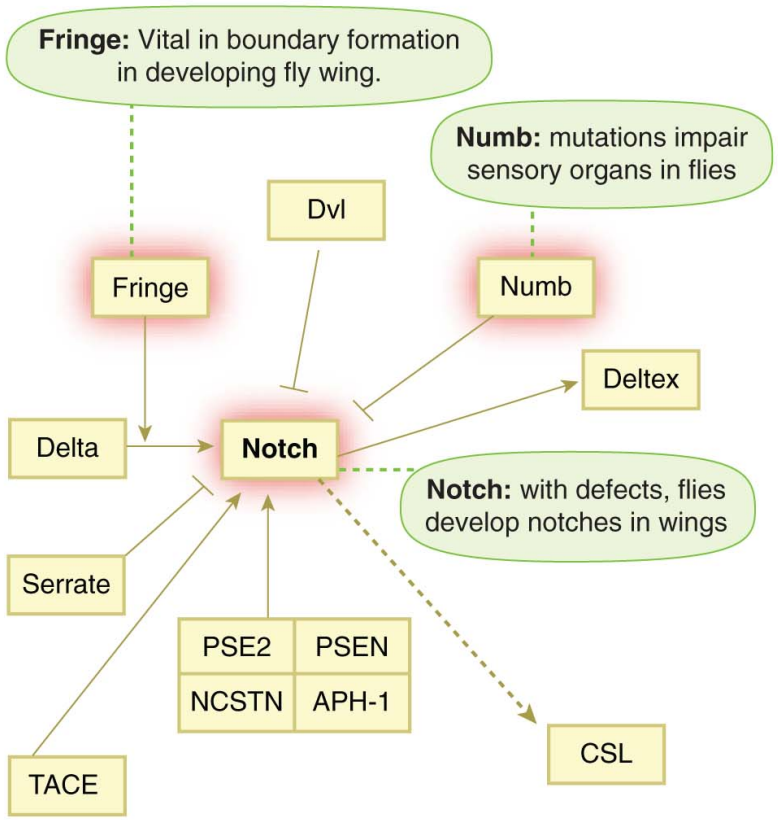
Understanding Protein Function on a Genome-scale through the Analysis of Molecular Networks

