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

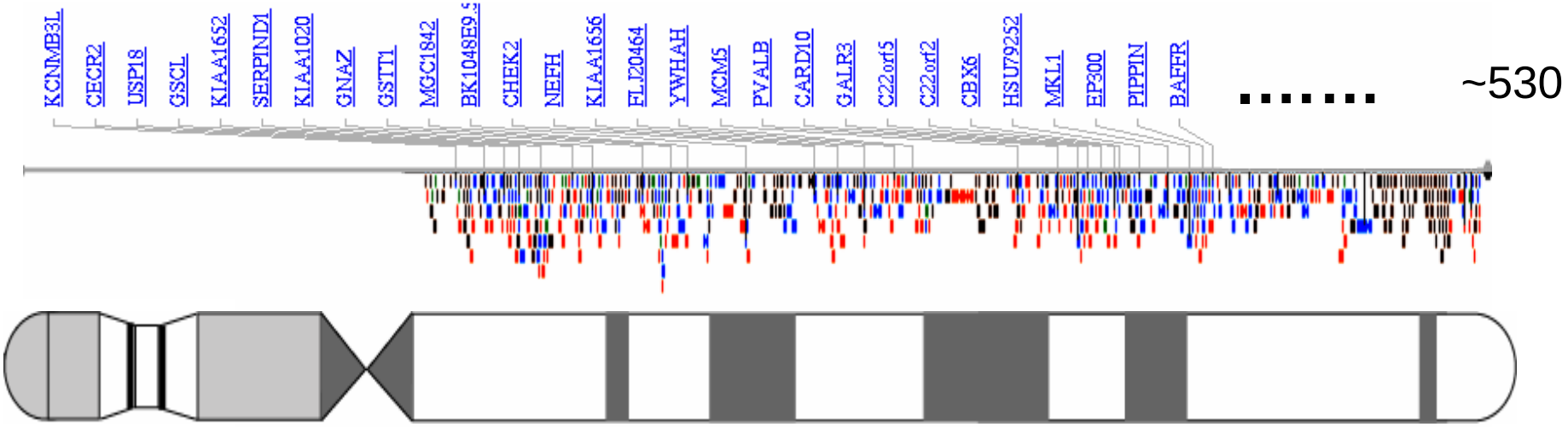
University of Chicago
2008.12.02, 12:00-13:00

Mark B Gerstein
Yale

Slides at Lectures.GersteinLab.org

(See Last Slide for References & More Info.)

The problem: Grappling with Function on a Genome Scale?



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

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

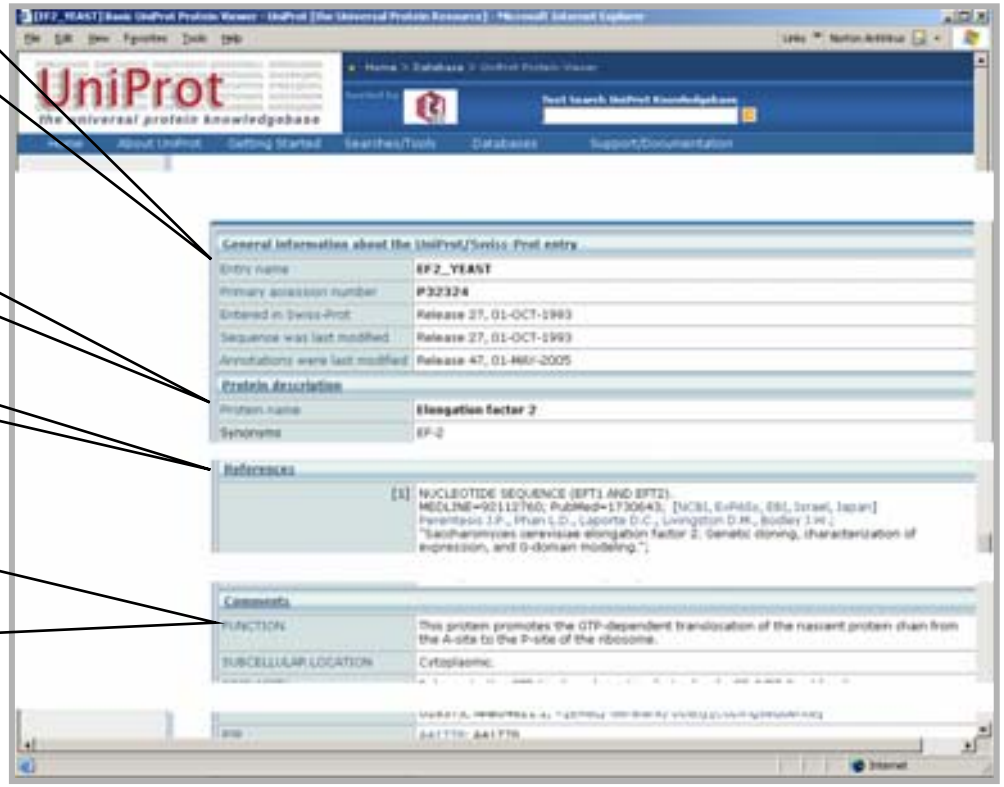
Traditional single molecule way to integrate evidence & describe function

EF2_YEAST

Descriptive Name:
Elongation Factor 2

Lots of references
to papers

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

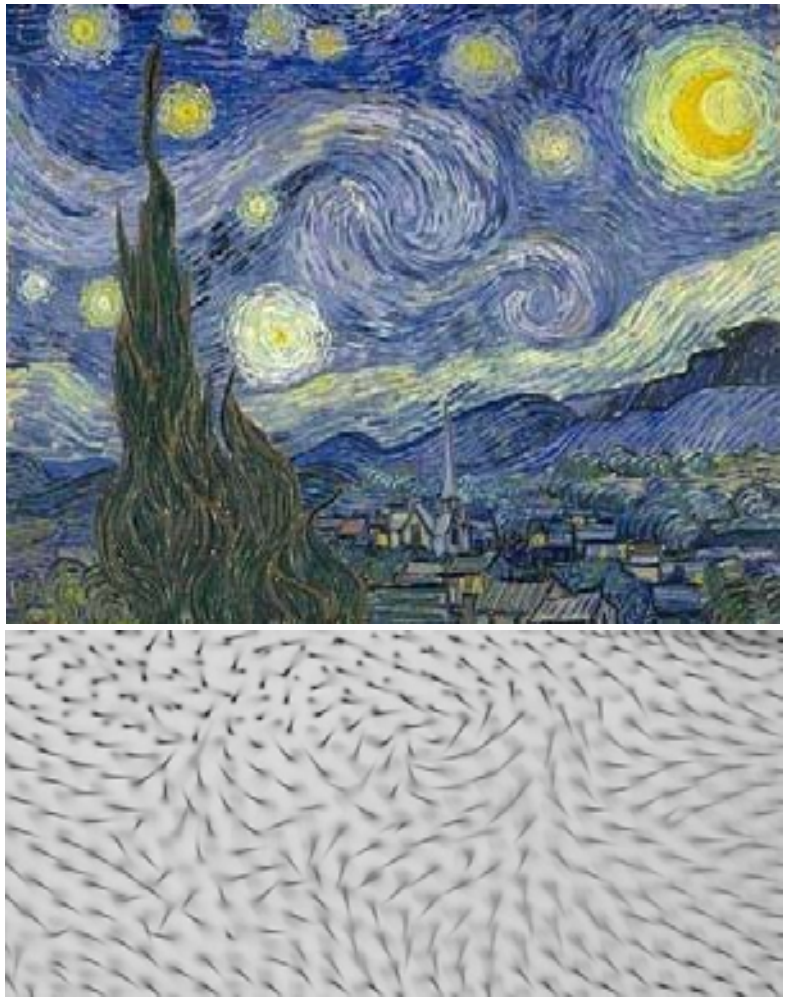


Some obvious issues in scaling single molecule definition to a genomic scale

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

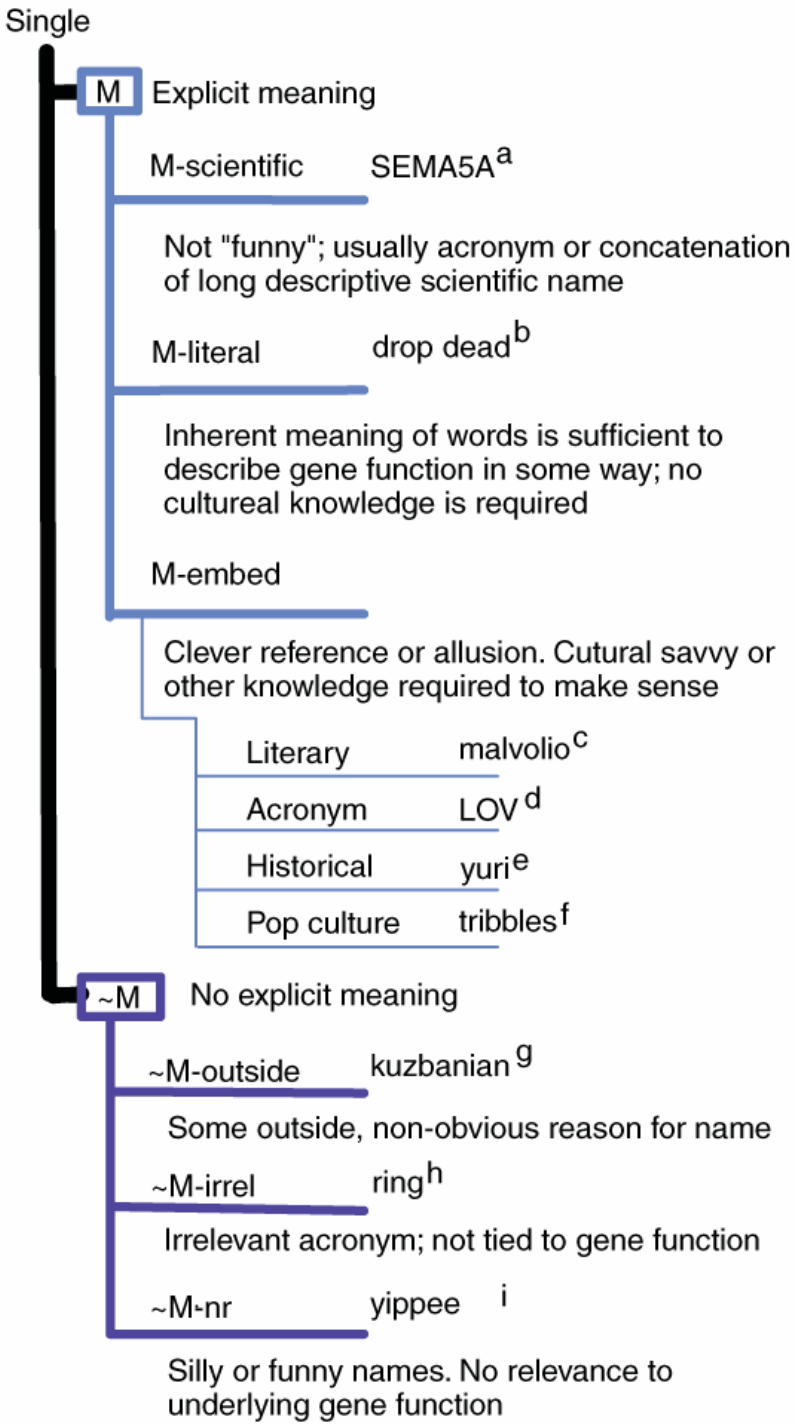
Some obvious issues in scaling single molecule definition to a genomic scale

- 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)]

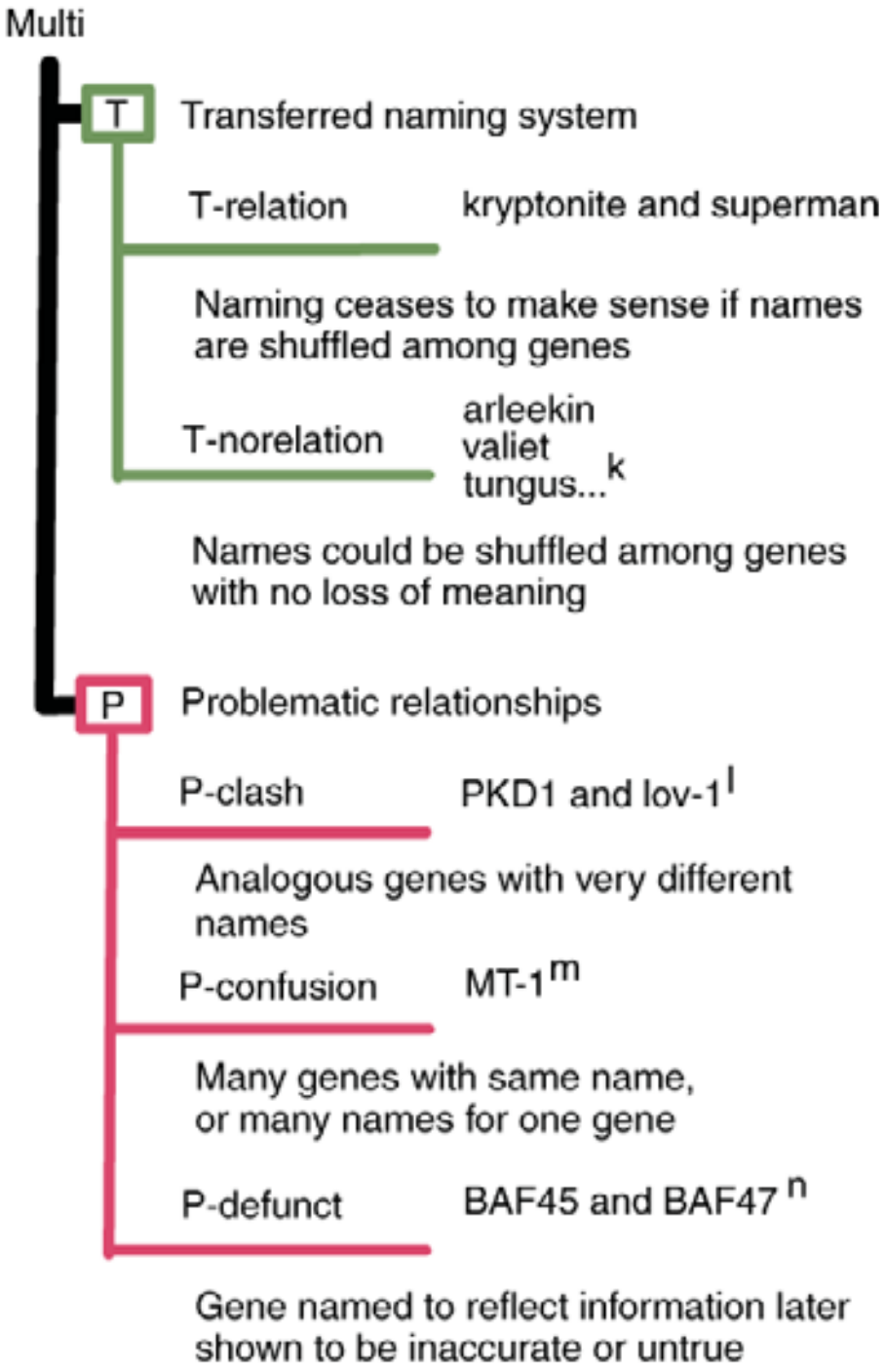
Naming Pathologies: Related to Single Genes



(b) drop dead: flies with mutations in drop dead die rapidly after their brain rapidly deteriorates. (c) malvolio: gene needed for normal taste behaviour. Malvolio in Shakespeare's Twelfth Night tasted "with distempered appetite". (d) LOV: light, oxygen, or voltage (LOV) family of blue-light photoreceptor domains. (e) yuri: this gene was discovered on the anniversary of Yuri Gagarin's space flight. Mutants have problems with gravitaxis and cannot stay aloft. (f) tribbles: cells divide uncontrollably, like the eponymous Star Trek characters. (g) kuzbanian: mutants have uncontrollable bristle growth. Koozbanians are alien Muppets with uncontrollable hair growth; spelling was changed to avoid copyright infringement. (h) ring: really interesting new gene. (i) yippee: a graduate student's reaction on cloning the gene

[Seringhaus et al. GenomeBiology (2008)]

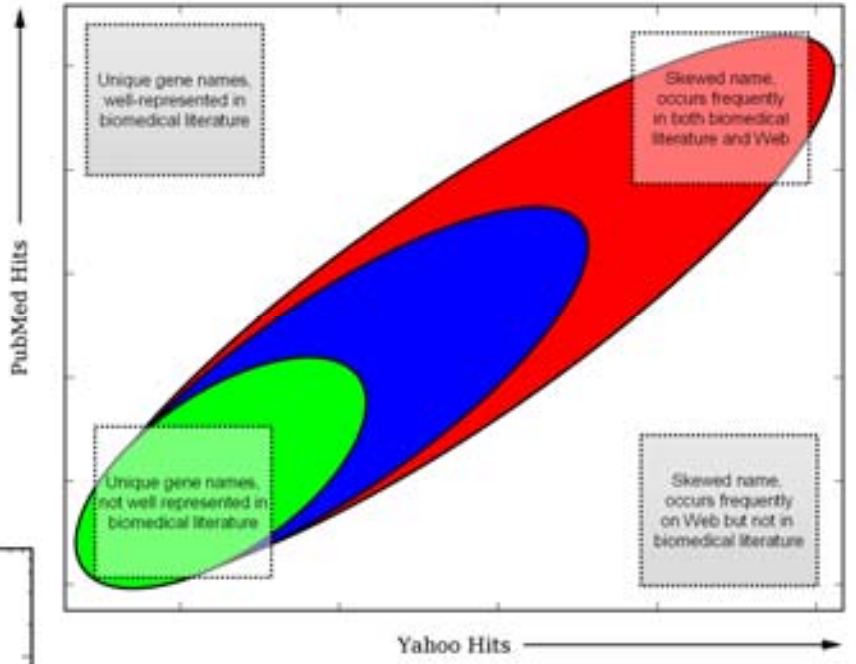
Naming Pathologies: Involving Multiple Gene Names

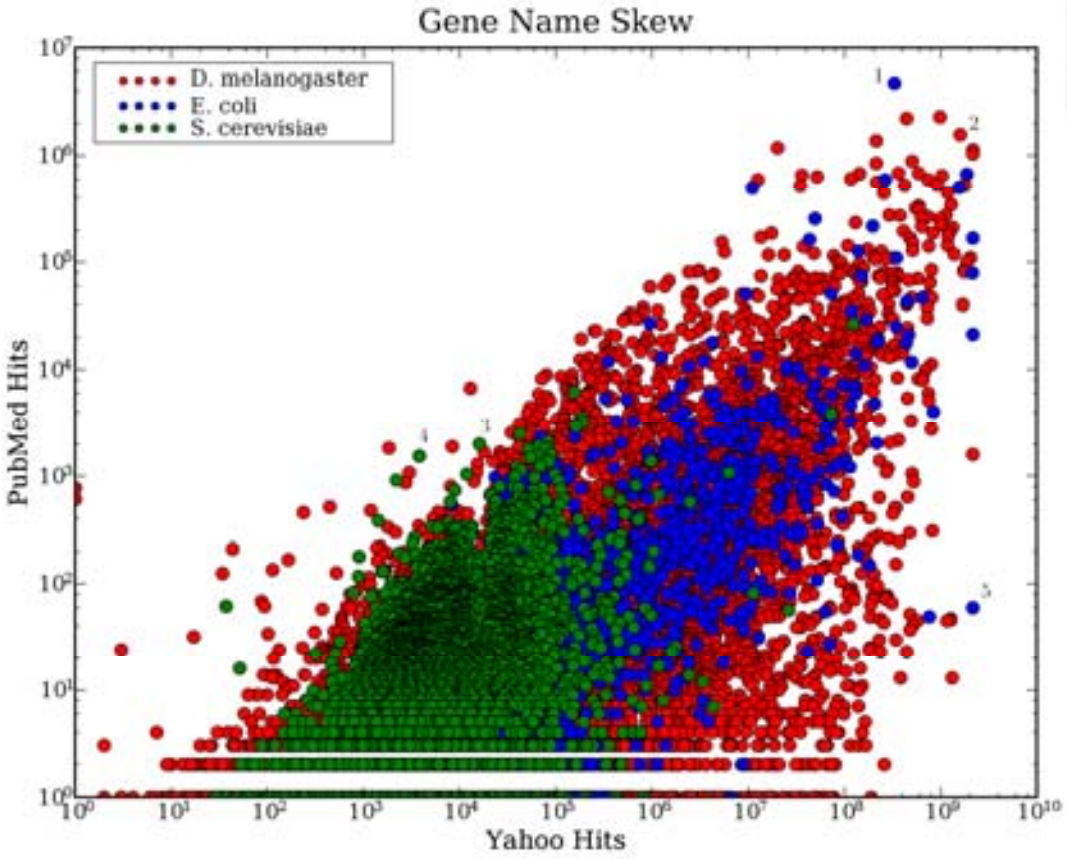


(j) kryptonite and superman: the kryptonite mutation suppresses the function of the SUPERMAN gene. (k) arleekin, valient, tungus: mutations in arleekin, valient, tungus and 29 other genes affect long-term memory. Named after Pavlov's dogs. (l) PKD1 (human) and lov-1 (worm): these are homologs, although their names do not suggest it. (m) MT-1: this label can refer to at least 11 different human genes. (n) BAF45 and BAF47: names for the same gene, reflecting a revision of the molecular weight of product.

[Seringhaus et al. GenomeBiology (2008)]

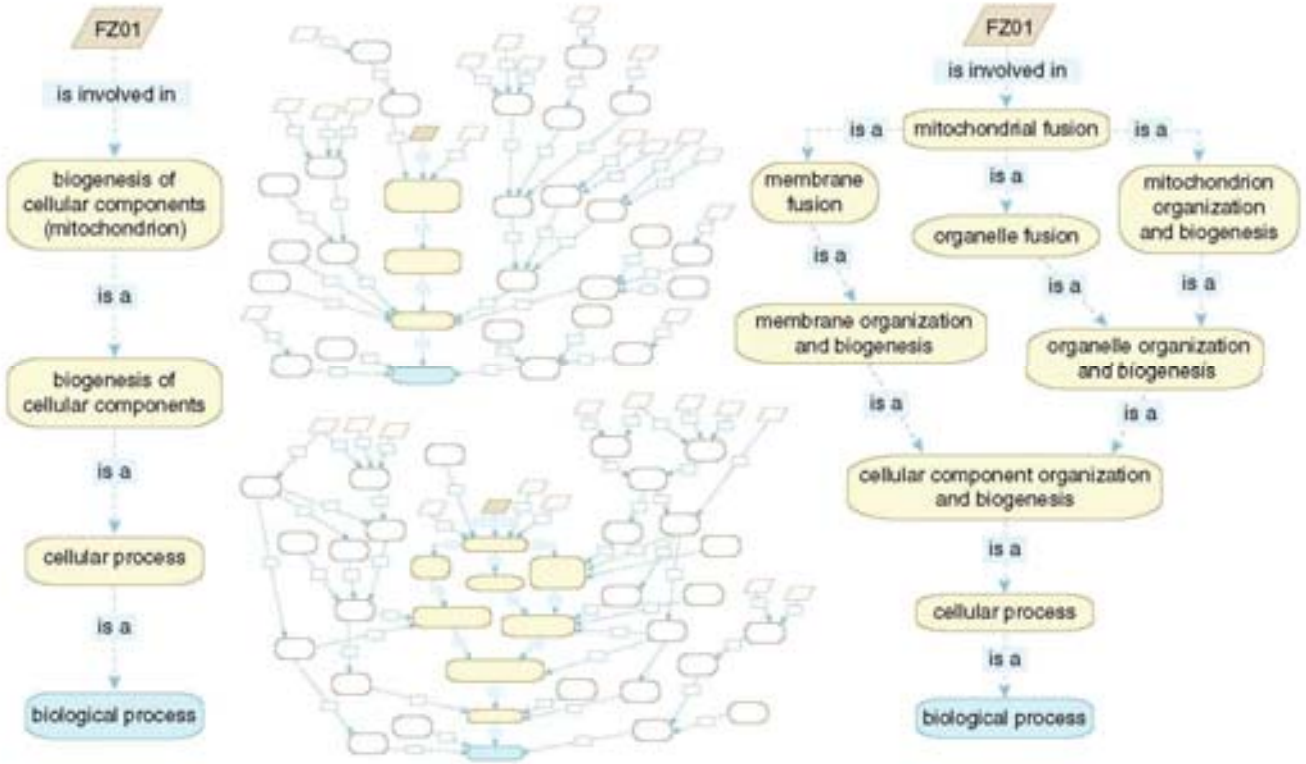
Gene Name Skew





[Seringhaus et al. GenomeBiology (2008)]

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



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

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

Towards Developing Standardized Descriptions of Function

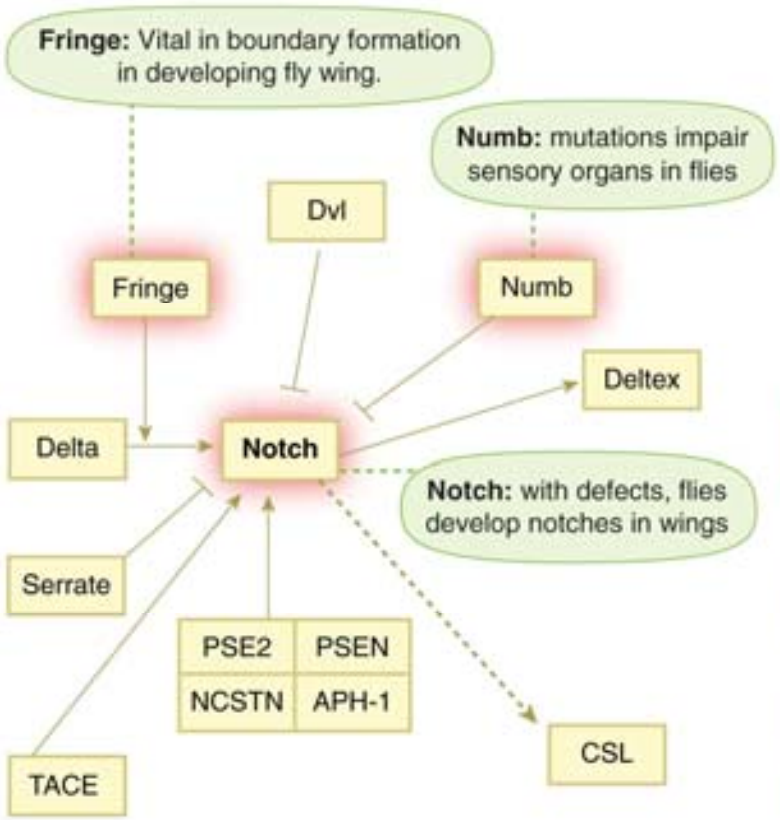
- Subjecting each gene to standardized expt. and cataloging effect
  - ◊ KOs of each gene in a variety of std. conditions => phenotypes
  - ◊ Std. binding expts for each gene (e.g. prot. chip)

- Function as a vector

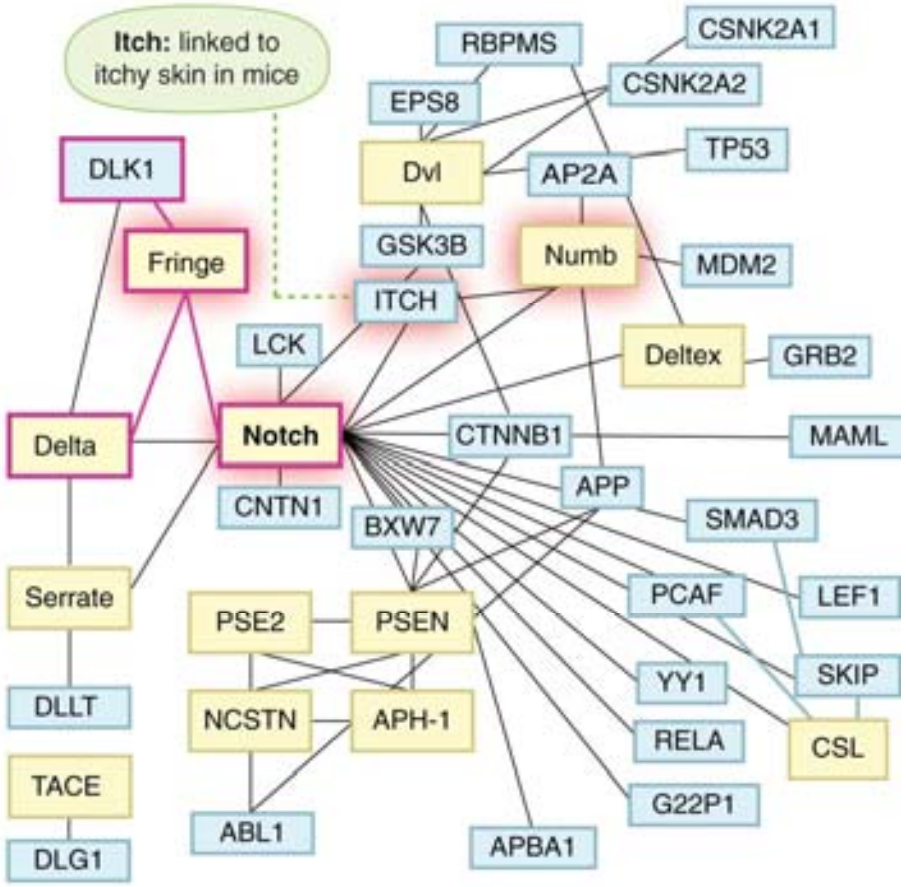
	nucleic acids		small molecules					proteins			
	DNA	RNA	ATP	Metal	CoA	NAD	.....	G protein	CDC28	Calmodulin	.....
protein 1	1.0	0	0	0	0	0	.....	0	0	0	.....
protein 2	0	0.9	0	0	0	0	.....	0	0	0	.....
protein 3	1.0	0	1.0	0	0	0	.....	0	0	0	.....
protein 4	0	0	0	0	0.8	0	.....	0	0	1.0	.....
protein 5	1.0	0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



