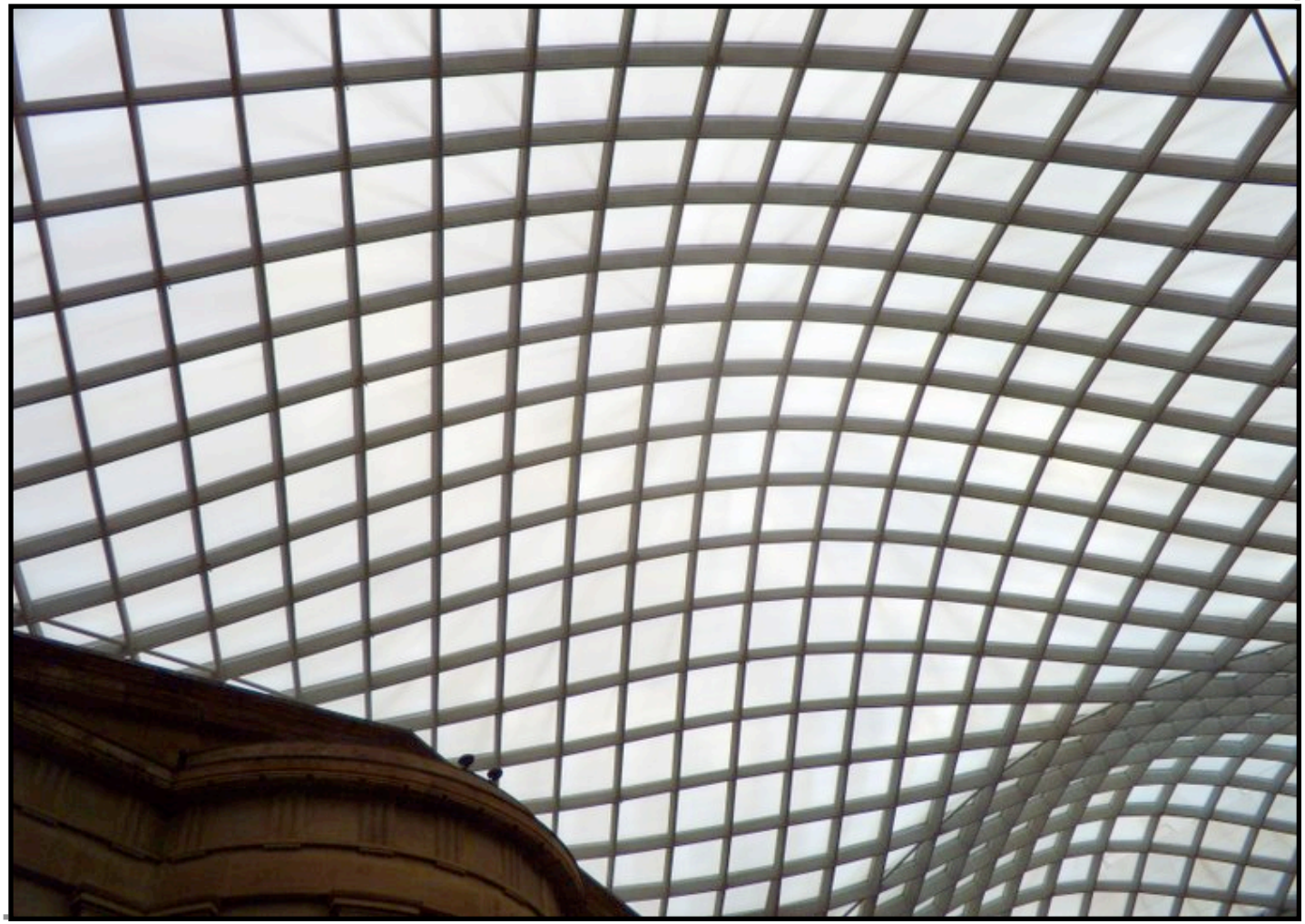
Understanding Protein Function on a Genome-scale through the Analysis of Molecular Networks

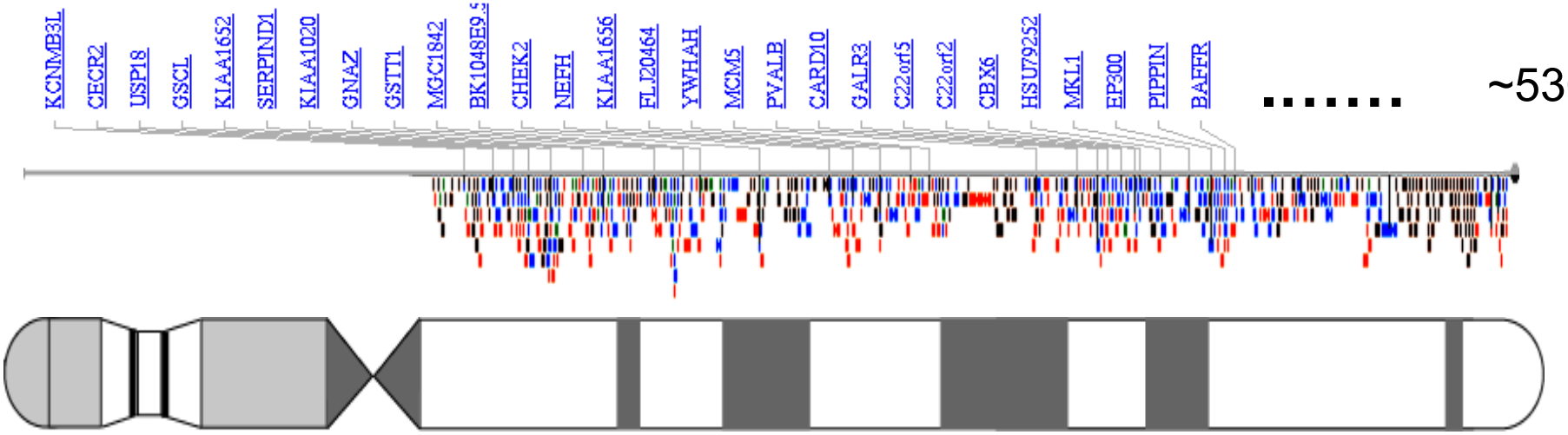
slides at **Lectures.GersteinLab.org**

Mark B Gerstein
Yale

(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)

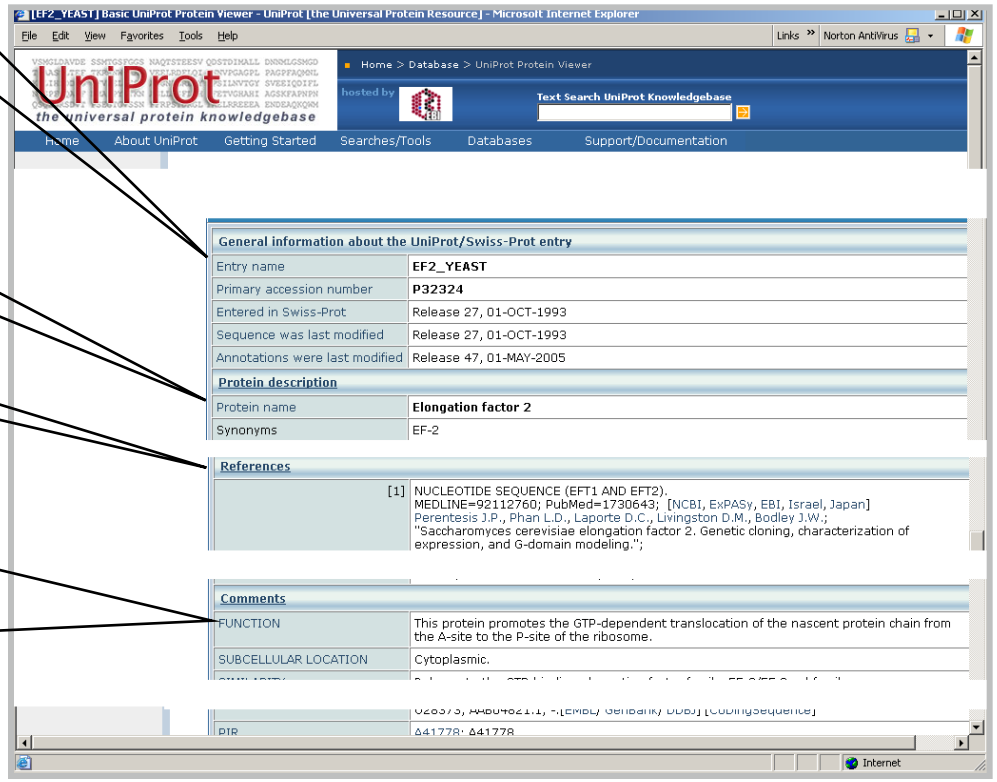
Traditional single molecule way to integrate evidence & describe function

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.

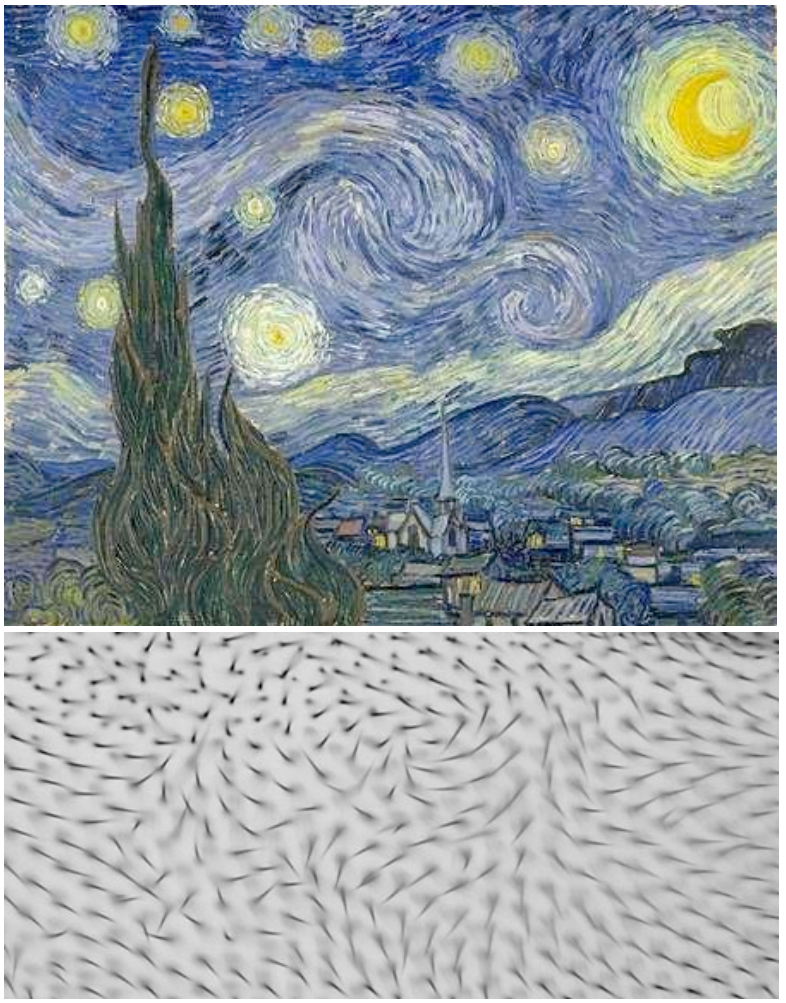


Some obvious issues in scaling single molecule definition to a genomic scale

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

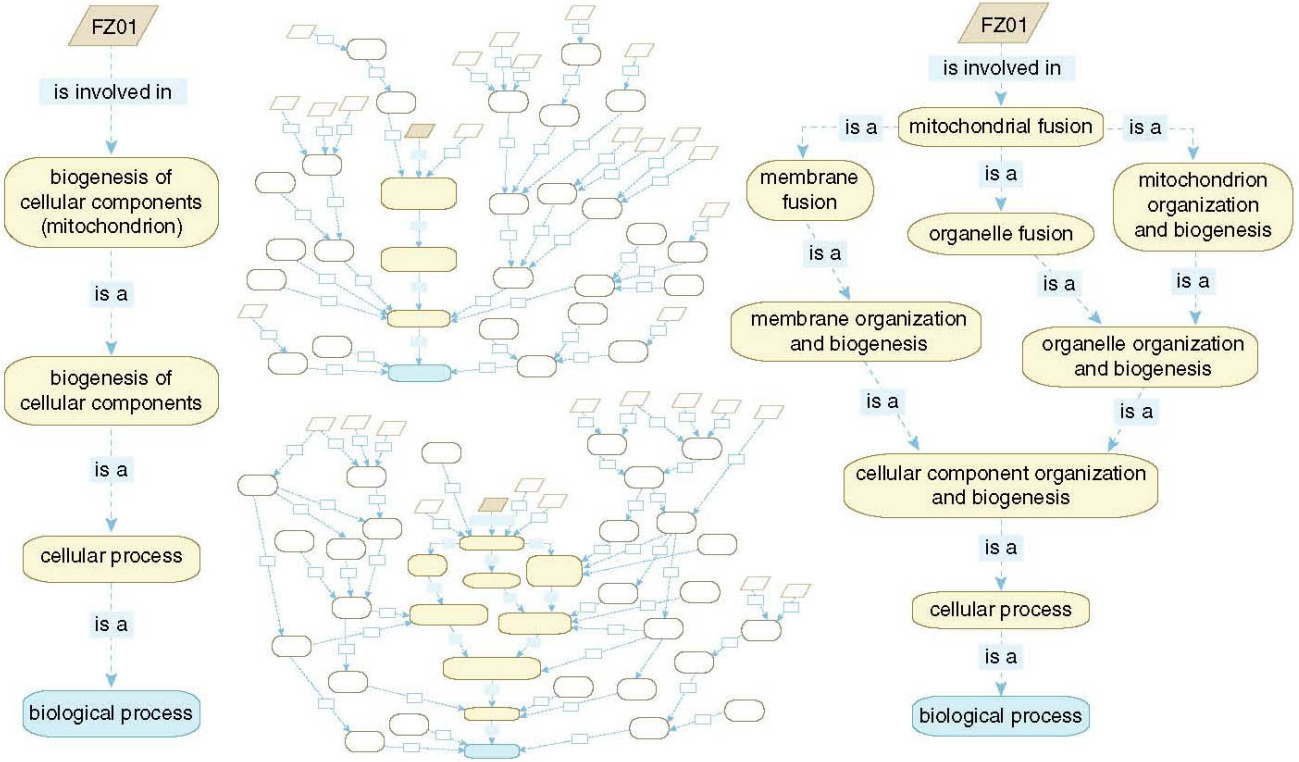
Some obvious issues in scaling single molecule definition to a genomic scale