Mark B Gerstein
Yale

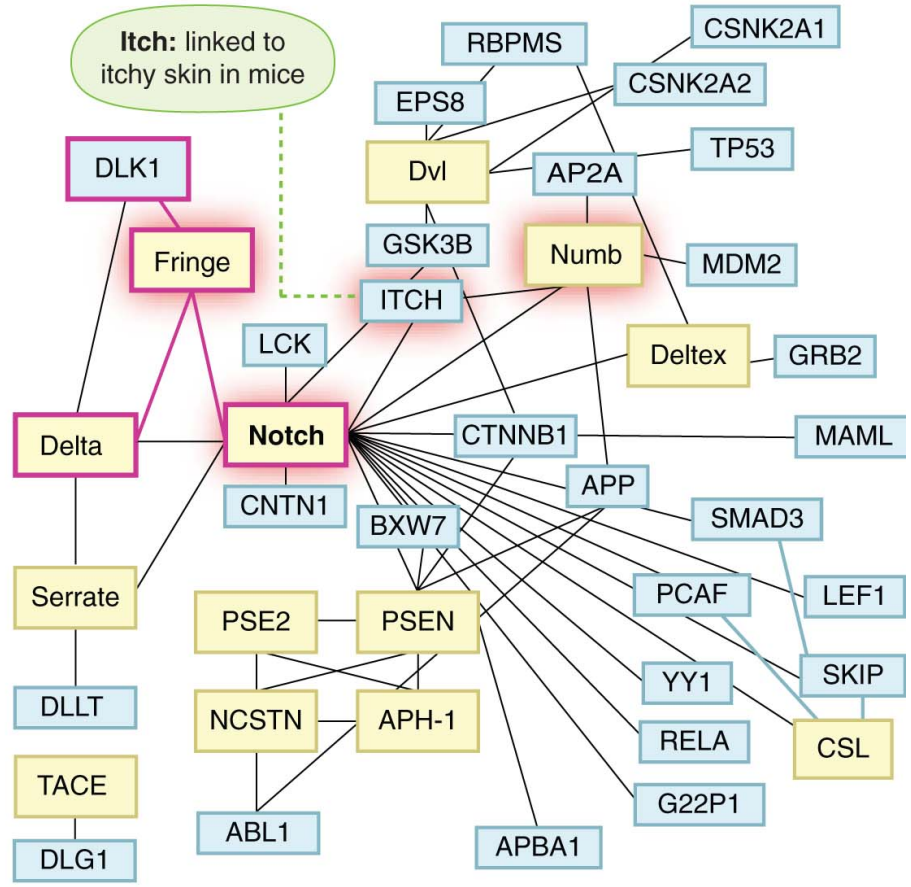
slides at
Lectures.GersteinLab.org

(See Last Slide for References & More Info.)

Networks (Old & New)



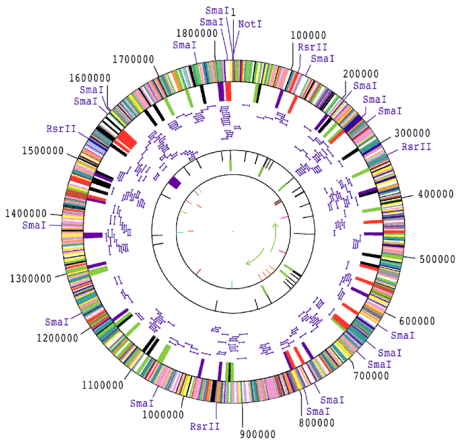
Classical KEGG pathway



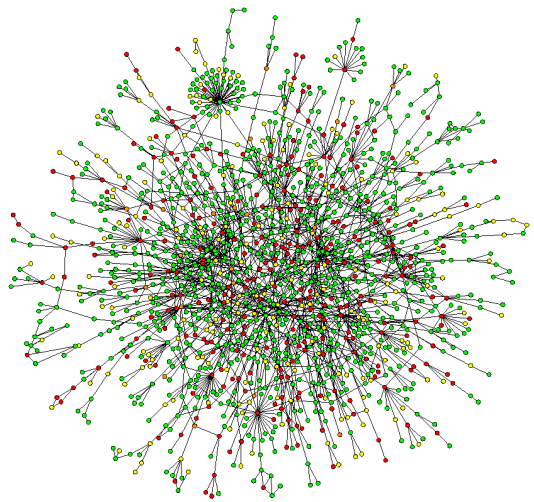
Same Genes in High-throughput Network

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

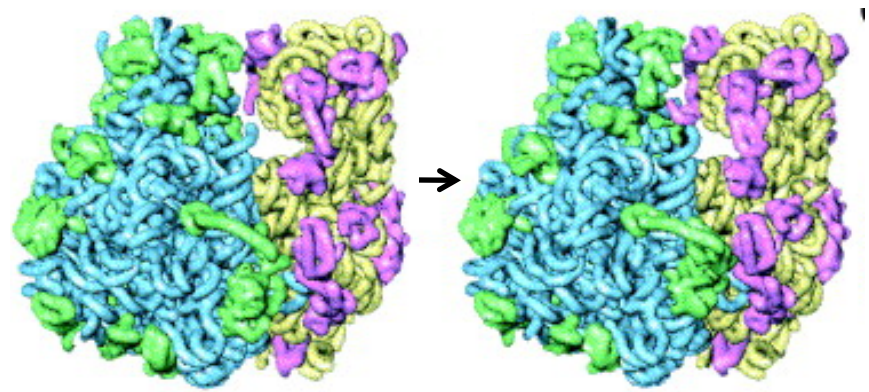
Networks occupy a midway point in terms of level of understanding



