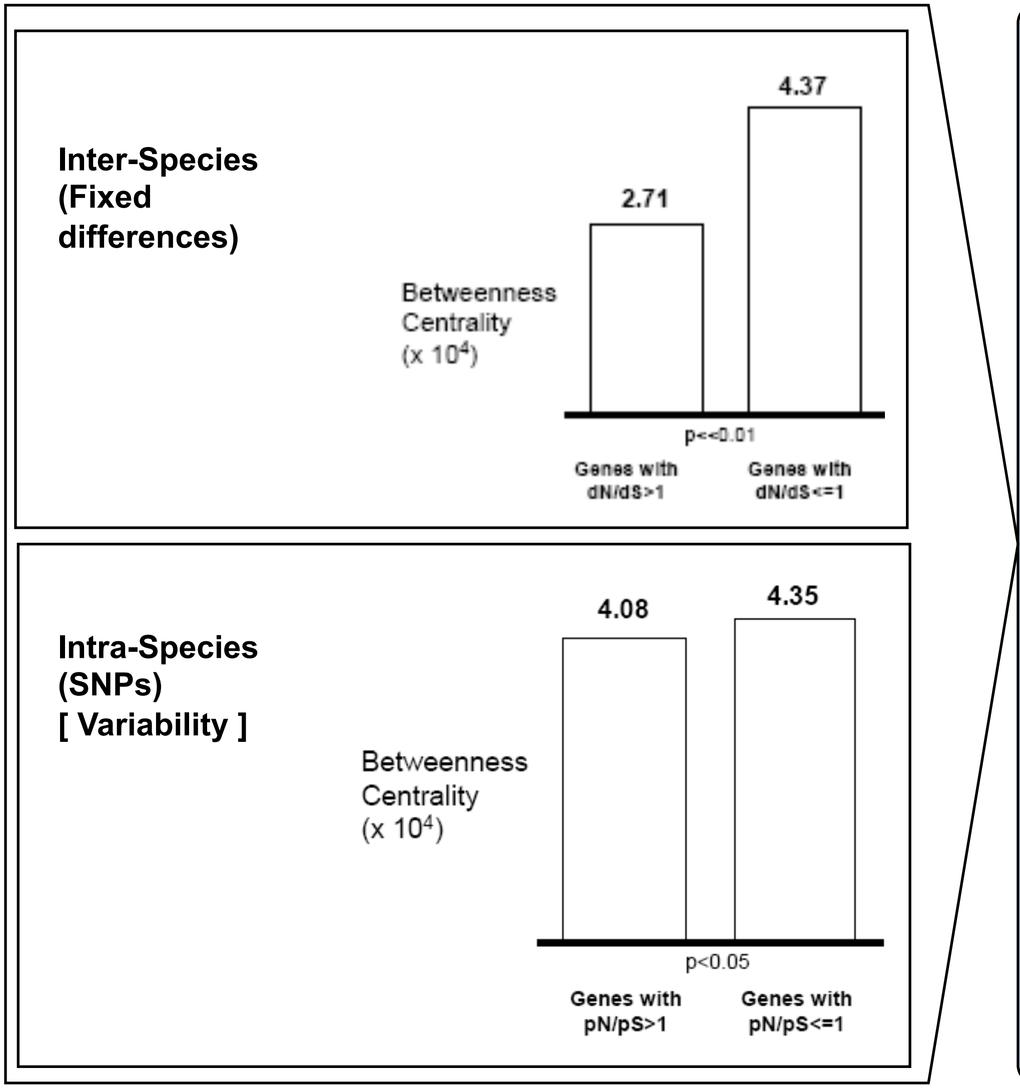
ILLUSTRATIVE

	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

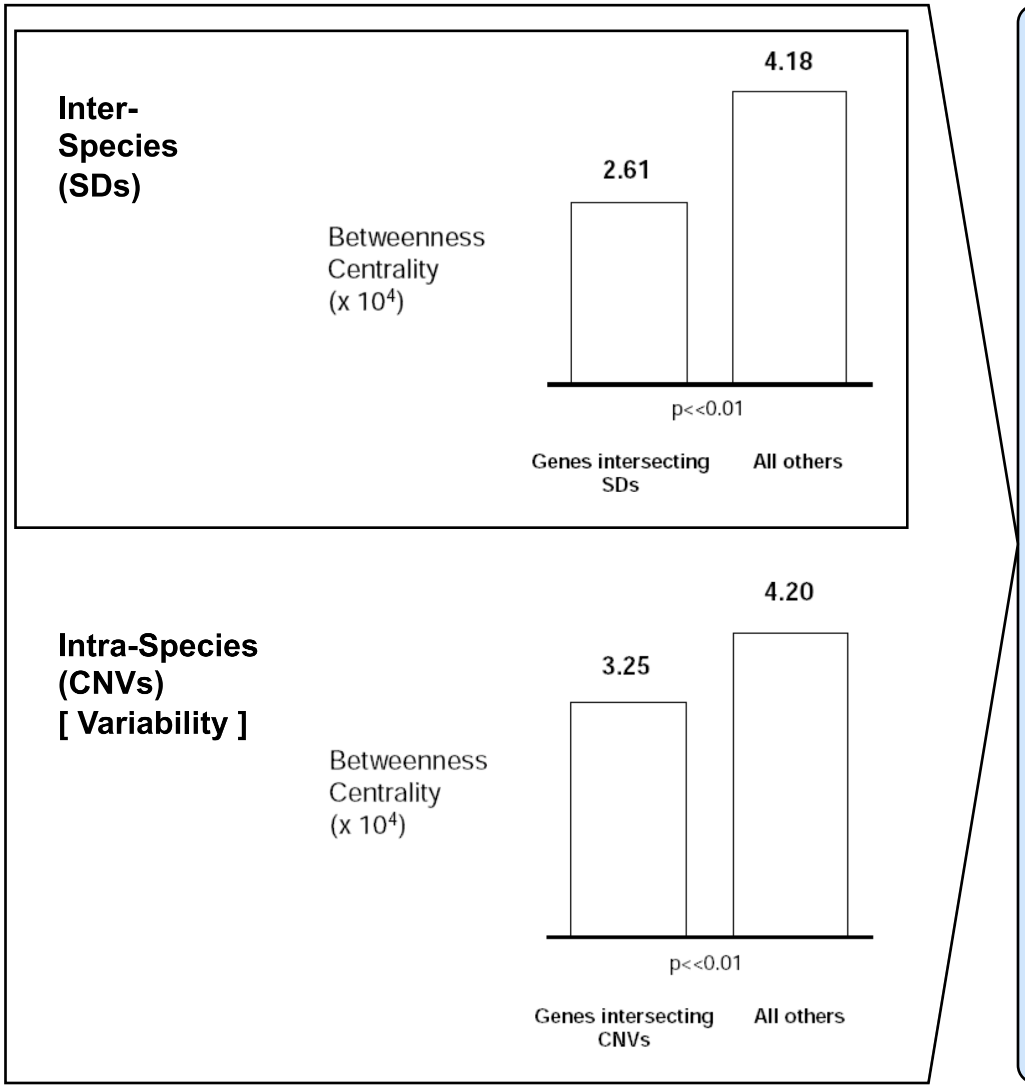
- There is a difference in **variability** (in terms of SNPs) between the network periphery and the center
- However, this difference is much smaller than the difference in **selection**
- This most likely means, that part of the effect we're seeing is due to relaxed constraint (and higher variability)
- 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

Conclusions: Net Intro. + Predicting Networks

- Developing Standardized Descriptions of Protein Function
  - ◊ Gene Naming
- Predicting Networks
  - ◊ Extrapolating from the Training Set
  - ◊ Principled ways of using the training set data in the fullest possible fashion