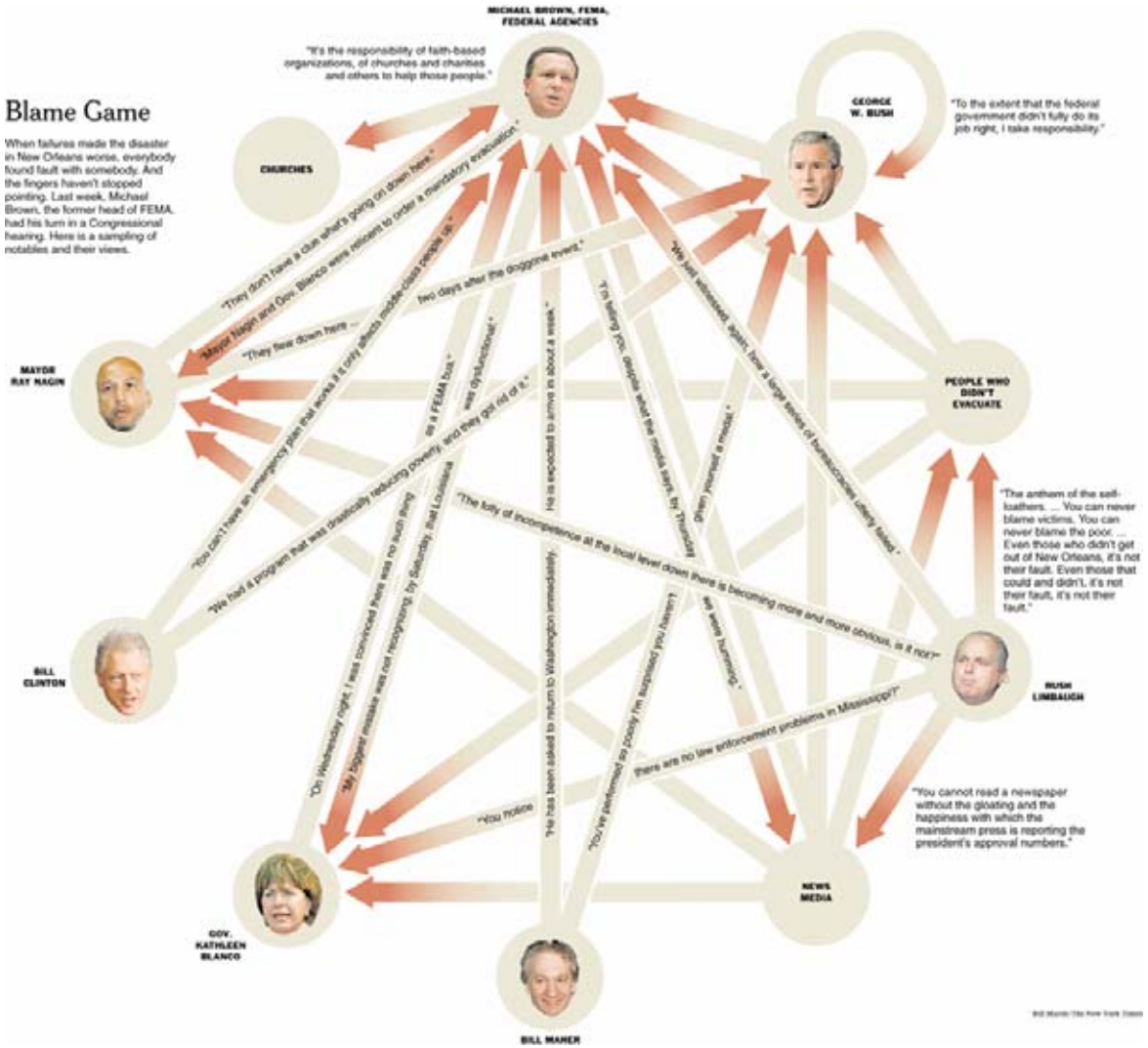
Classical KEGG pathway



Same Genes in High-throughput Network

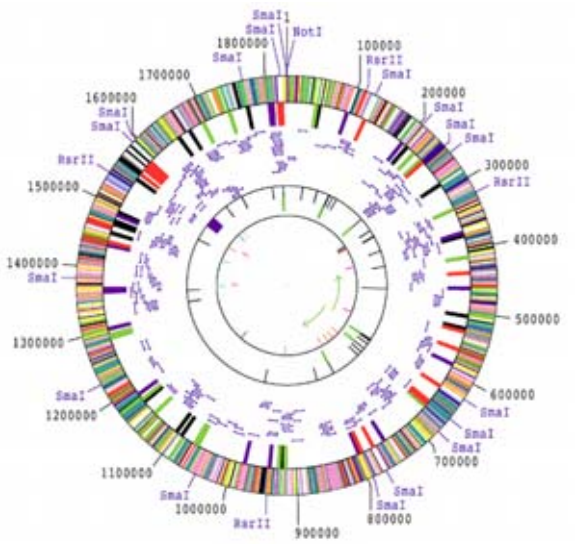
[Seringhaus & Gerstein, Am. Sci. '08]

Using Networks to Describe Function



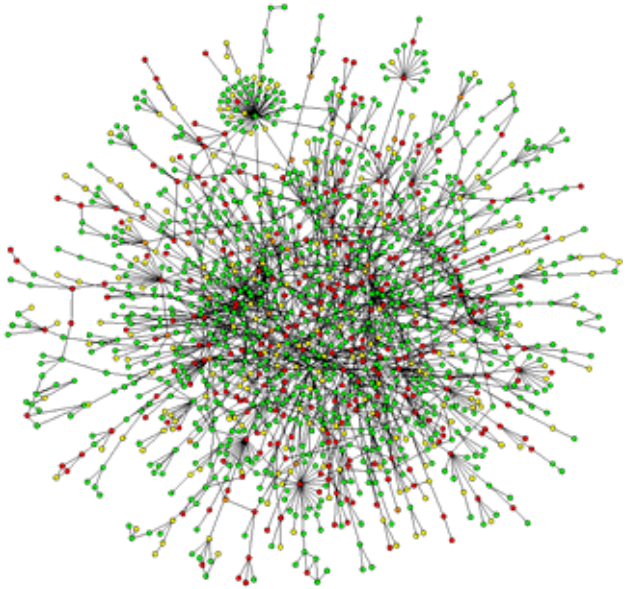
[NY Times, 2-Oct-2005]

Networks occupy a midway point in terms of level of understanding



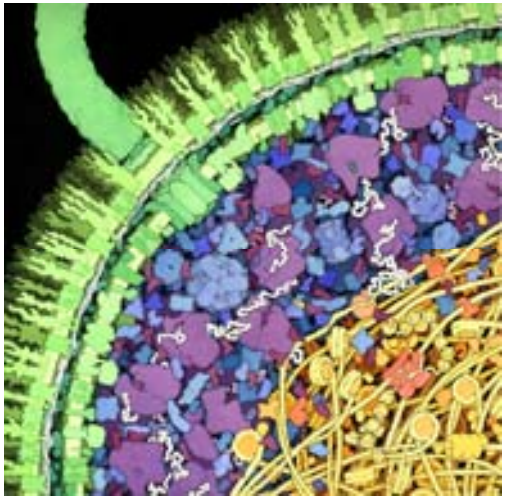
1D: Complete Genetic Partslist

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



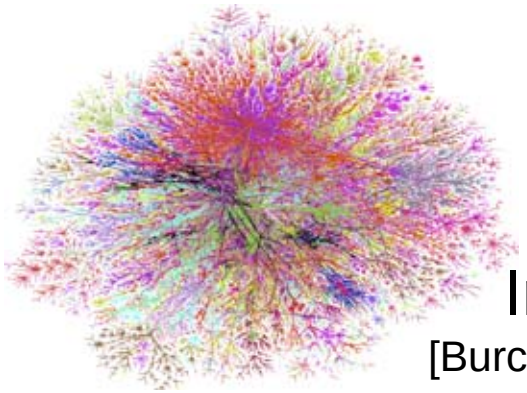
~2D: Bio-molecular Network Wiring Diagram

[Jeong et al. Nature, 41:411]

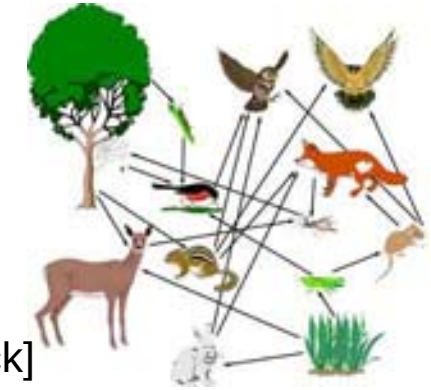


3D: Detailed structural understanding of cellular machinery

Networks as a universal language



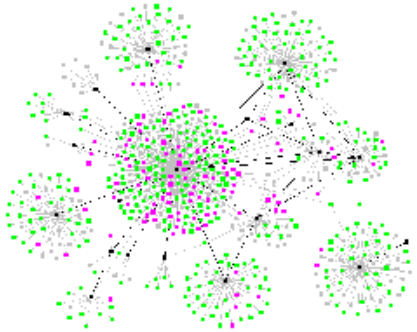
Internet
[Burch & Cheswick]



Food Web

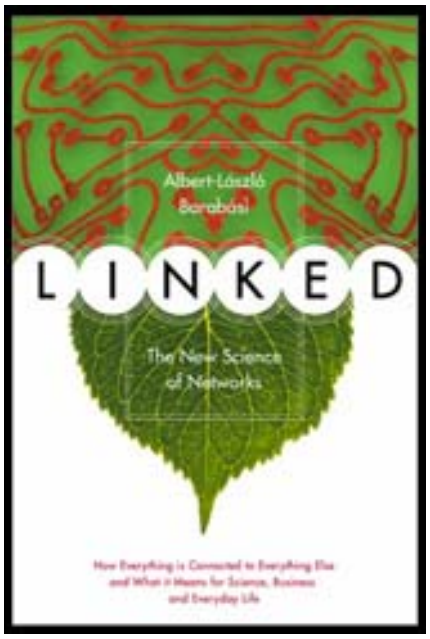
Electronic Circuit

Neural Network
[Cajal]



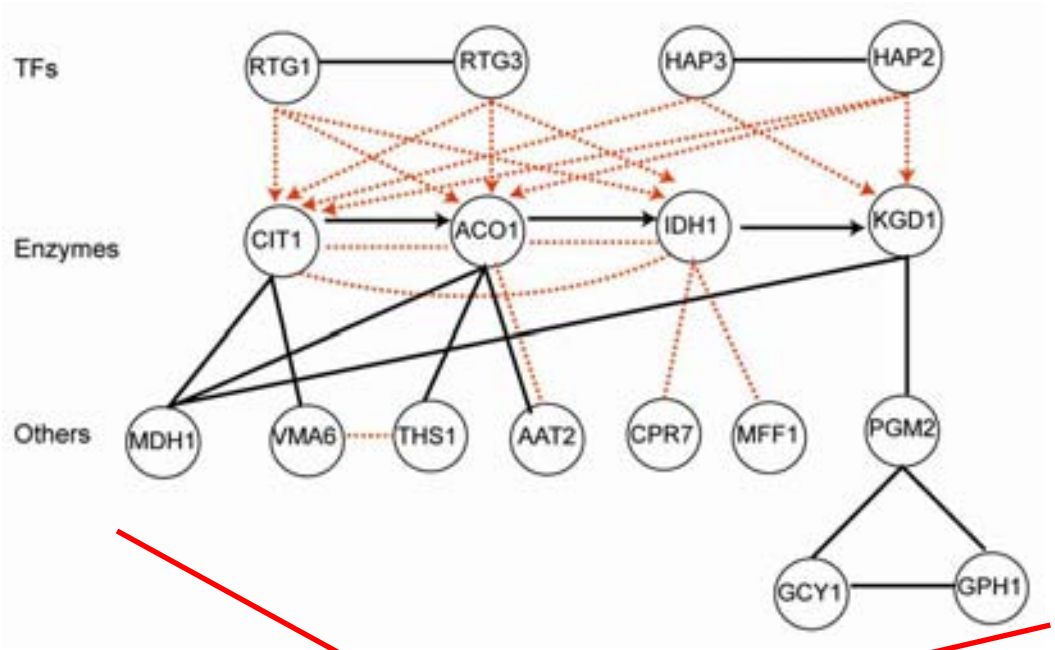
Disease Spread
[Krebs]

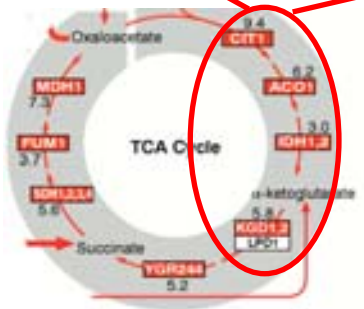
Protein Interactions
[Barabasi]



Social Network

Combining networks forms an ideal way of integrating diverse information





Part of the TCA cycle

Metabolic pathway

Transcriptional regulatory network

Physical protein-protein Interaction

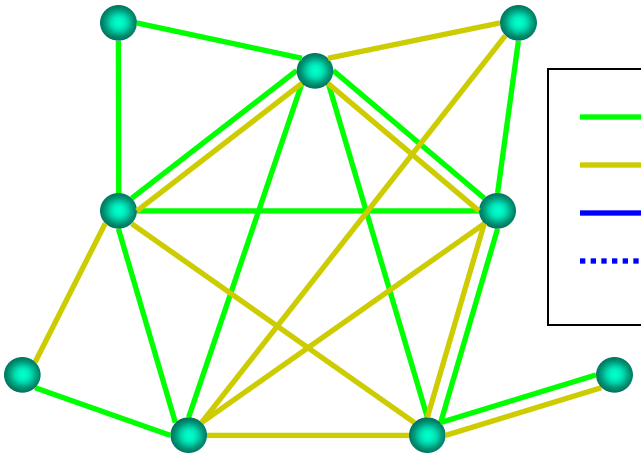
Co-expression Relationship

Genetic interaction (synthetic lethal)
Signaling pathways

Predicting Networks

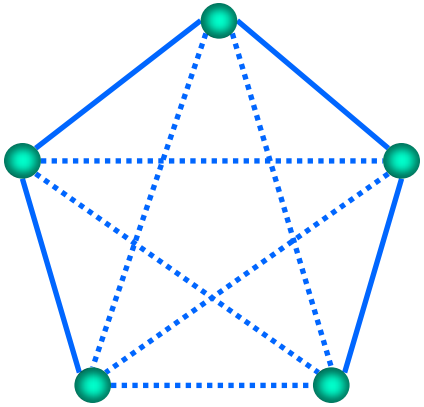
How do we construct large molecular networks? From extrapolating correlations between functional genomics data with fairly small sets of known interactions. Making best use of the known training data.

Prediction of protein interactions: Bayesian integration



Network

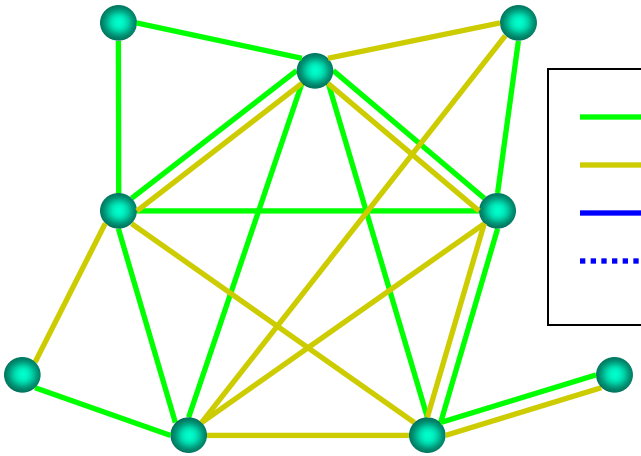
Feature 1, e.g. co-expression
Feature 2, e.g. Y2H
Gold-standard +
Gold-standard –



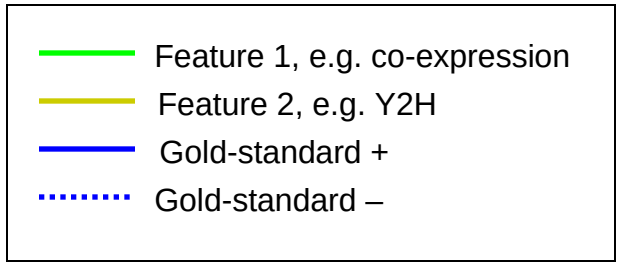
Gold-Standards

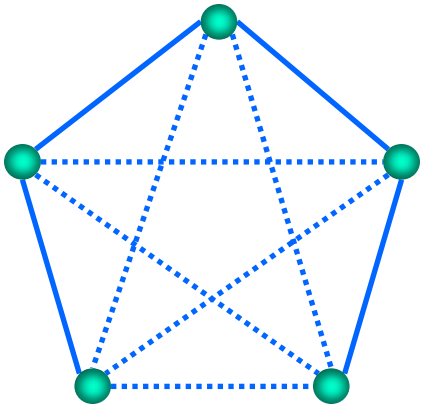
[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]

Prediction of protein interactions: Bayesian integration



Network



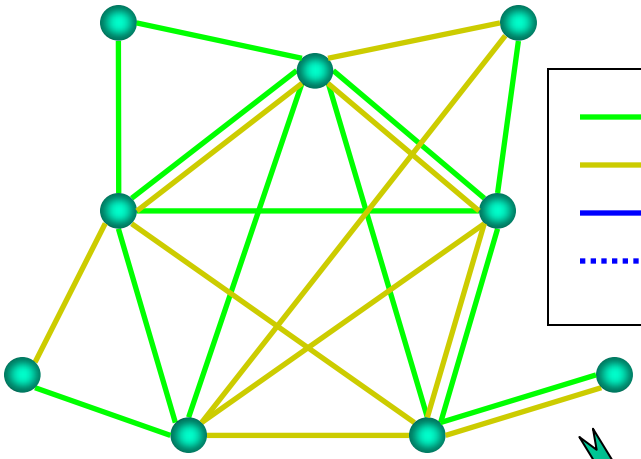


Gold-Standards

$$\text{"Quality Score"} = \frac{\text{Frac. of Gold-Std Positives with Feature}}{\text{Frac. of Gold-Std Negatives with Feature}}$$

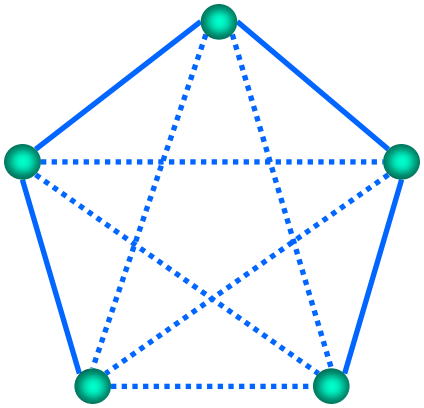
[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]

Prediction of protein interactions: Bayesian integration

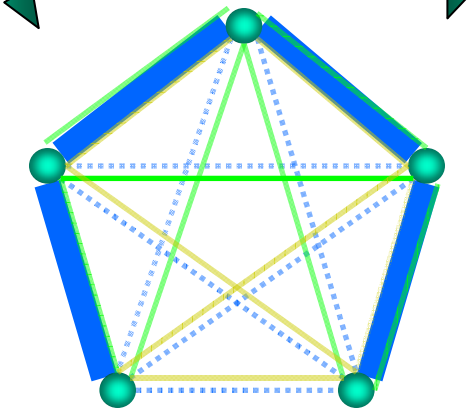


Network

Feature 1, e.g. co-expression
Feature 2, e.g. Y2H
Gold-standard +
Gold-standard –



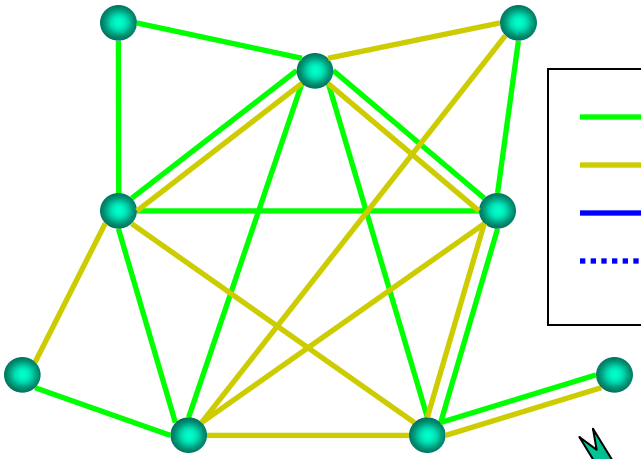
Gold-Standards



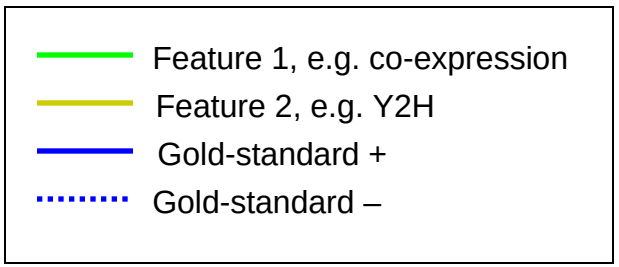
$L_1 = (4/4)/(3/6) = 2$

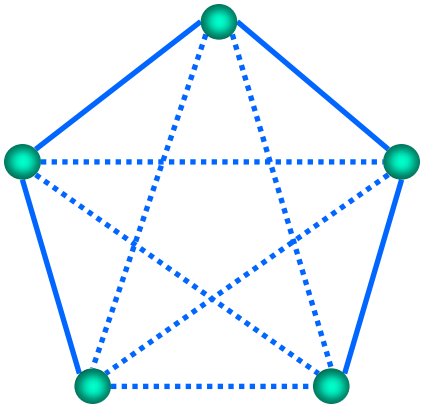
[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]

Prediction of protein interactions: Bayesian integration

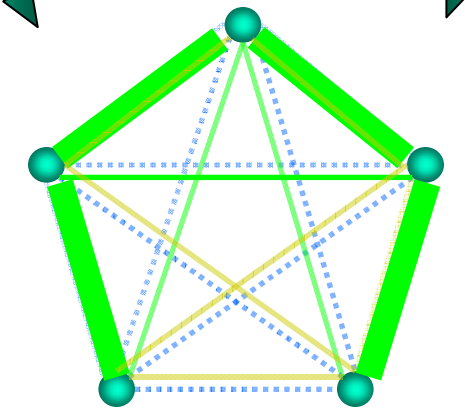


Network





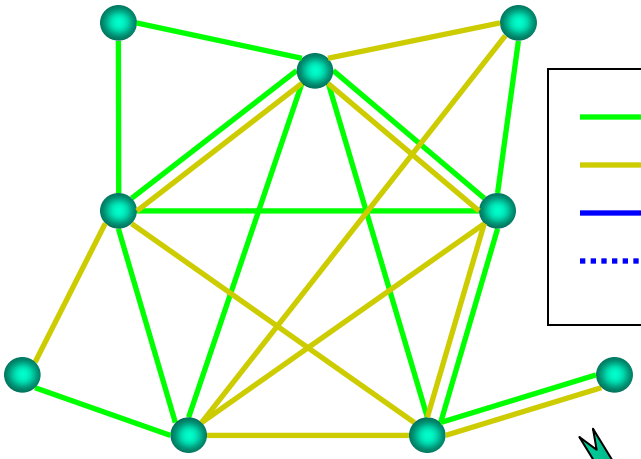
Gold-Standards



$L_1 = (4/4)/(3/6) = 2$

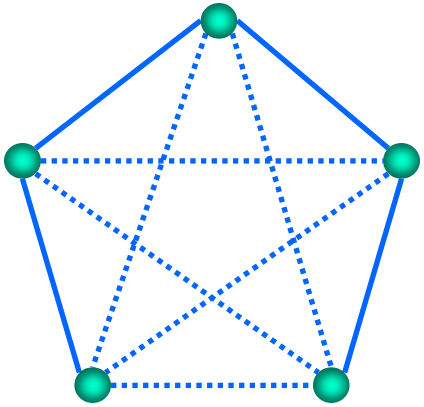
[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]

Prediction of protein interactions: Bayesian integration

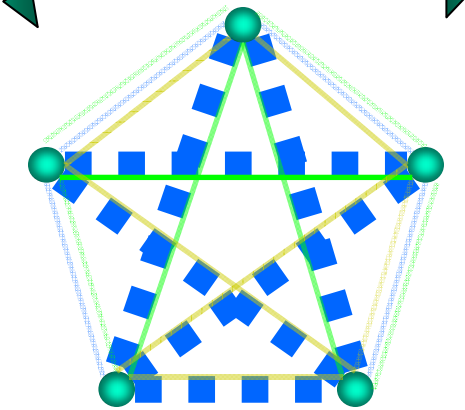


Network

Feature 1, e.g. co-expression
Feature 2, e.g. Y2H
Gold-standard +
Gold-standard –



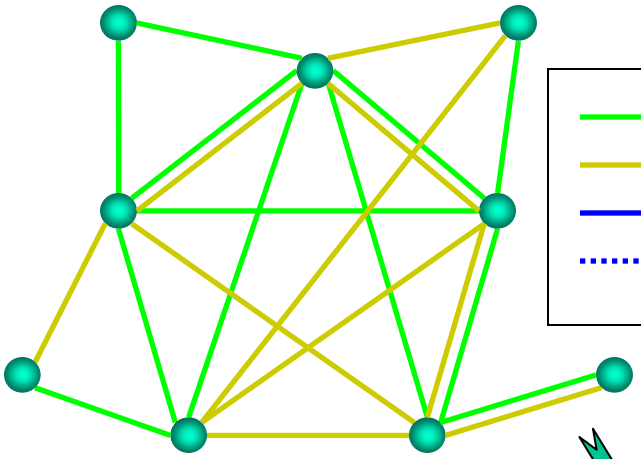
Gold-Standards



$L_1 = (4/4)/(3/6) = 2$

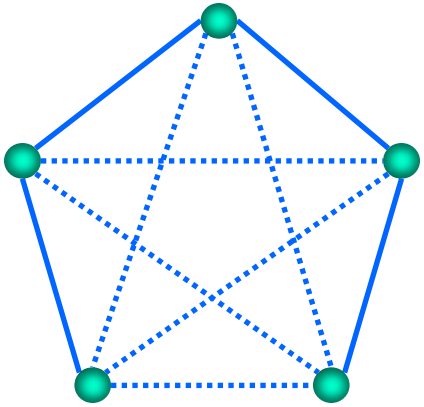
[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]

Prediction of protein interactions: Bayesian integration

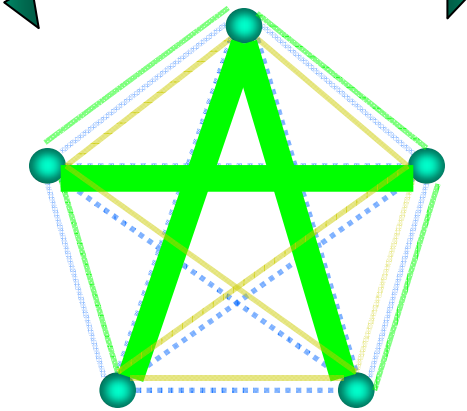


Network

Feature 1, e.g. co-expression
Feature 2, e.g. Y2H
Gold-standard +
Gold-standard –



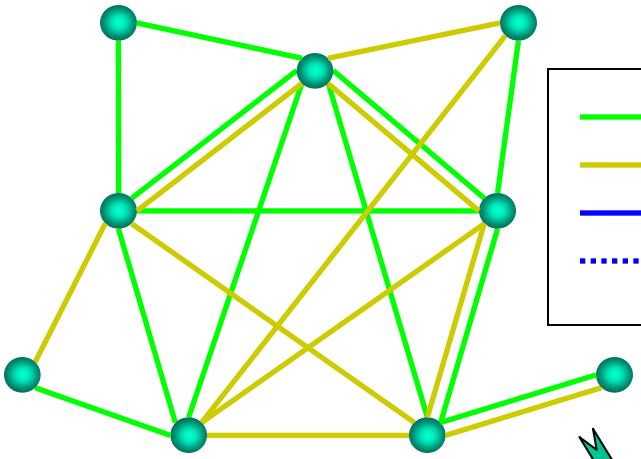
Gold-Standards



$L_1 = (4/4)/(3/6) = 2$

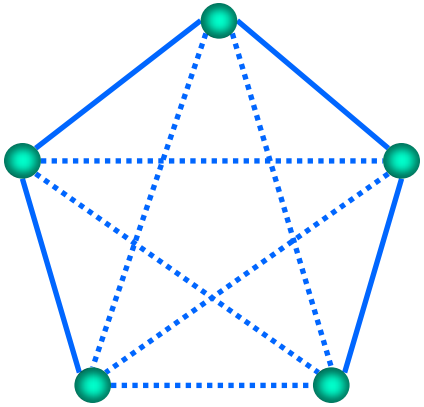
[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]

Prediction of protein interactions: Bayesian integration

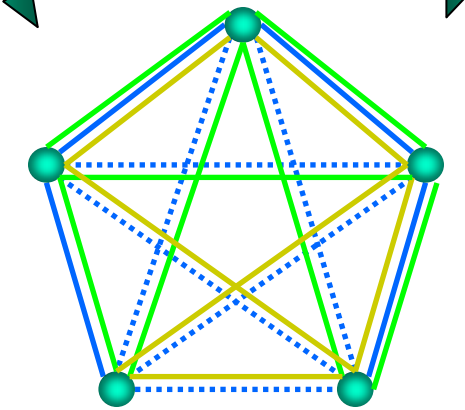


Network

Feature 1, e.g. co-expression
Feature 2, e.g. Y2H
Gold-standard +
Gold-standard –



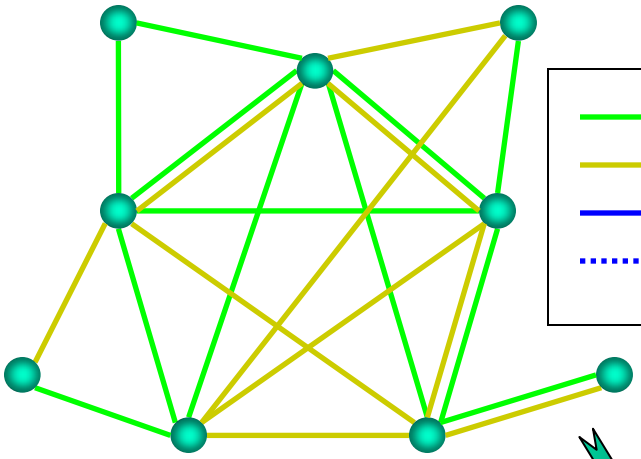
Gold-Standards



$L_1 = (4/4)/(3/6) = 2$

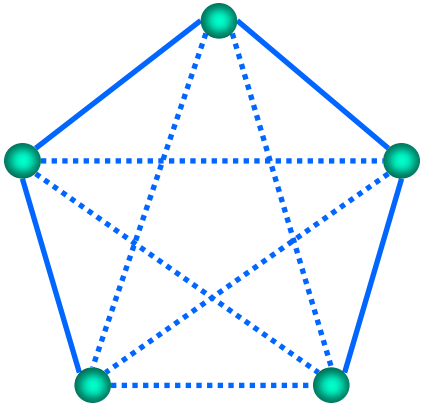
[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]