1D: Complete Genetic Partslist



~2D: Bio-molecular Network Wiring Diagram



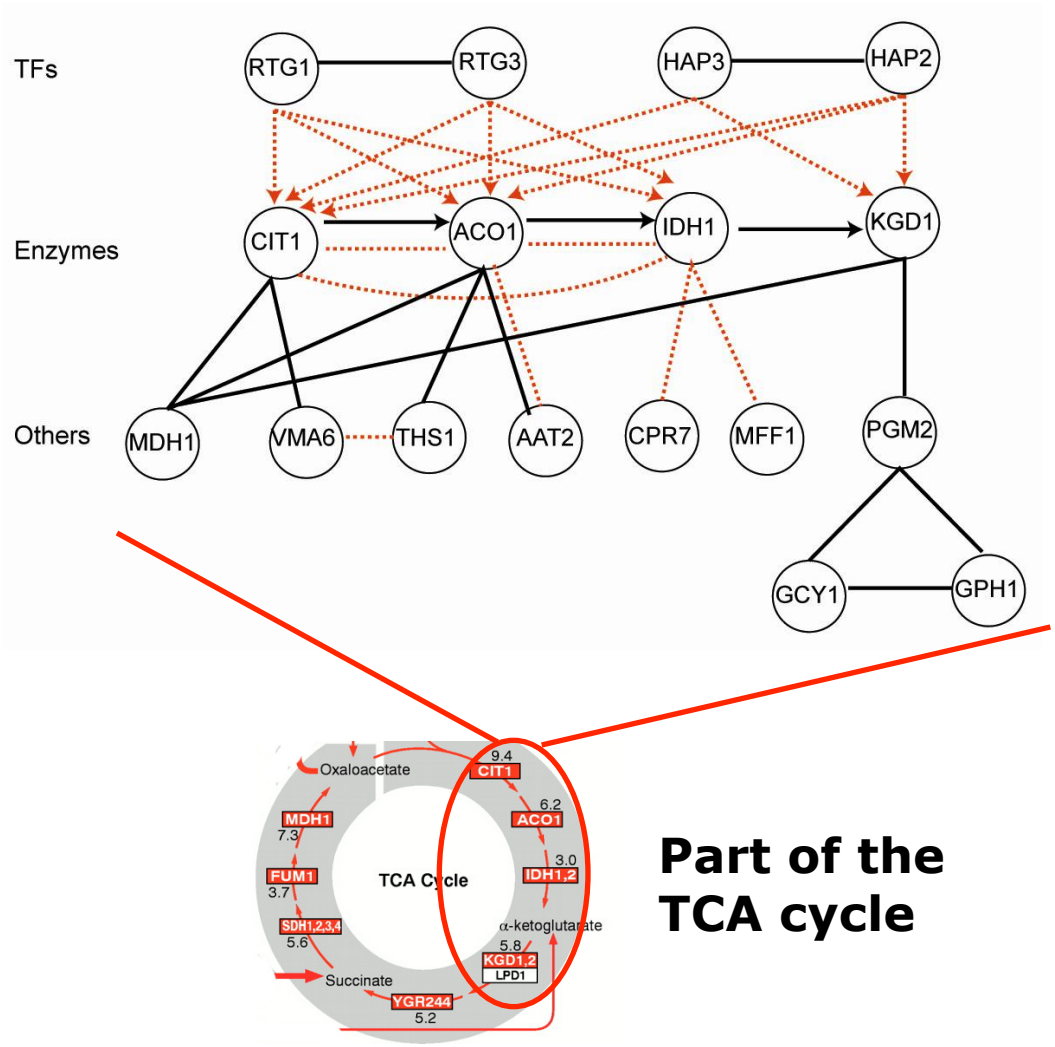
3D and 4D: Detailed structural understanding of cellular machinery (e.g. ribosome in different functional states)

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

[Jeong et al. Nature, 41:411]

[Chiu et al. Trends in Cell Biol, 16:144]