    - Prediction Propagation
    - Kernel Initialization

Conclusions: Network Dynamics across Cellular States

- Merge expression data with Networks
- Active network markedly different in different conditions
- Identify transient hubs associated with particular conditions
- Use these to annotate genes of unknown function

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 & Human Variation

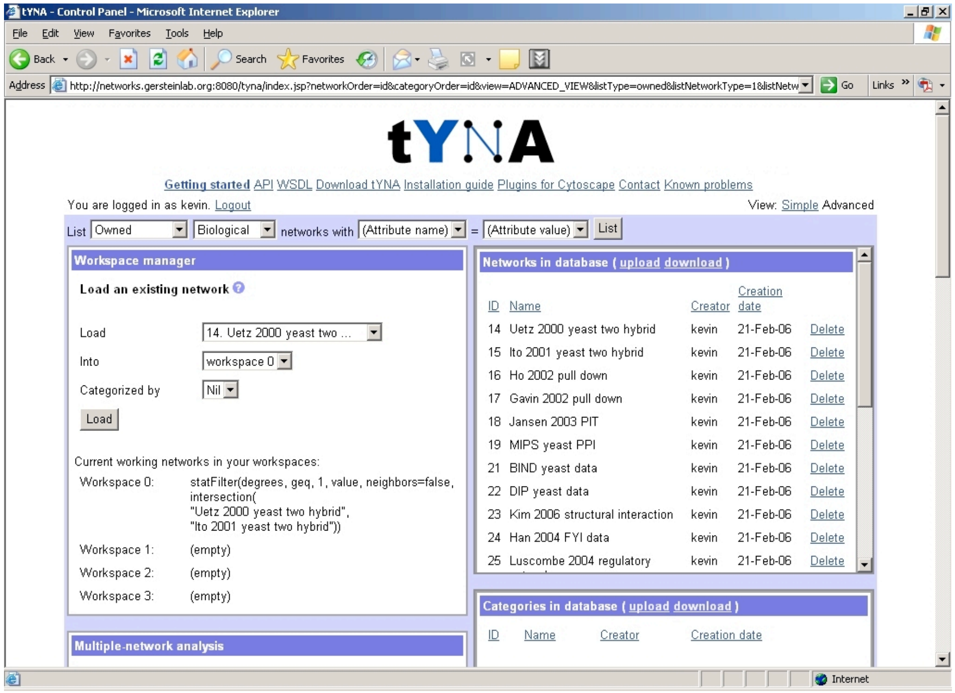
- We find ongoing evolution (positive selection) at the network periphery.
  - ◊ This trend is present on two levels:
    - On a sequence level, it can be seen as positive selection of peripheral nodes
    - On a structural level, it can be seen as the pattern of SDs that display significantly higher 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,,,