- 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)]

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

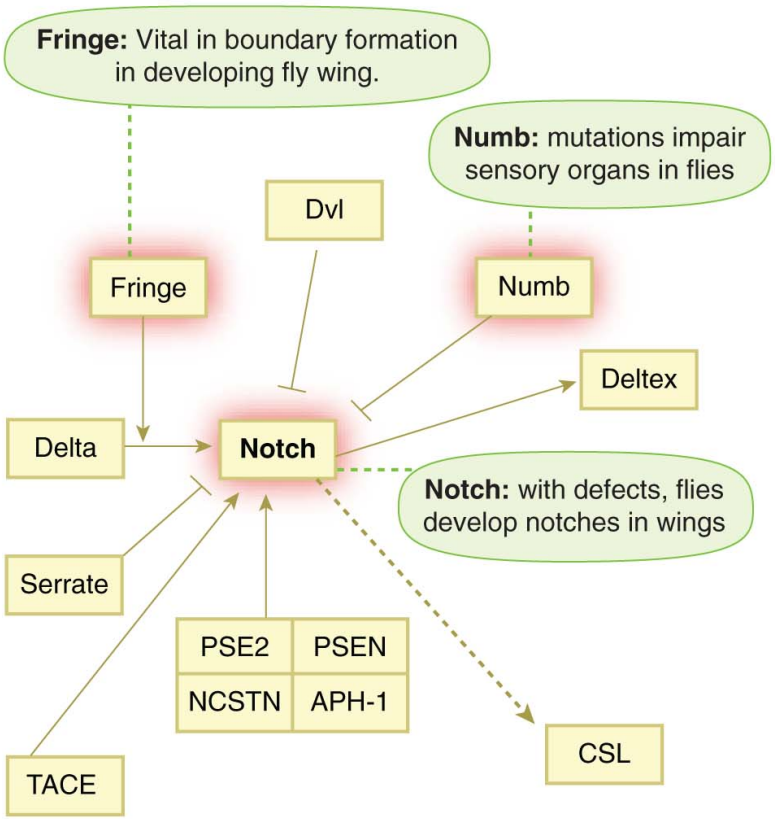
- Subjecting each gene to standardized expt. and cataloging effect
  - ◊ KOs of each gene in a variety of std. conditions => phenotypes
  - ◊ Std. binding expts for each gene (e.g. prot. chip)

- Function as a vector

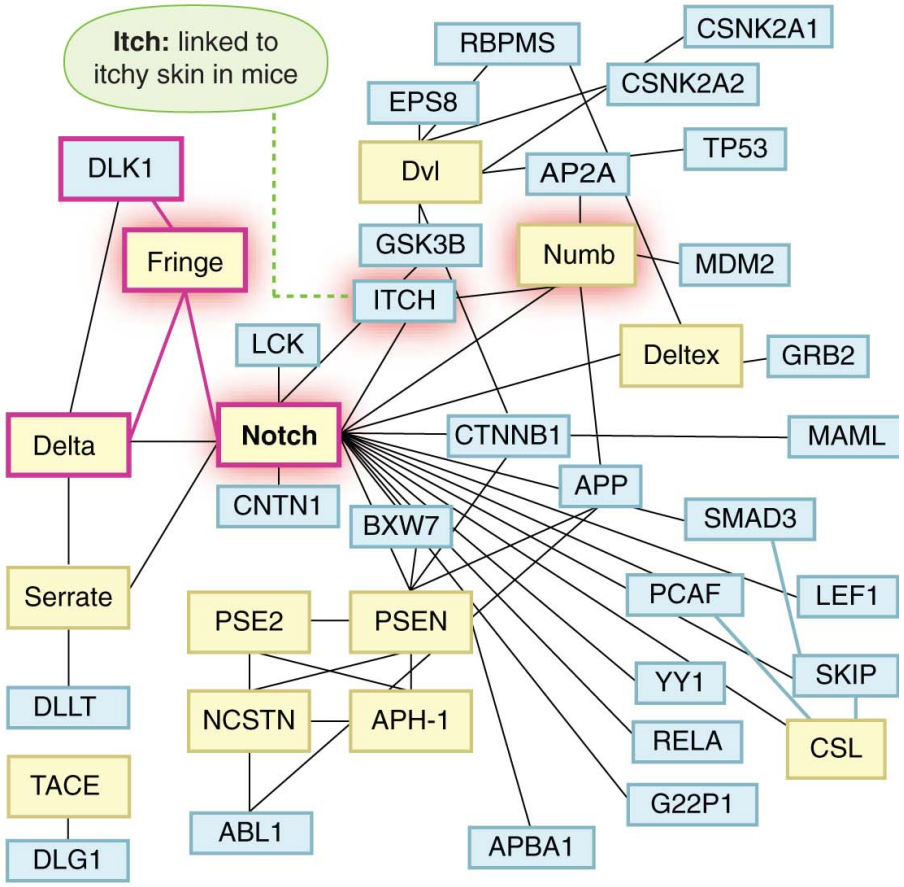
	nucleic acids		small molecules					proteins			
	DNA	RNA	ATP	Metal	CoA	NAD	.....	G protein	CDC28	Calmodulin	.....
protein 1	1.0	0	0	0	0	0	.....	0	0	0	.....
protein 2	0	0.9	0	0	0	0	.....	0	0	0	.....
protein 3	1.0	0	1.0	0	0	0	.....	0	0	0	.....
protein 4	0	0	0	0	0.8	0	.....	0	0	1.0	.....
protein 5	1.0	0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



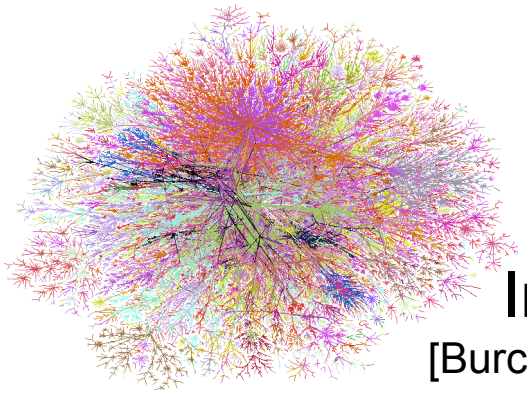
Classical KEGG pathway



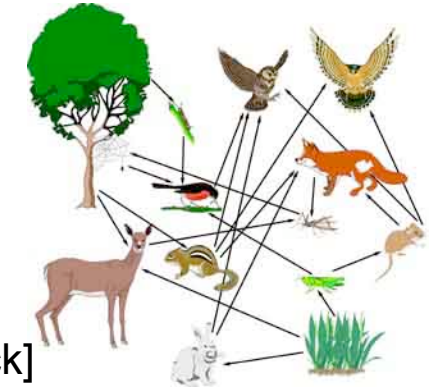
Same Genes in High-throughput Network

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

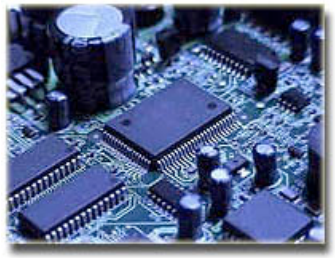
Networks as a universal language



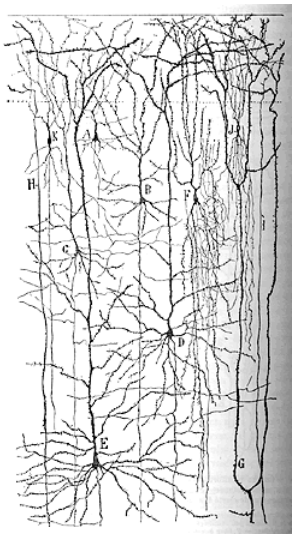
Internet
[Burch & Cheswick]



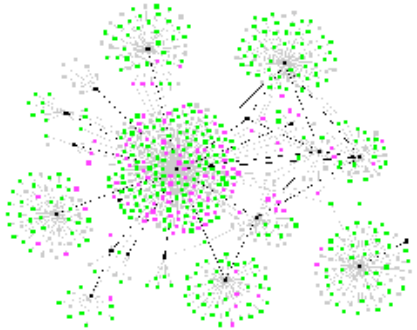
Food Web



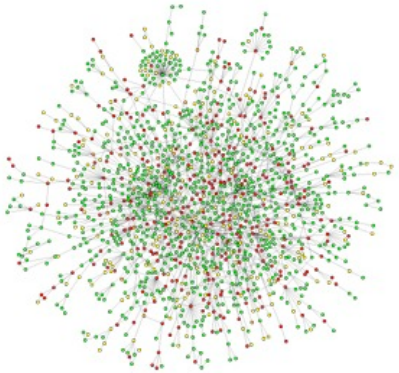
Electronic Circuit



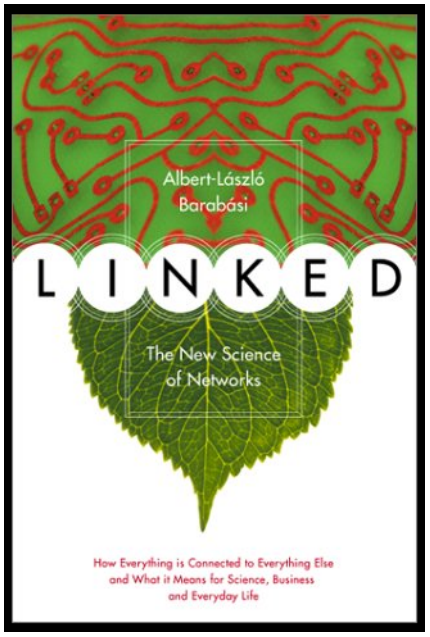
Neural Network
[Cajal]

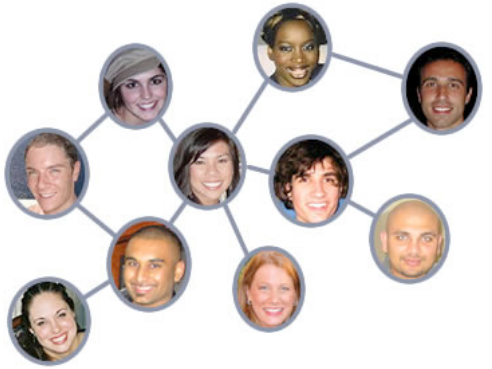


Disease Spread
[Krebs]



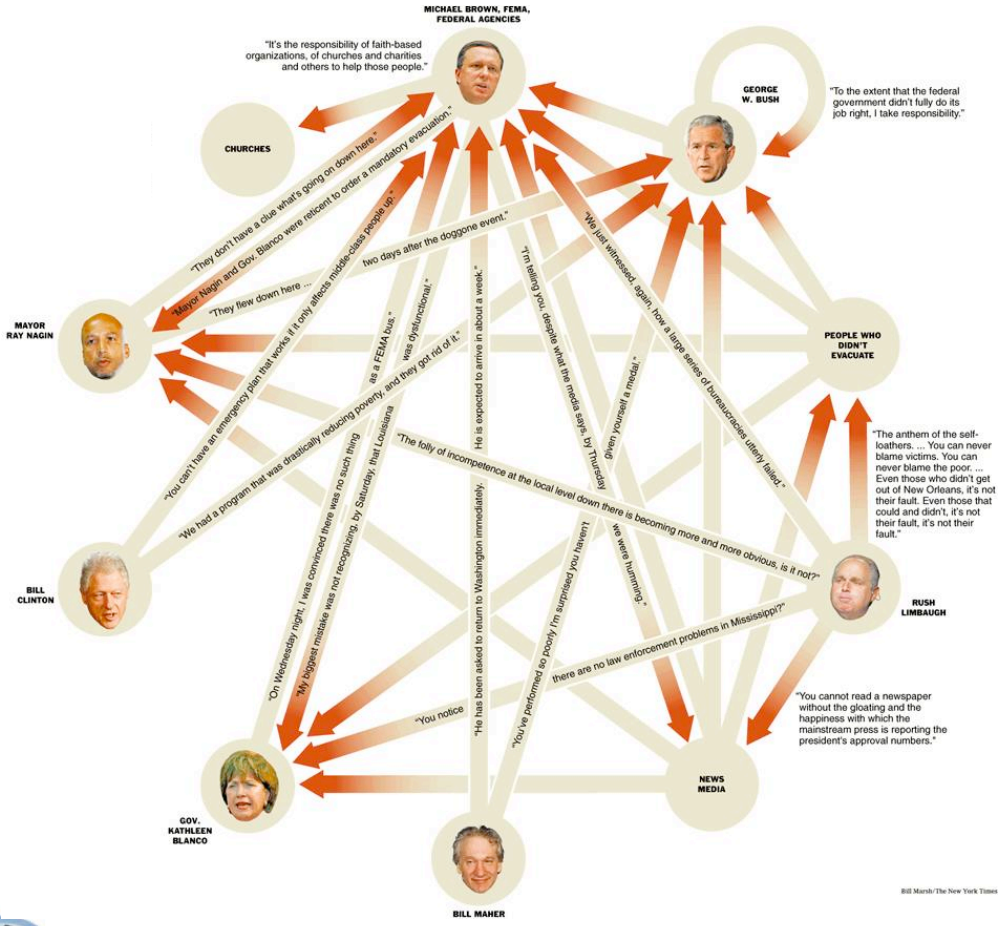
Protein Interactions
[Barabasi]