Prediction of protein interactions: Bayesian integration

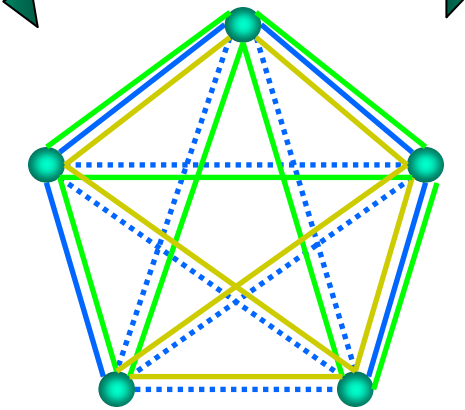


Network

Feature 1, e.g. co-expression
Feature 2, e.g. Y2H
Gold-standard +
Gold-standard –



Gold-Standards



$L_1 = (4/4)/(3/6) = 2$
$L_2 = (3/4)/(3/6) = 1.5$

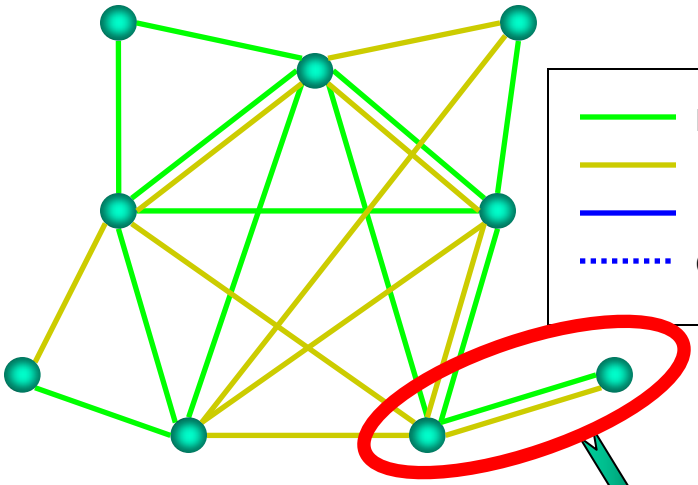
For each protein pair:
$LR = L_1 \times L_2$
$\log(LR) = \log(L_1) + \log(L_2)$

Weighted Voting

(Assuming uncorrelated features and Naïve Bayes)

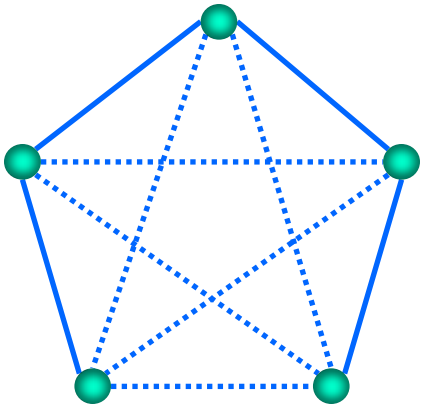
[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]

Prediction of protein interactions: Bayesian integration

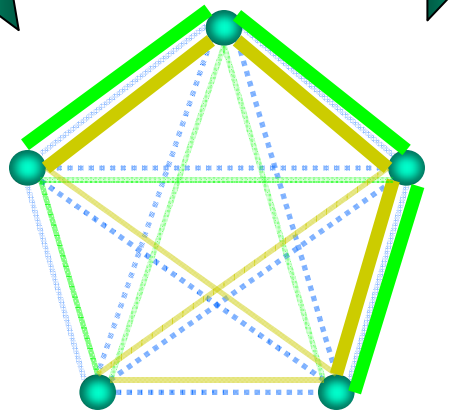


Network

Feature 1, e.g. co-expression
Feature 2, e.g. Y2H
Gold-standard +
Gold-standard –



Gold-Standards



$L_1 = (4/4)/(3/6) = 2$
$L_2 = (3/4)/(3/6) = 1.5$

For each protein pair:
$LR = L_1 \times L_2$
$\log(LR) = \log(L_1) + \log(L_2)$

[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]

Protein-protein interaction (PPI) network

Network reconstruction

Model organism: baker's yeast

- Size:
  - ~6,000 for yeast
    - → Computational cost: ~18M pairs
  - ~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)

Many Previous approaches in predicting PPI

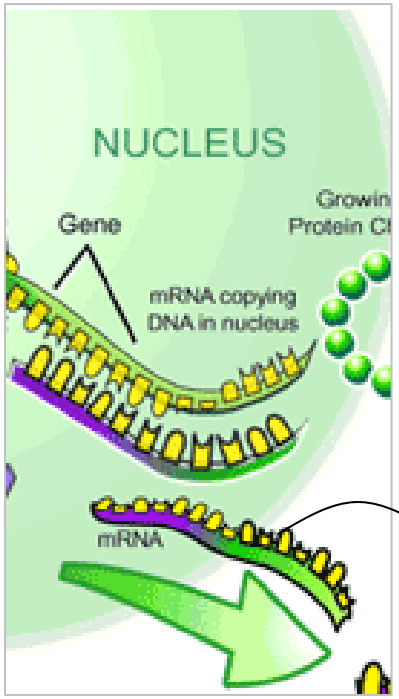
Network reconstruction

- 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)

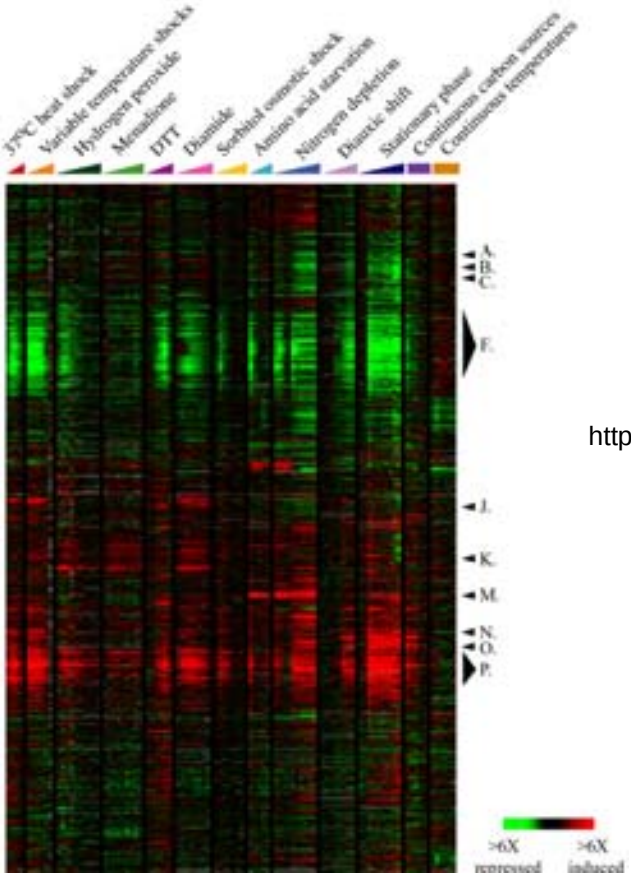
- **DREAM**

Features for predicting PPI – functional genomics

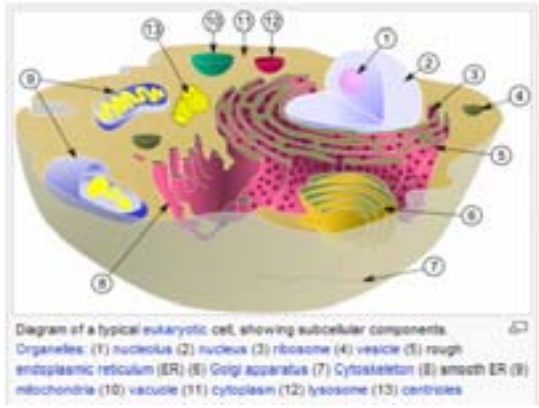
Network reconstruction



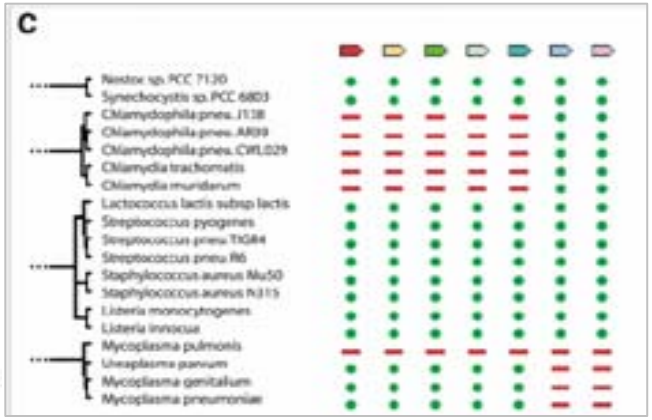
Microarray gene expression



Gasch et al., 2000



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

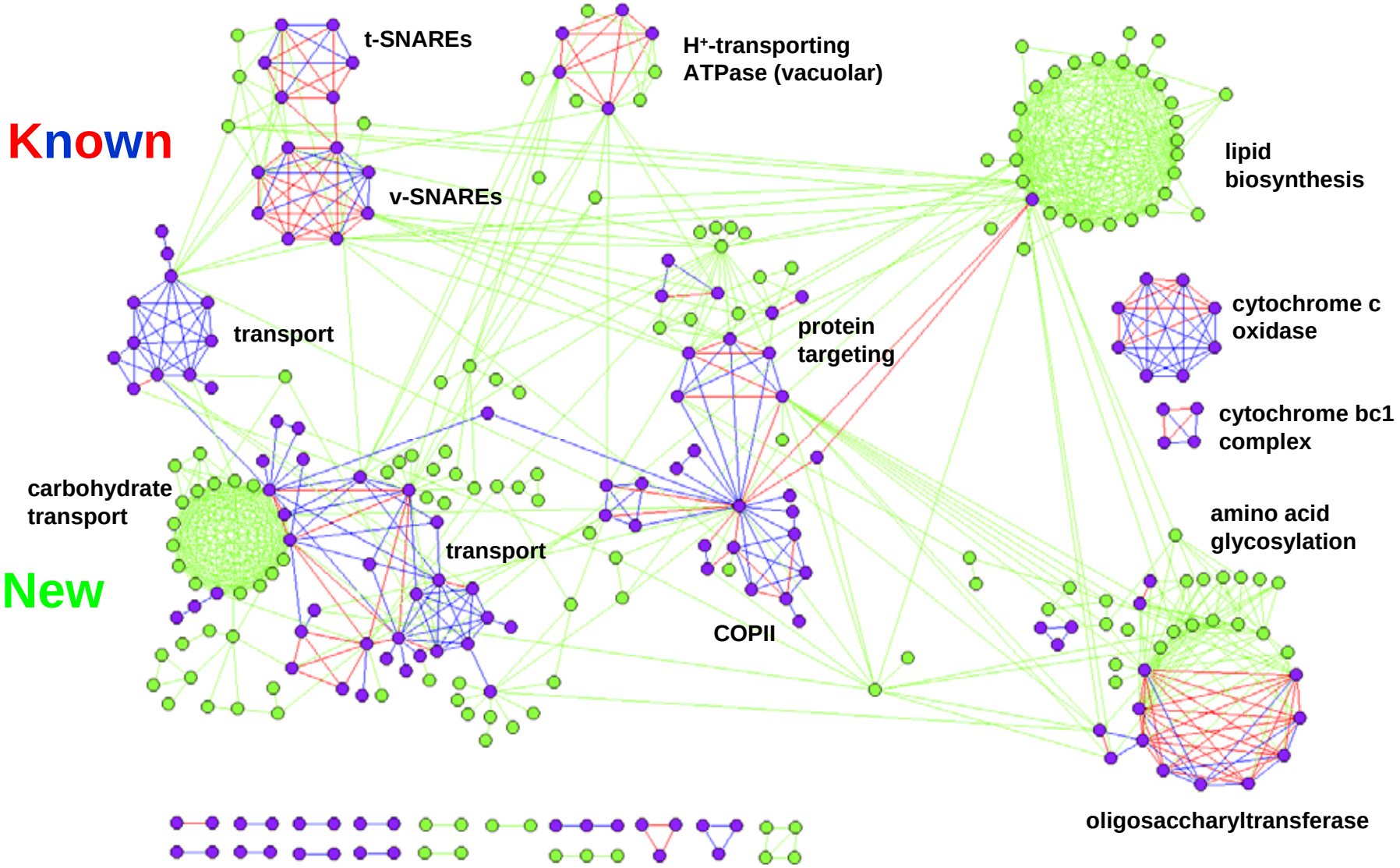


Phylogenetic profiles
Von Mering et al., 2003

$x = (0.23, 2.41, 1.52, \ldots)$

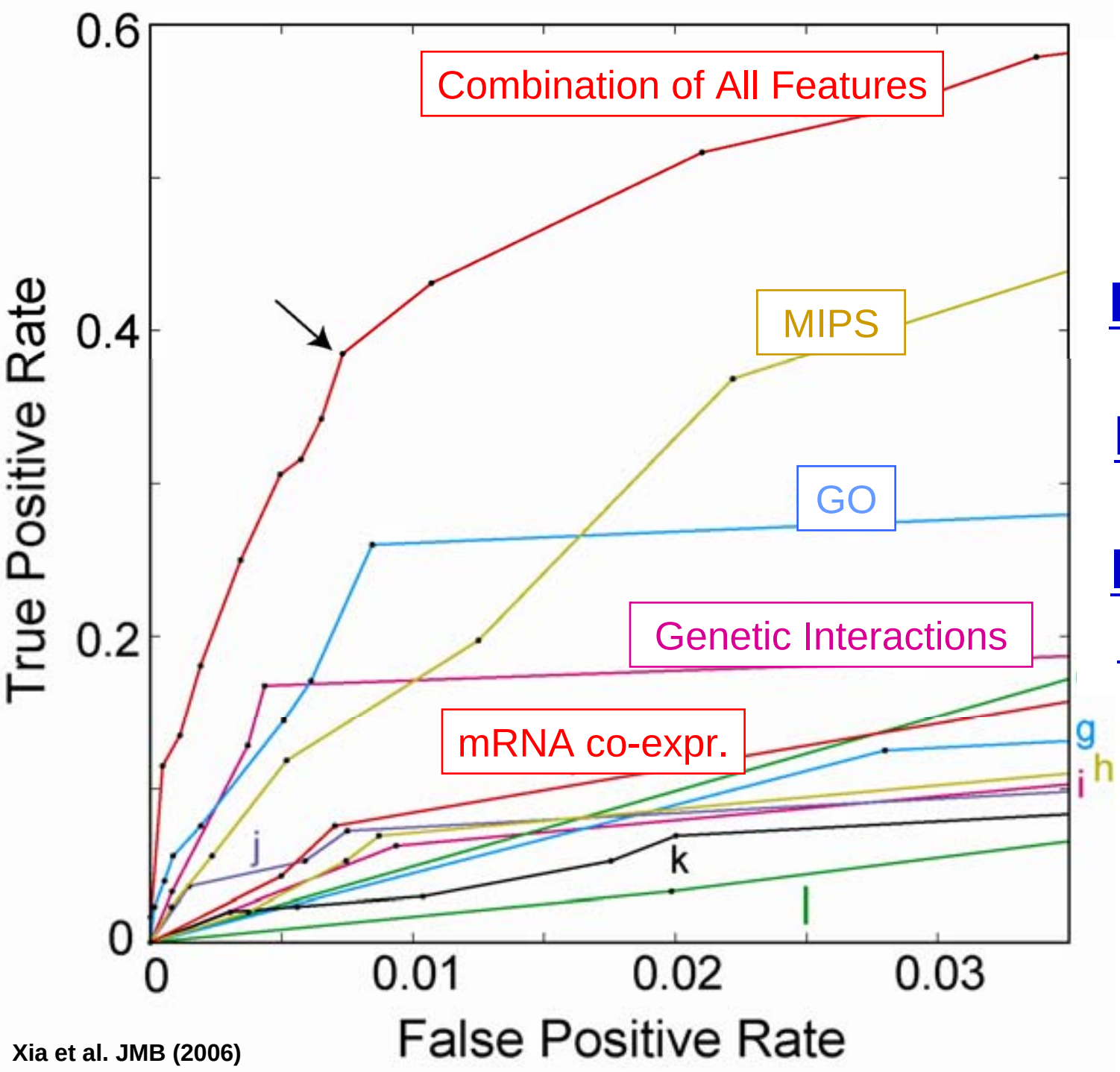
$x = (1, 1, 0, 0, 0, 0, 0, 1, 1, \ldots)$

Map of Known and Predicted Membrane Protein Interactome in Yeast



Xia et al. JMB (2006)

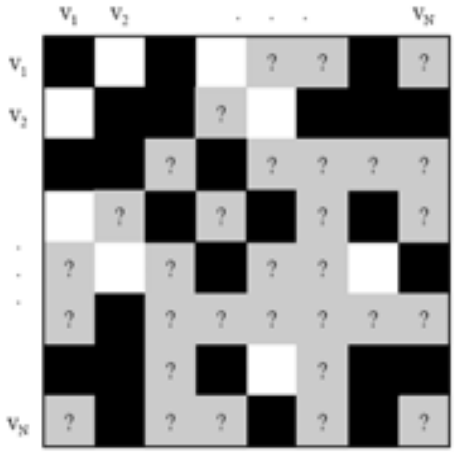
Individual Features and their Integration for Yeast Membrane Protein Interaction Prediction



Xia et al. JMB (2006)

Problem with Network Prediction

- Training sets too small
- Known examples are unevenly spread amongst space one is doing prediction on
- Particularly afflicts kernel methods

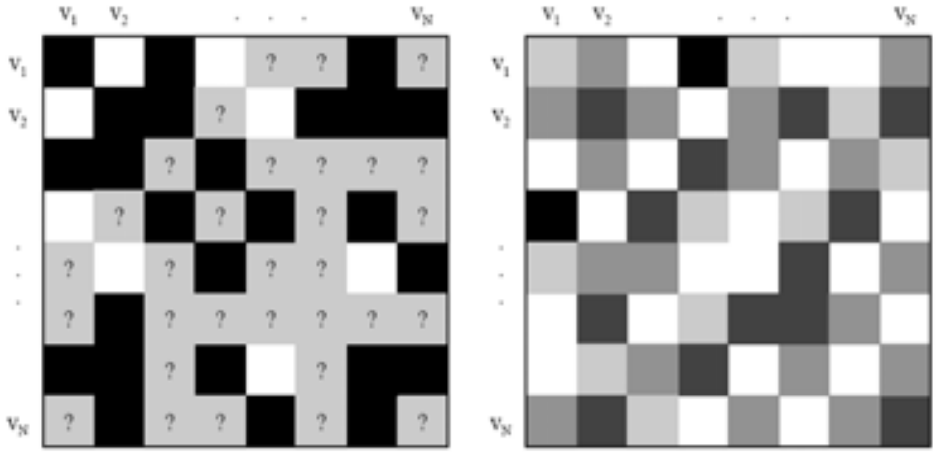


[Yip et al., Bioinformatics ('09, in press)]

Kernel Methods

Network reconstruction

- Kernel: similarity matrix
- Positive semi-definiteness of kernel → similarity values correspond to inner products in an embedded space
- Good for integrating different kinds of data
  - DNA sequences: strings
  - Gene expression: real numbers
  - Phylogenetic profiles: binary numbers



[Yip et al., Bioinformatics ('09, in press)]

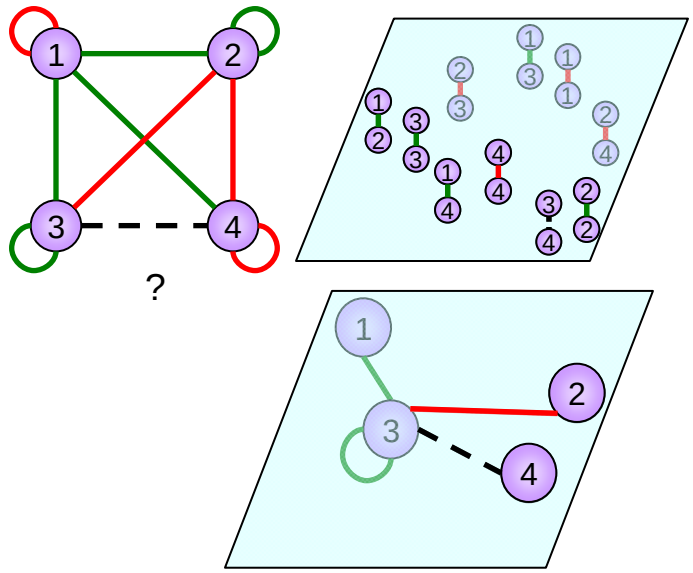
Local v Global Modelling

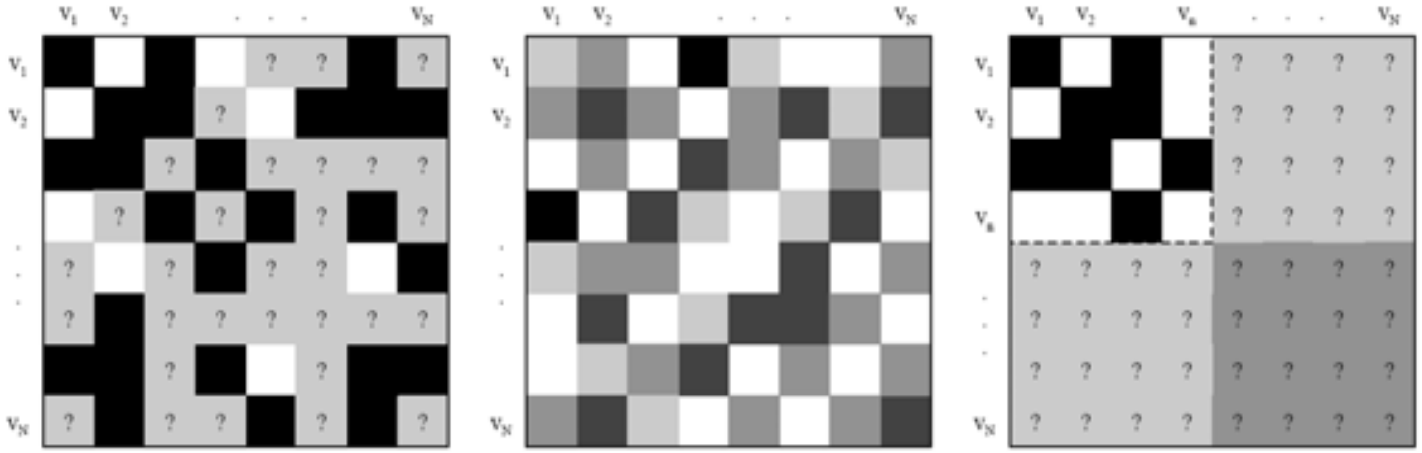
Global modeling

- Pairwise kernel (Ben-Hur and Noble, 2005)
  - $O(n^2)$ instances, $O(n^4)$ kernel elements

Local modeling

- Bleakley et al., 2007: global model may not fit sub-classes well → learn one local model per protein
  - Flexible
  - Lack of training data





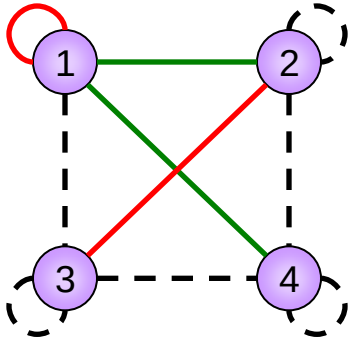
[Yip et al., Bioinformatics ('09, in press)]

Our method: 1. prediction propagation

Network reconstruction – Training set expansion

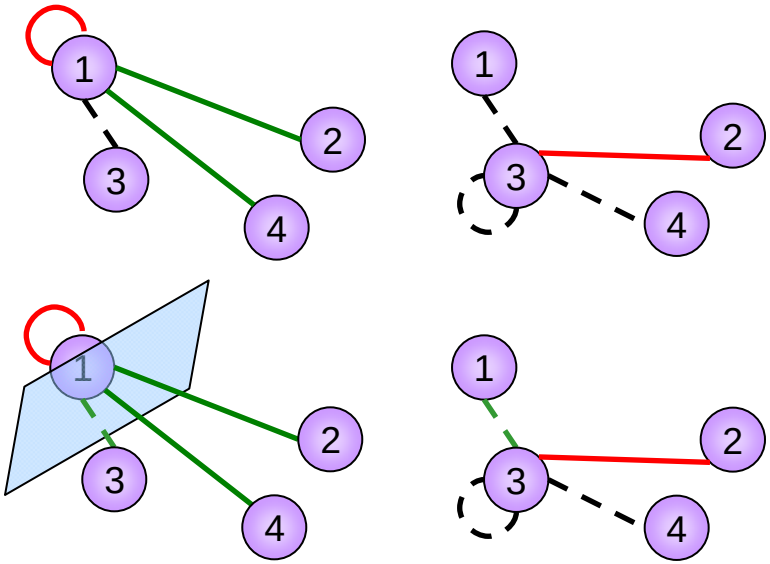
Goals:

- Preserve the flexibility of local modeling
- Tackle the issue of insufficient training examples



Idea 1: prediction propagation

- Motivation: some objects have more examples than others
- Learn models for proteins with more examples first
- Use distance to separating hyperplane to measure confidence
- Propagate the most confident predictions



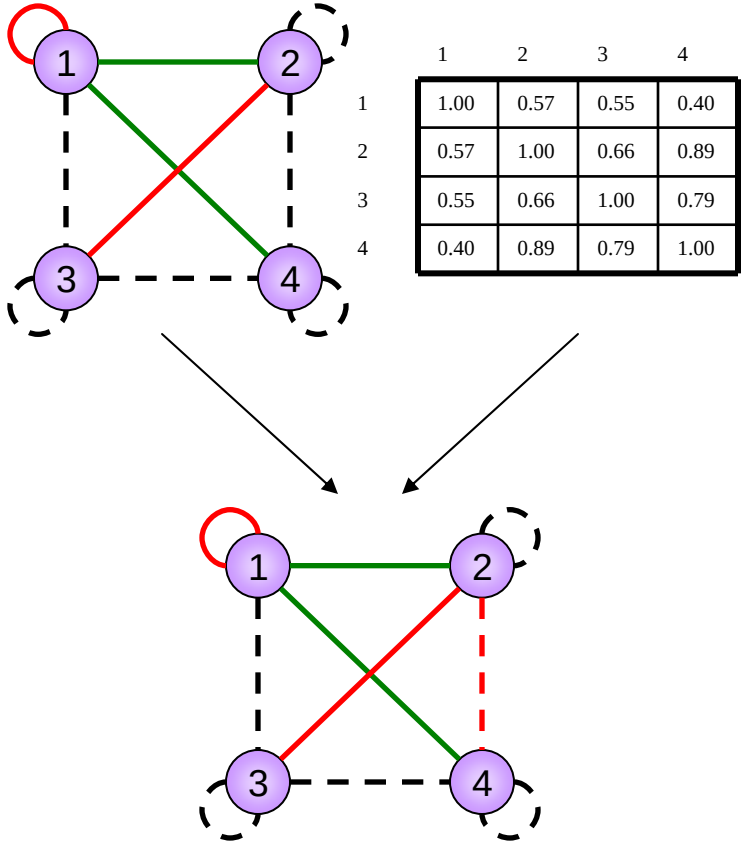
[Yip et al., Bioinformatics ('09, in press)]

Our method: 2. kernel Initialization

Network reconstruction – Training set expansion

Idea 2: kernel initialization

- Motivation: what if most objects have very few examples?
- Add the most similar pairs to training set



	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

Remarks

Network reconstruction – Training set expansion

- Can use 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

Experiments

Network reconstruction – Training set expansion

Predicting the BioGRID-10 dataset

- Gold-standard: all physical interactions in 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 et al., Bioinformatics ('09, in press)]

Results

Network reconstruction – Training set expansion

Accuracy – %AUC (area under receiver operator curve):

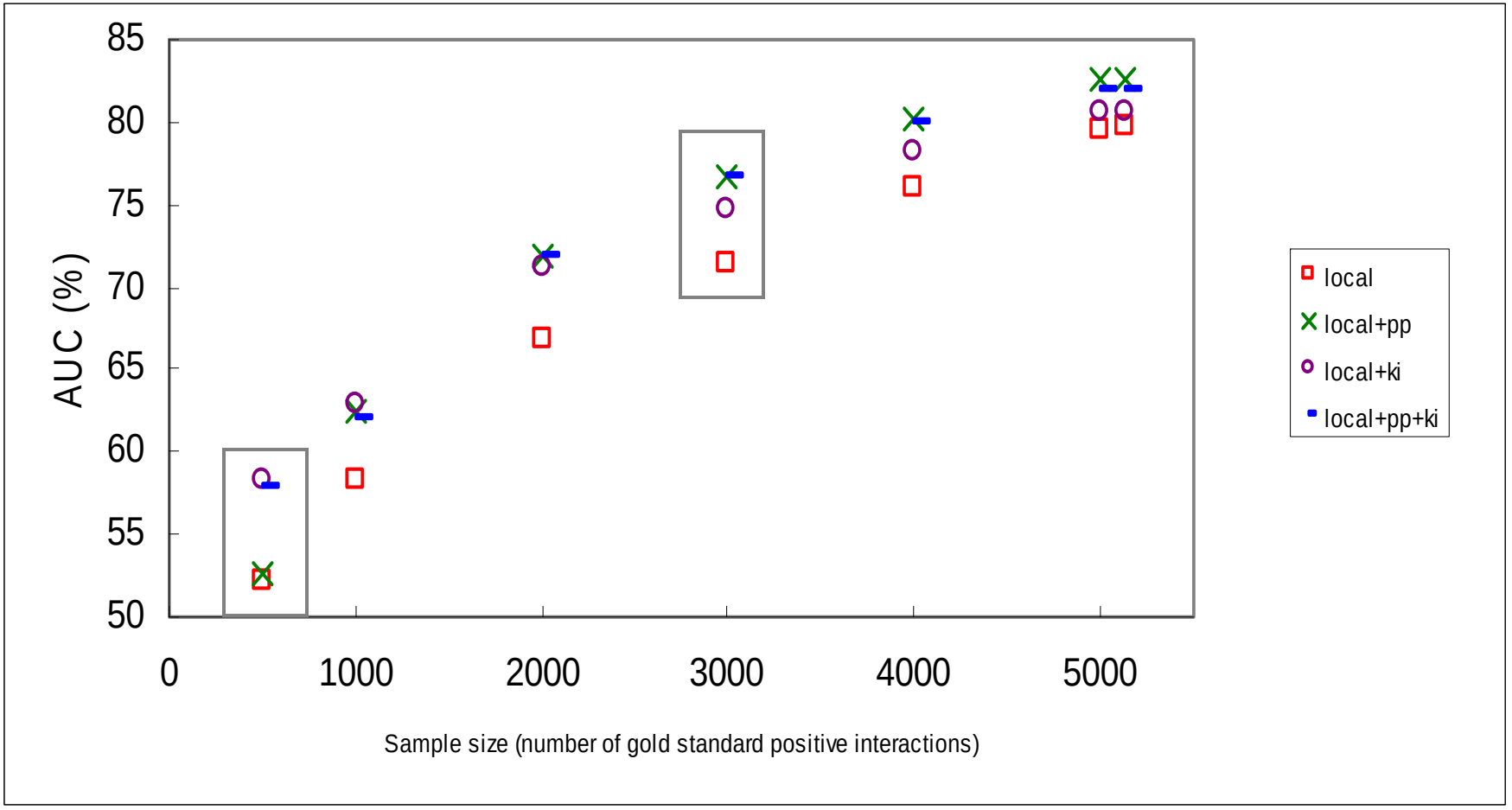
	phy	loc	exp-gasch	exp-spellman	y2h-ito	y2h-uetz	tap-gavin	tap-krogan	int
Mode 1									
direct	58.04	66.55	64.61	57.41	51.52	52.13	59.37	61.62	70.91
kCCA	65.80	63.86	68.98	65.10	50.89	50.48	57.56	51.85	80.98
kML	63.87	68.10	69.67	68.99	52.76	53.85	60.86	57.69	73.47
em	71.22	75.14	67.53	64.96	55.90	53.13	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