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)

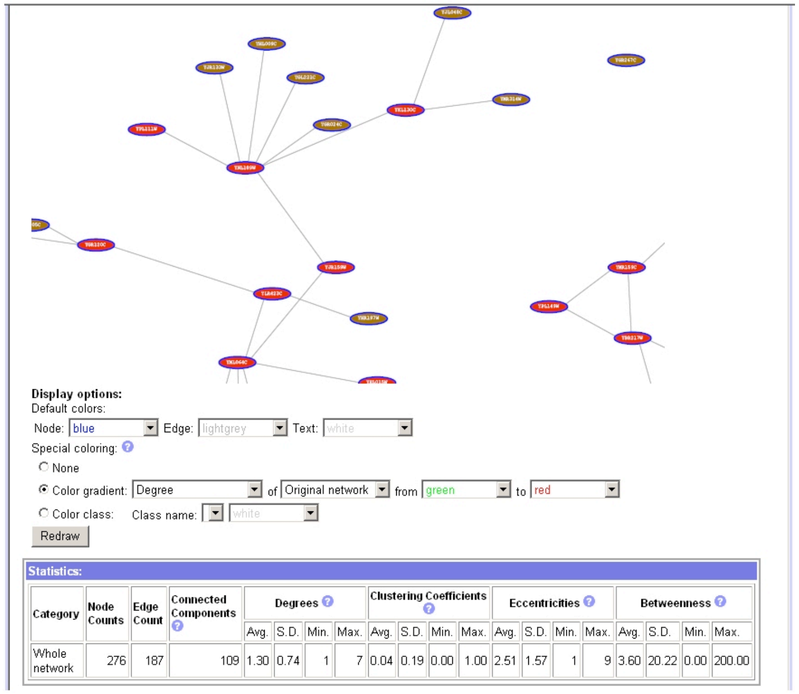
Multiple-network analysis

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:

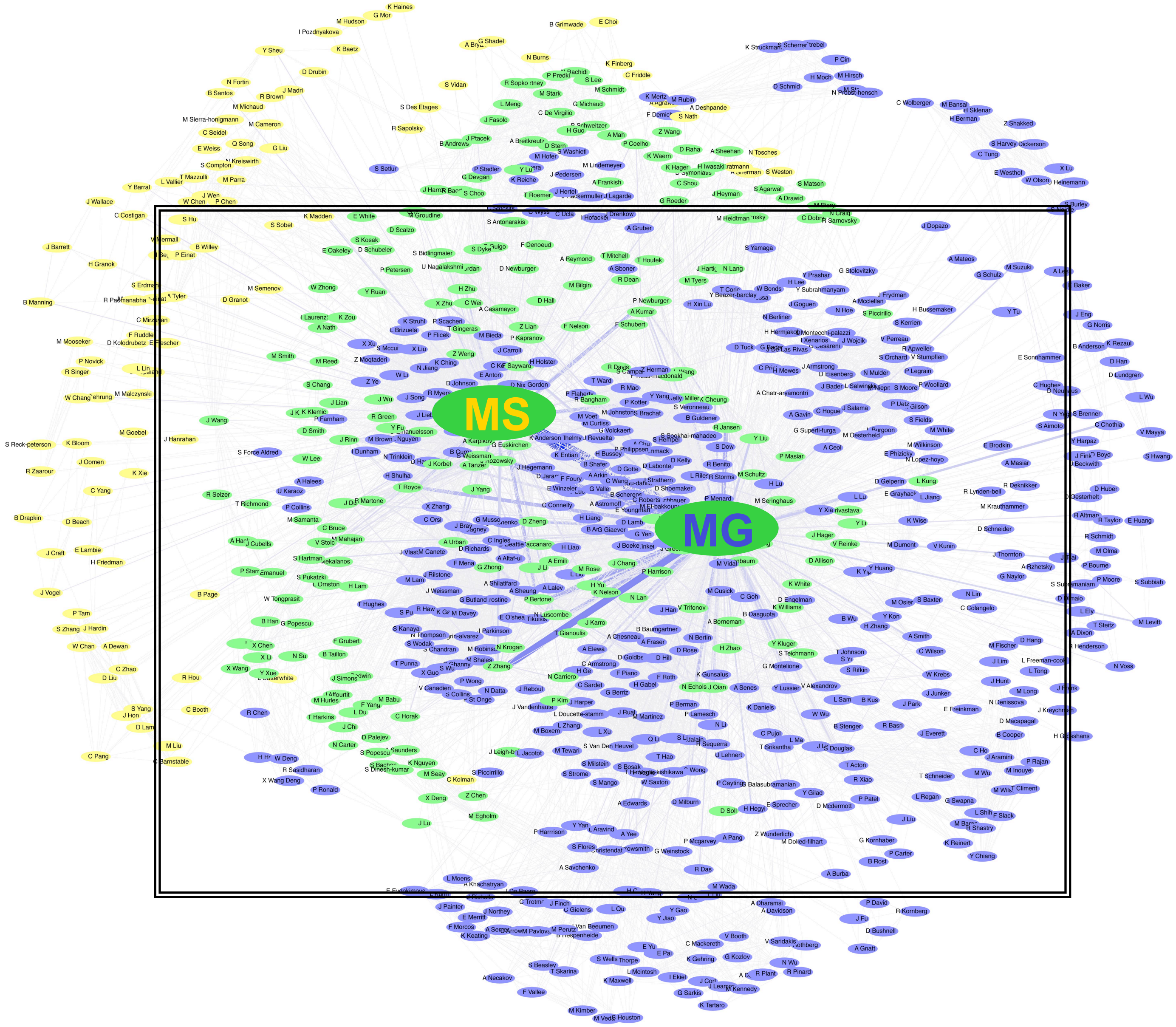
Category	Node Counts	Edge Count	Connected Components	Degrees				Clustering Coefficients				Eccentricities				Betweenness			
				Avg.	S.D.	Min.	Max.	Avg.	S.D.	Min.	Max.	Avg.	S.D.	Min.	Max.	Avg.	S.D.	Min.	Max.
Whole network	276	187	109	1.30	0.74	1	7	0.04	0.19	0.00	1.00	2.51	1.57	1	9	3.60	20.22	0.00	200.00