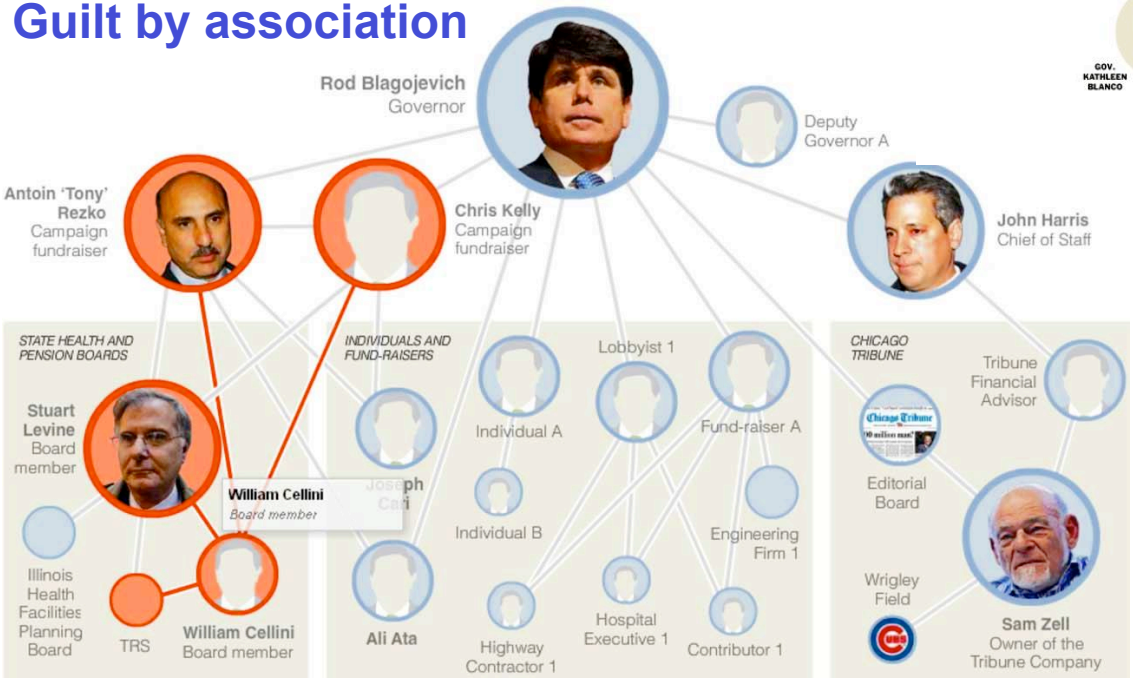


Social Network

Using the position in networks to describe function



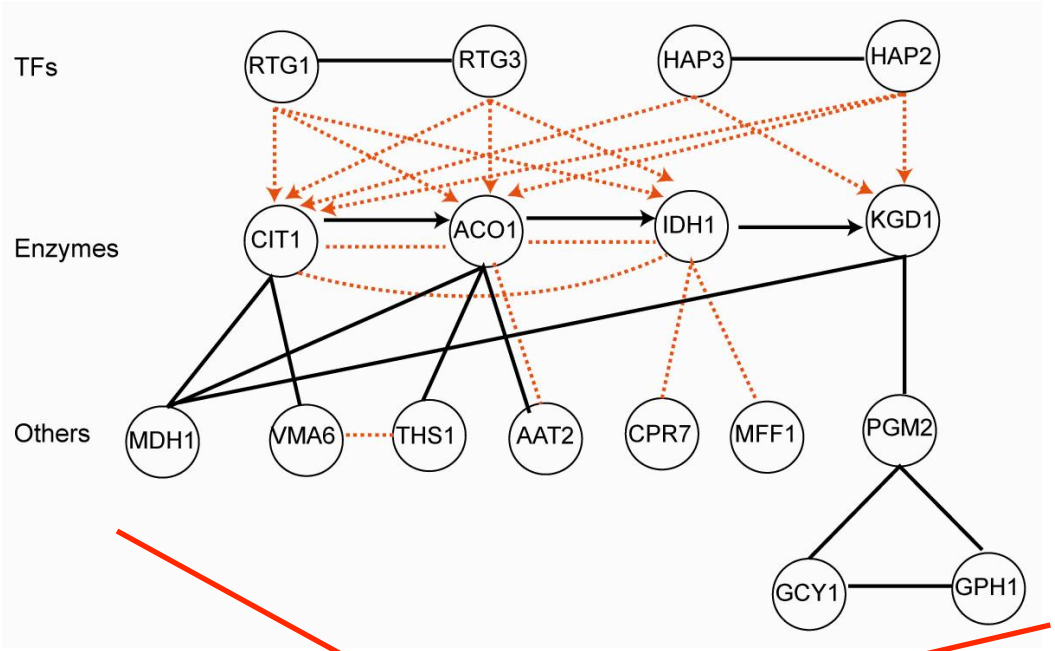
Guilt by association

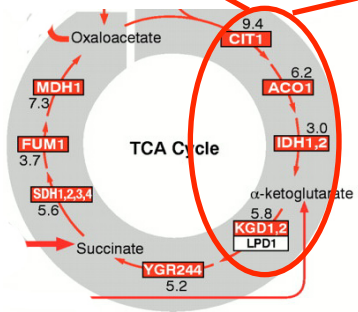


Finding the causal regulator (the "Blame Game")

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

Combining networks forms an ideal way of integrating diverse information





Part of the TCA cycle

Metabolic pathway

Transcriptional regulatory network

Physical protein-protein Interaction

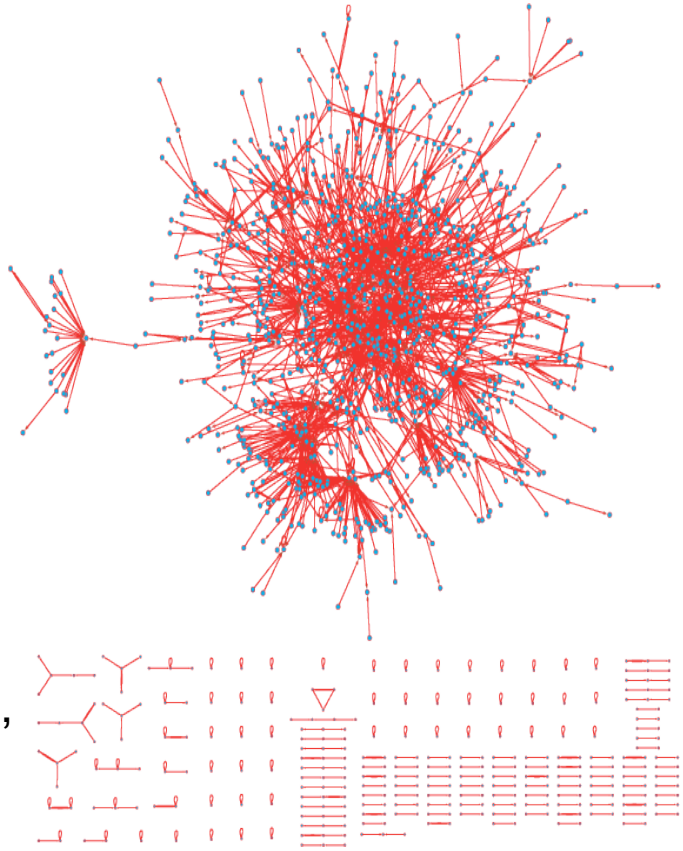
Co-expression Relationship

Genetic interaction (synthetic lethal)
Signaling pathways

Outline: Molecular Networks

- Why Networks?
- Predicting Networks (yeast)
  - ◊ Propagating known information
- Dynamics & Variation of Networks
  - ◊ Across environments (in prokaryotes)

Example: yeast PPI network



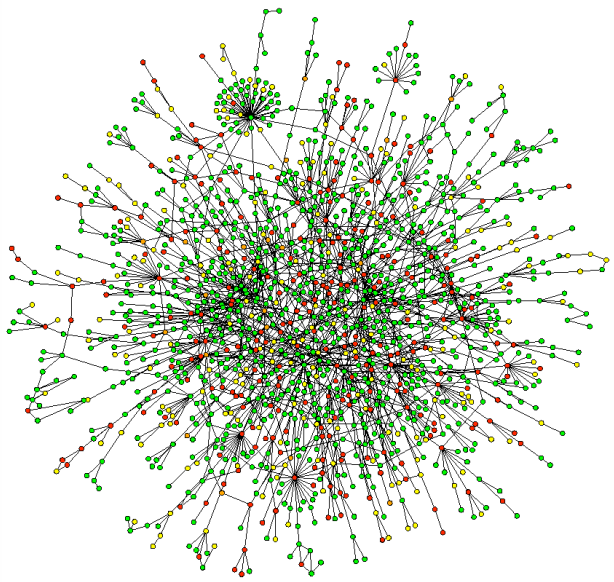
Actual size:

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

Known interactions:

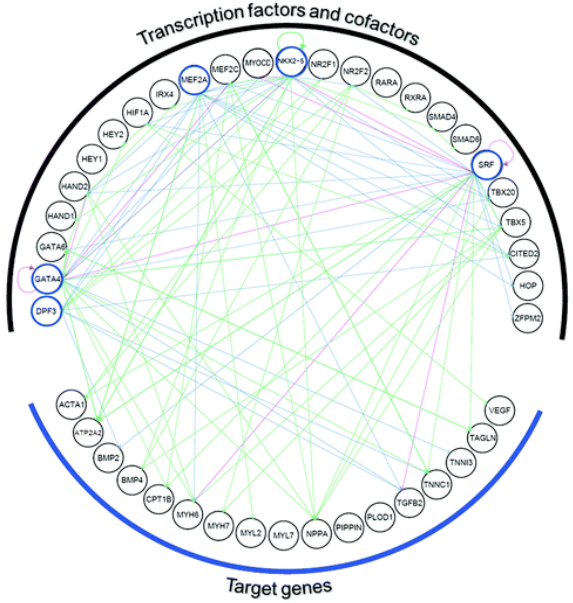
- Small-scale experiments: accurate but few
  → Overfitting: ~5,000 in BioGRID, involving ~2,300 proteins
- Large-scale experiments: abundant but noisy
  → Noise: false +ve/-ve for yeast two-hybrid data up to
  45% and 90% (Huang et al., 2007)

Types of Networks

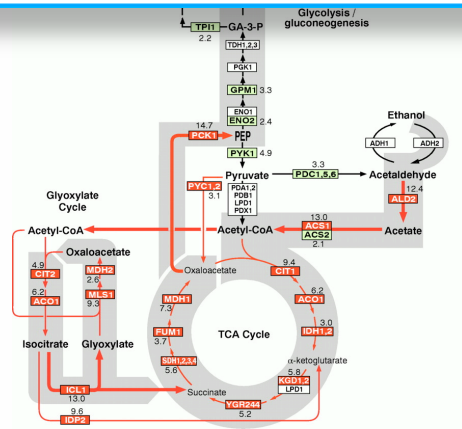


Interaction networks

Nodes: proteins or genes
Edges: interactions



Regulatory networks



Metabolic networks

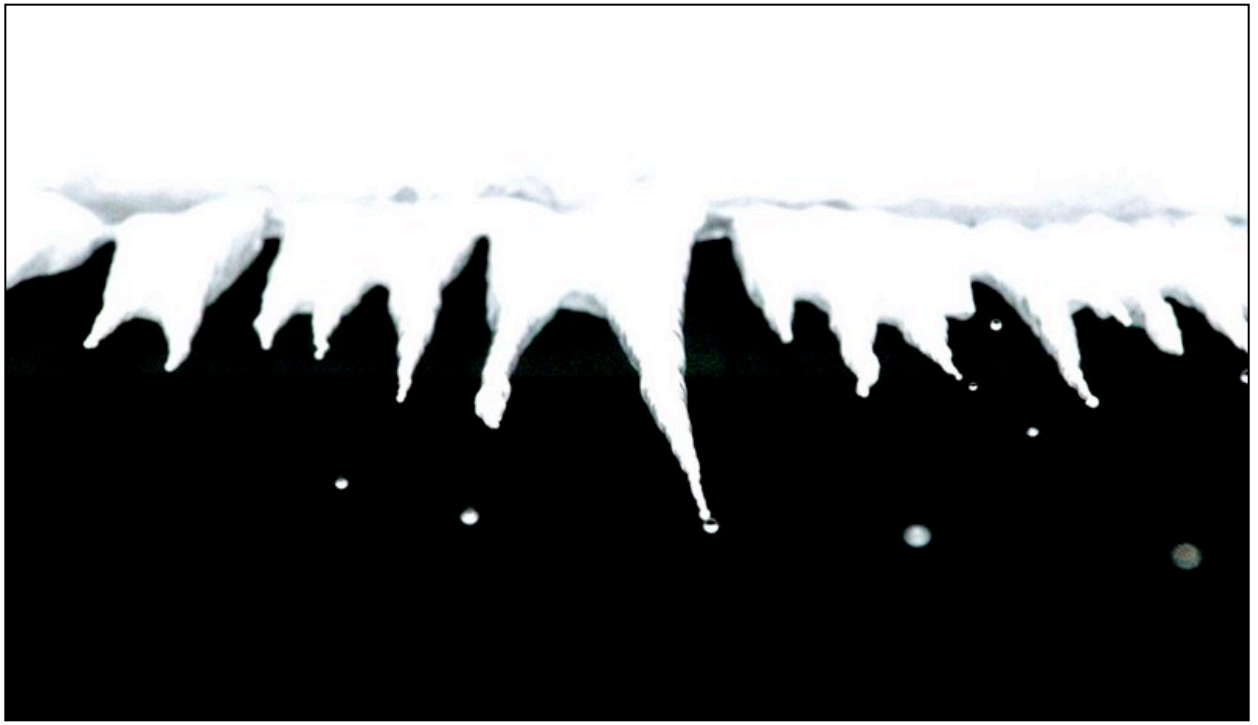
[Horak, et al, Genes & Development, 16:3017-3033]

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

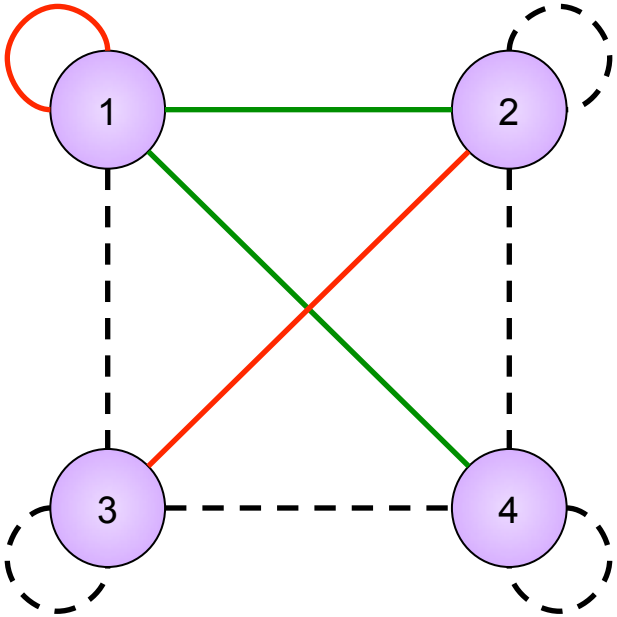
[Jeong et al, Nature, 41:411]