- Highest accuracy by training set expansion
- Overfitting of local modeling without training set expansion
- Comparing prediction propagation and kernel initialization

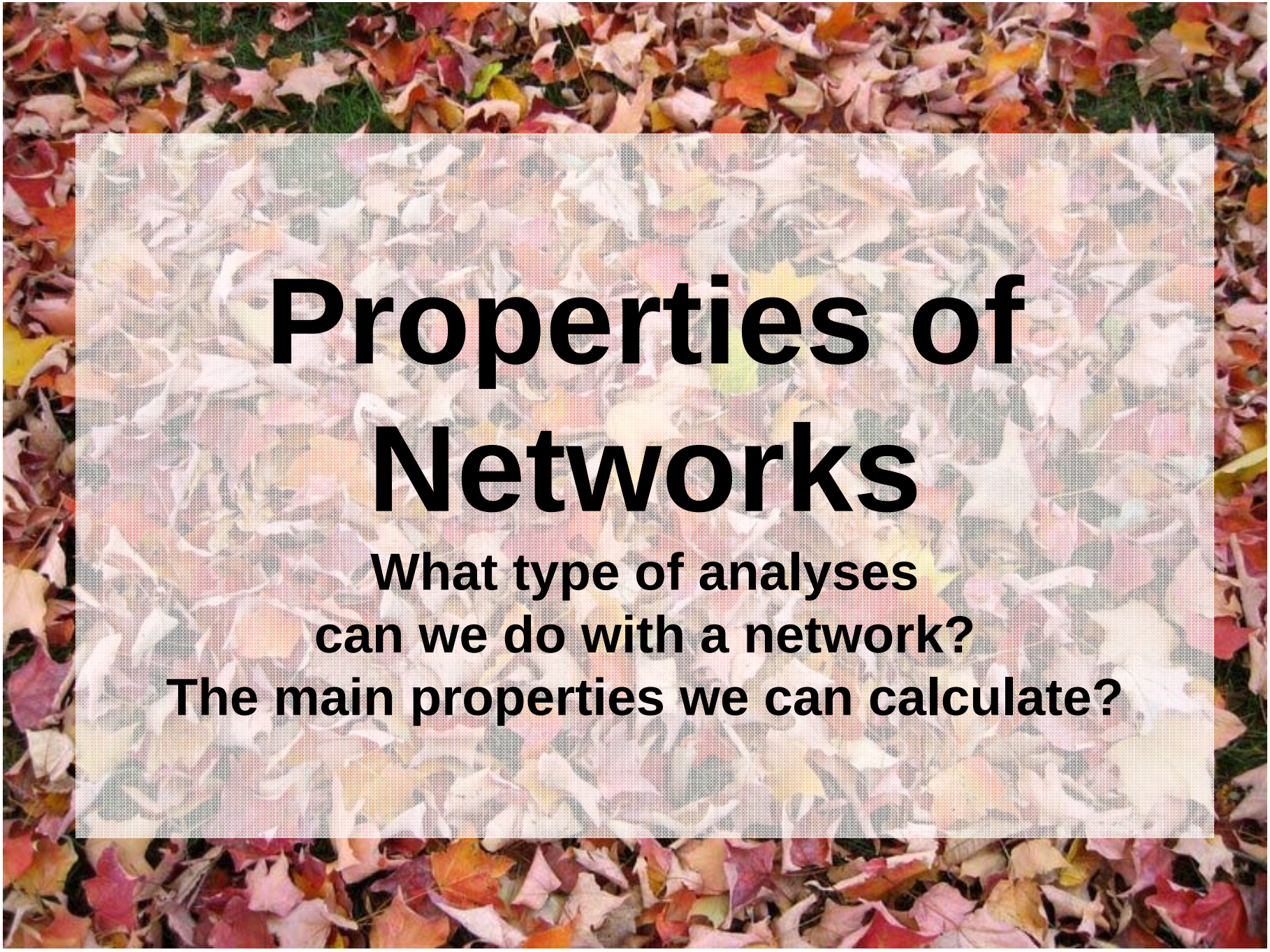
[Yip et al., Bioinformatics ('09, in press)]

Complementarity of the two methods

Network reconstruction – Training set expansion

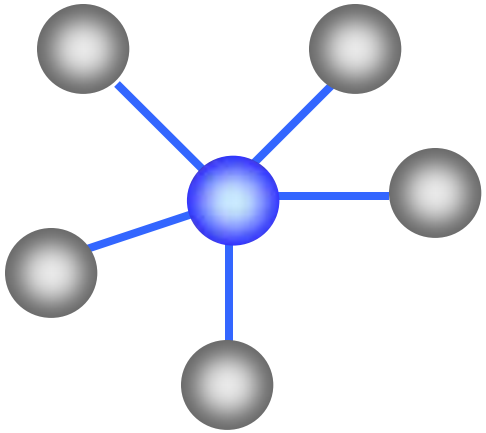


[Yip et al., Bioinformatics ('09, in press)]


Properties of Networks
What type of analyses
can we do with a network?
The main properties we can calculate?

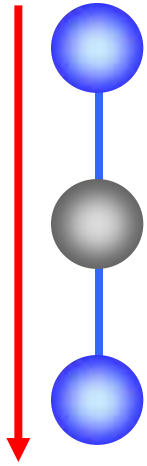
Global topological measures

Indicate the gross topological structure of the network



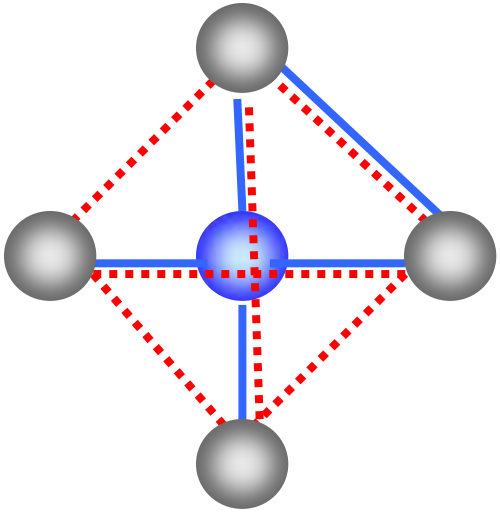
Degree (K)

5



Path length (L)

2



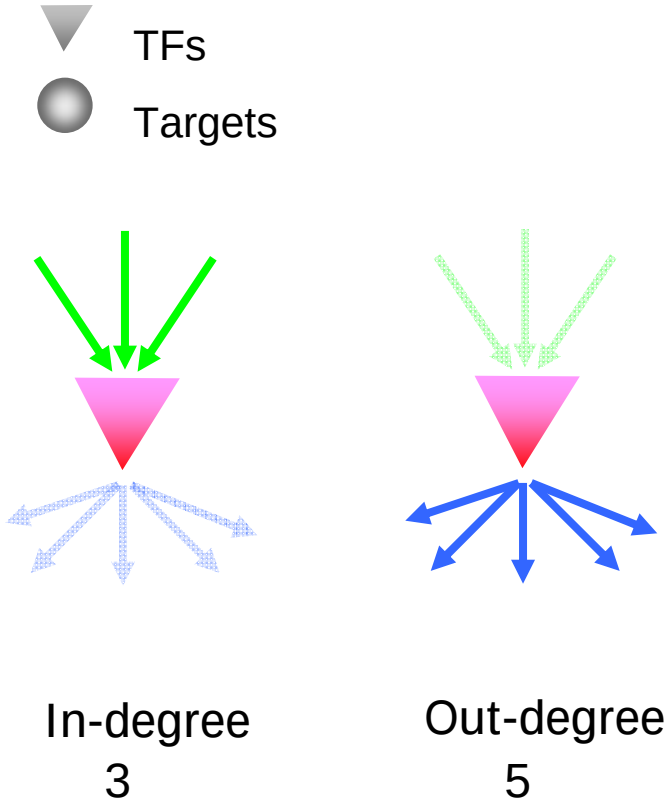
Clustering coefficient (C)

1/6

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

[Barabasi]

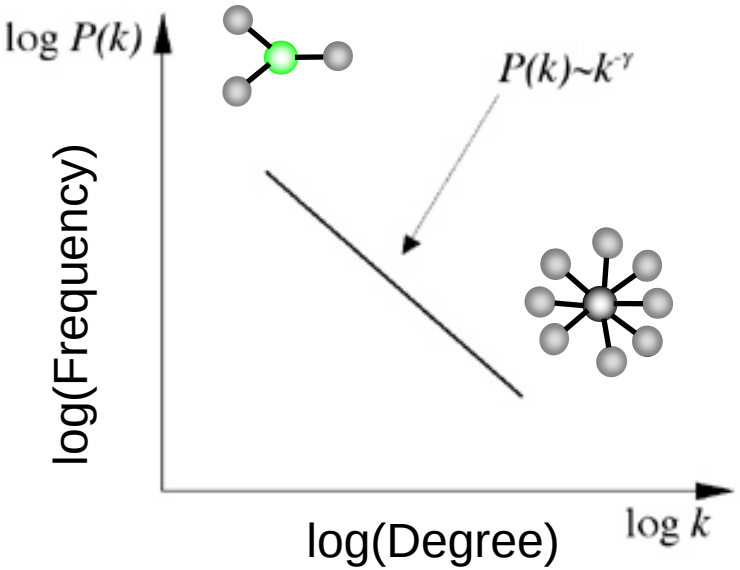
Global topological measures for directed networks

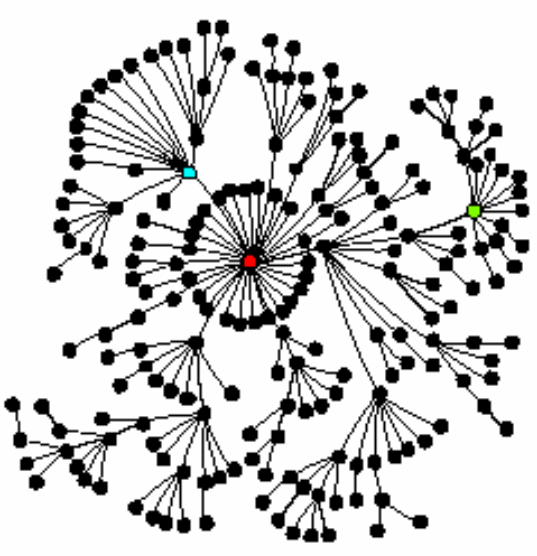


Regulatory and metabolic networks are ***directed***

Scale-free networks

Power-law distribution





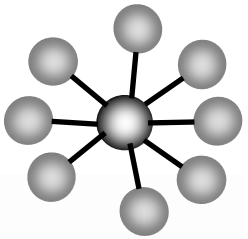
Hubs dictate the structure of the network

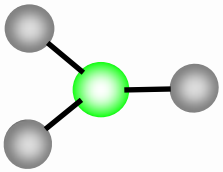
[Barabasi]

Hubs tend to be Essential

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

[Lauffenburger, Barabasi]



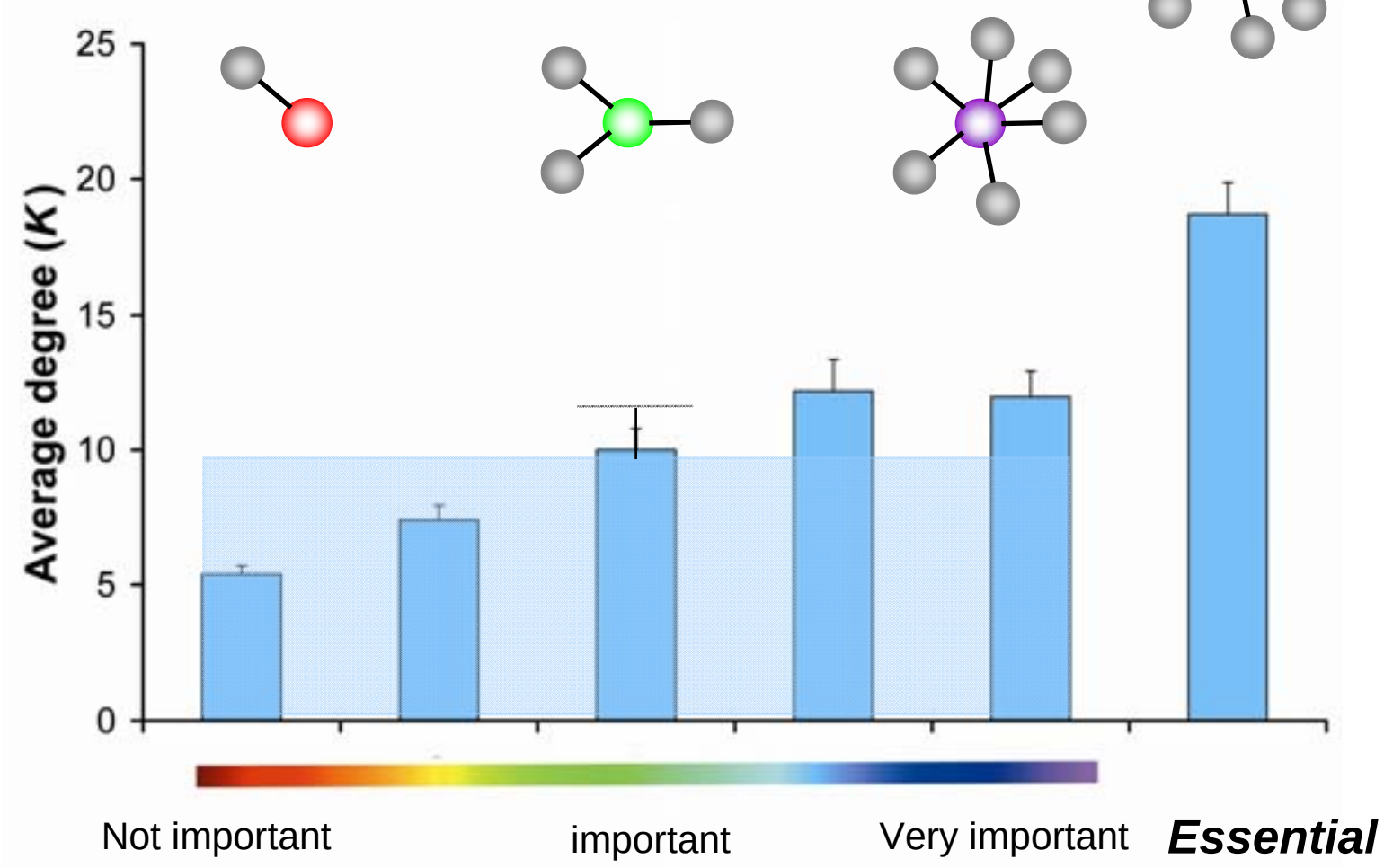


"hubbiness"

[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"

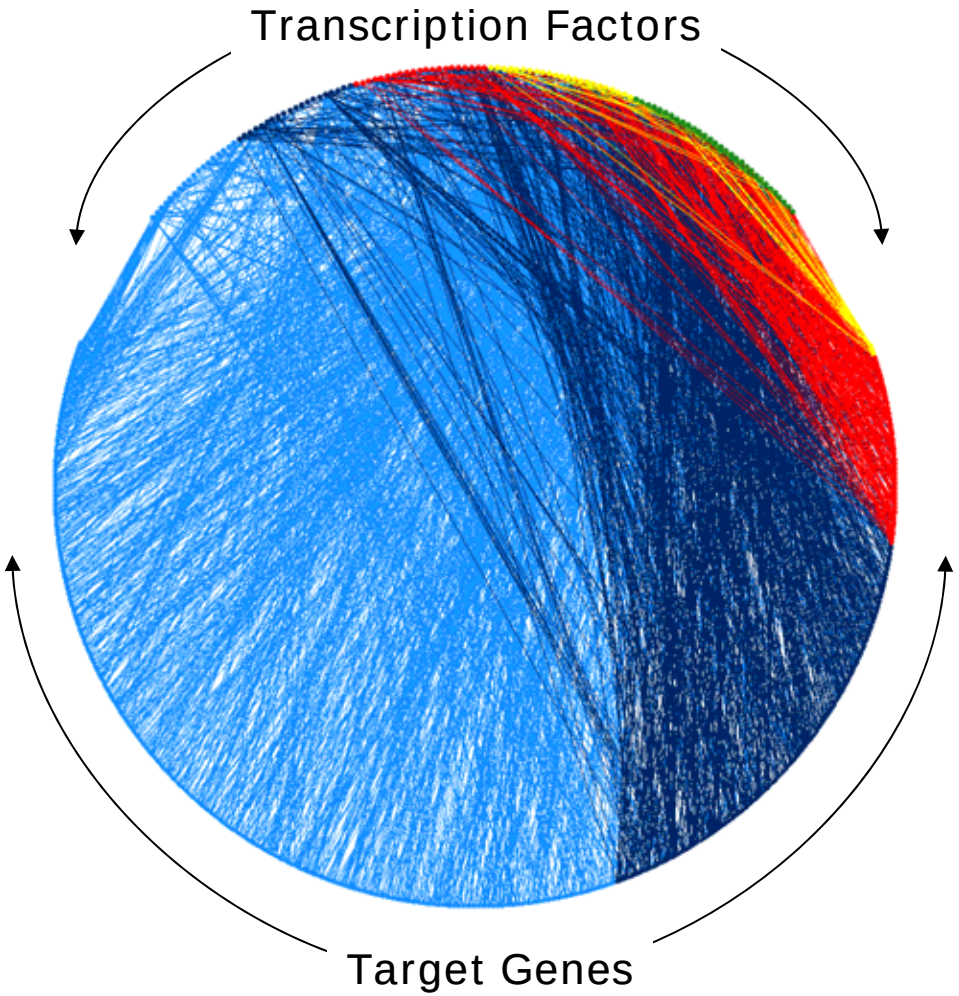


[Yu et al., 2003, TIG]

Network Dynamics #1: Cellular States

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

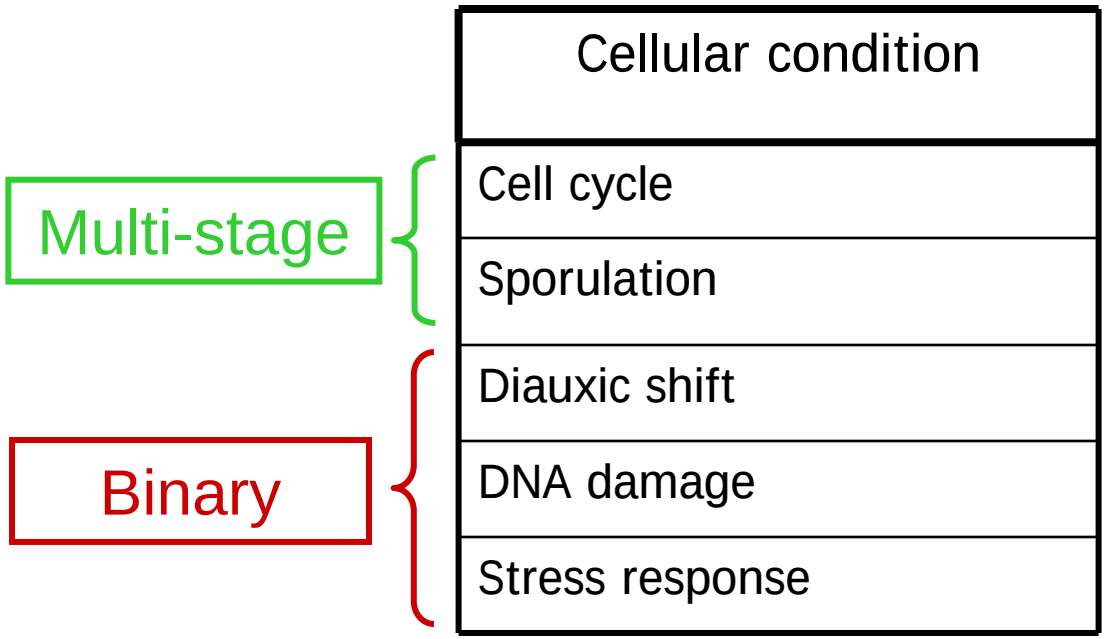
Dynamic Yeast TF network



- 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

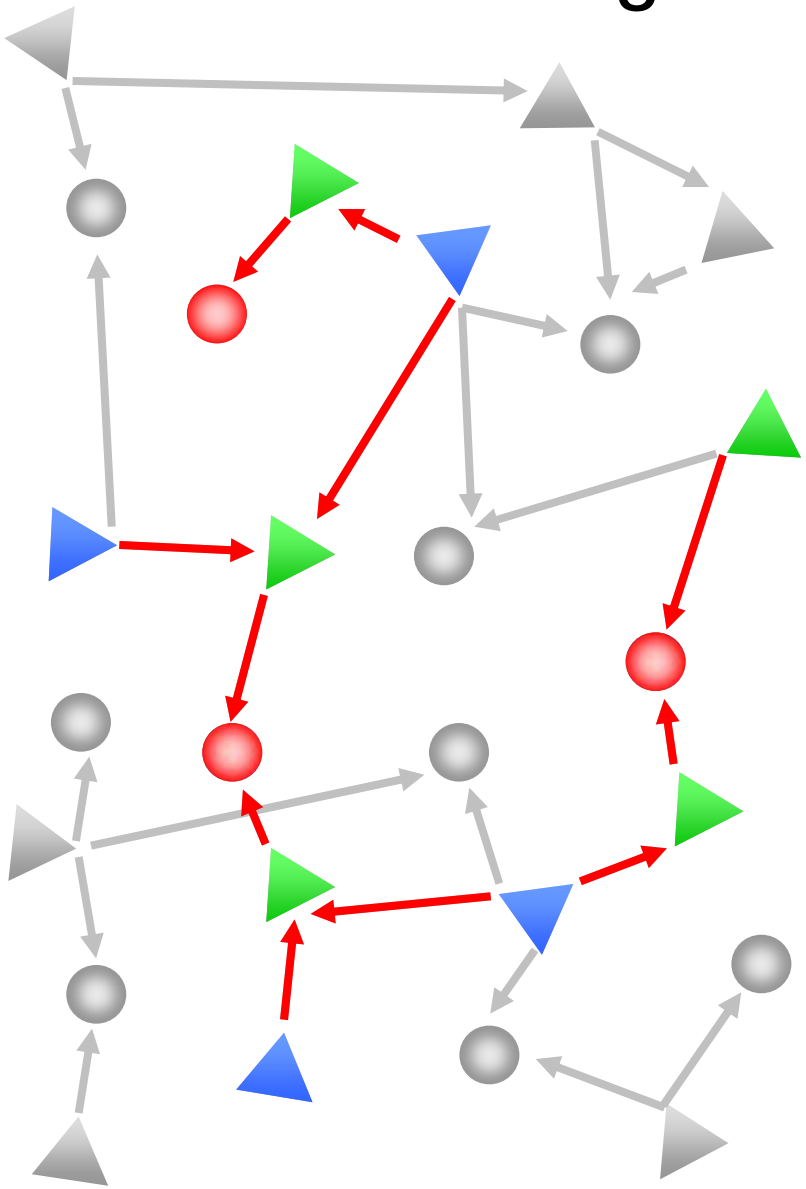
Luscombe et al. Nature 431: 308

Gene expression data for five cellular conditions in yeast



[Brown, Botstein, Davis….]

Backtracking to find active sub-network

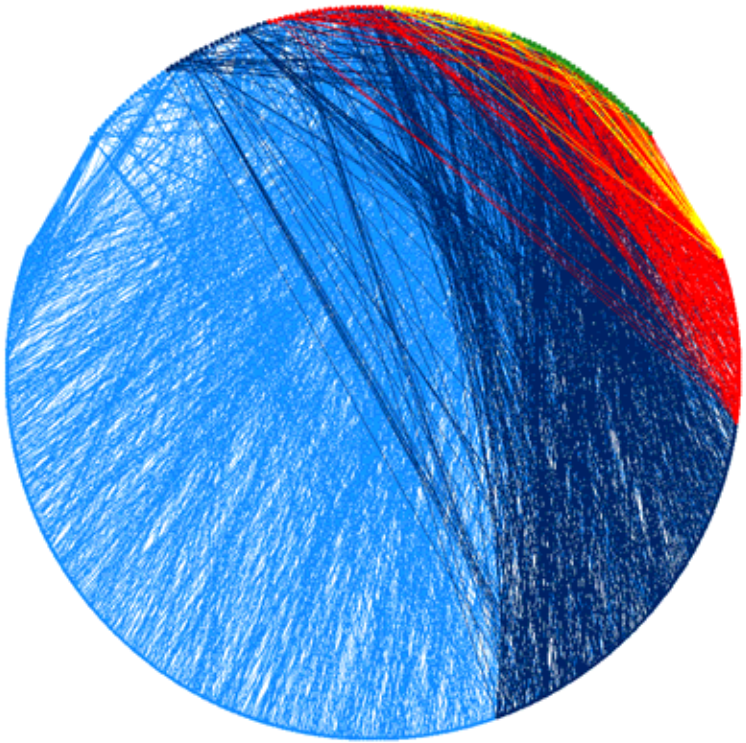


- Define differentially expressed genes
- Identify TFs that regulate these genes
- Identify further TFs that regulate these TFs

Active regulatory sub-network

Network usage under different conditions

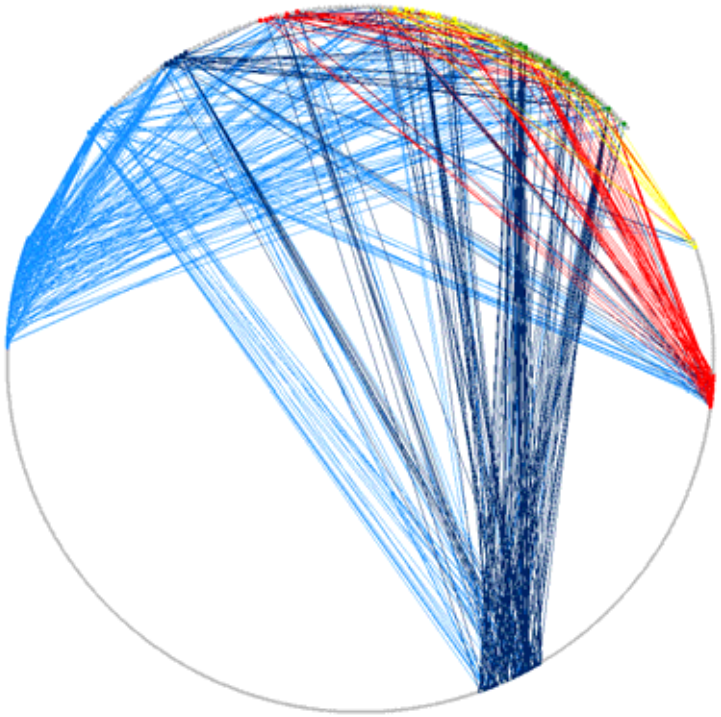
static



Luscombe et al. Nature 431: 308

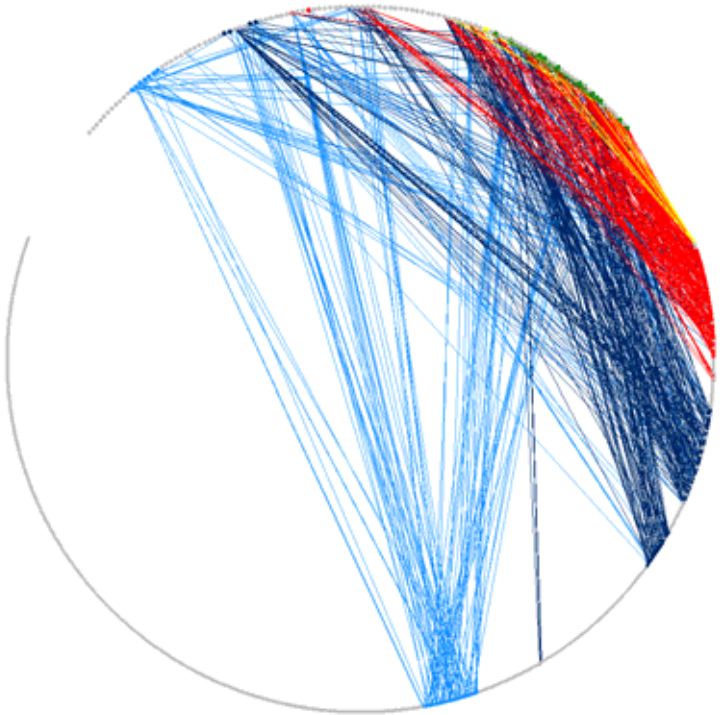
Network usage under different conditions

cell cycle



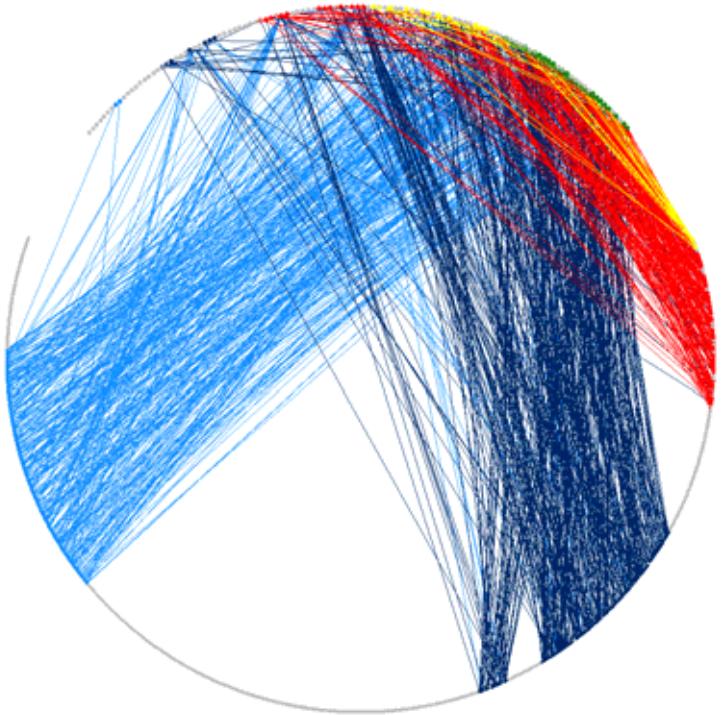
Network usage under different conditions

sporulation



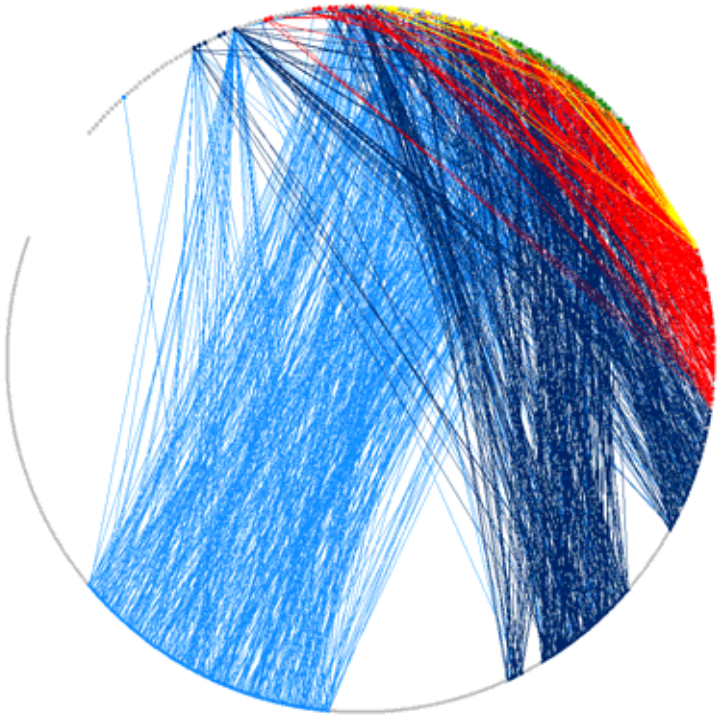
Network usage under different conditions

diauxic shift



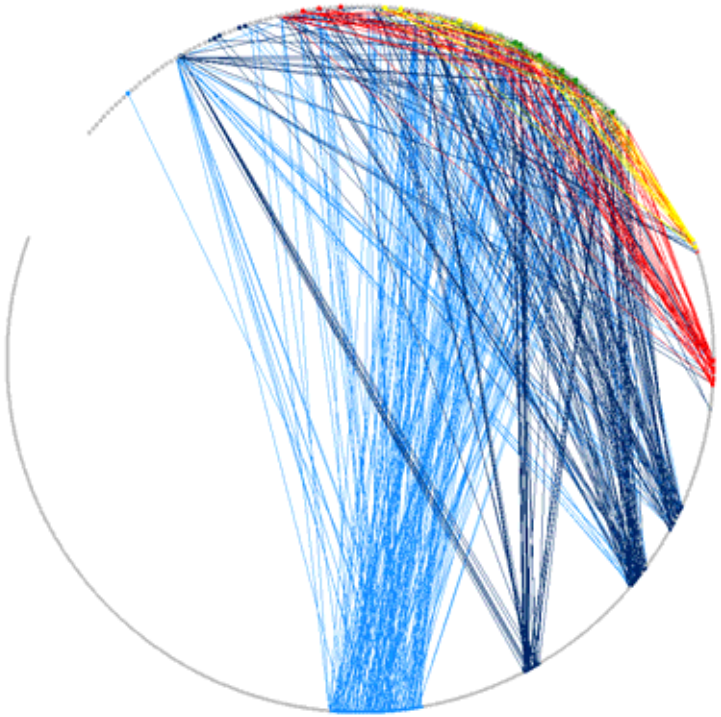
Network usage under different conditions