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]

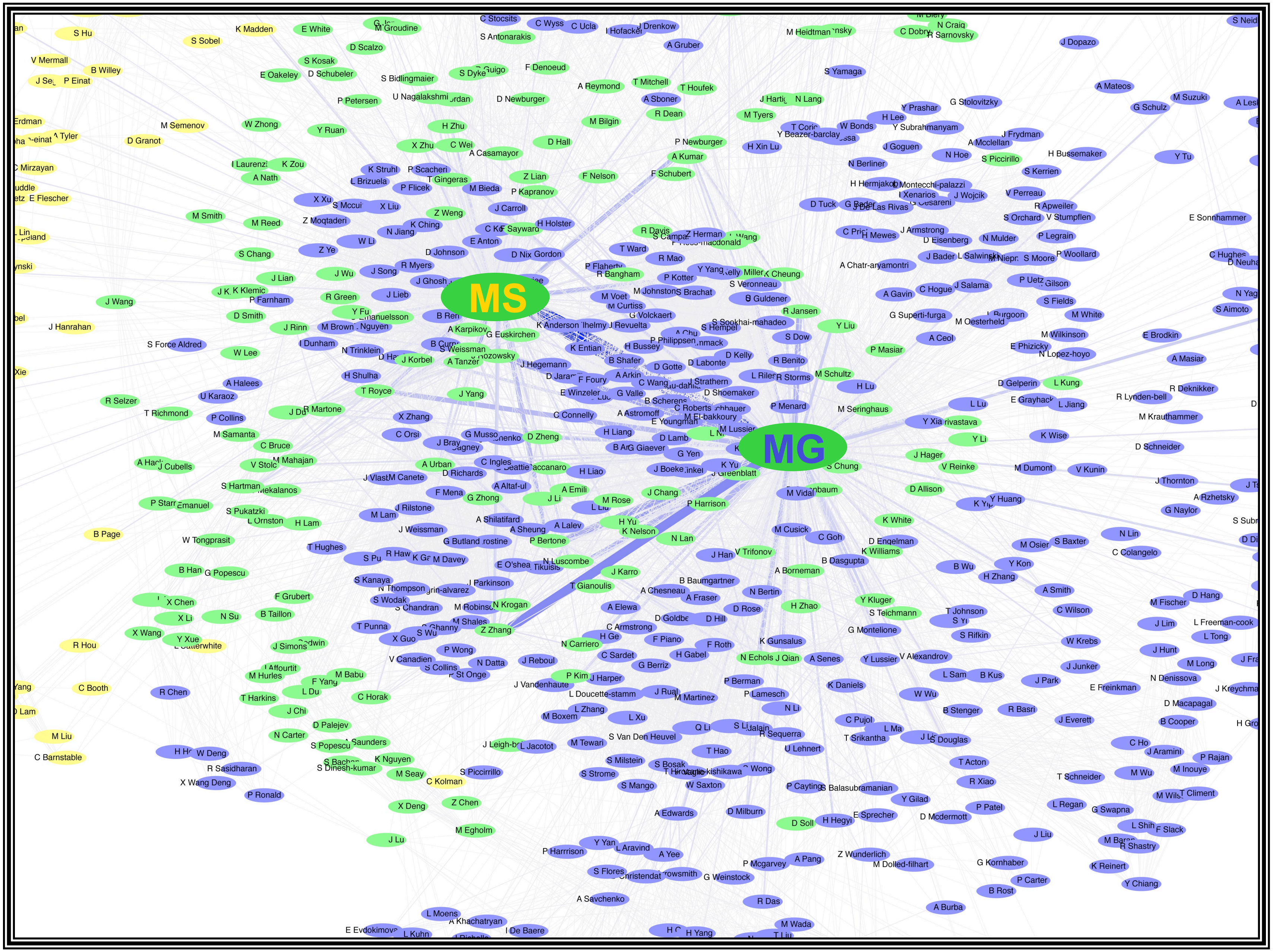
Acknowledgements

TopNet.GersteinLab.org



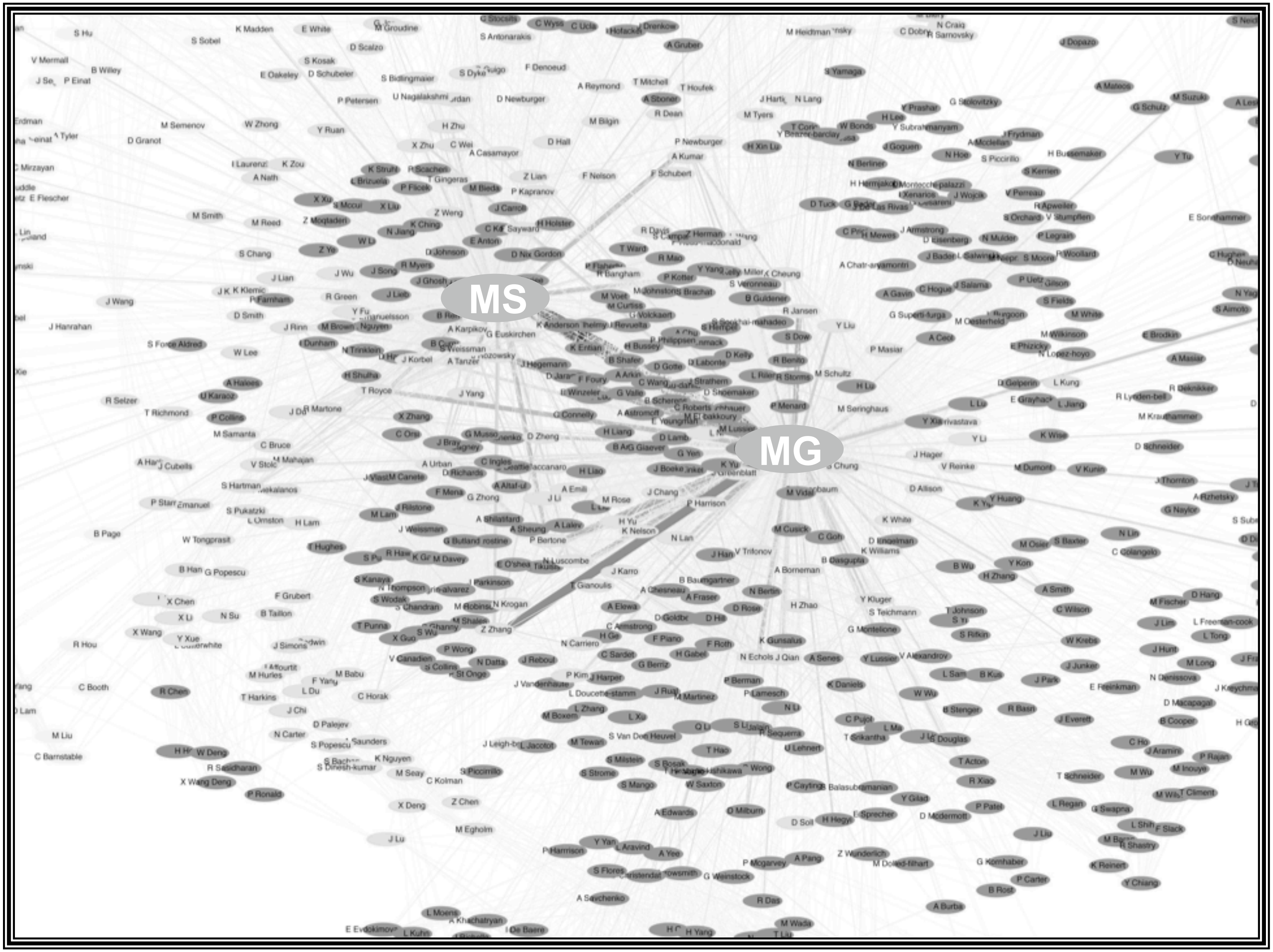
Acknowledgements

TopNet.GersteinLab.org



Acknowledgements