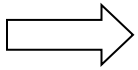
Predicting Networks

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



Training sets





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

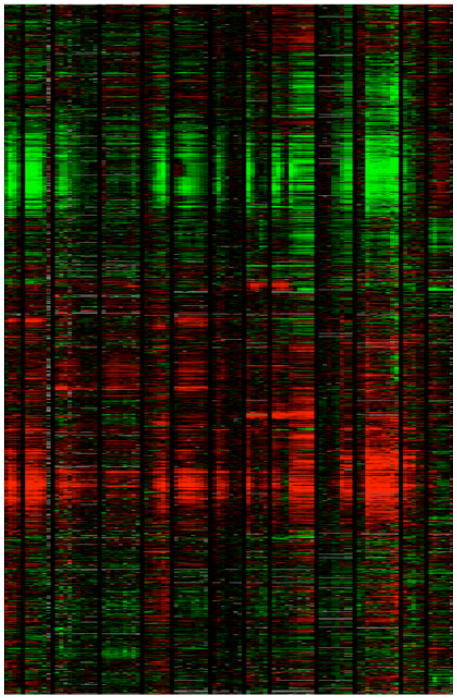
Known interactions

Known non-interactions

Unknown

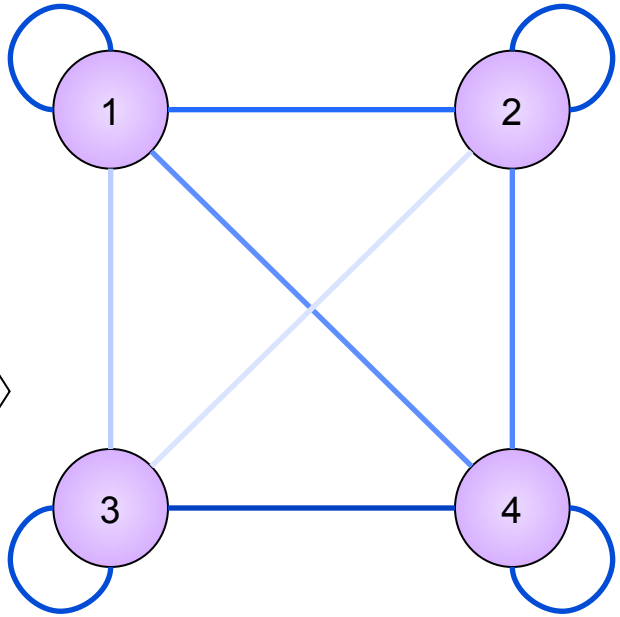
Network prediction: features

- Example 1: gene expression



Gasch et al., 2000

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

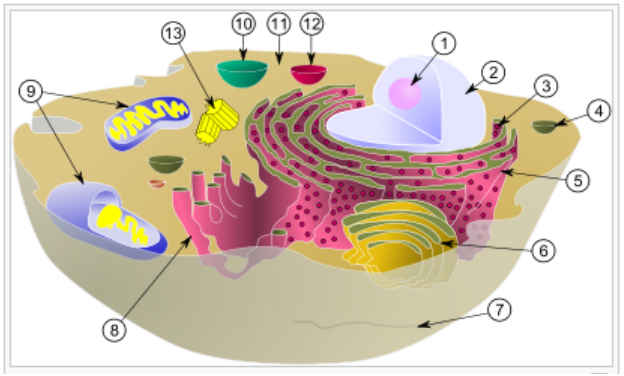


Similarity scale:

1 -1

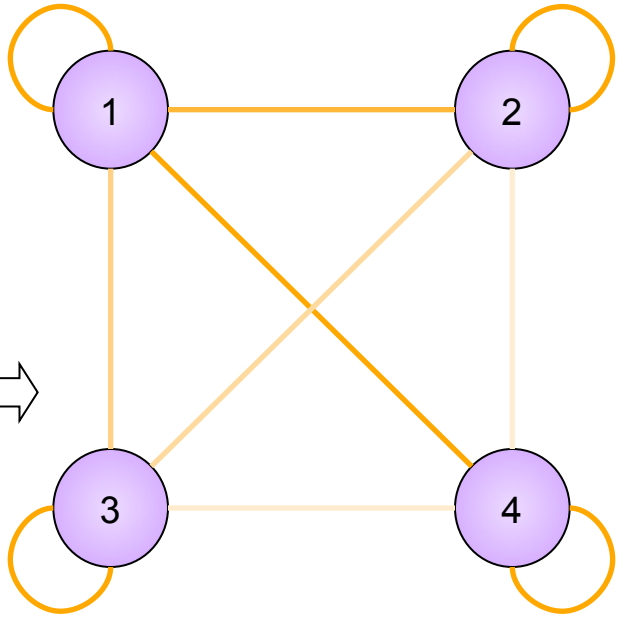
Network prediction: features

- Example 2: sub-cellular localization



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

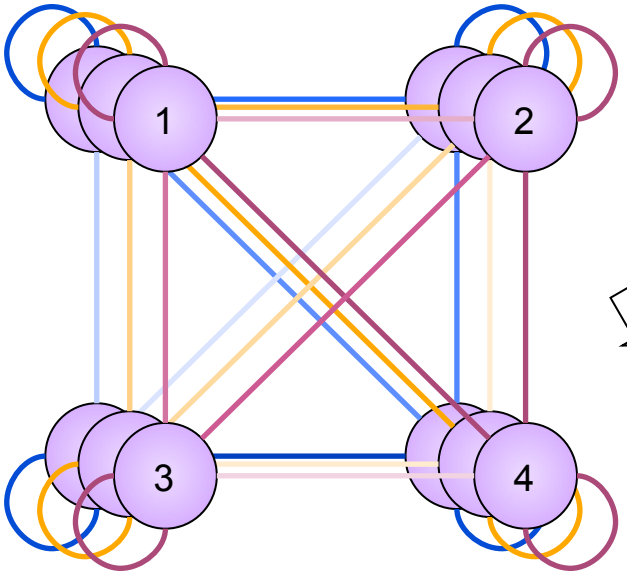
$x_1 = (1, 1, 0, 0, \ldots)$
$x_2 = (1, 1, 1, 0, \ldots)$
$x_3 = (1, 0, 1, 0, \ldots)$
…
$sim(x_1, x_2) = 0.81$
$sim(x_1, x_3) = 0.12$
…

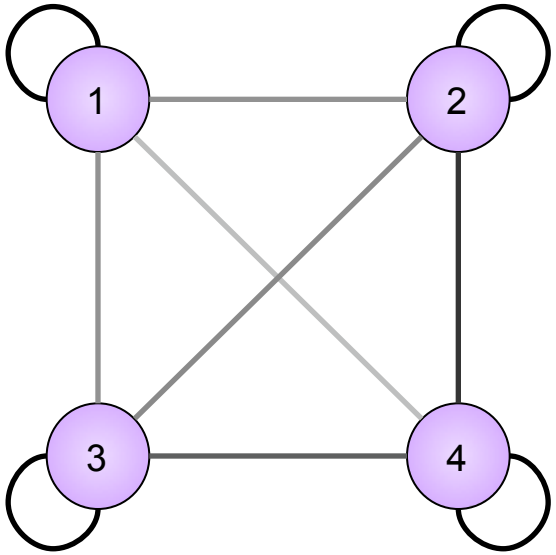


Similarity scale:

1 **-1**

Data integration & Similarity Matrix





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

Learning methods

An endless list:

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

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

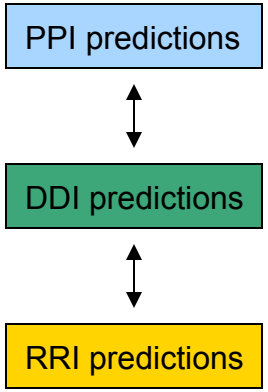
Our work: efficiently propagating known information

Training set expansion

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

Multi-level learning

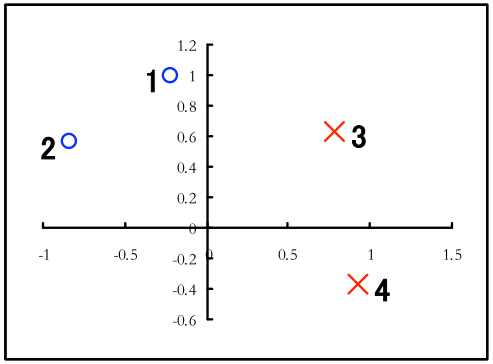
- Motivation: hierarchical nature of interaction
- Expand training sets vertically



DREAM3 *in silico* regulatory network reconstruction challenge

Kernels

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



Objects in an feature space

Compute inner products

p.s.d. implies

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

Similarity matrix

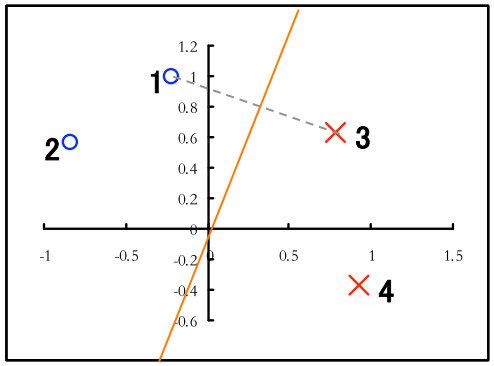
Good for integrating heterogeneous datasets (protein sequences, PSSM, gene expression, …)

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

Kernel methods

Use the kernel as proxy to work in the feature space

Example: SVM (finding the best separating hyperplane)



Equivalent to

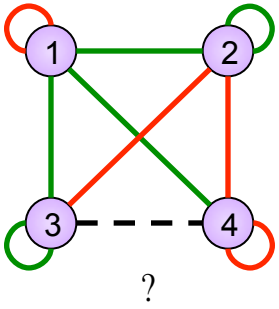
Maximize $\sum_i \lambda_i - \frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle$

Subject to $\lambda \geq 0$

$\sum_i \lambda_i y_i = 0$

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

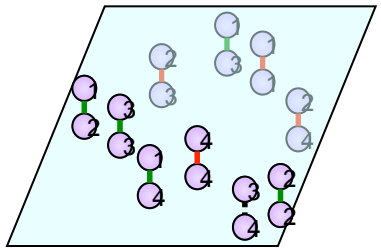
Kernel methods for predicting networks: local vs. global modeling



Global modeling: build one model for the whole network

Pairwise kernel: consider object pairs instead of individual objects

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



Direct methods: threshold the kernel to make predictions

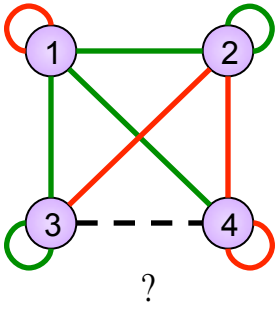
Problem: One single global model, may not be able to handle subclasses

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

Threshold: 0.7

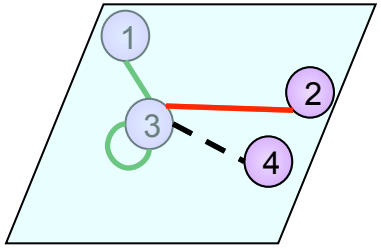
	1	2	3	4
1	1.00	0.57	0.55	0.40
2	0.57	1.00	0.66	0.89
3	0.55	0.66	1.00	0.79
4	0.40	0.89	0.79	1.00

Kernel methods for predicting networks: local vs. global modeling



Local modeling: build one model for each node

Model for node 3:



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

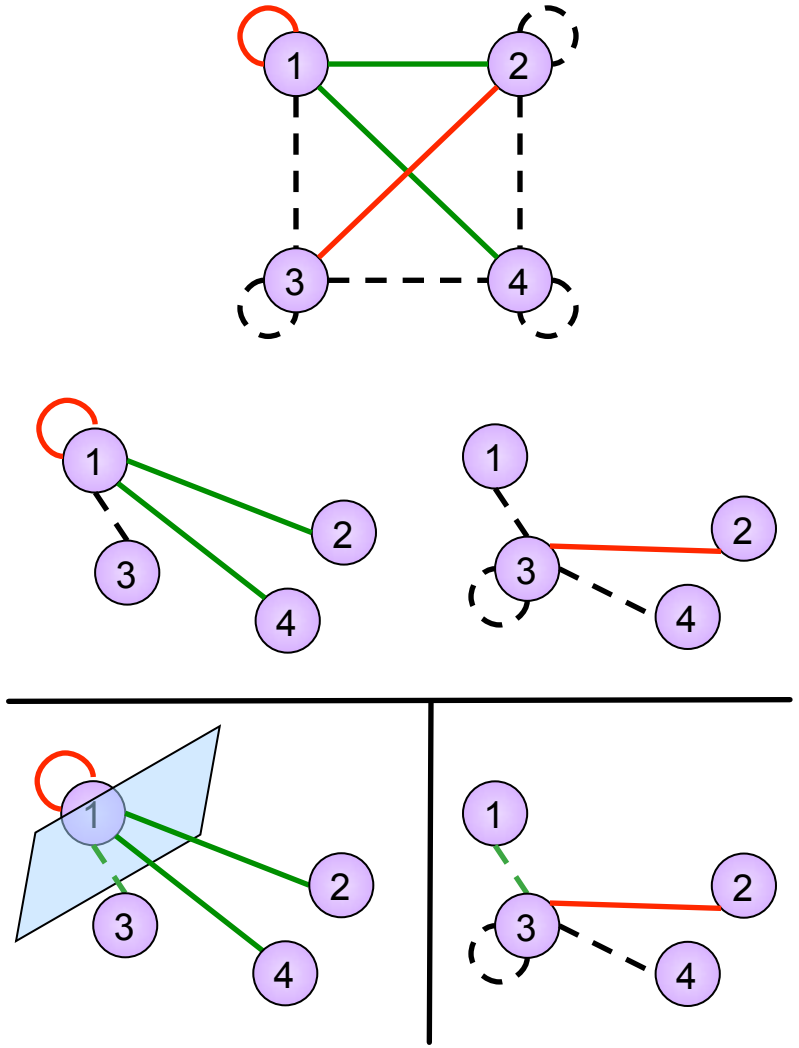
Our work: training set expansion

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

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

Prediction propagation

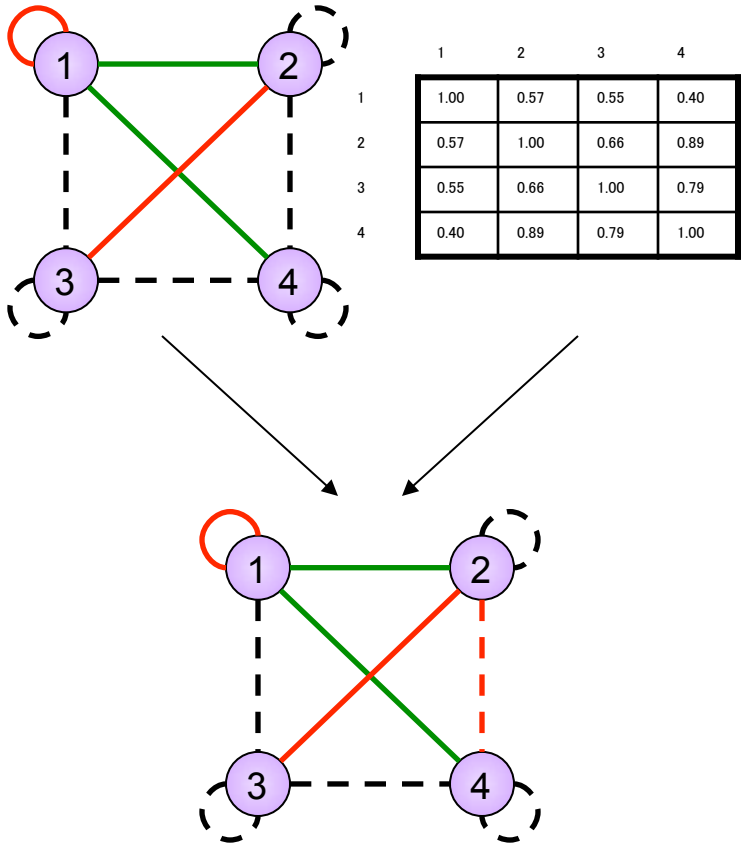
- Motivation: some objects have more examples than others
- Our approach:
  - ◊ Learn models for objects with more examples first
  - ◊ Propagate the most confident predictions as auxiliary examples of other objects