DNA damage



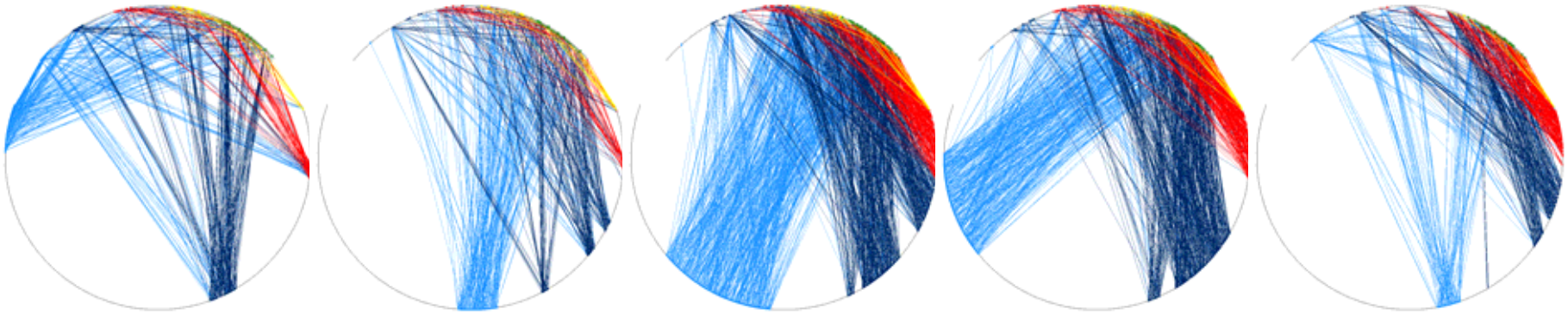
Network usage under different conditions

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:

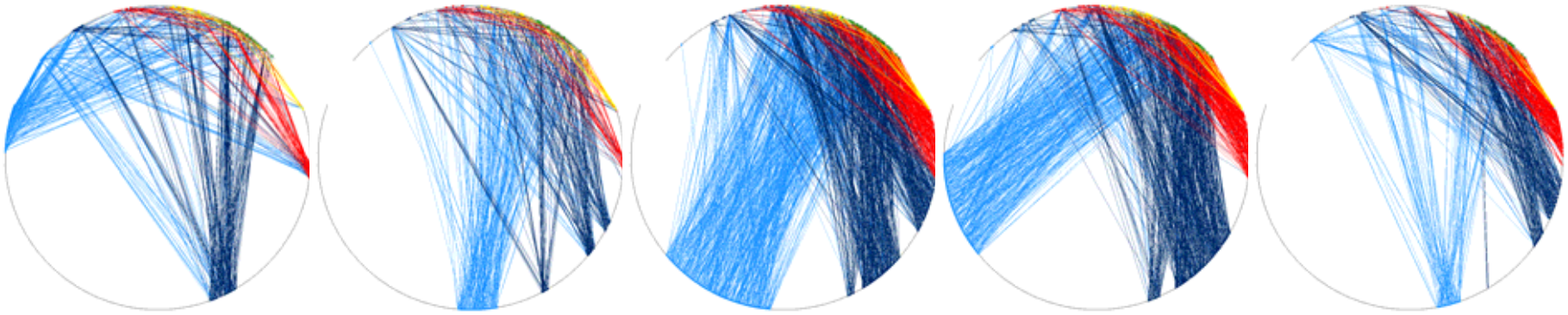
- Hub usage
- Interaction rewiring

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:

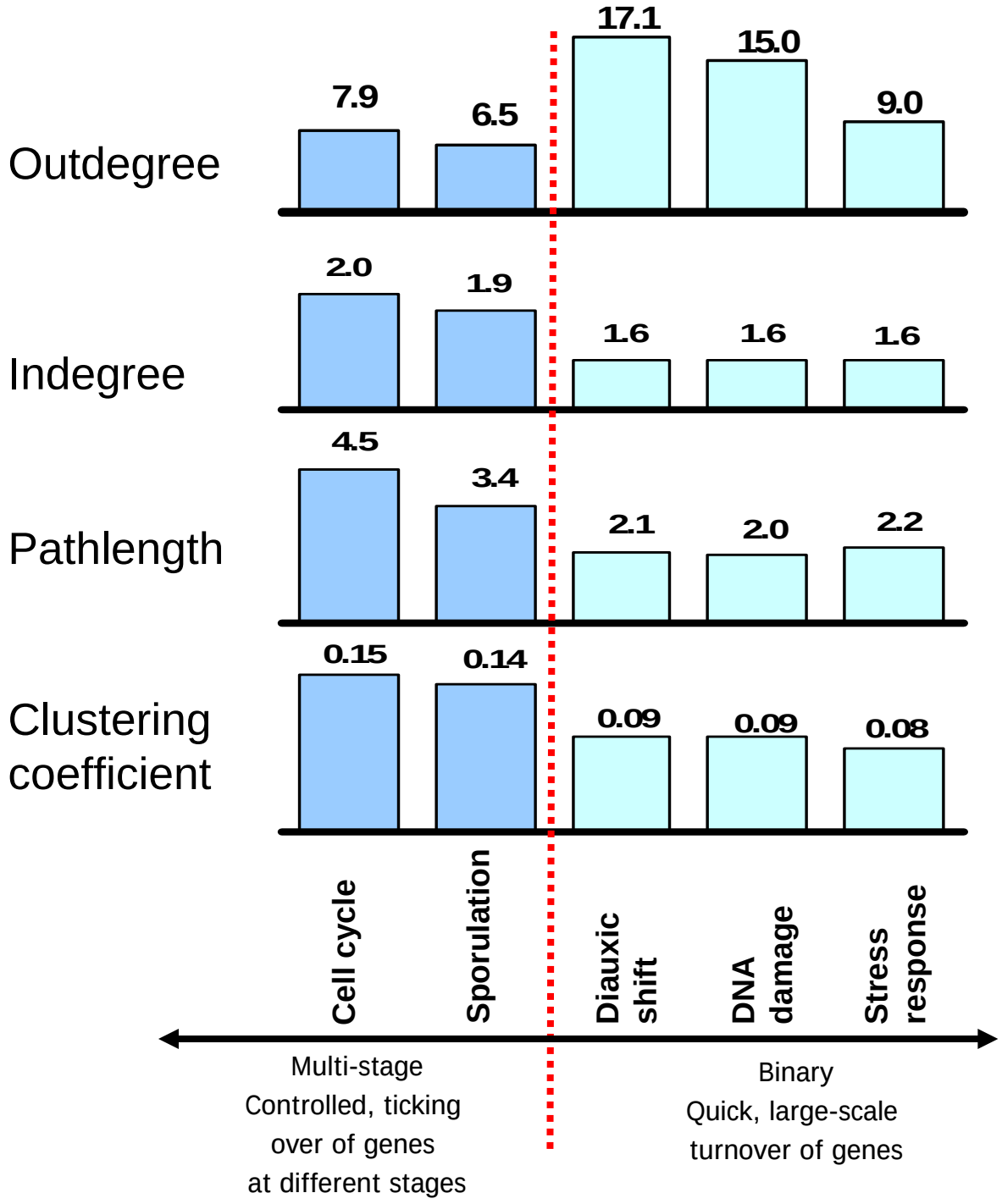
- Global topological measures
- Local network motifs

2. Newly derived follow-on statistics:

- Hub usage
- Interaction rewiring

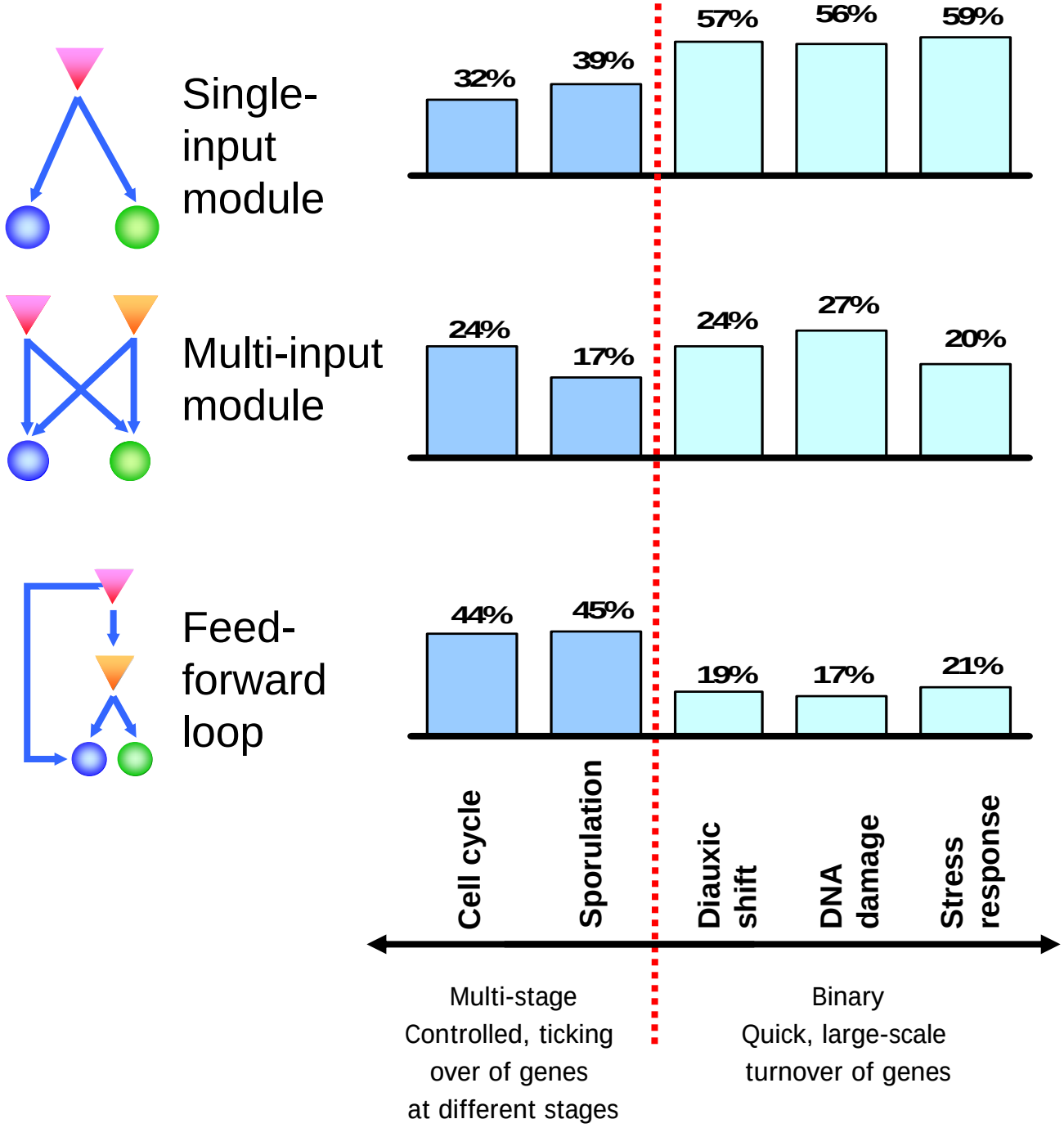
3. Statistical validation of results

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



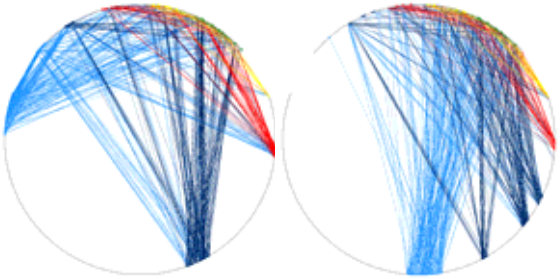
Luscombe et al. Nature 431: 308

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

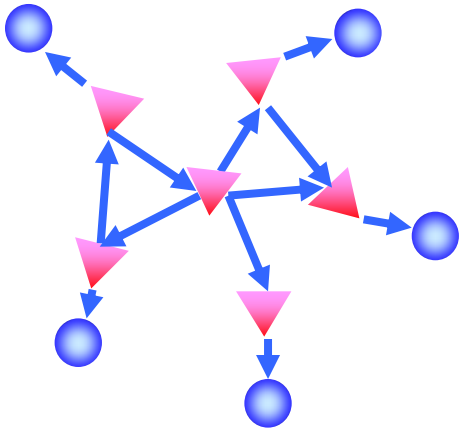


Luscombe et al. Nature 431: 308

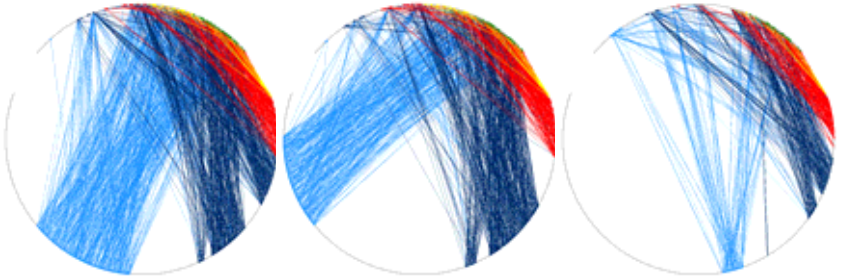
Cell cycle Sporulation



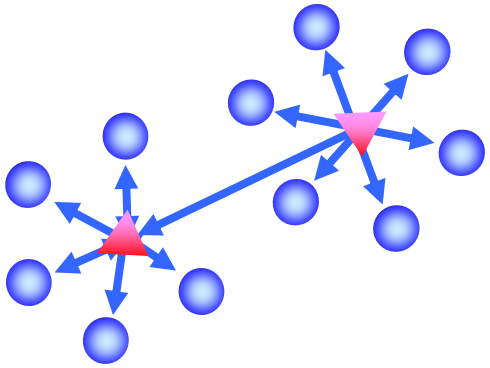
multi-stage conditions



Diauxic shift DNA damage Stress



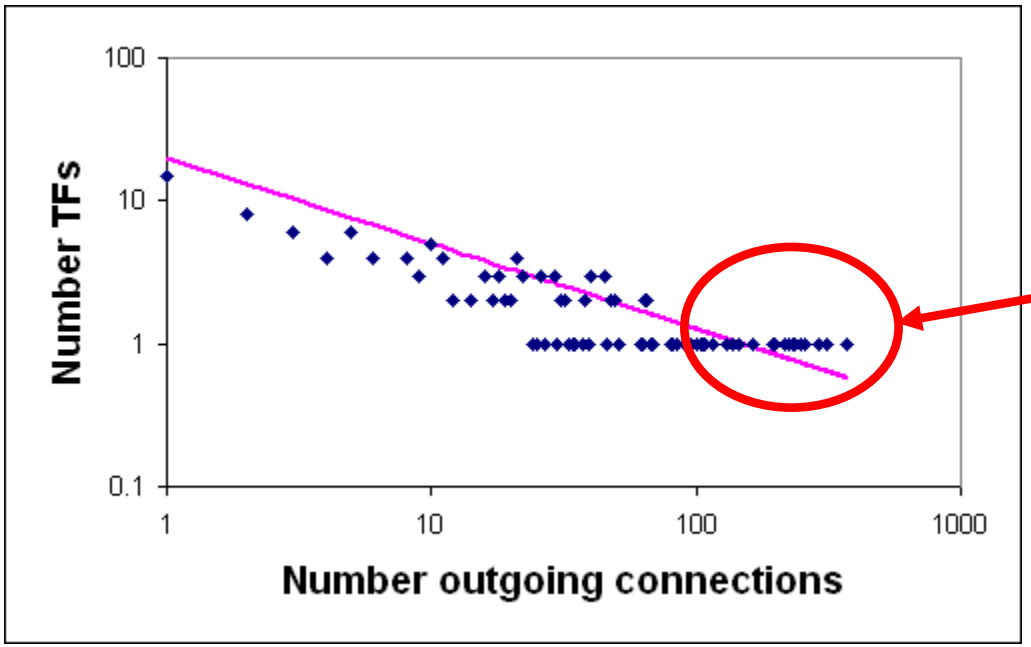
binary conditions

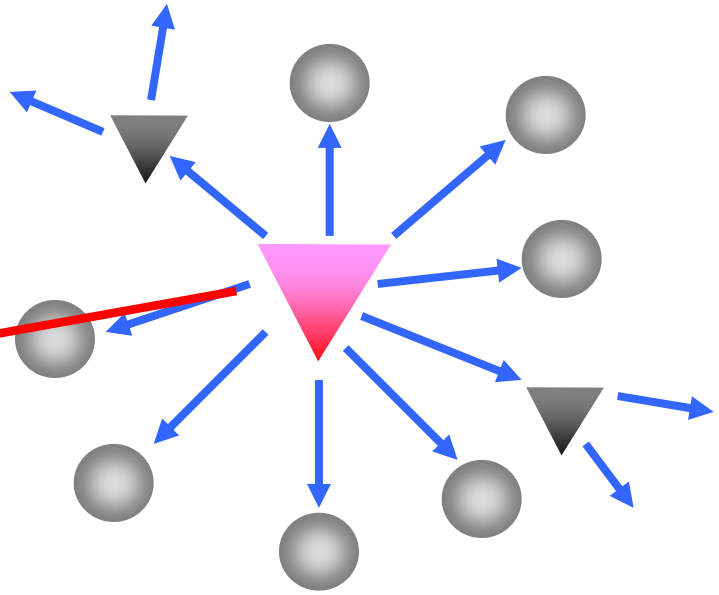


Summary

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

Transient Hubs

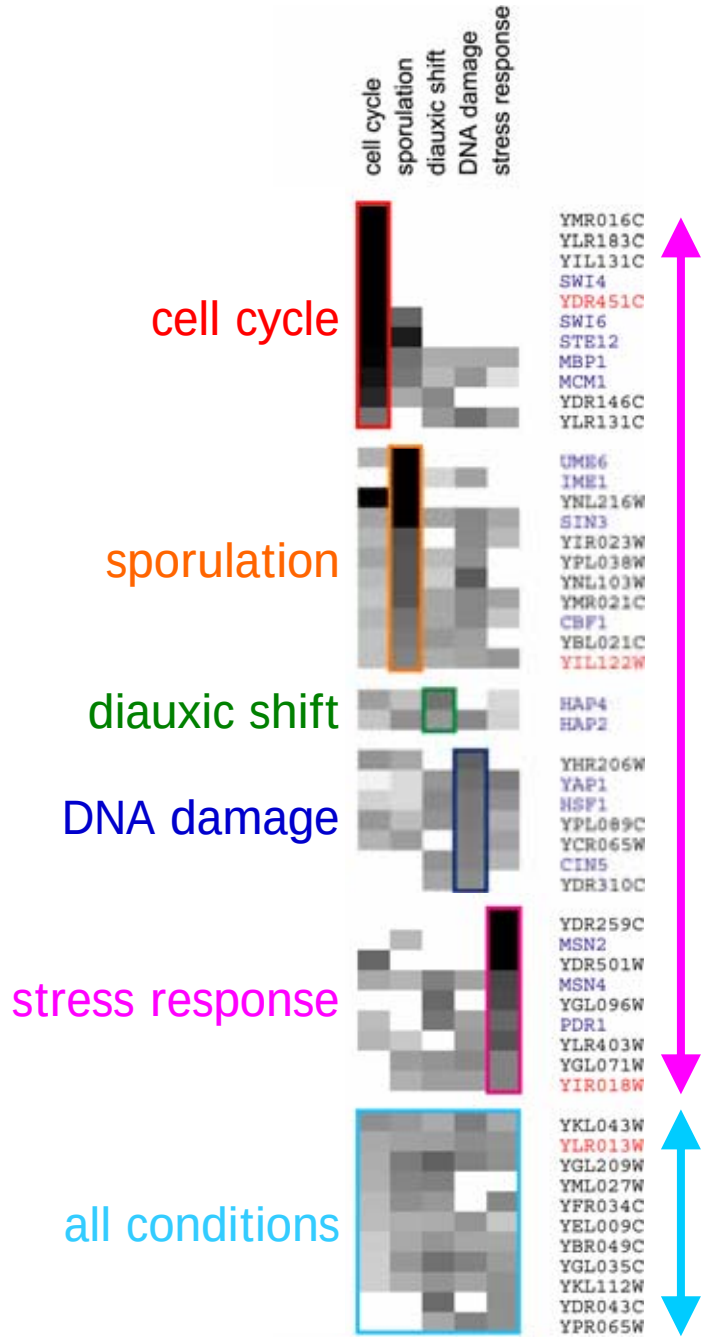




Regulatory hubs

- 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

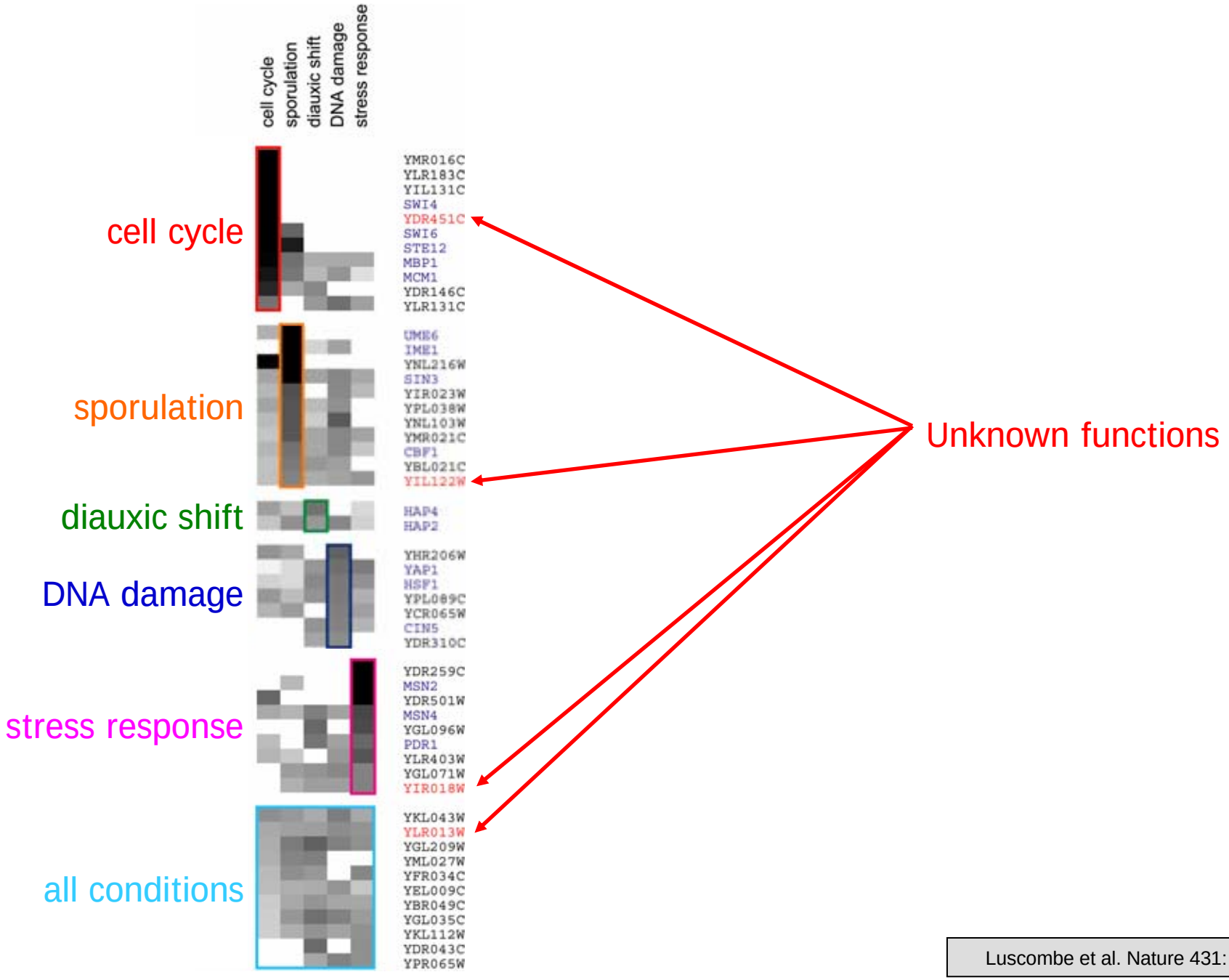
Luscombe et al. Nature 431: 308

Swi4, Mbp1

Ime1, Ume6

Msn2, Msn4

Luscombe et al. Nature 431: 308



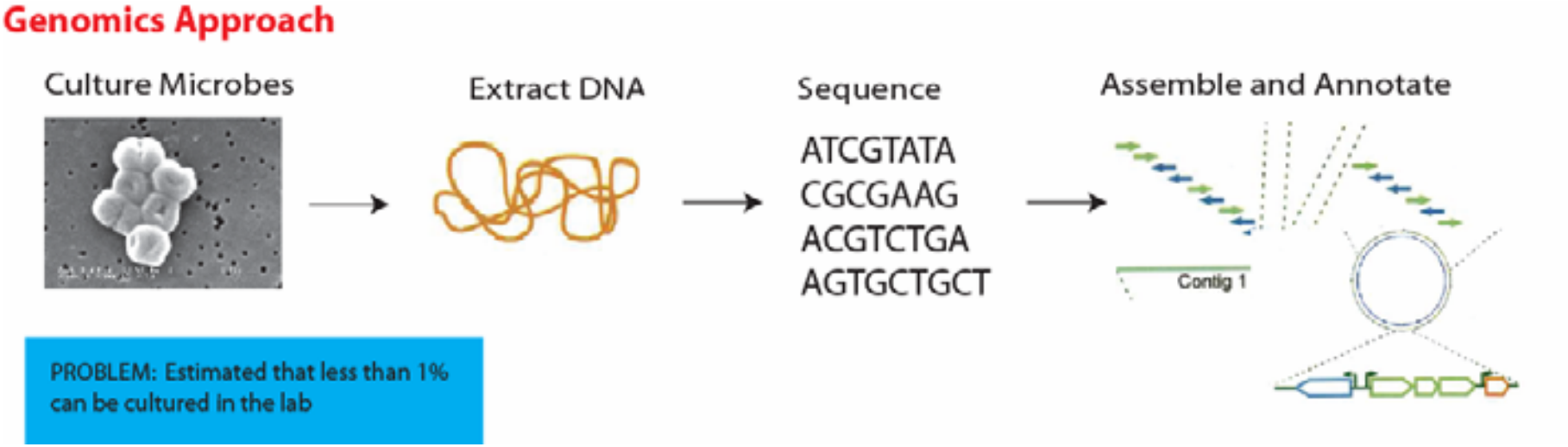
Luscombe et al. Nature 431: 308

Network Dynamics #2: Environments

How do molecular networks change across different environments? What pathways are used more or less? Can this be used as a biosensor ?