Combining networks forms an ideal way of integrating diverse information



Part of the TCA cycle

Metabolic pathway

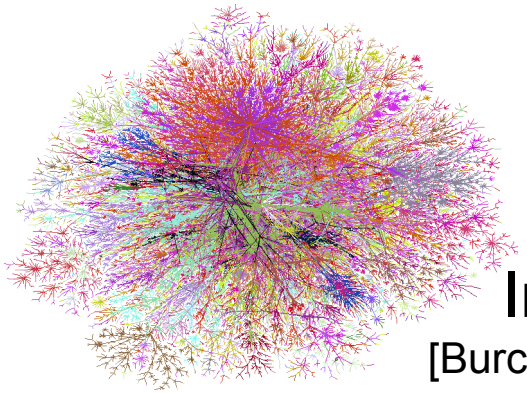
Transcriptional regulatory network

Physical protein-protein Interaction

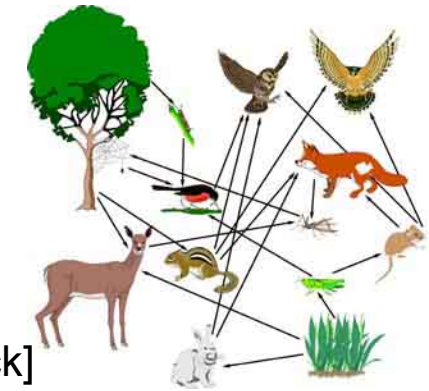
Co-expression Relationship

Genetic interaction (synthetic lethal)
Signaling pathways

Networks as a universal language

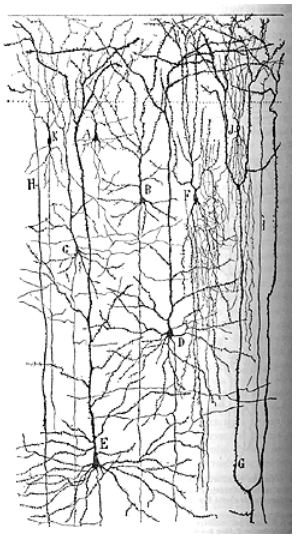


Internet
[Burch & Cheswick]

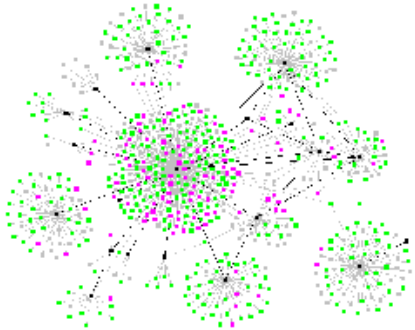


Food Web

Electronic Circuit

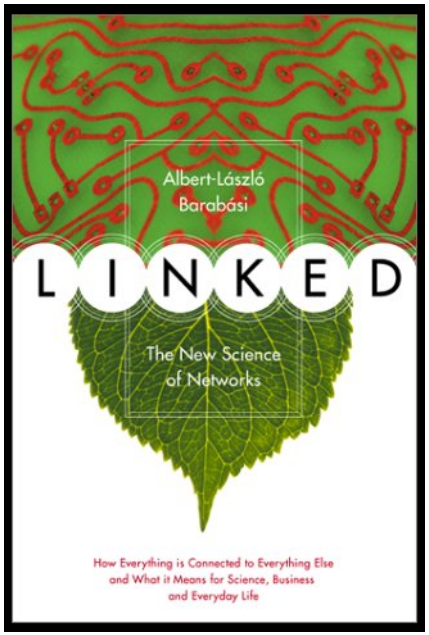


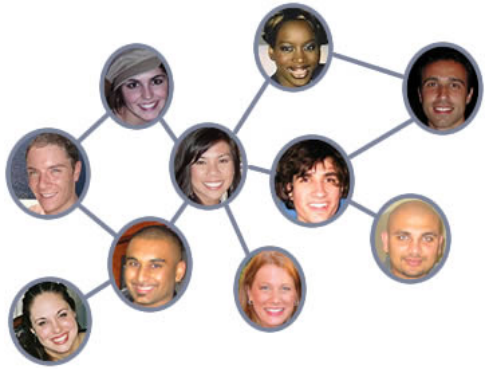
Neural Network
[Cajal]