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

Kernel initialization

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



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

Remarks

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

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

Prediction accuracy (AUC)

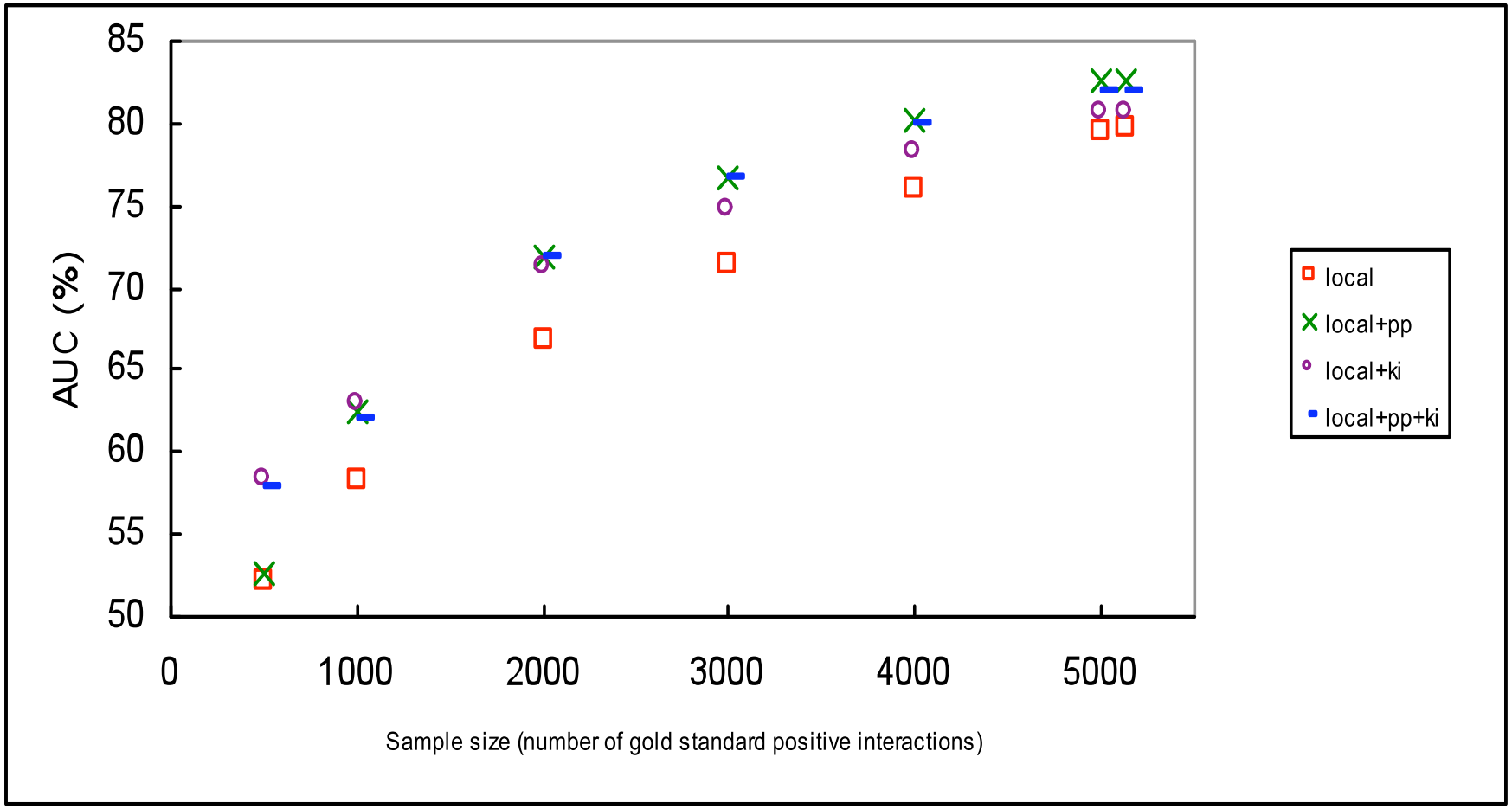
	phy	loc	exp-gasch	exp-spellman	y2h-ito	y2h-uetz	tap-gavin	tap-krogan	int
Mode 1									
direct	58.04	66.55	64.61	57.41	51.52	52.13	59.37	61.62	70.91
kCCA	65.80	63.86	68.98	65.10	50.89	50.48	57.56	51.85	80.98
kML	63.87	68.10	69.67	68.99	52.76	53.85	60.86	57.69	73.47
em	71.22	75.14	67.53	64.96	55.90	53.13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

Observations:

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

[Yip and Gerstein, Bioinformatics ('09)]

Complementarity of the two methods



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

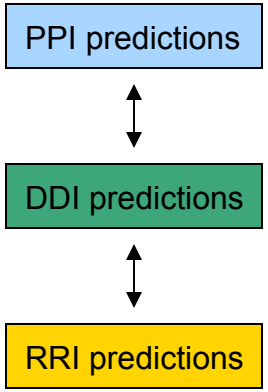
From horizontal to vertical

Training set expansion

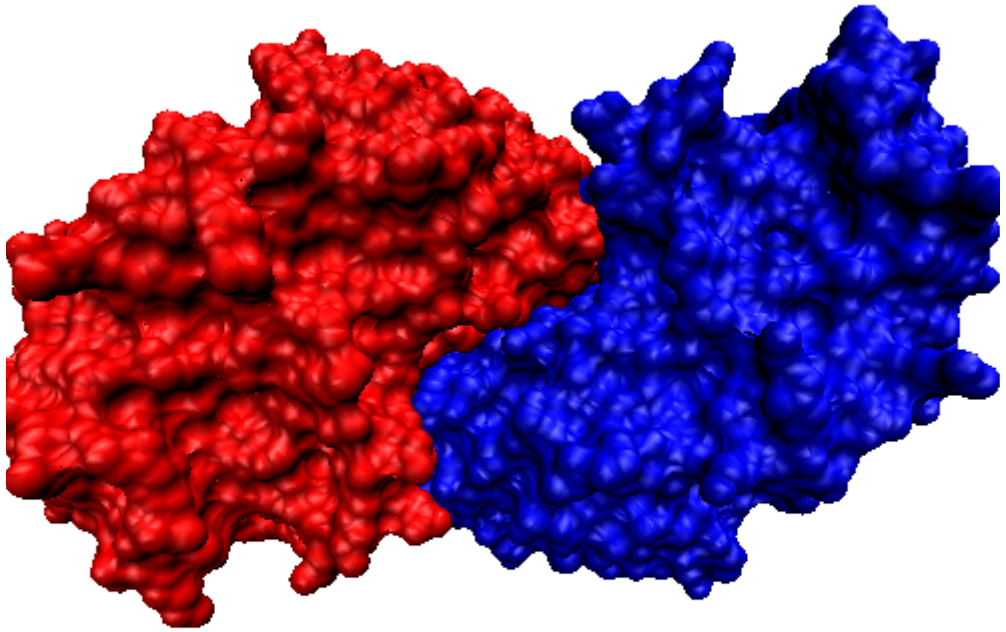
- Motivation: lack of training examples
- Expand training sets horizontally

Multi-level learning

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



Protein interaction

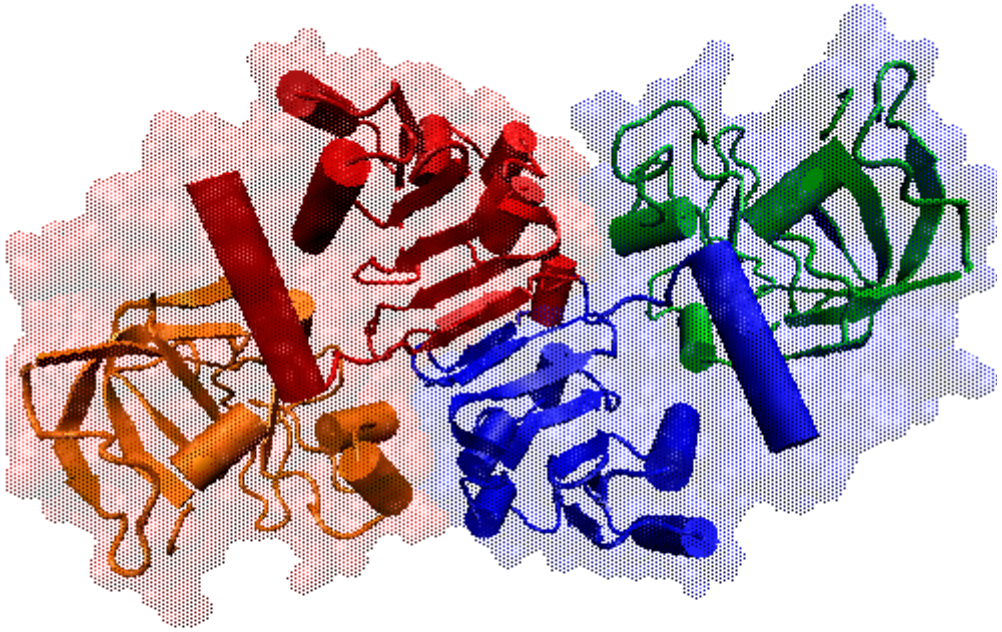


Yeast NADP-dependent alcohol dehydrogenase 6 (PDB: 1piw)

Protein-level features for interaction prediction: functional genomic information

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

Domain interaction

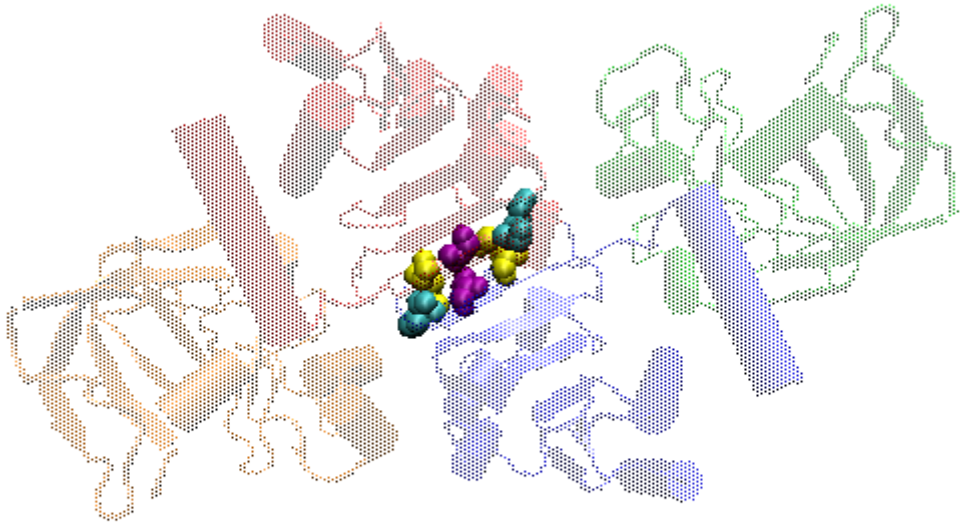


Pfam domains: PF00107 (inner) and PF08240 (outer)

Domain-level features for interaction prediction: evolutionary information

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

Residue interaction

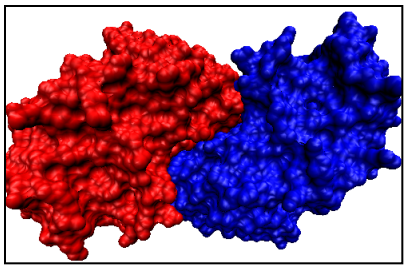


Interacting residues: 283 (yellow) with 287 (cyan), and 285 (purple) with 285

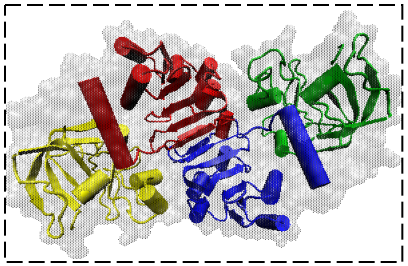
Residue-level features for interaction prediction: physical-chemical information