TopNet.GersteinLab.org

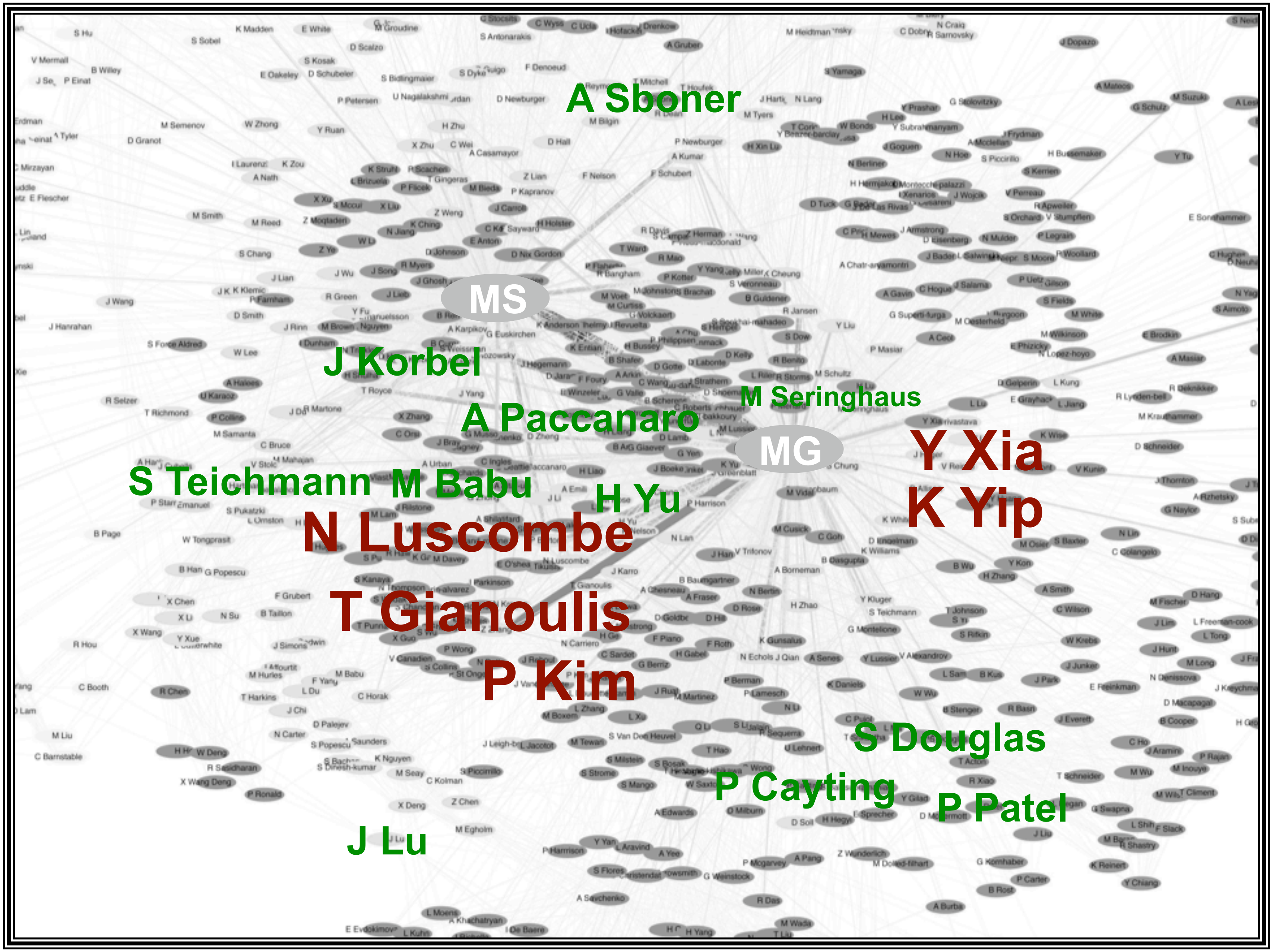


P Bork, J Raes

Acknowledgements

TopNet.GersteinLab.org

Job opportunities currently for postdocs & students



A Sboner

J Korbel

M Seringhaus

A Paccanaro

Y Xia