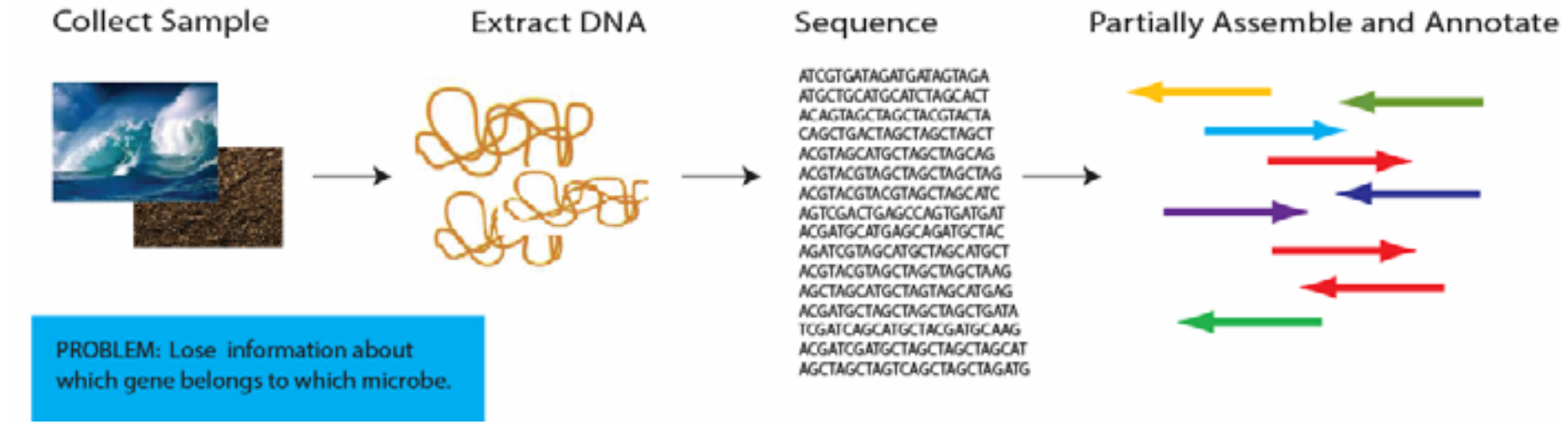
What is metagenomics?

Genomics Approach



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

Metagenomics Approach

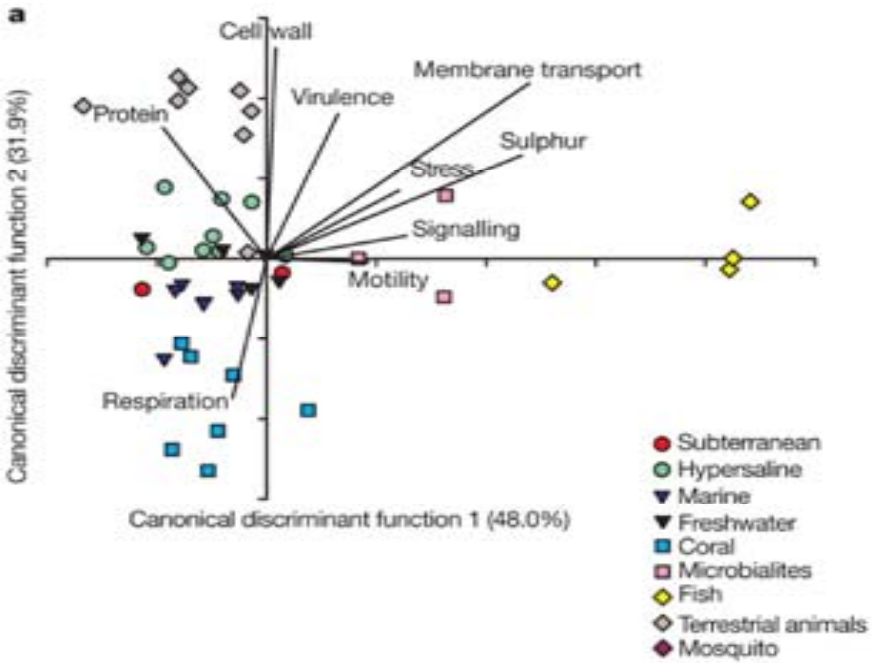


PROBLEM: Lose information about which gene belongs to which microbe.

Comparative Metagenomics

Water

Soil



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

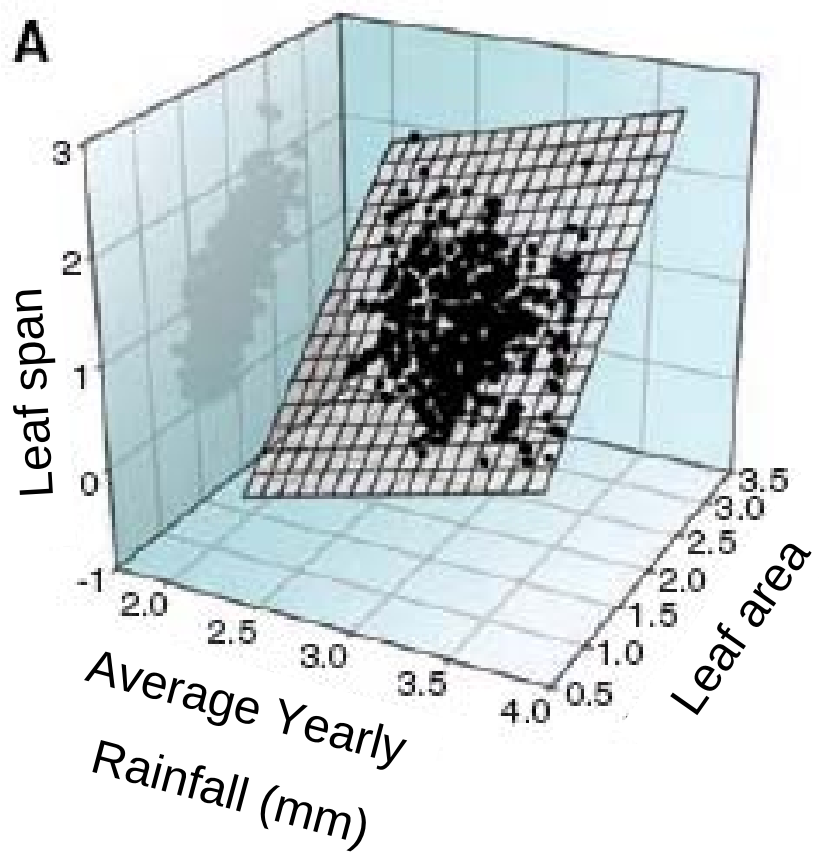
Dinsdale et. al., Nature 2008

Trait-based Biogeography

Charles River, MA

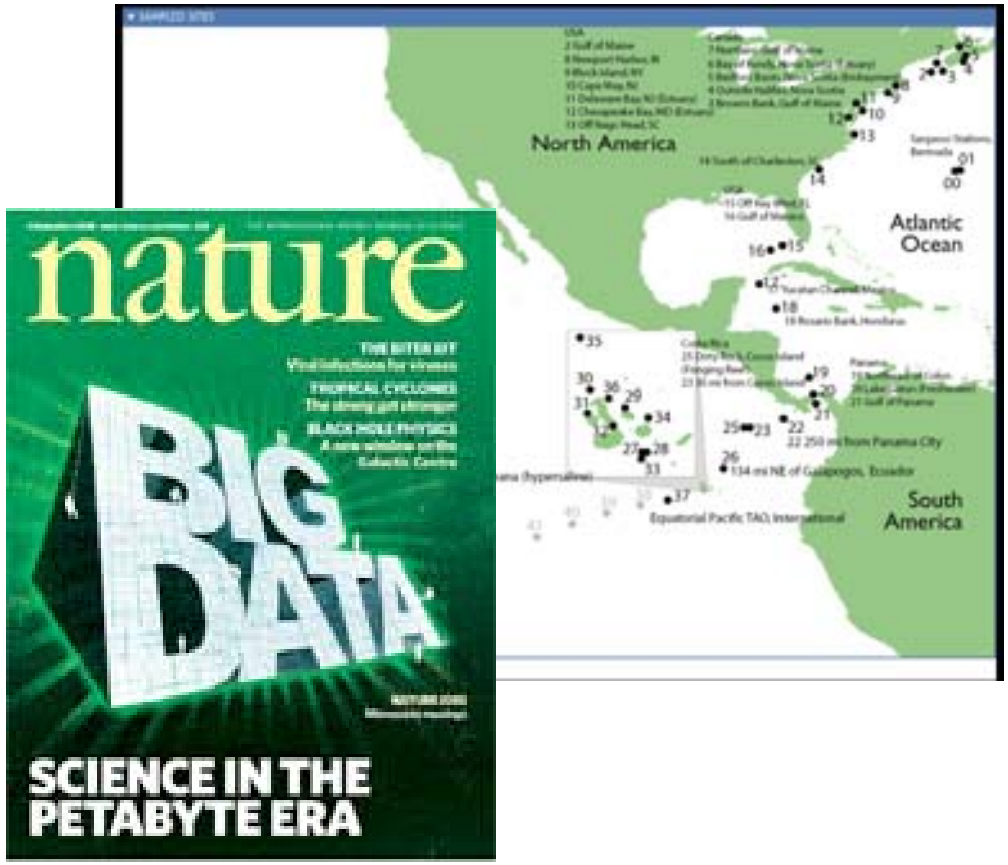
Long Island Sound, CT

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)

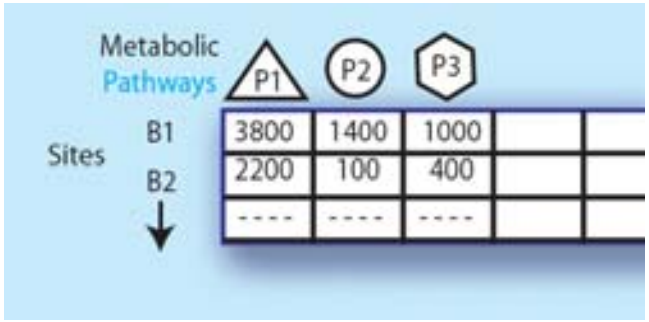


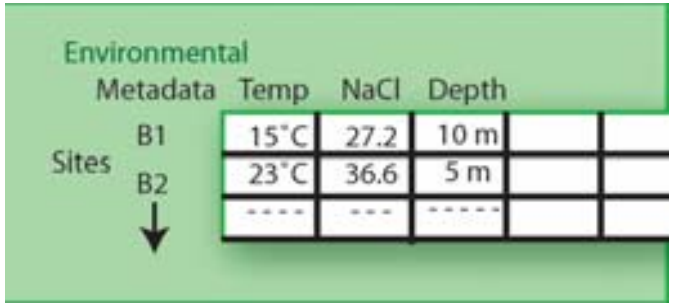
6.25 GB of data
7.7M Reads
1 million CPU hours
to process

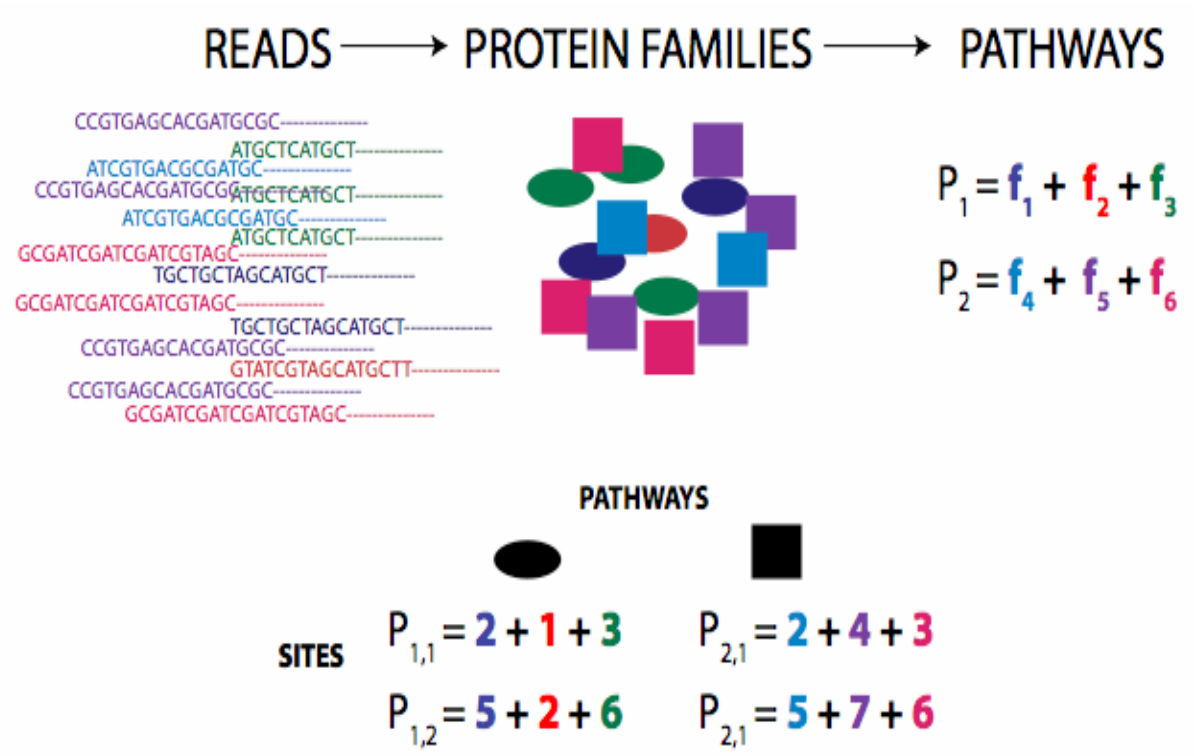
Rusch, et al., PLOS Biology 2007

Pathway Sequences
(Community Function)

Environmental
Features



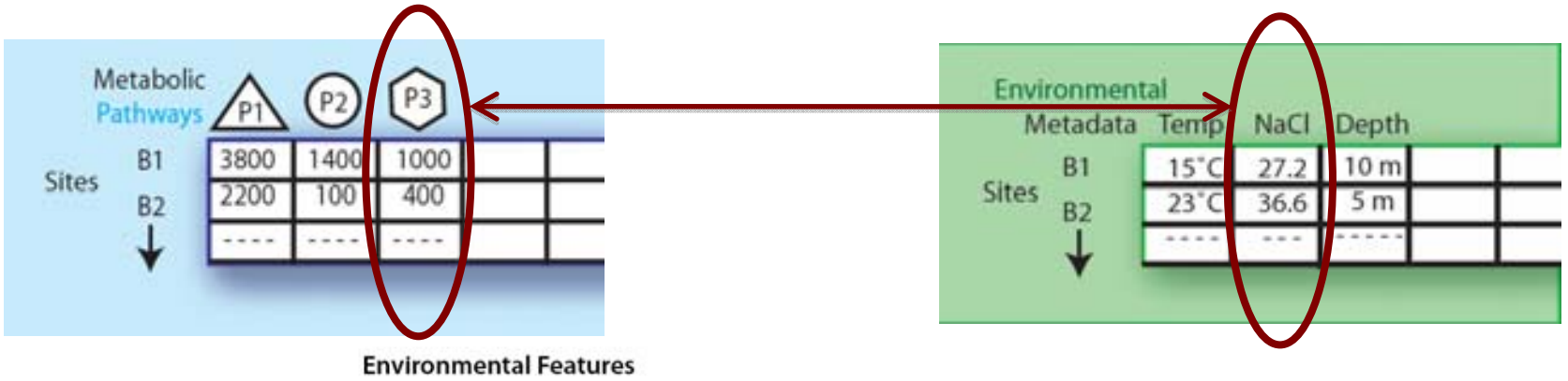




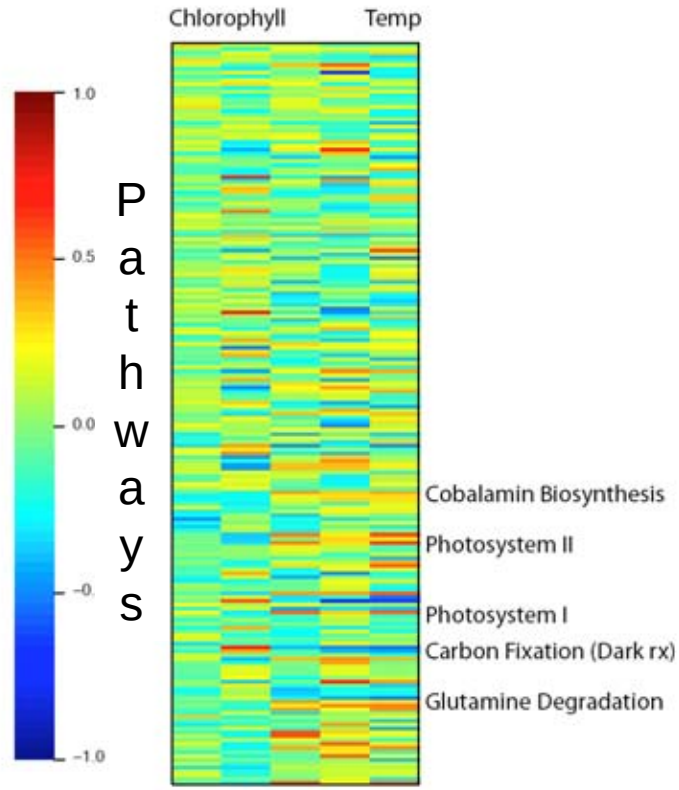
Expressing data as matrices indexed by site, env. var., and pathway usage

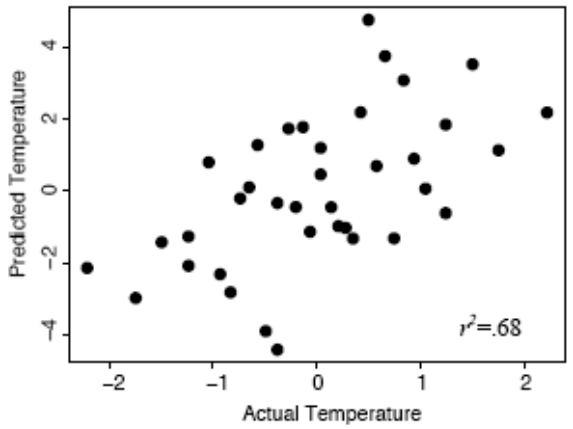
[Rusch et. al., (2007) PLOS Biology;
Gianoulis et al., PNAS (in press, 2009]

Simple Relationships: Pairwise Correlations



Environmental Features





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

Canonical Correlation Analysis: Simultaneous weighting

Score	# of papers published
GRE	

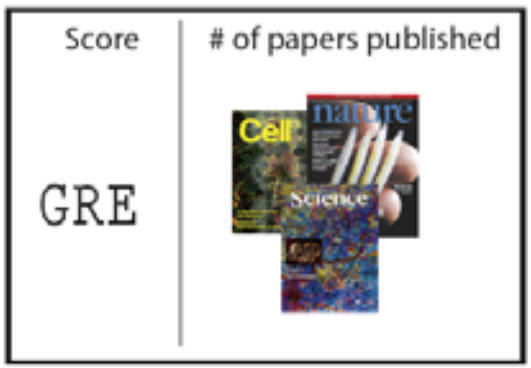
Undergraduate Performance Index (UPI)	Graduate School Performance Index (GPI)
GRE GPA 	

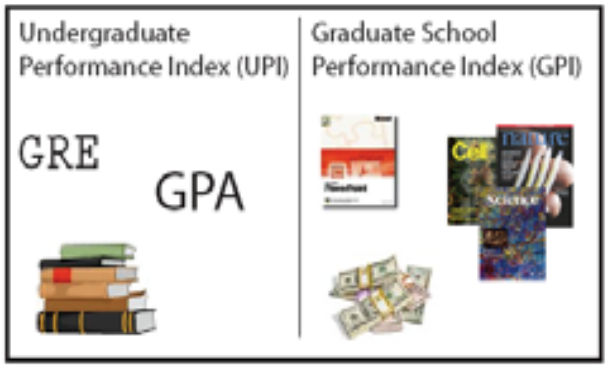
UPI = a GRE + b  + c GPA

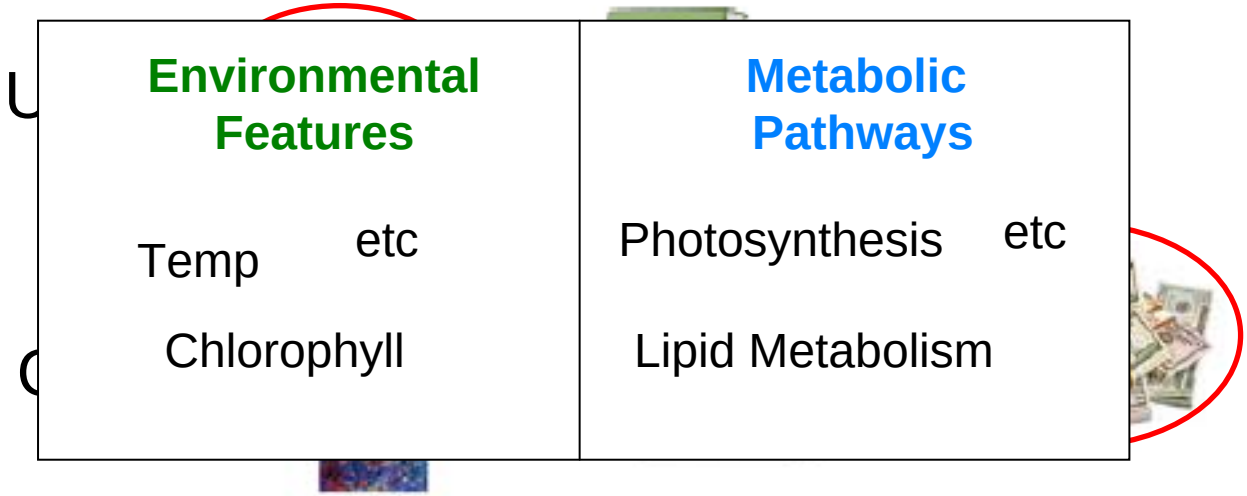
GPI = a'  + b'  + c' 

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

Canonical Correlation Analysis: Simultaneous weighting

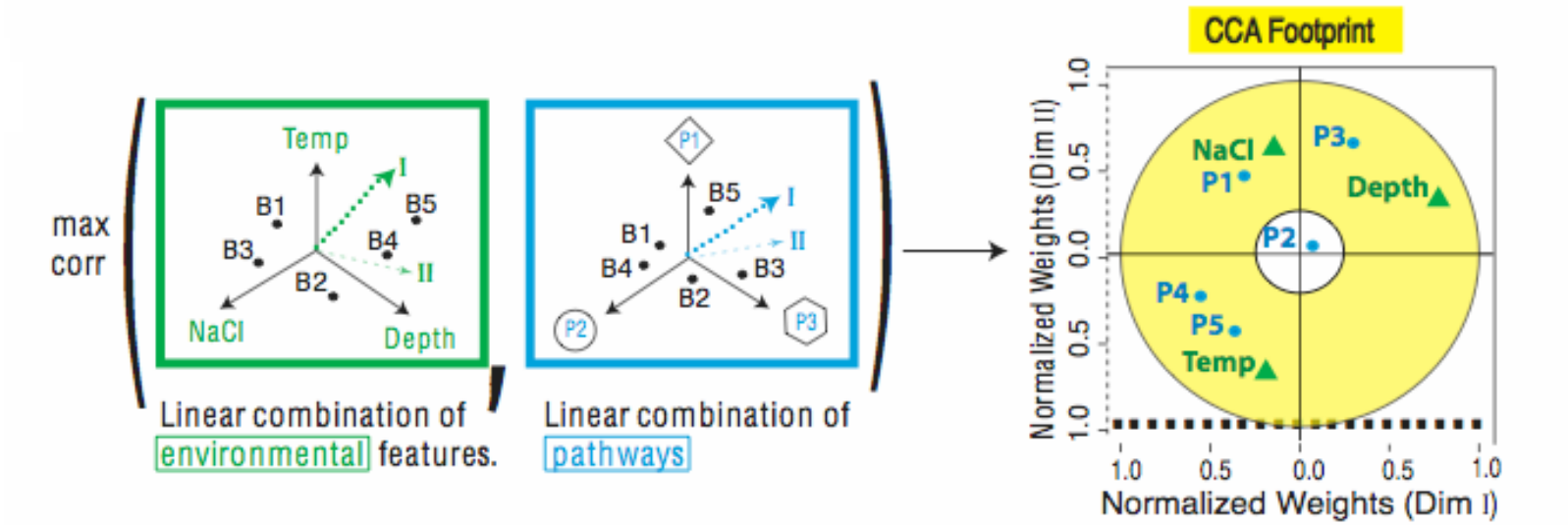






[Gianoulis et al., PNAS (in press, 2009)]

Environmental-Metabolic Space



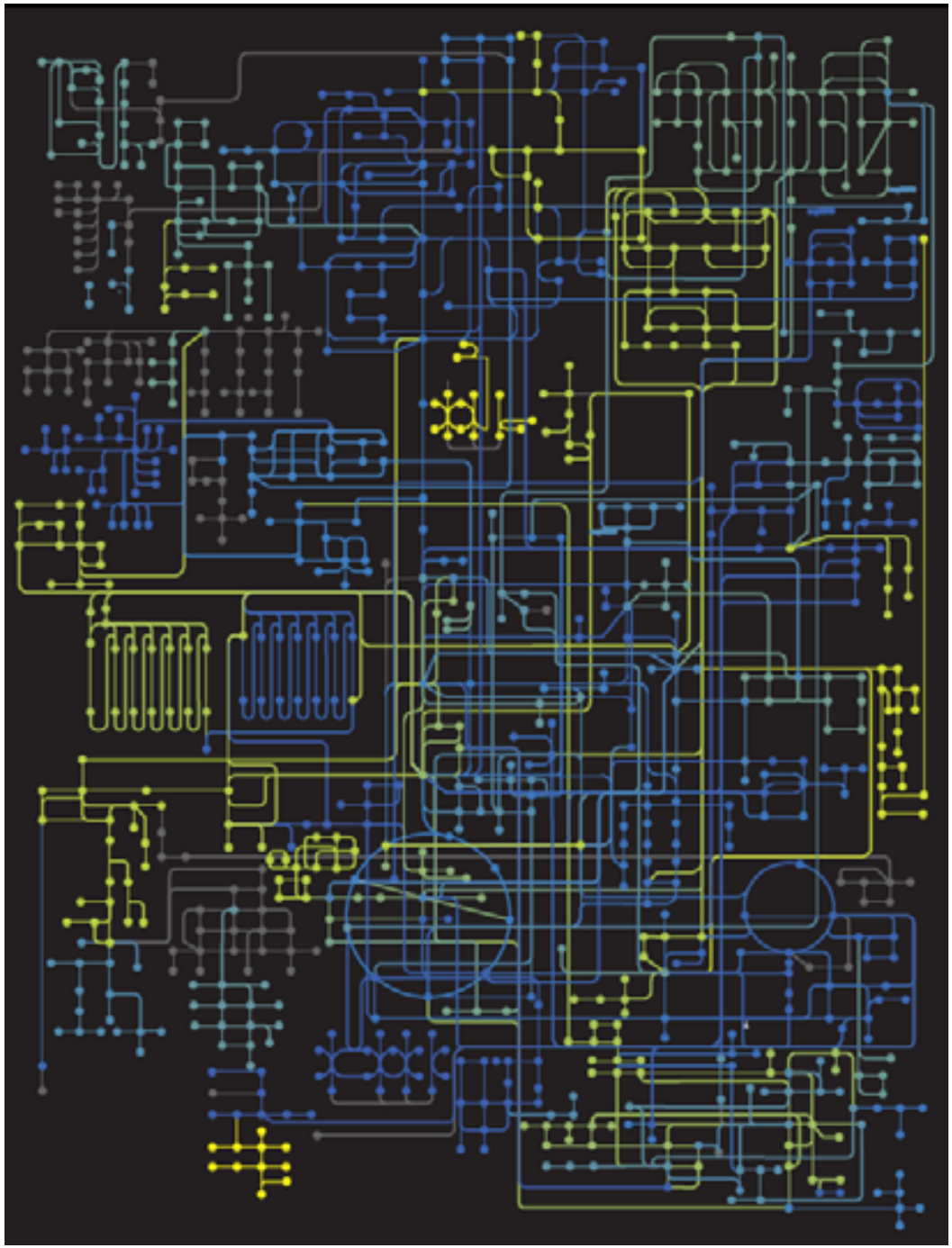
The goal of this technique is to interpret cross-variance matrices
We do this by defining a change of basis.