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

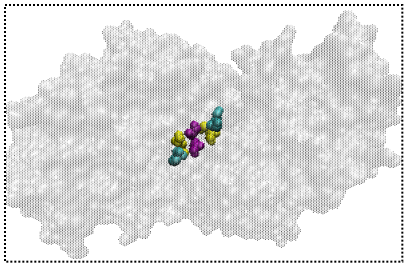
Combining the three problems



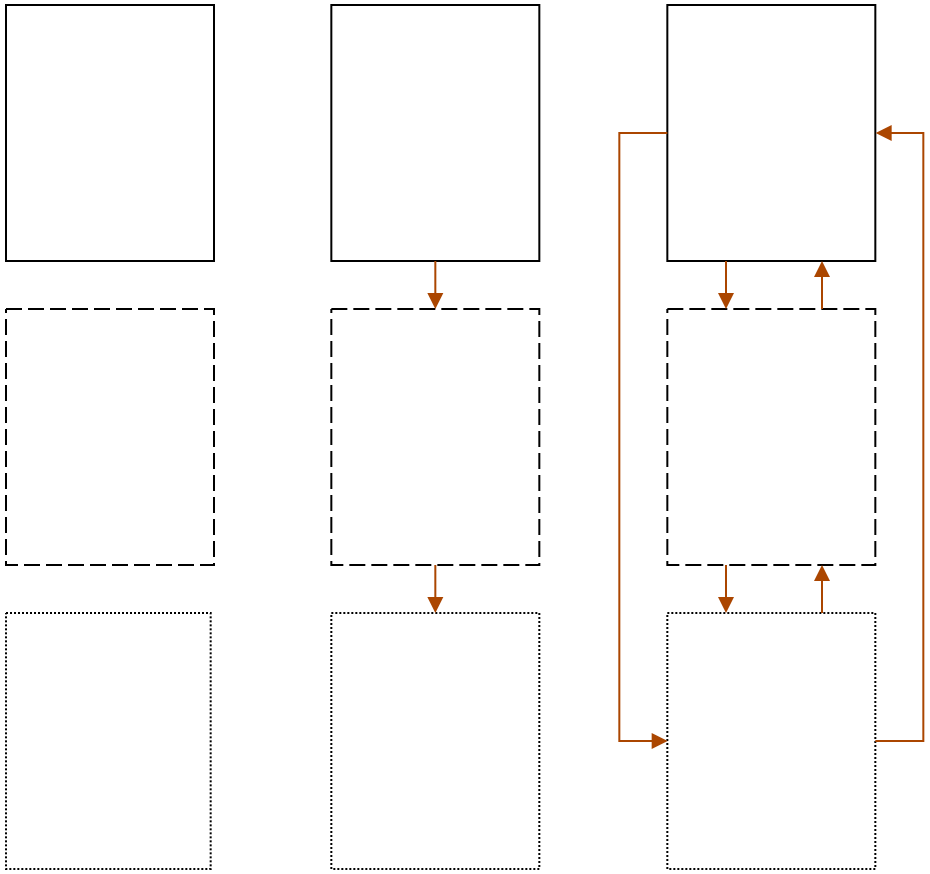
Protein interactions



Domain interactions



Residue interactions

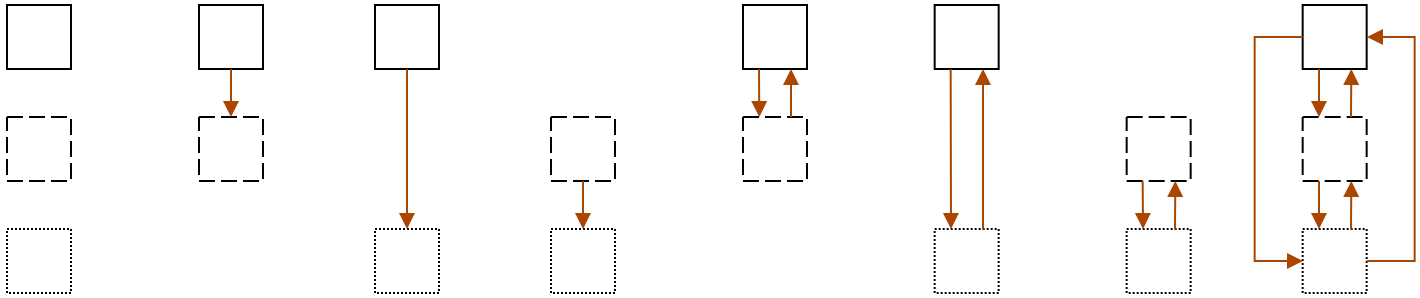


i. Independent levels ii. Unidirectional flow iii. Bidirectional flow

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

Empirical results (AUCs)

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

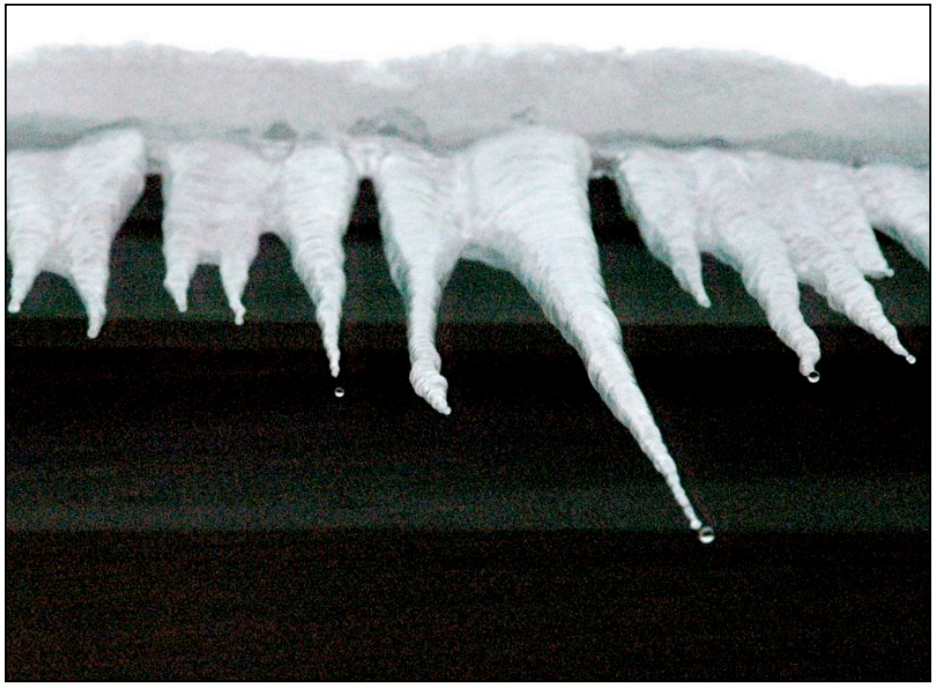


- Highest accuracy by bidirectional flow
- Additive effect: 2 vs. 3 levels

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

Network Dynamics #2: Environments

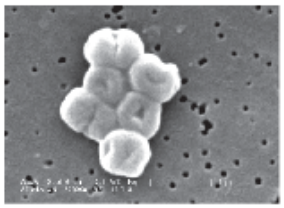
How do molecular networks change across environments?
What pathways are used more ?
Used as a biosensor ?



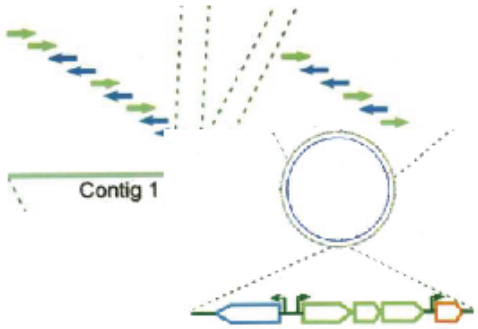
What is metagenomics?

Genomics Approach

Culture Microbes → Extract DNA → Sequence → Assemble and Annotate



ATCGTATA
CGCGAAG
ACGTCTGA
AGTGCTGCT

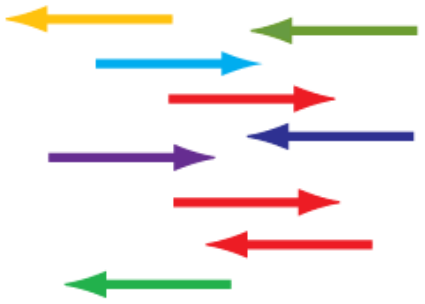


PROBLEM: Estimated that less than 1% can be cultured in the lab

Metagenomics Approach

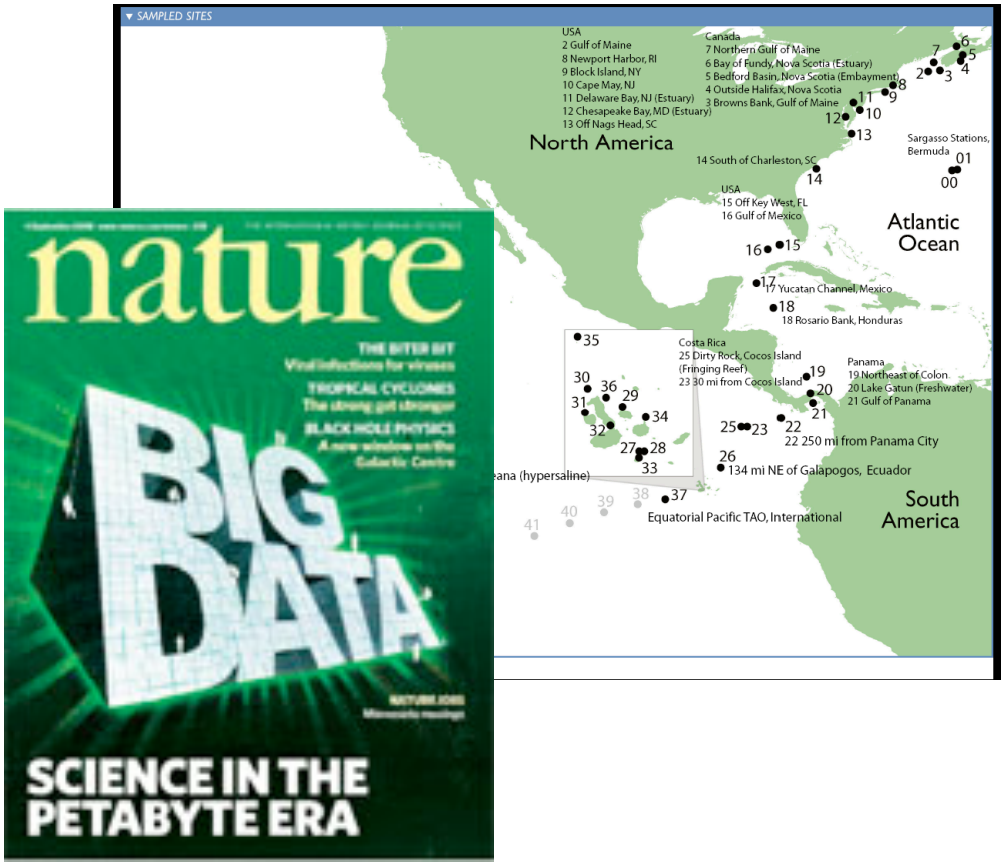
Collect Sample → Extract DNA → Sequence → Partially Assemble and Annotate

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



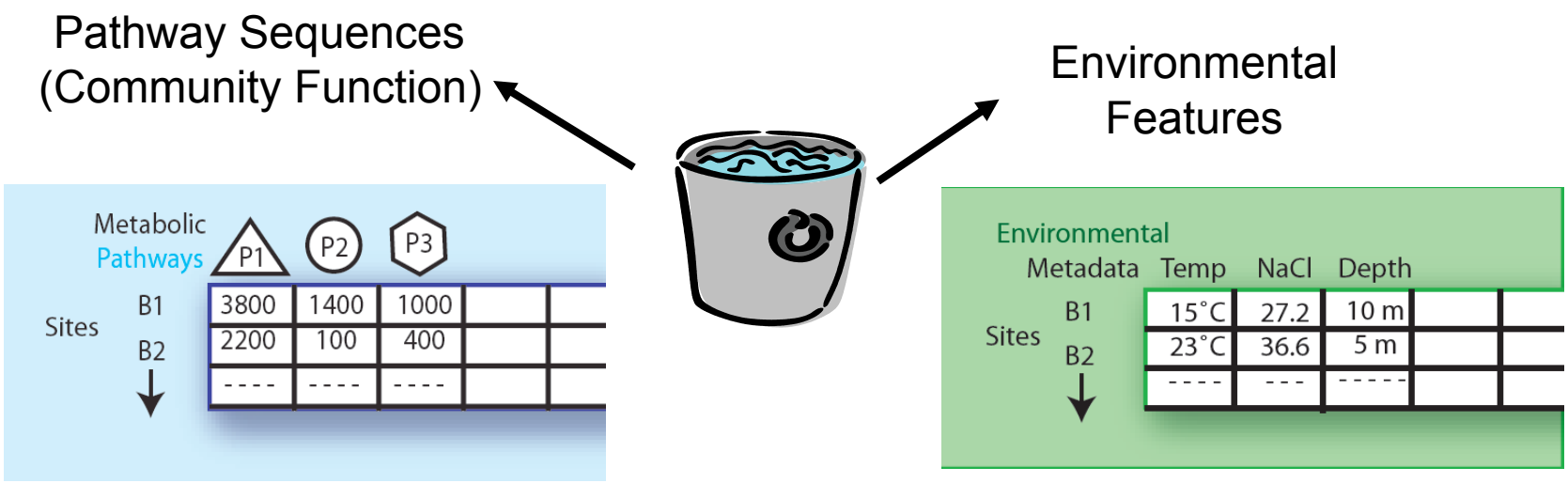
PROBLEM: Lose information about which gene belongs to which microbe.