K Yip

S Teichmann

M Babu

H Yu

N Luscombe

T Gianoulis

P Kim

S Douglas

P Cayting

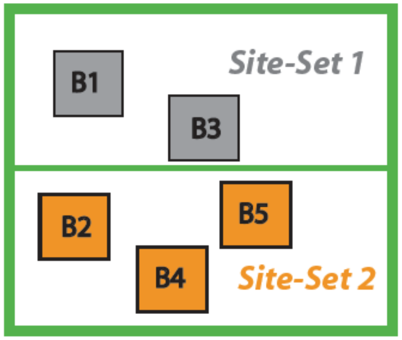
P Patel

J Lu

Extra

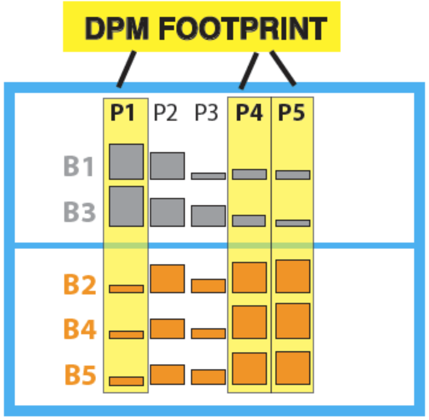
DPM: Discriminative Partition Matching

Environment