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

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

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

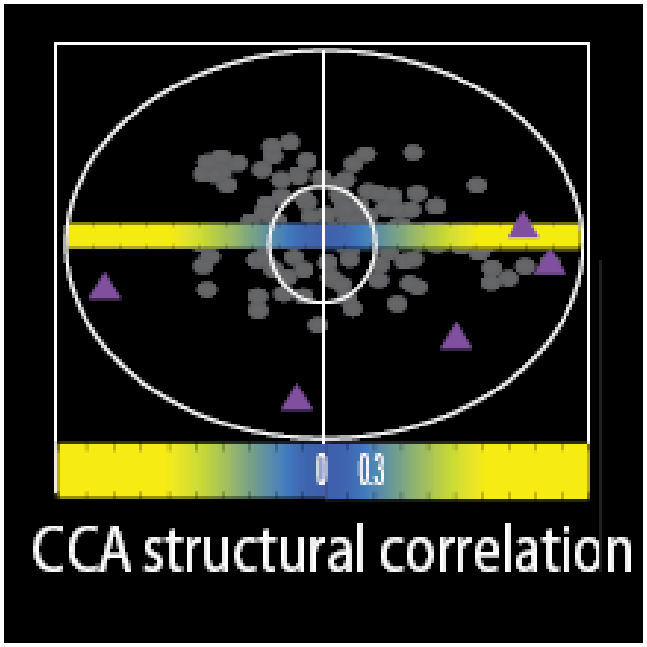
[Gianoulis et al., PNAS (in press, 2009)]



Strength of Pathway co-variation with environment

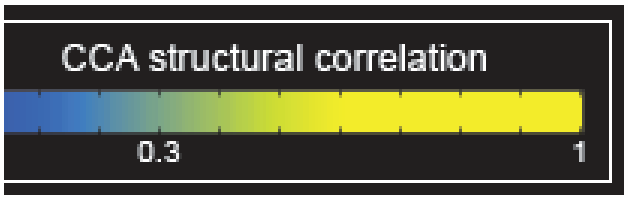
Environmentally invariant

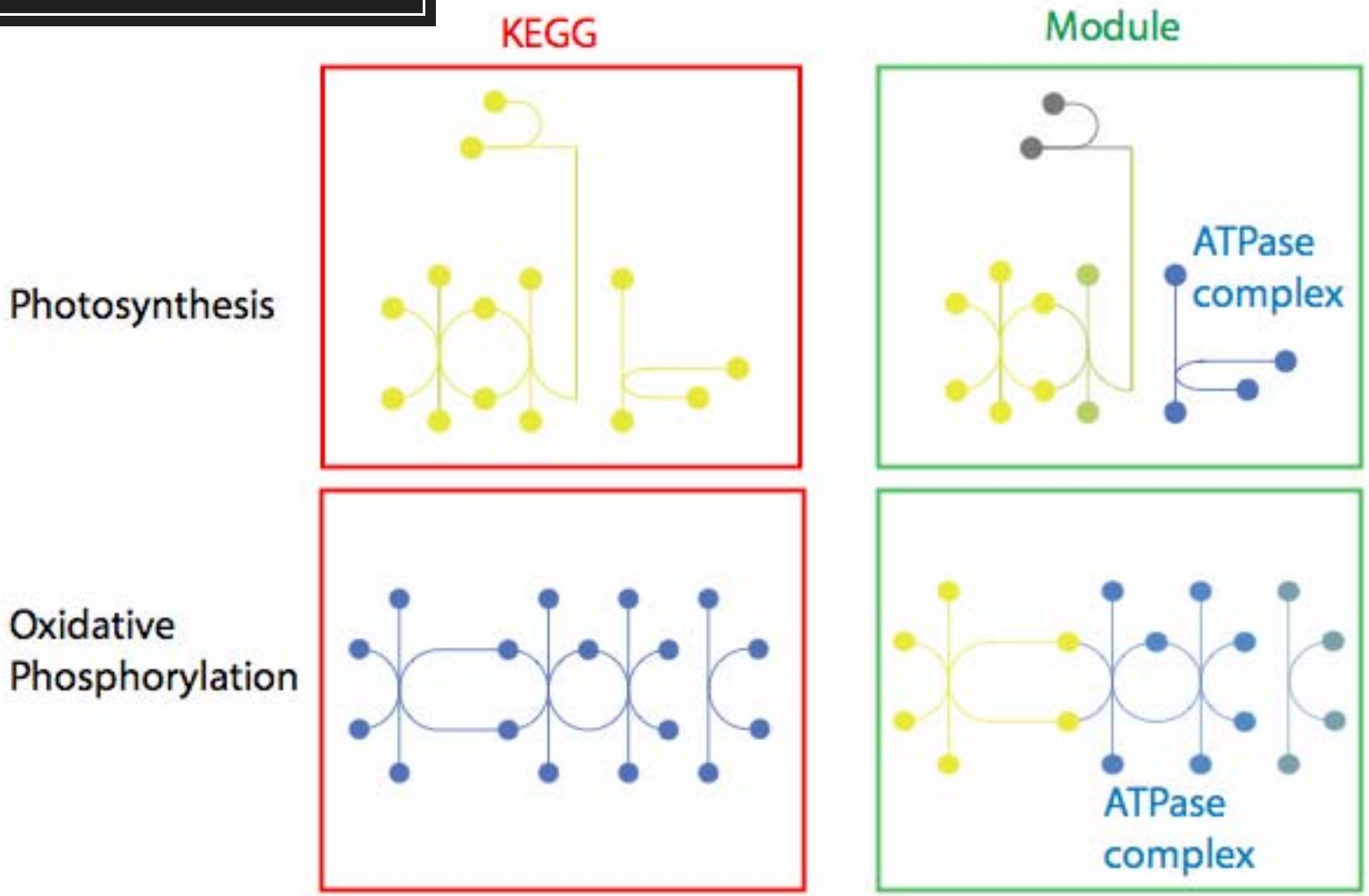
Environmentally variant



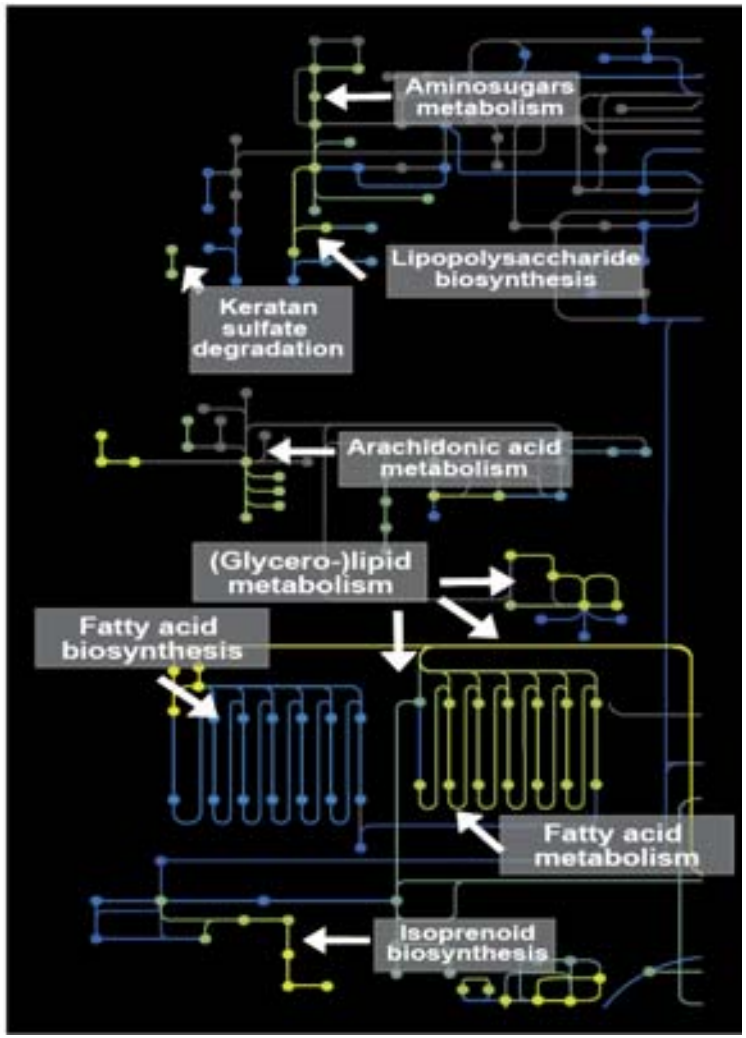
[Gianoulis et al., PNAS (in press, 2009)]

Conclusion #1: energy conversion strategy, temp and depth



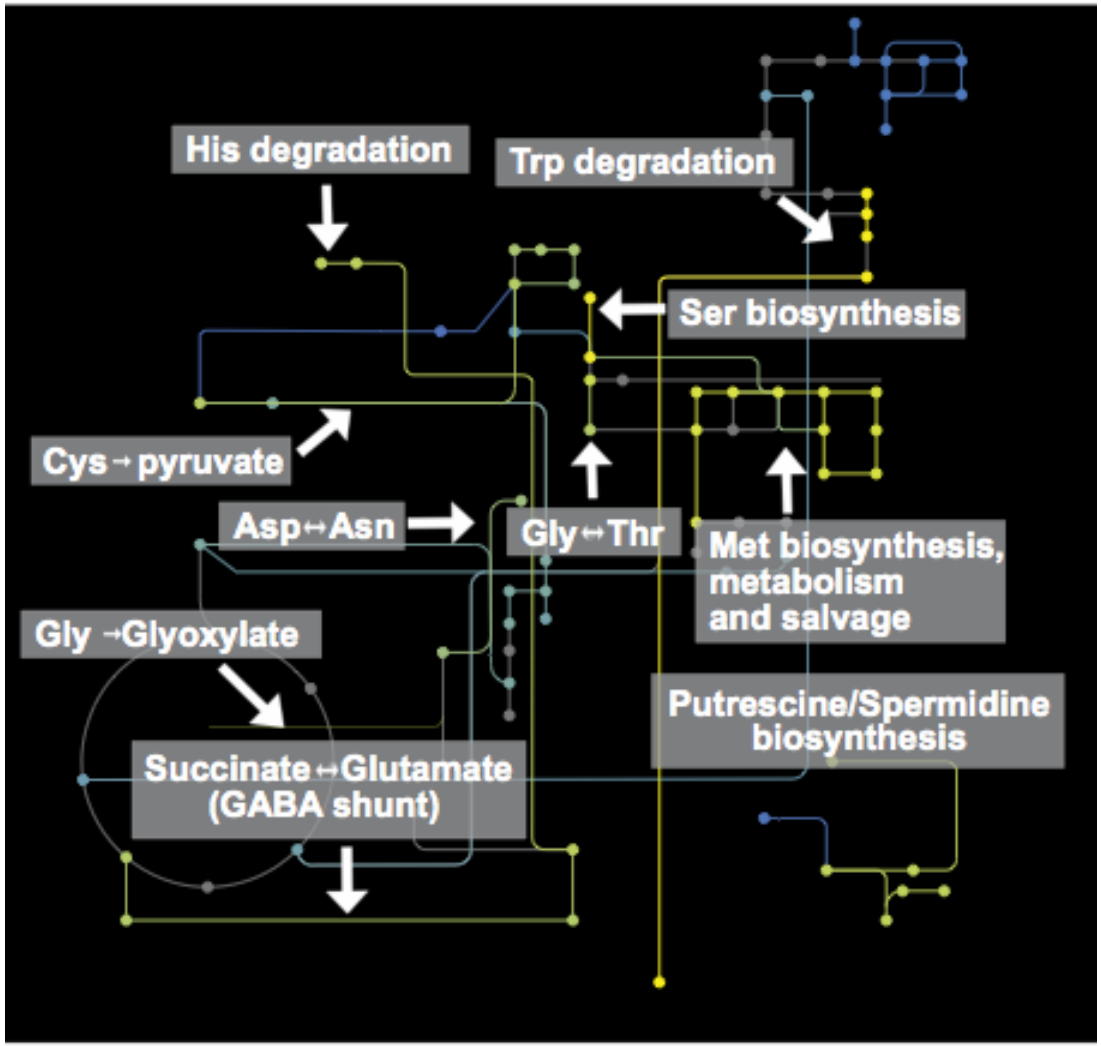


Conclusion #2: Outer Membrane components vary the environment



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

Conclusion #3: Covariation of AA biosynthesis and Import



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

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

Conclusion #4: Cofactor (Metal) Optimization

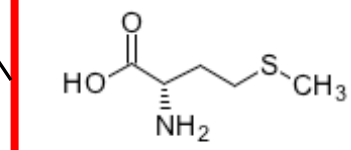
Methionine salvage, synthesis, and uptake, transport

IS DEPENDENT-ON

Methionine synthesis

Cobalamin biosynthesis

Cobalt transporters



C00073

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 & Human 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

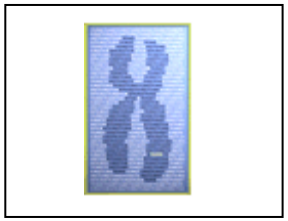
ILLUSTRATIVE

Hapmap/Perlegen

International HapMap Project

SNPs

Database of Genomic Variants



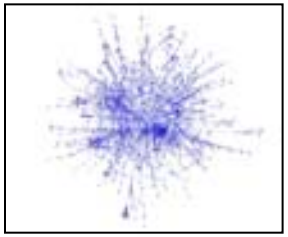
CNVs + SDs

Map to ENSEMBL genes

Ensembl Genes

ENSG000XXXX:
rsSNP00XXX
CNV_XXX
DN/DS XXXX
Recombination rate

Interactome



~30000 interactions from HPRD and Y2H screens

Map to proteins in the interaction network

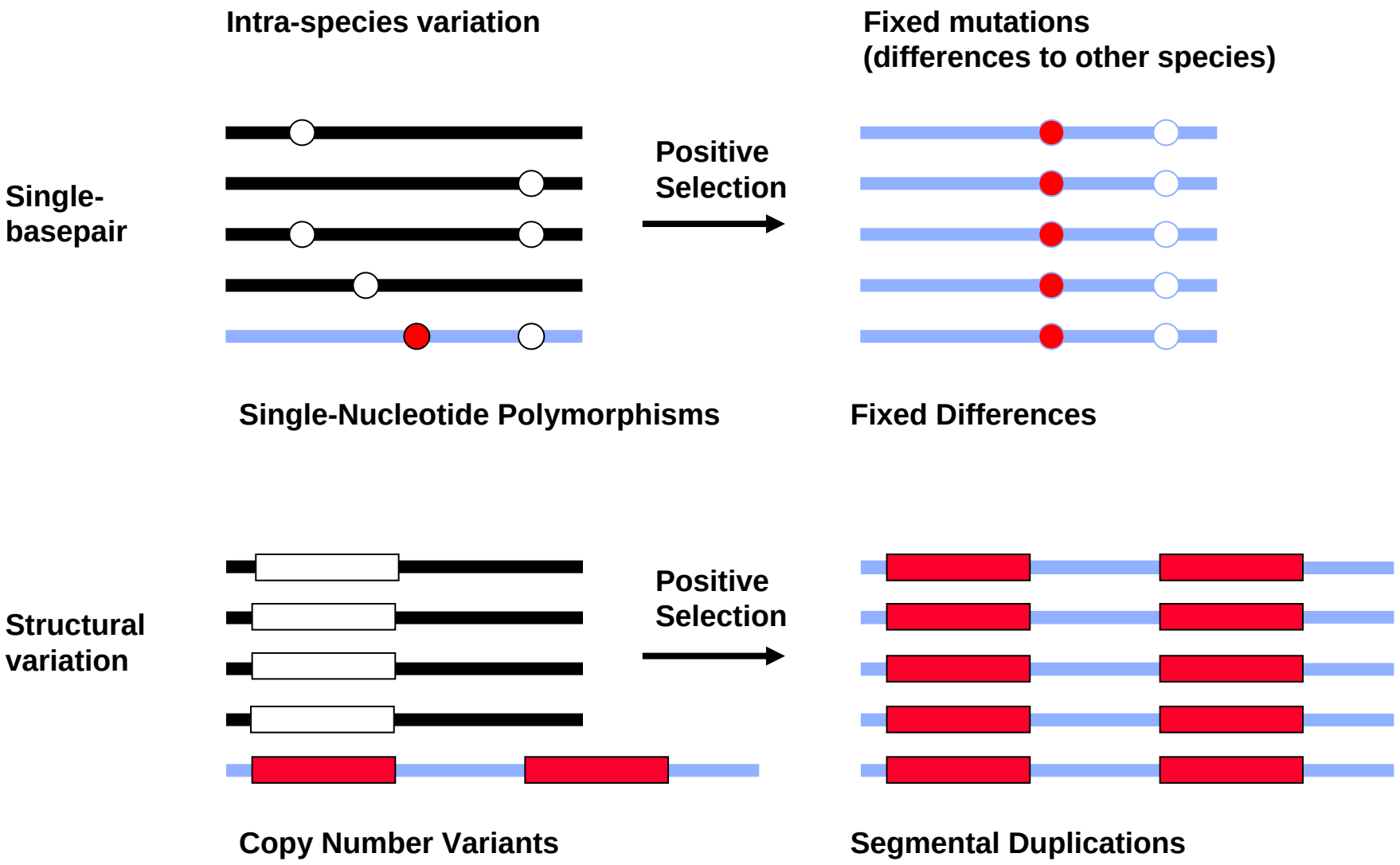
Result

- 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

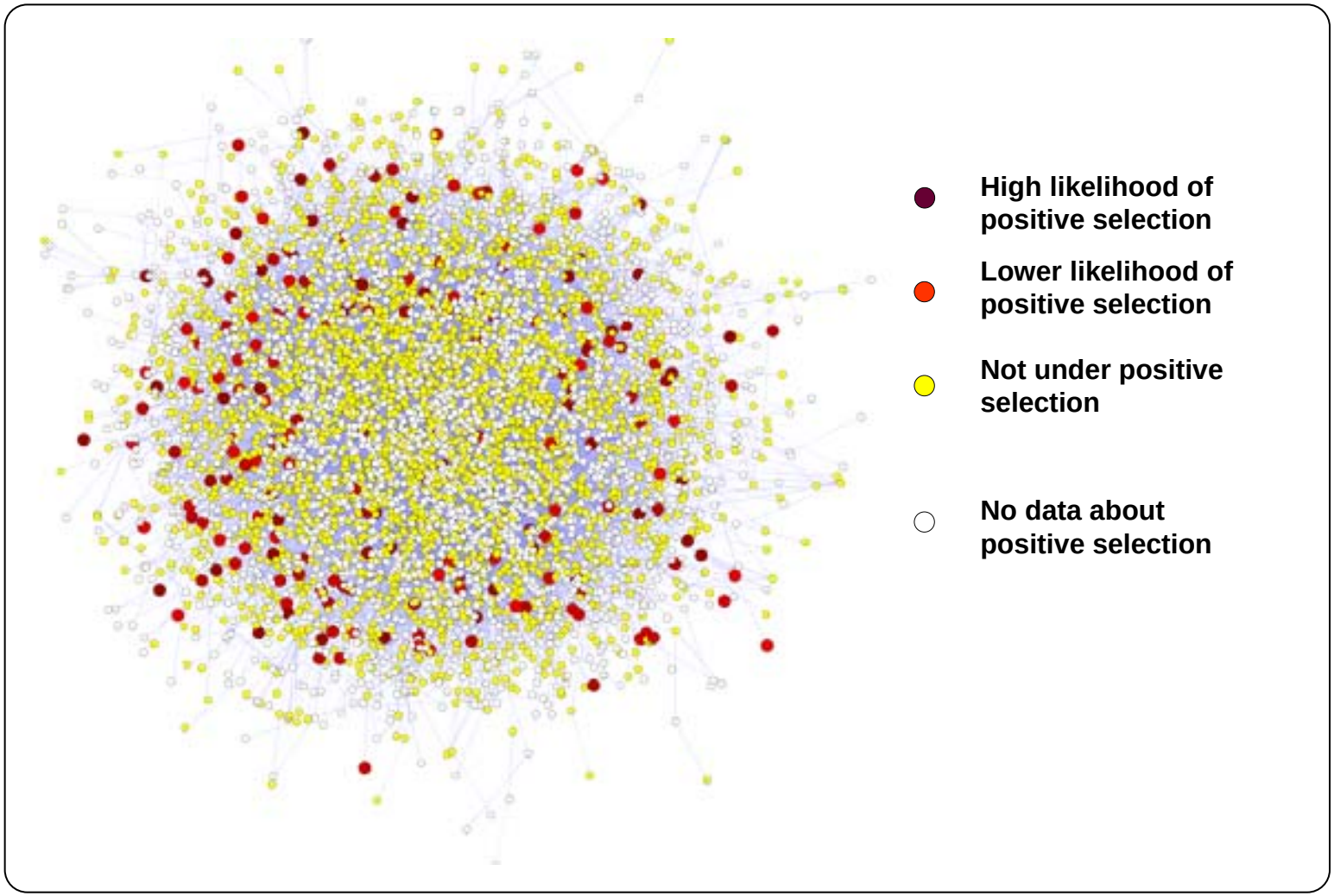
ADAPTIVE EVOLUTION CAN BE SEEN ON TWO DIFFERENT LEVELS



Source: PMK

POSITIVE SELECTION LARGELY TAKES PLACE AT THE NETWORK PERIPHERY

Positive selection in the human interactome

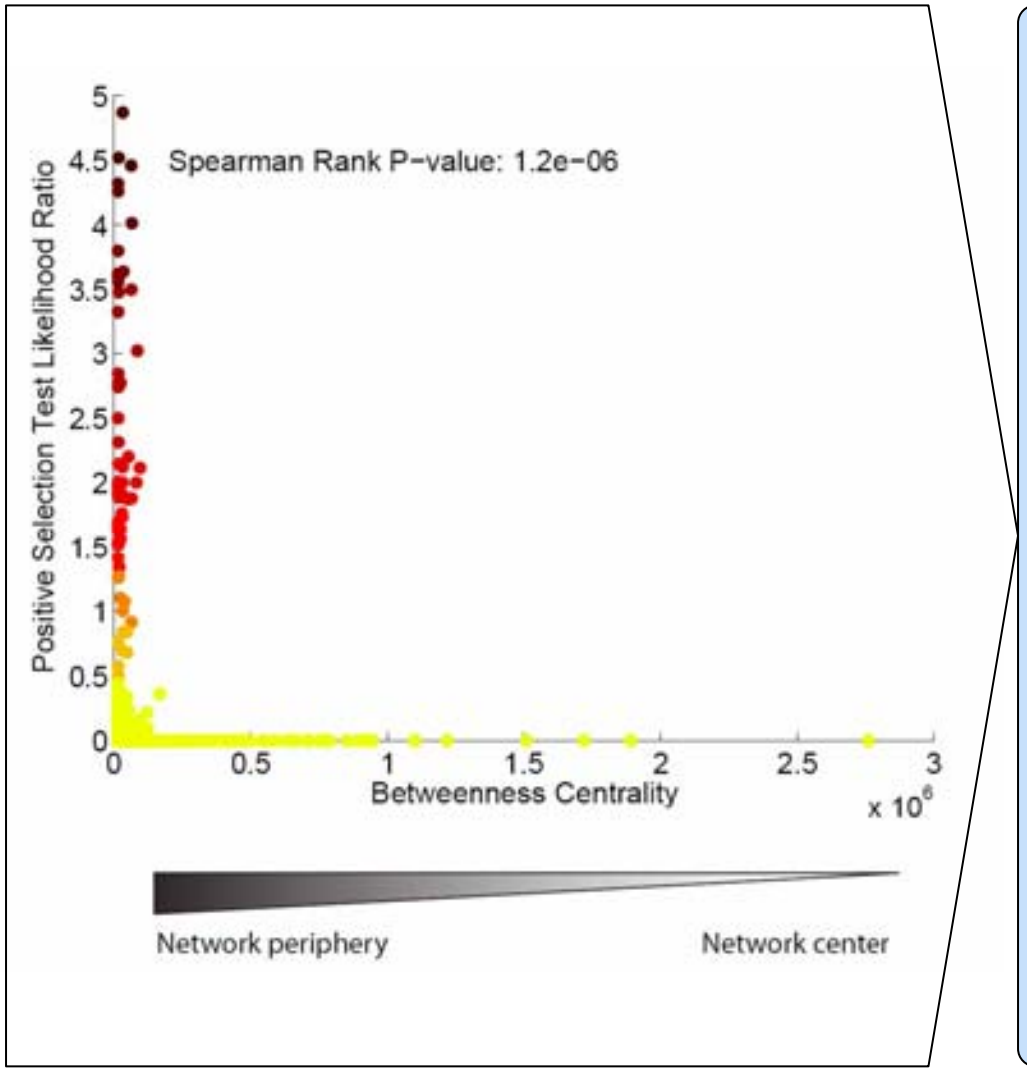


Source: Nielsen et al. *PLoS Biol.* (2005), HPRD, and Kim et al. PNAS (2007)

CENTRAL PROTEINS ARE LESS LIKELY TO BE UNDER POSITIVE SELECTION

Hubs

Degree vs. Positive Selection



Reasoning

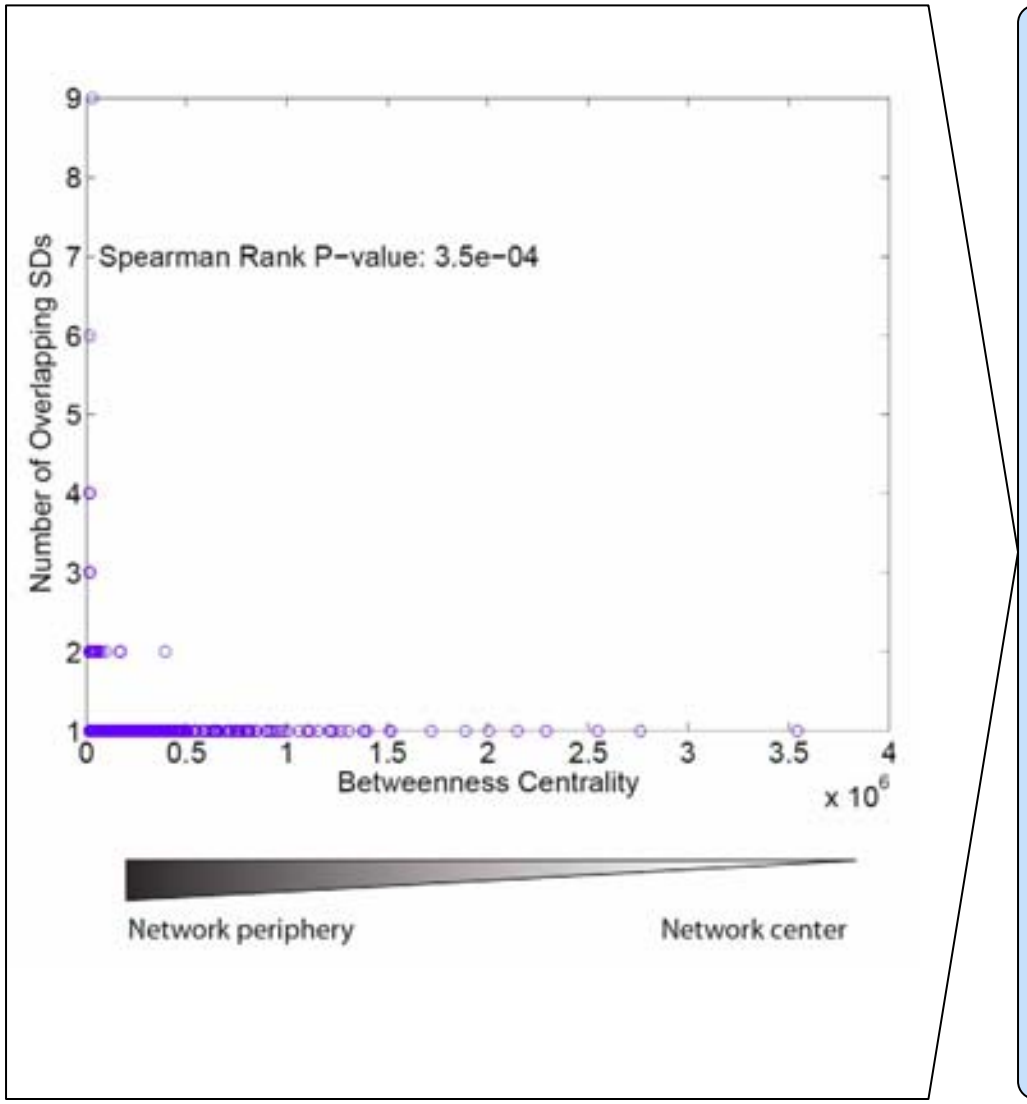
- Peripheral genes are likely to under positive selection, whereas hubs aren't
- This is likely due to the following reasons:
  - Hubs have stronger structural constraints, the network periphery doesn't
  - Most recently evolved functions (e.g. "environmental interaction genes" such as sensory perception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

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

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

CENTRAL NODES ARE LESS LIKELY TO LIE INSIDE OF SDs

Centrality vs. SD occurrence



Reasoning

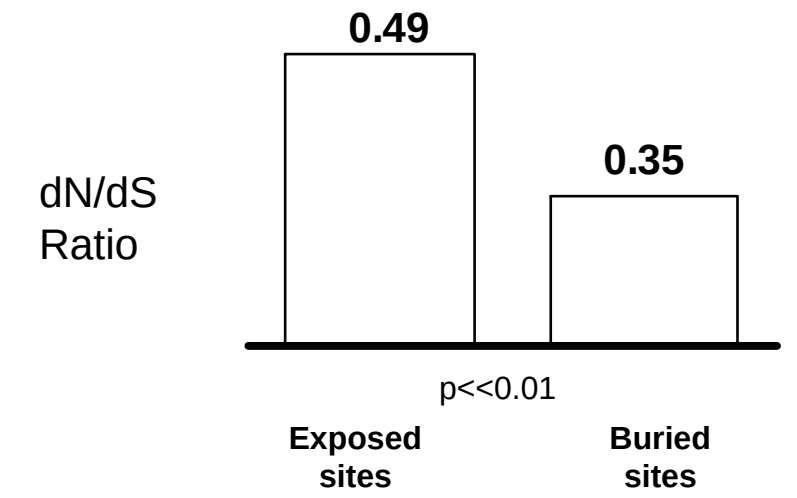
- This result also confirms our initial hypothesis – peripheral nodes tend to lie in regions rich in SDs.
- Since segmental duplications are a different mechanism of ongoing evolution, the less constrained peripheral proteins are enriched in them.
- Note that despite the small size of our dataset for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

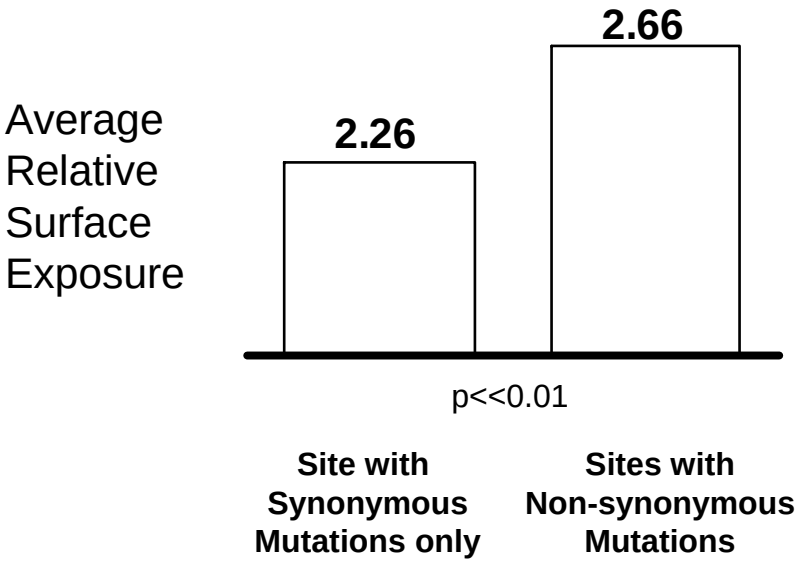
* Specifically, a number of the SDs are likely not fixed, but rather common CNVs in the reference genome

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.