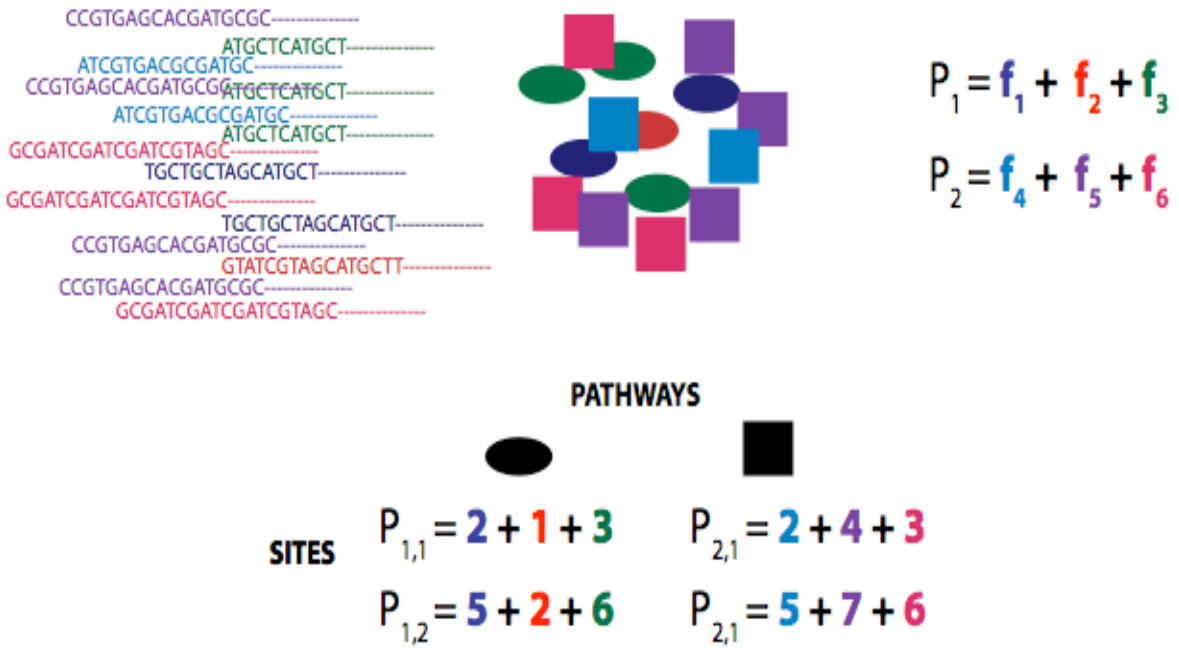
Global Ocean Survey Statistics (GOS)



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

Rusch, et al., PLOS Biology 2007



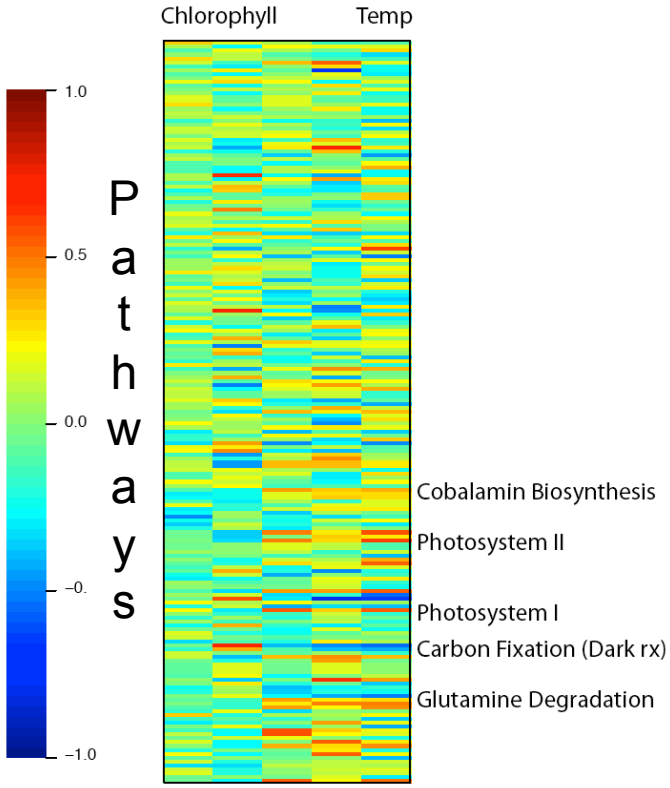


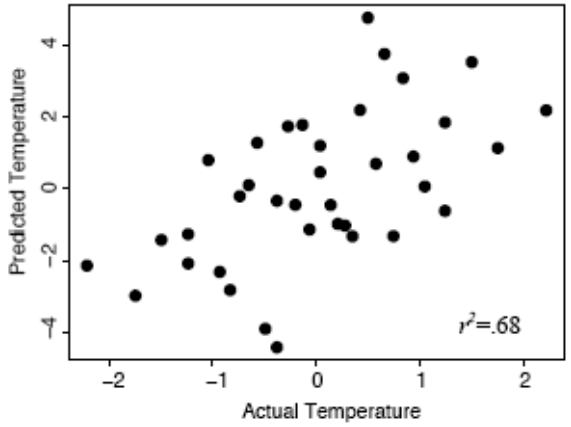
Expressing data as matrices indexed by site, env. var., and pathway usage

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

Simple Relationships: Pairwise Correlations

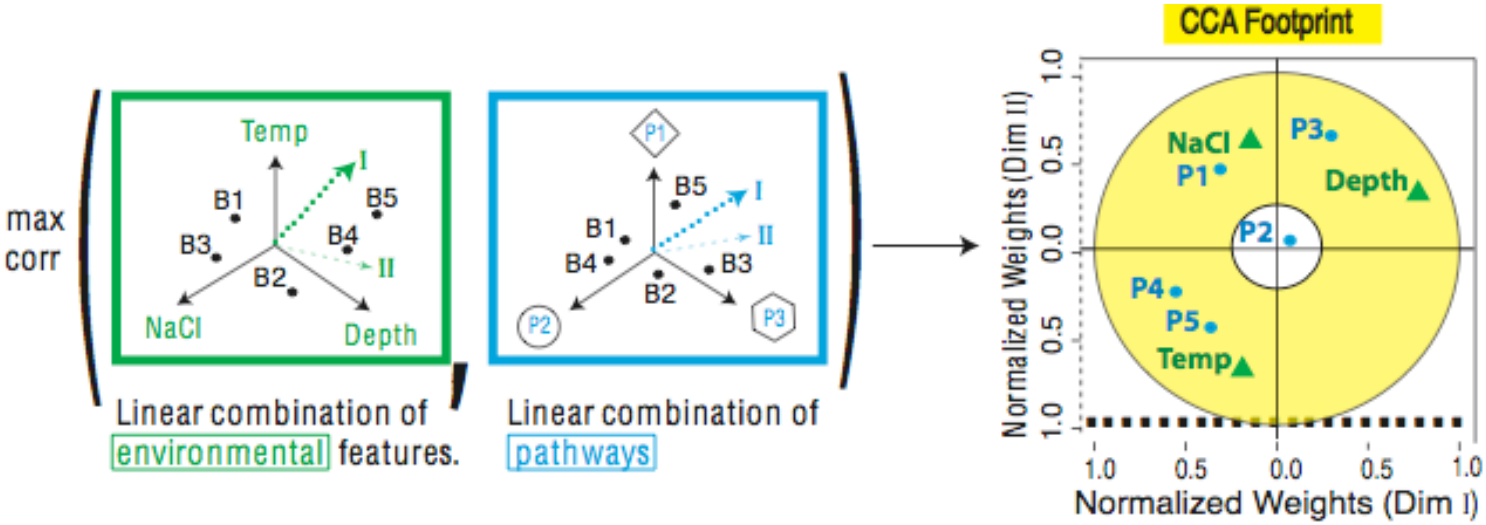
Environmental Features





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

Environmental-Metabolic Space



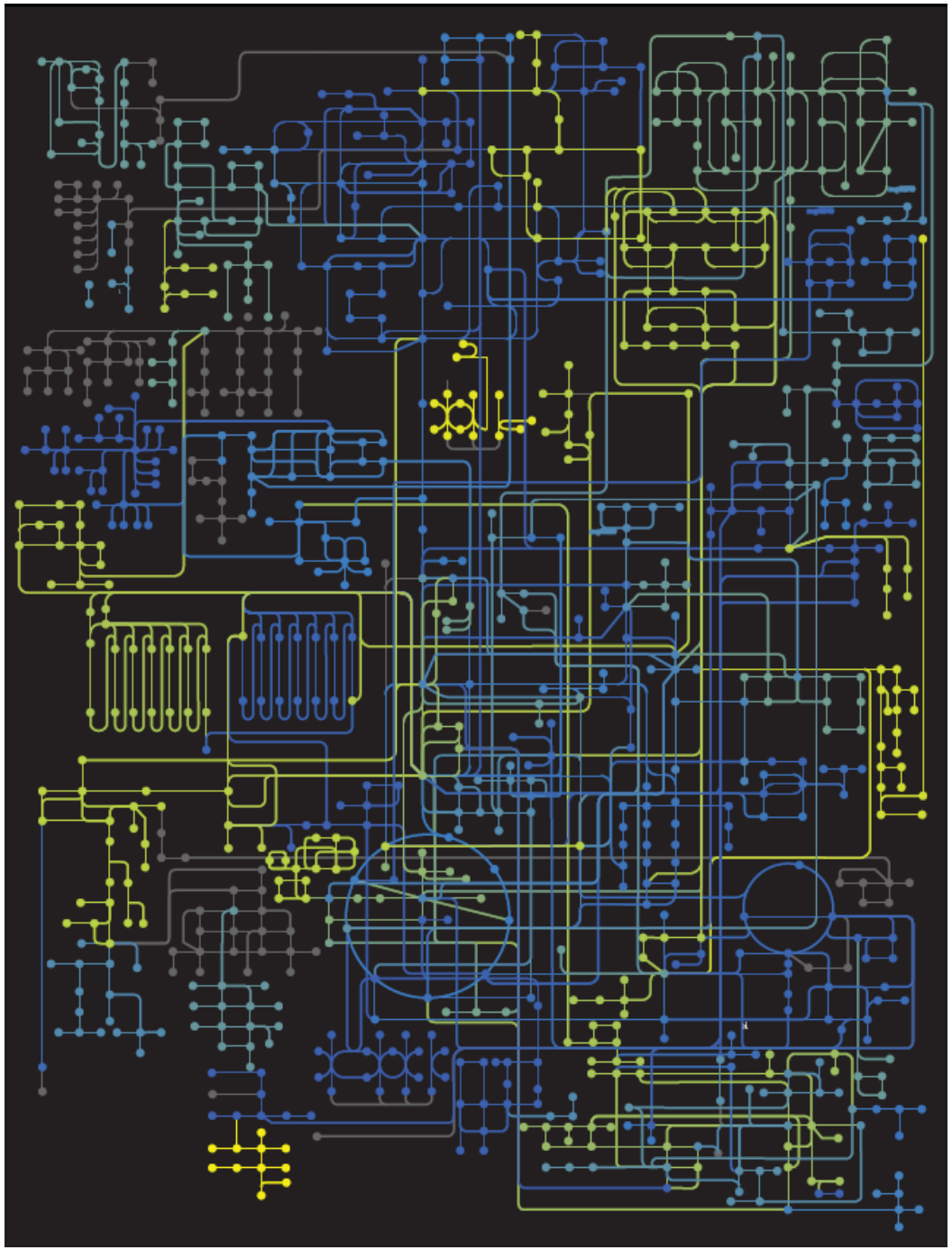
The goal of this technique is to interpret cross-variance matrices
We do this by defining a change of basis.

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

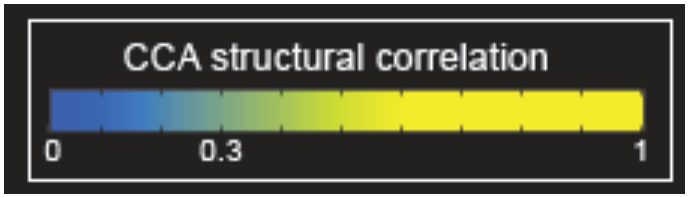
$$C = \begin{matrix} \Sigma_X & \Sigma_{X,Y} \\ \Sigma_Y & \Sigma_{Y,X} \end{matrix}$$