Cluster (Partition)

Metabolism



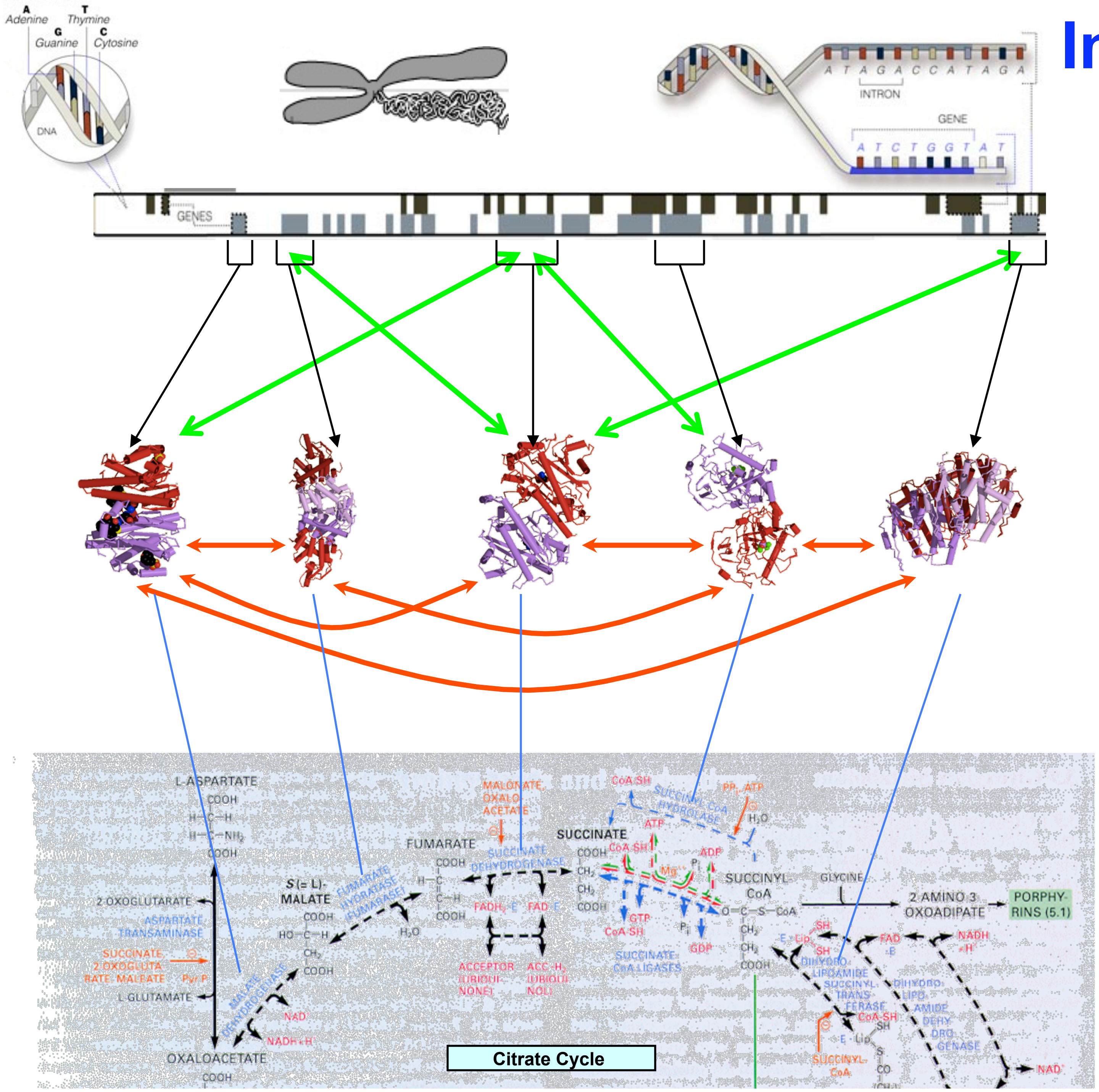
Test

Taurine biosynthesis
Heme biosynthesis
Asparagine degradation
Nitrogen fixation
Acylglycerol degradation
Asparagine biosynthesis
Cysteine Metabolism