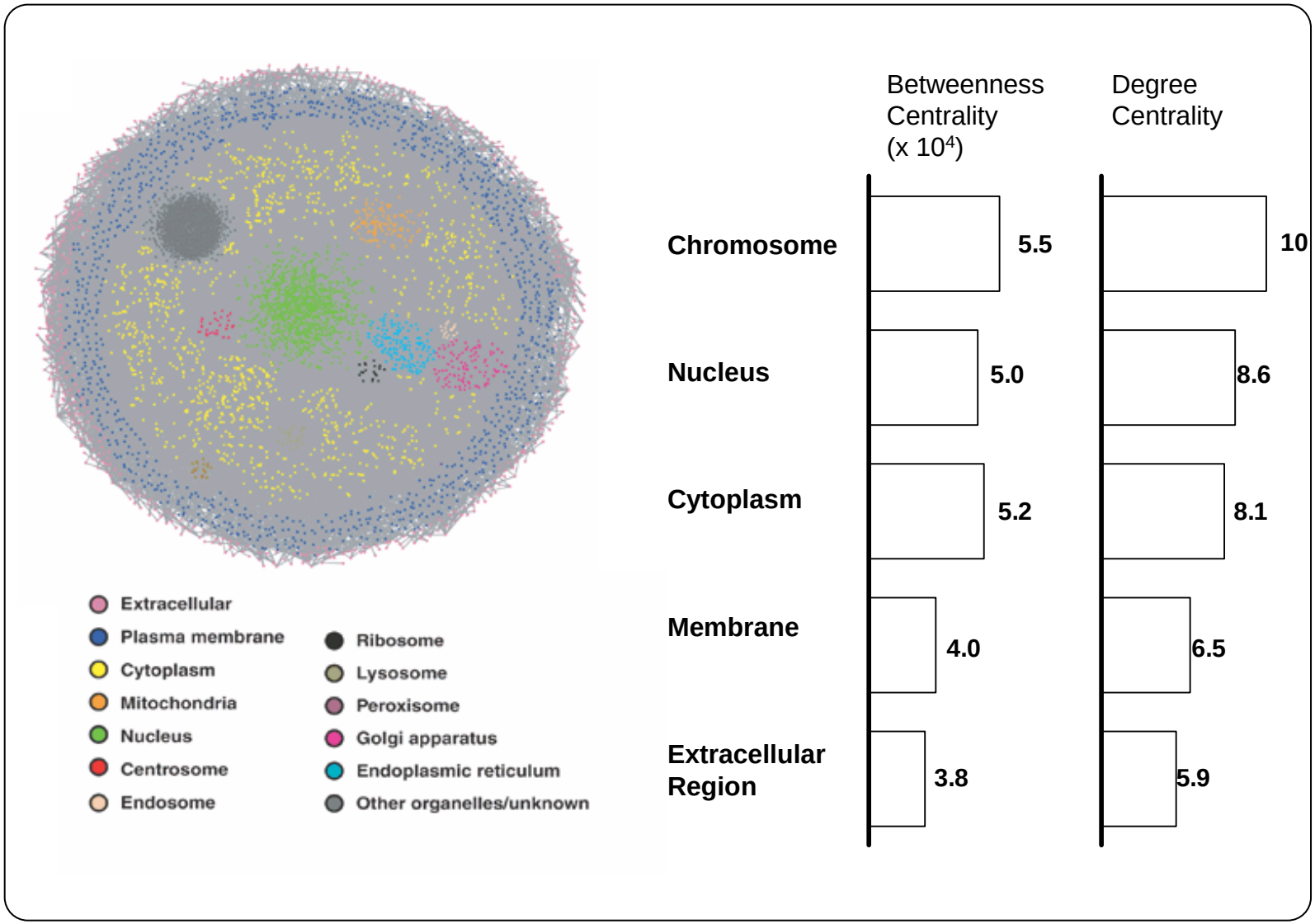
BURIED SITES ARE CONSERVED AND MUCH LESS LIKELY TO HARBOR NON-SYNONYMOUS MUTATIONS





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?

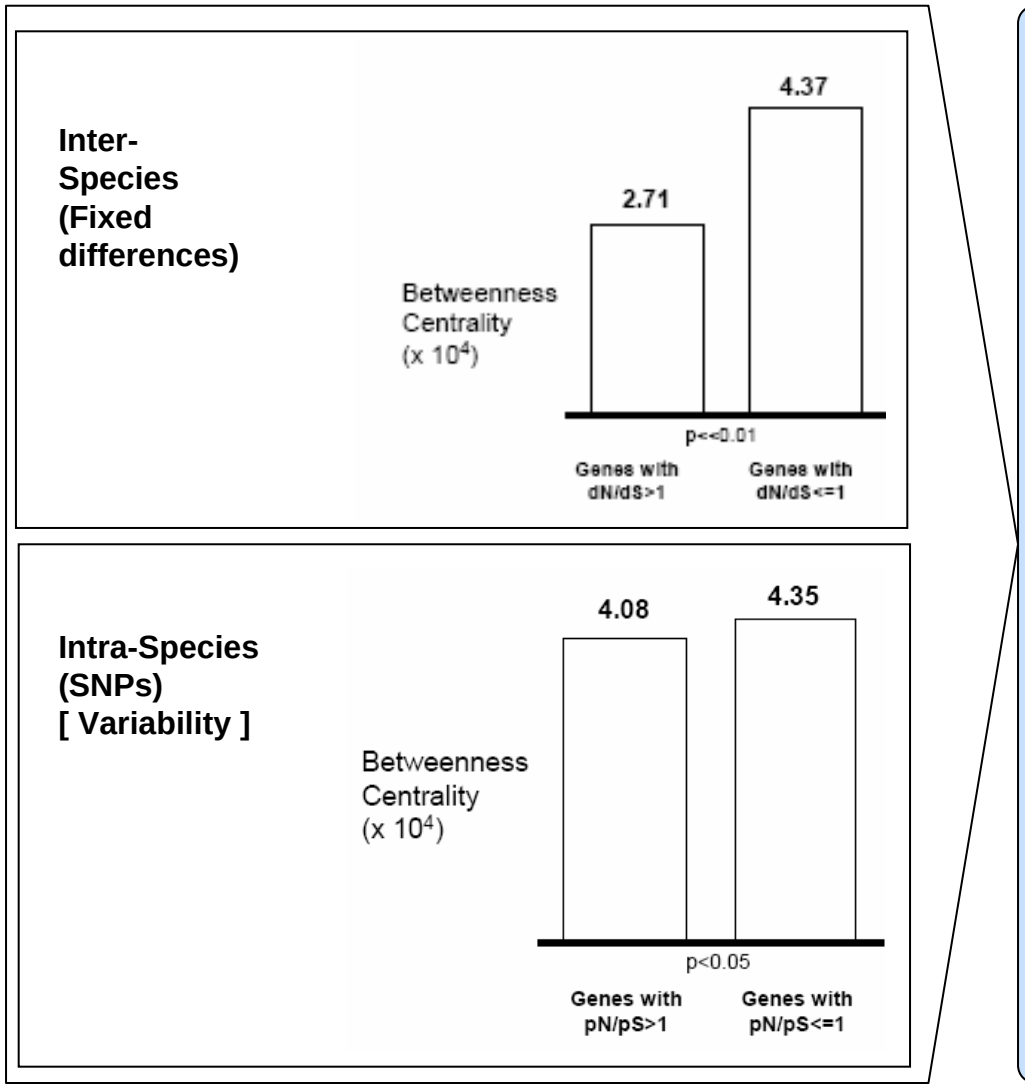
ILLUSTRATIVE

	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

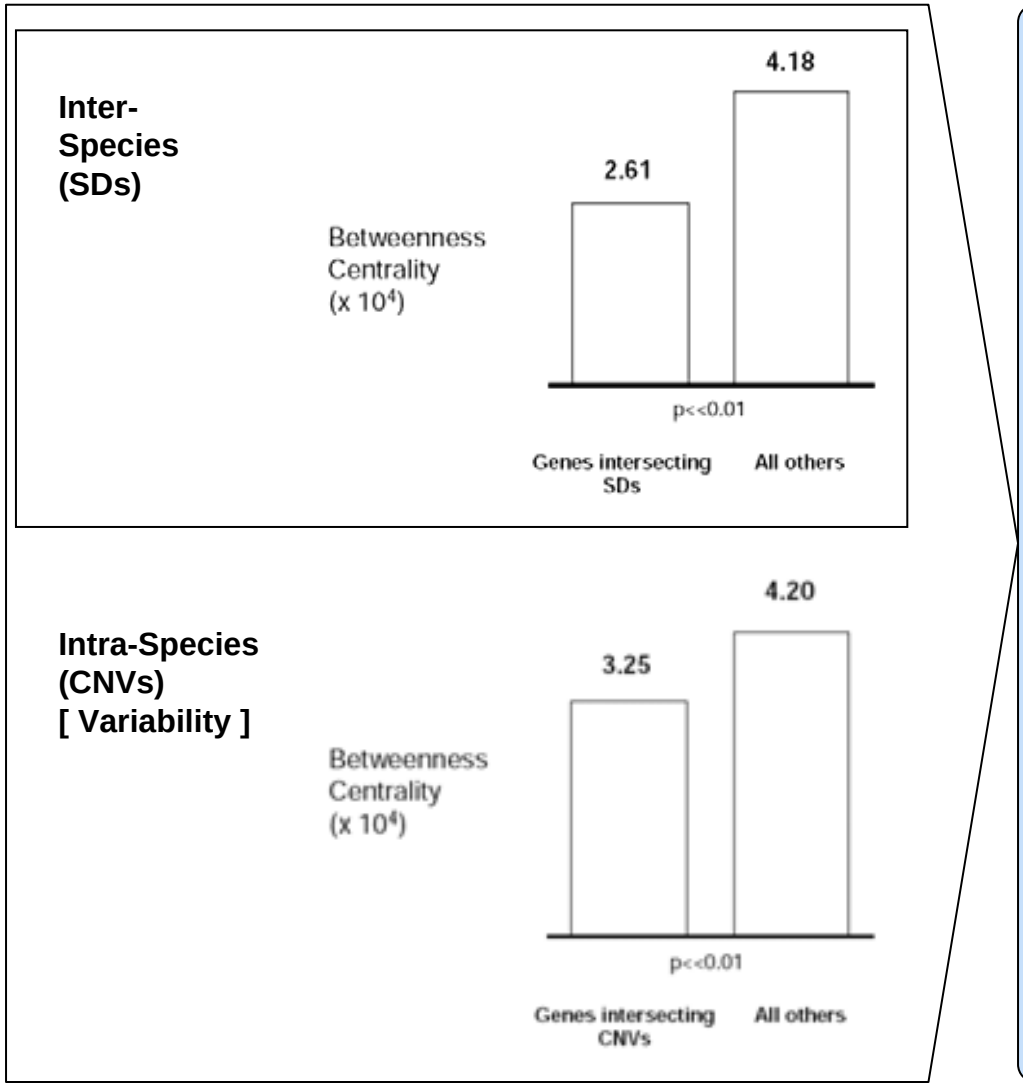
- There is a difference in **variability** (in terms of SNPs) between the network periphery and the center
- However, this difference is much smaller than the difference in **selection**
- This most likely means, that part of the effect we're seeing is due to relaxed constraint (and higher variability)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

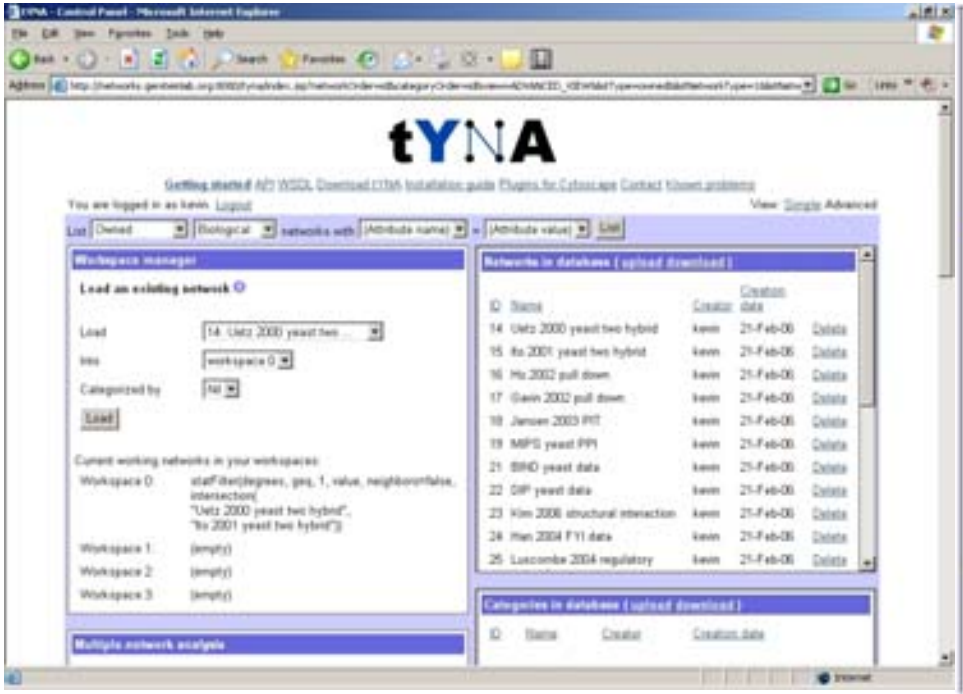
Conclusions: Connecting Networks & Human Variation

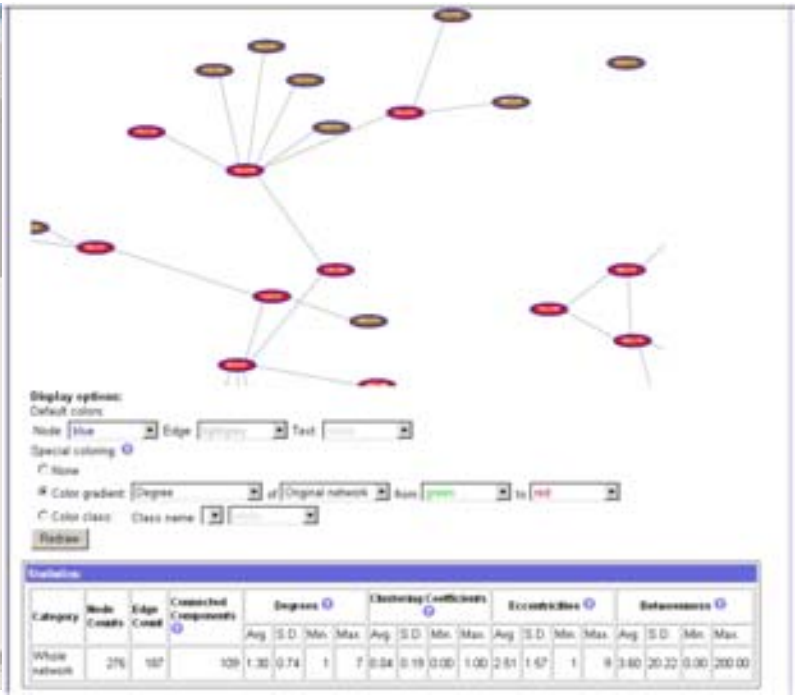
- We find ongoing evolution (positive selection) at the network periphery.
  - ◊ This trend is present on two levels:
    - On a sequence level, it can be seen as positive selection of peripheral nodes
    - On a structural level, it can be seen as the pattern of SDs that display significantly higher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

- an automated web tool

tYNA

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



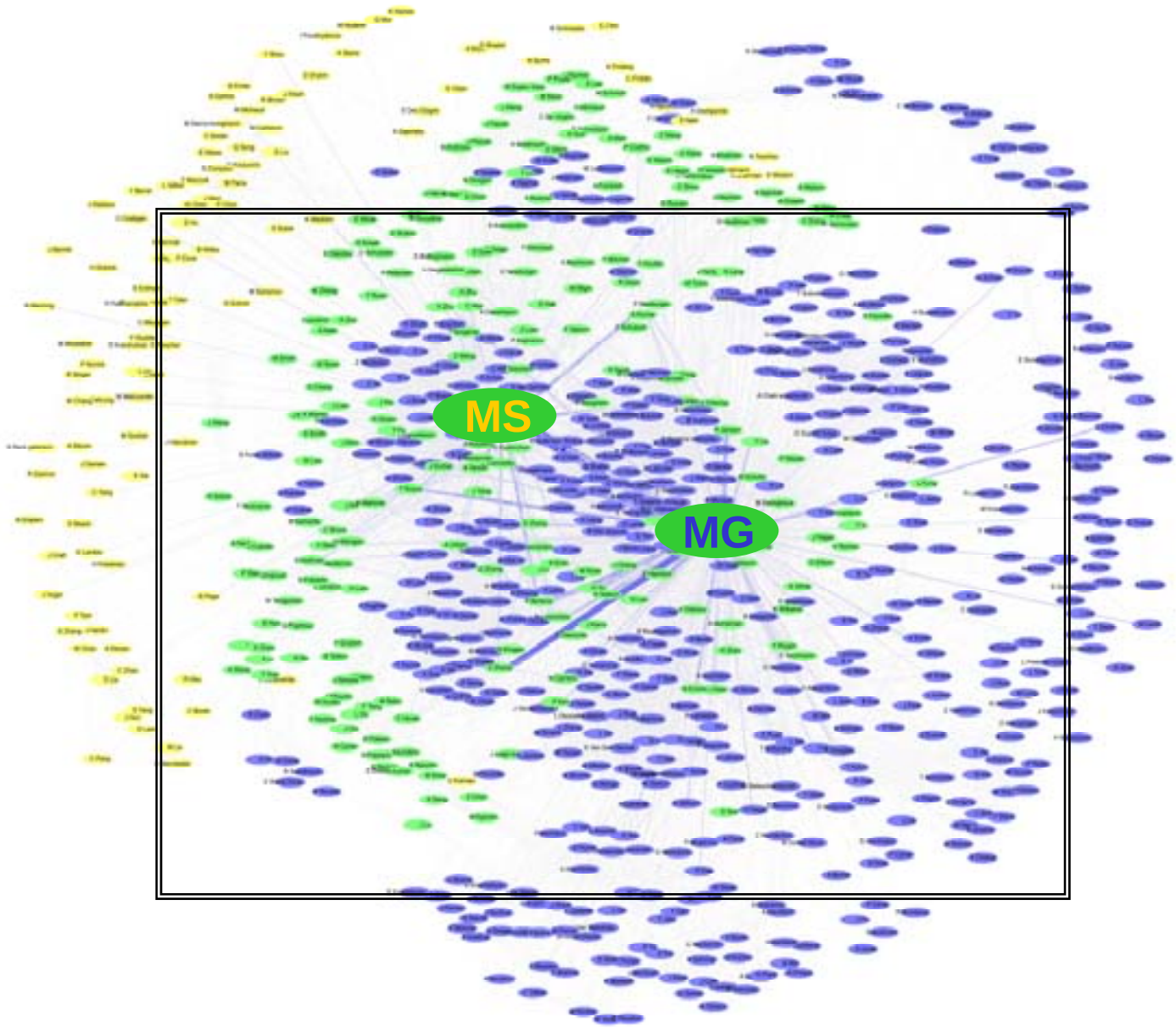


Normal website + Downloaded code (JAVA)
+ Web service (SOAP) with Cytoscape plugin

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

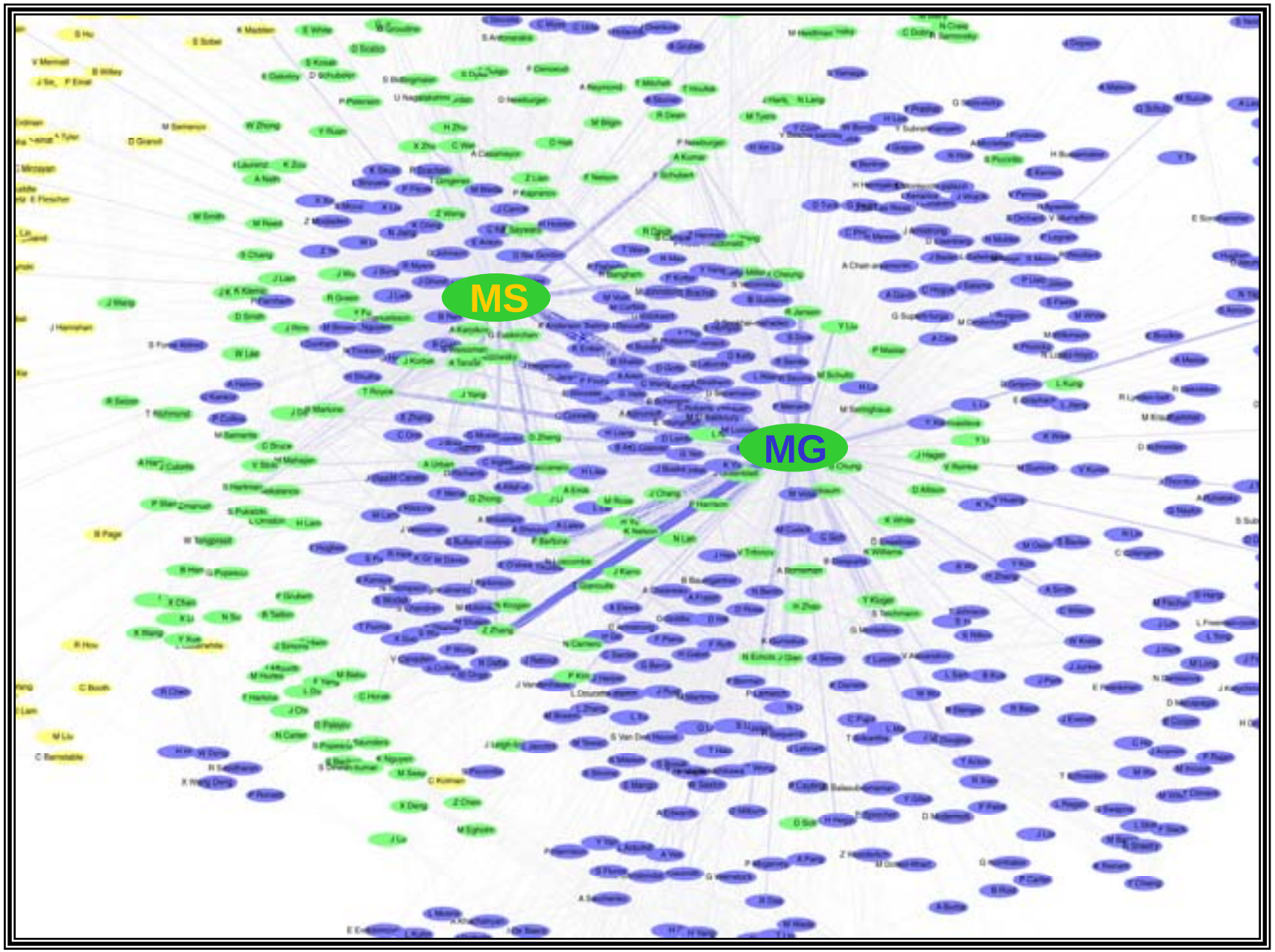
Acknowledgements

TopNet.GersteinLab.org



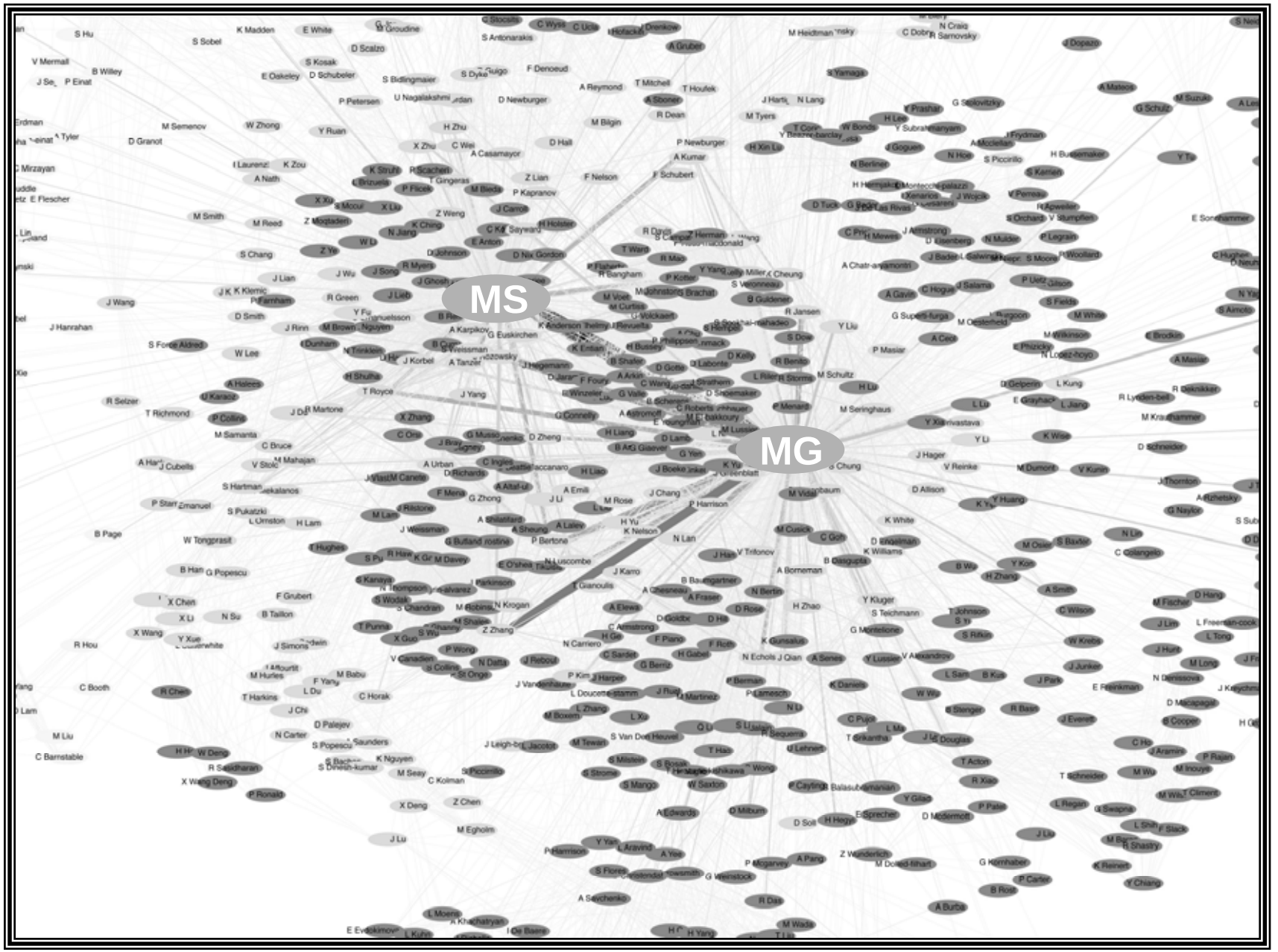
Acknowledgements

TopNet.GersteinLab.org



Acknowledgements

TopNet.GersteinLab.org

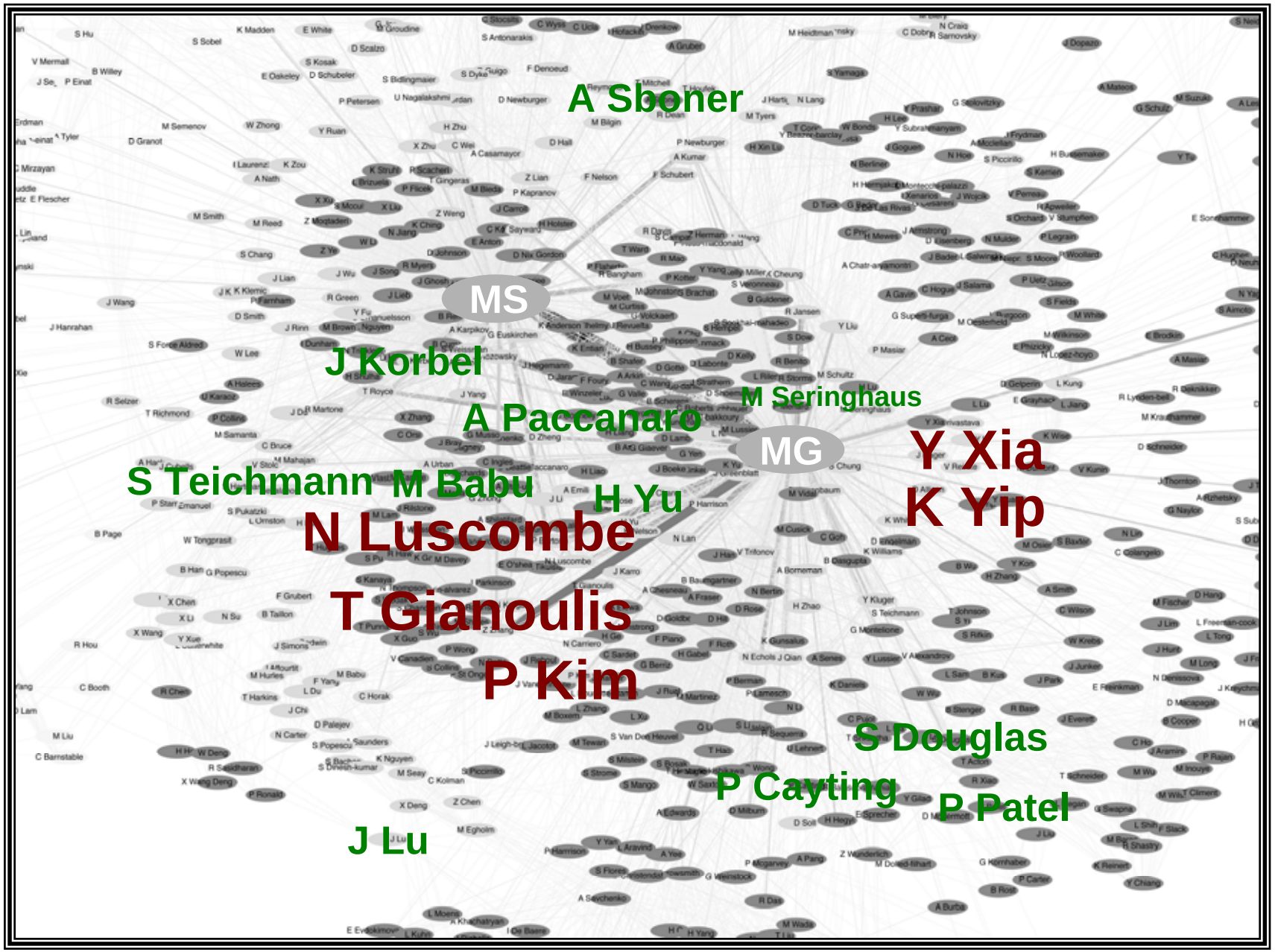


NIH, NSF, Keck

Acknowledgements

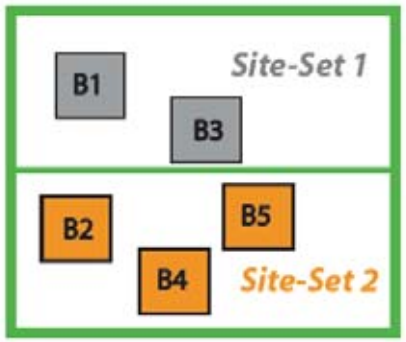
TopNet.GersteinLab.org

Job opportunities currently for postdocs & students



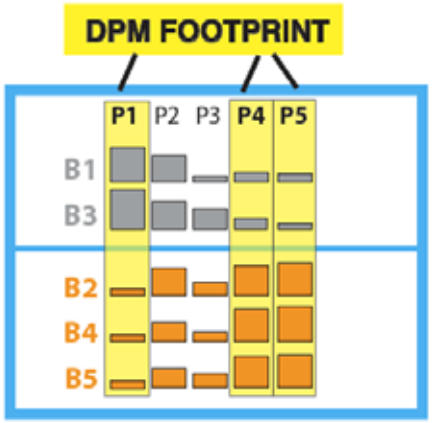
DPM: Discriminative Partition Matching

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)]

More Information on this Talk

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

SUBJECT: Networks

DESCRIPTION:
University of Chicago, Inst. of Biophysical Dynamics,
2008.12.02, 12:00-13:00; [I:CHICAGOBIOPHYS]
(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 **kwchicagobiophys**.)

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

.

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