Disease Spread
[Krebs]

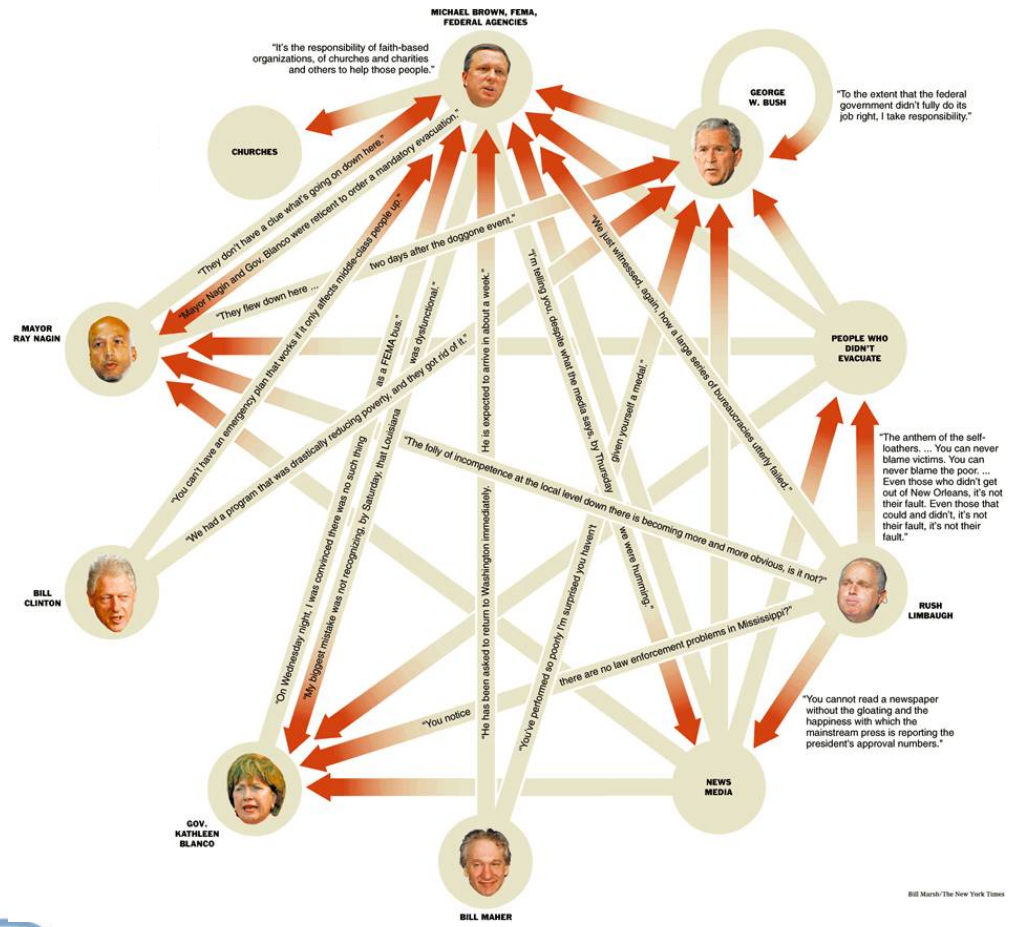
Protein Interactions
[Barabasi]