Functional class	**pval**
InfoStorage & Processing	.07
Cellular Process	.08
Metabolism	4x10-14

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

Interactome



GENOME

protein-DNA interactions

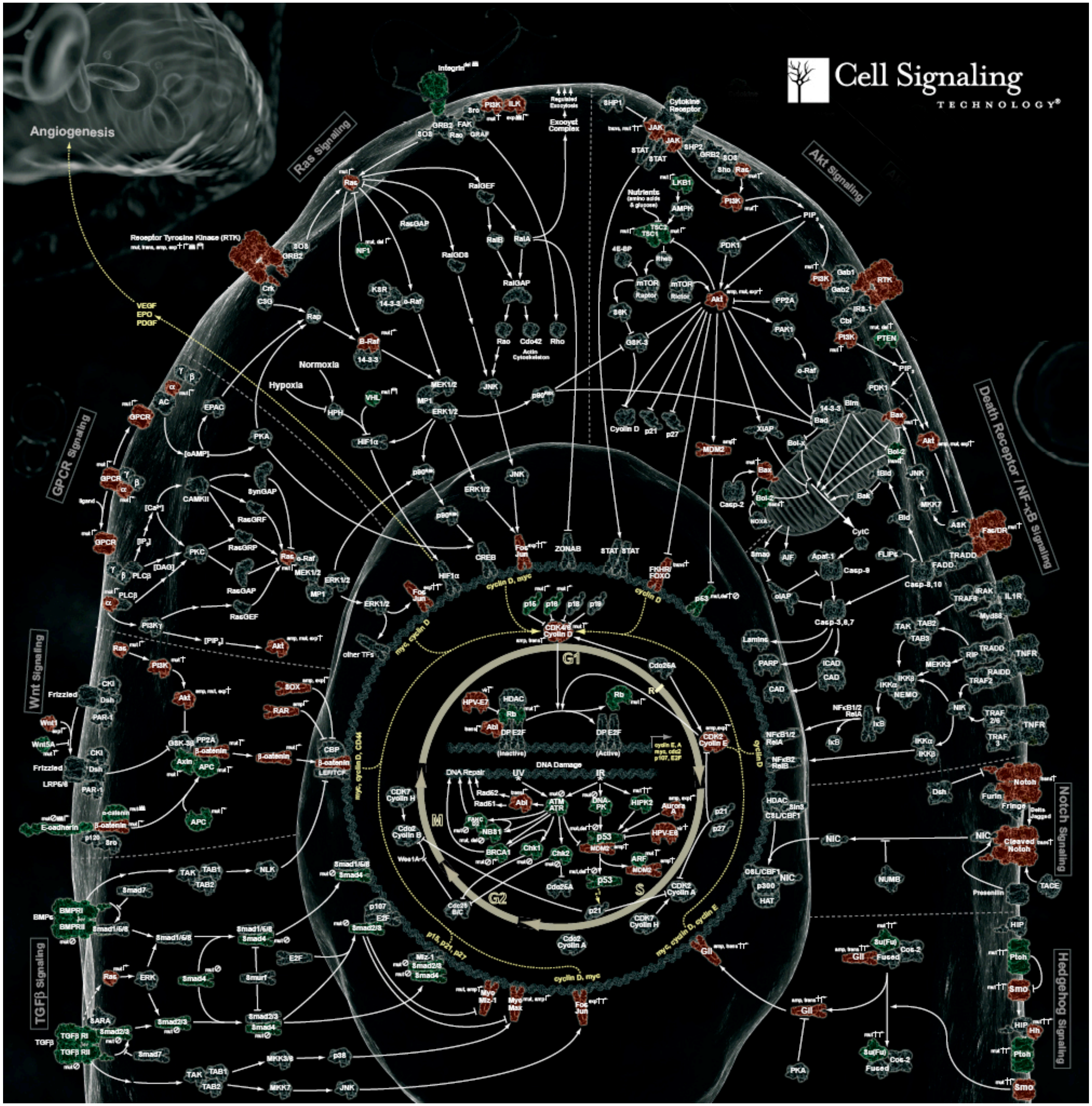
PROTEOME

protein-protein interactions

METABOLISM

Protein-small molecule interactions

[From H Yu]



Networks help us understand biological processes

[From H Yu]

More Information on this Talk

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

SUBJECT: Networks

DESCRIPTION:
Cornell Medical School, Physiology, Biophysics and Systems Biology (PBSB) graduate program, 2009.01.26, 16:00-17:00; [I:**CORNELL-PBSB**]
(Long networks talk, incl. the following topics:
why networks w. **amsci*, funnygene***, net. prediction intro, **memint*, tse*, essen*, sandy*, metagenomics*, netpossel*, tyna*+ topnet***, & **pubnet*** . Fits easily into 60′ w. 10′ questions. PPT works on mac & PC and has many photos w. EXIF tag **kwcornellpbsb** .)

(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** .