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

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

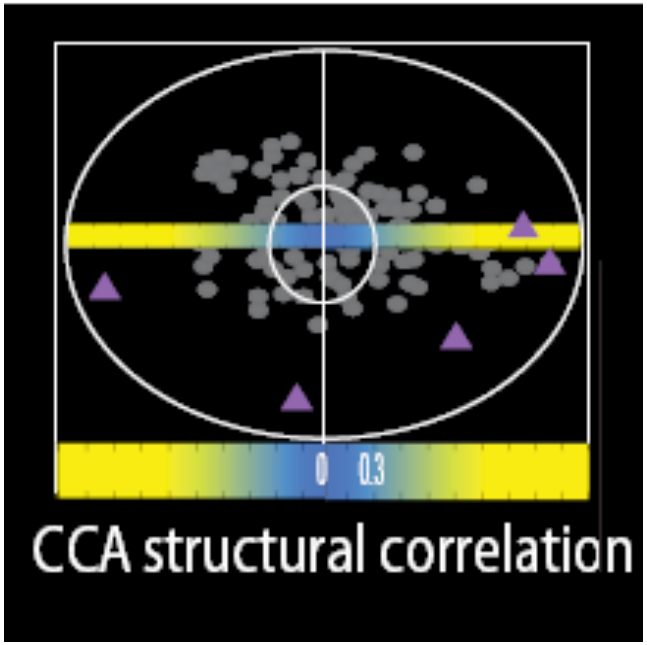


Strength of Pathway co-variation with environment



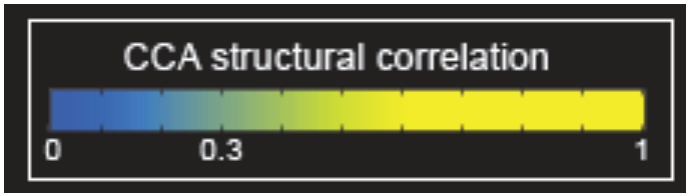
Environmentally invariant

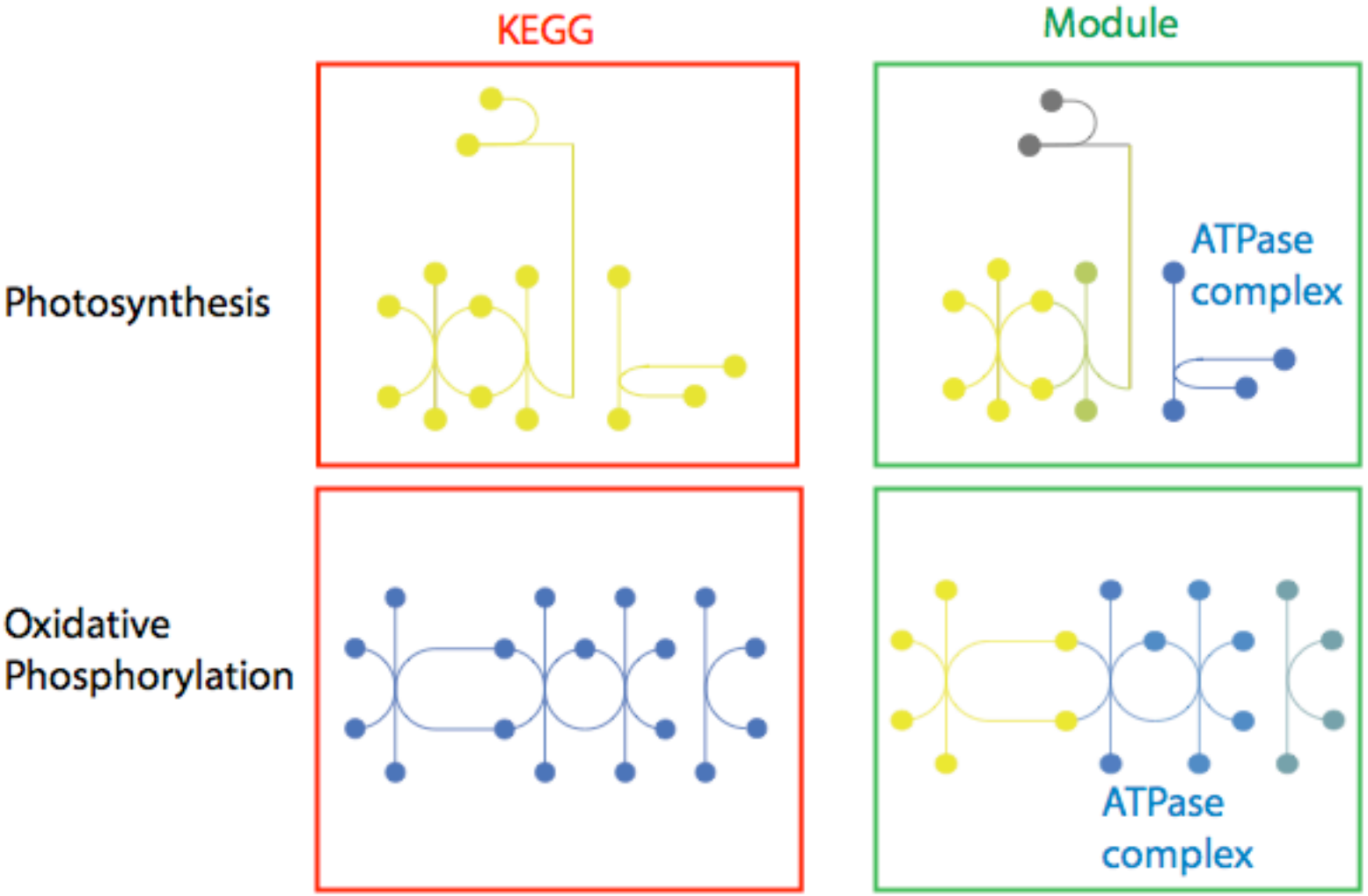
Environmentally variant



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

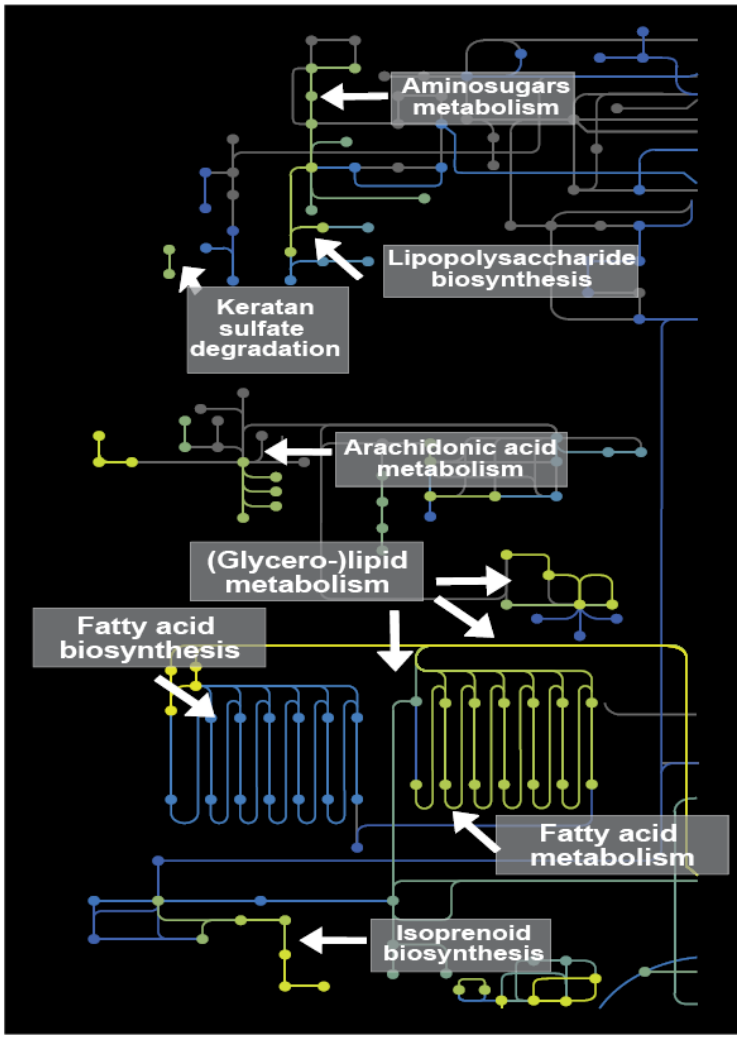
Conclusion #1: energy conversion strategy, temp and depth

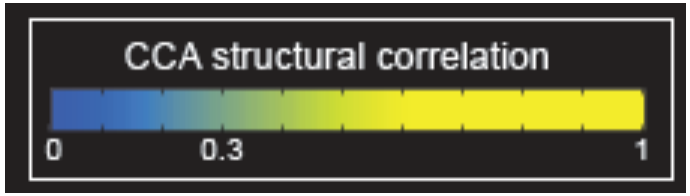




[Gianoulis et al., PNAS (in press, 2009)]

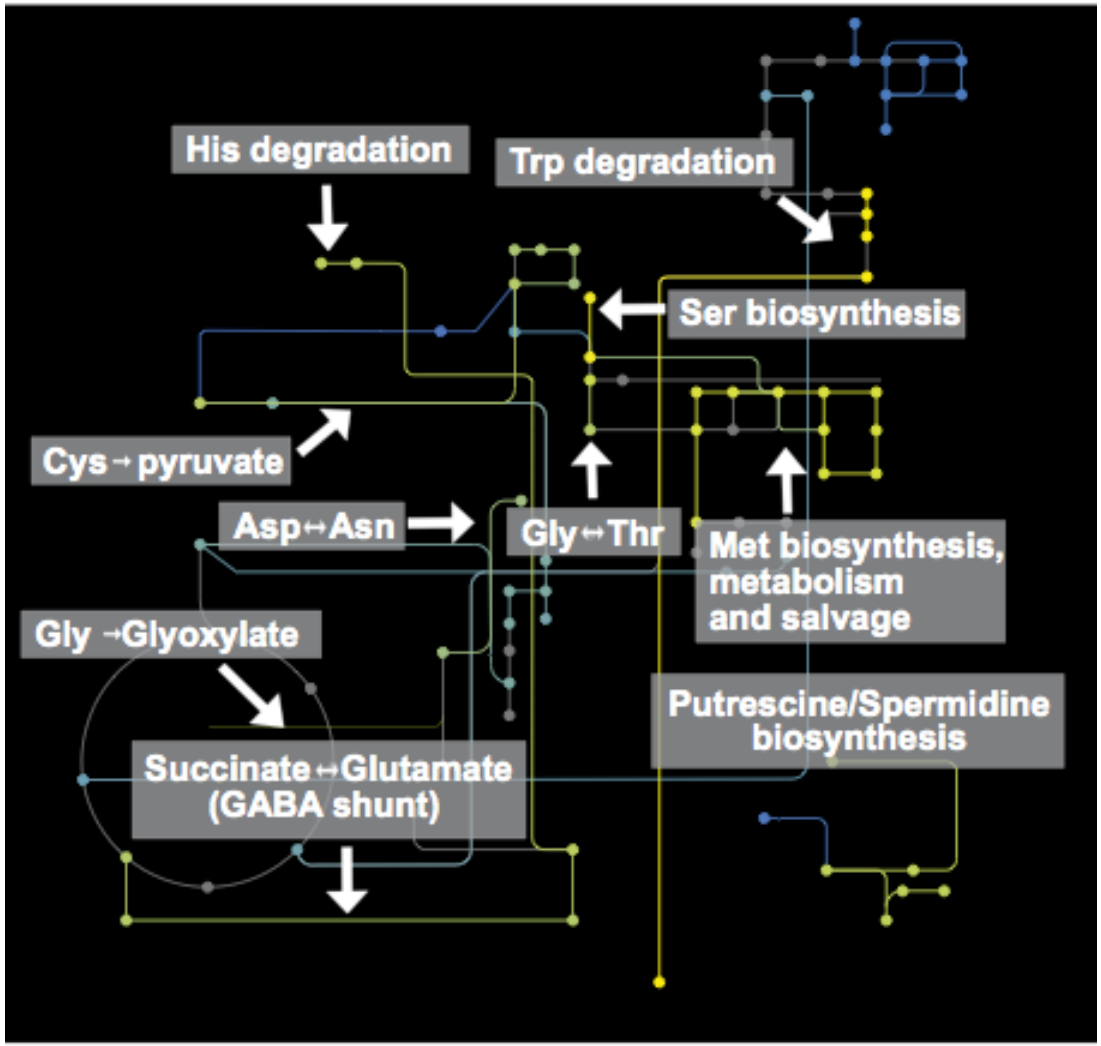
Conclusion #2: Outer Membrane components vary the environment





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

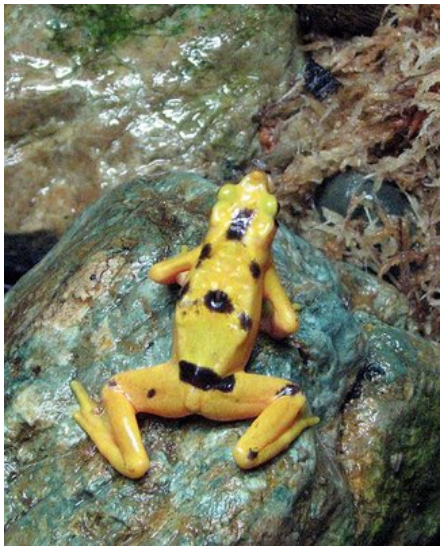
Conclusion #3: Covariation of AA biosynthesis and Import



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

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

Biosensors:
Beyond Canaries in a Coal Mine



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

Outline: Molecular Networks

- Why Networks?
- Predicting Networks (yeast)
  - ◊ Propagating known information
- Dynamics & Variation of Networks
  - ◊ Across environments (in prokaryotes)

Conclusions on Networks: Predictions

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

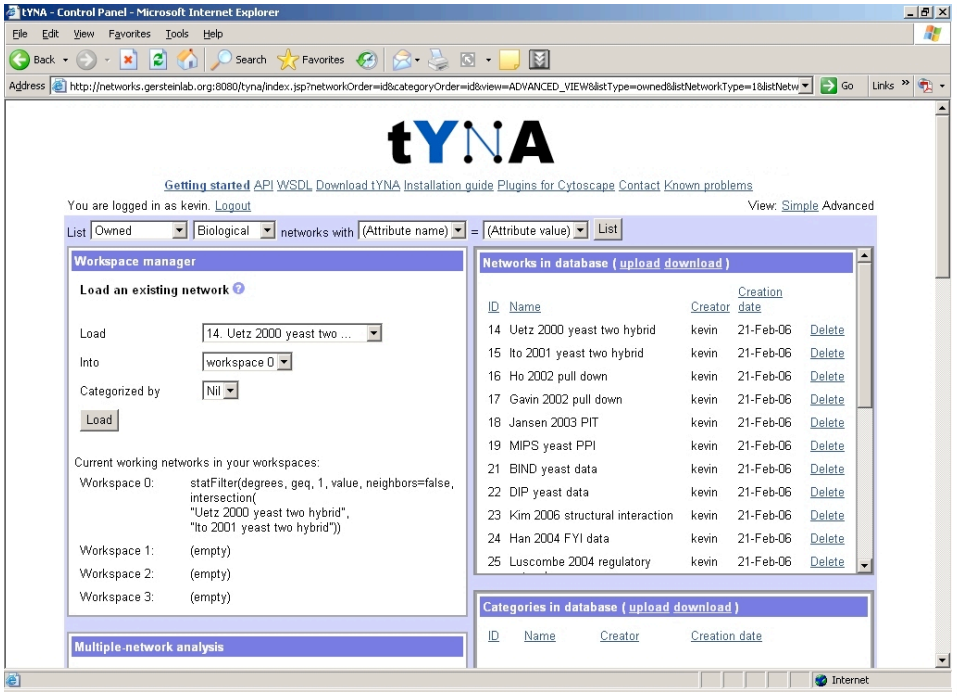
Conclusions: Networks Dynamics across Environments

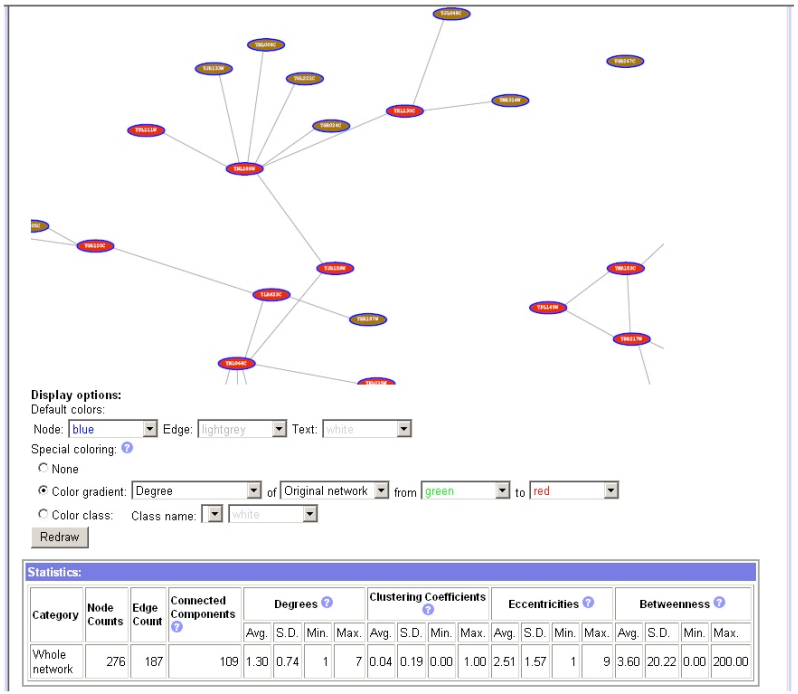
- Developed and adapted techniques to connect quantitative features of environment to metabolism.
- Applied to available aquatic datasets, we identified footprints that were predictive of their environment (potentially could be used as biosensor).
- Strong correlation exists between a community's energy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

- an automated web tool

tYNA

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





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

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

Acknowledgements

Networks.GersteinLab.org

Job opportunities currently

for postdocs & students

K Yip

T Gianoulis

H Yu

M Seringhaus
M Snyder

A Paccanaro

P Kim

P Cayting

P Patel

P Bork
J Raes

More Information on this Talk

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

SUBJECT: Networks

DESCRIPTION:
Joint Statistical Meetings 2009, Washington, DC, 2009.08.02, 14:00-14:20; [I:**JSM**] (Very short networks talk, just with **tse***, **multilevel***, **and** **metagenomics***. Justs fits into 19' w. 1 questions (with some skipping of slides). PPT works on mac & PC and has many photos.)

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

.

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