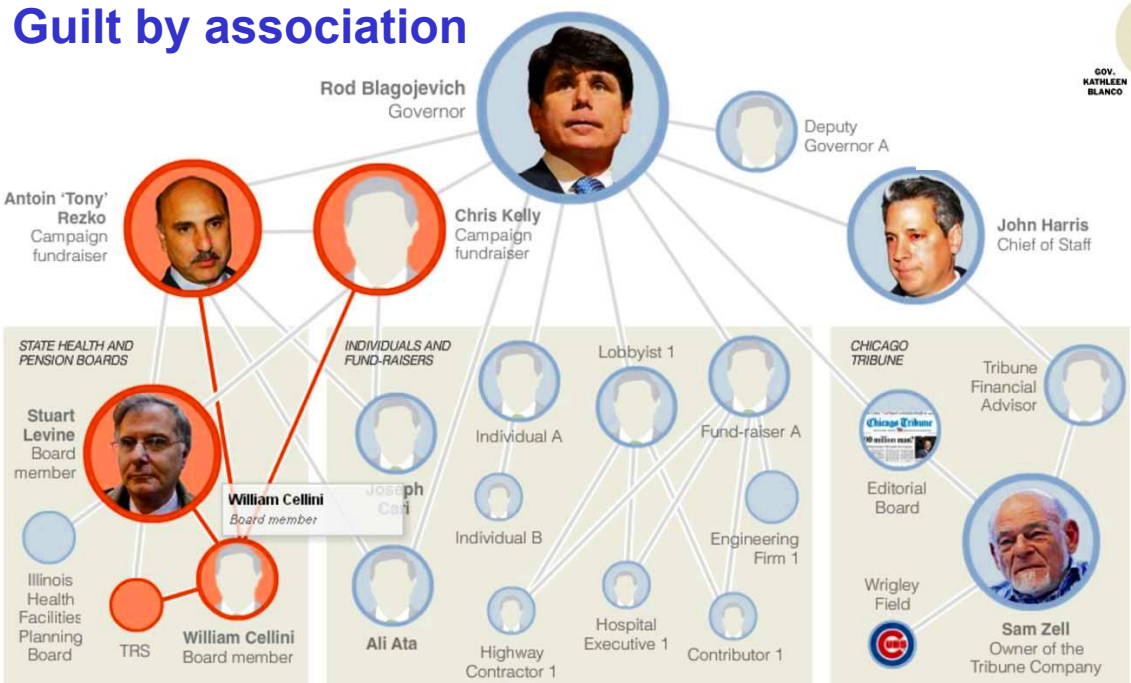


Social Network

Using the position in networks to describe function



Guilt by association



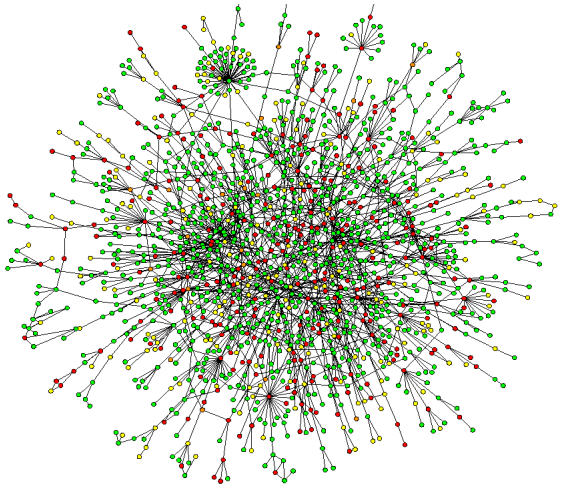
Finding the causal regulator (the "Blame Game")

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

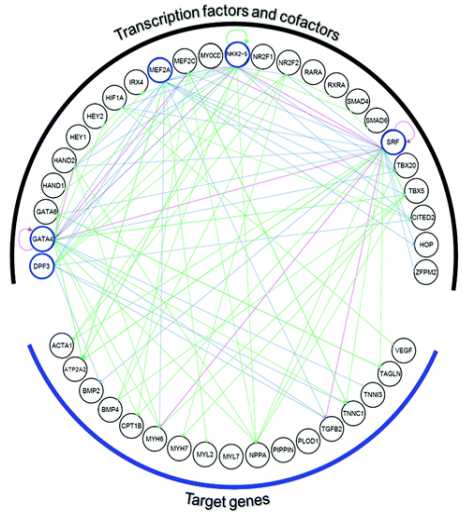
Outline: Molecular Networks

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

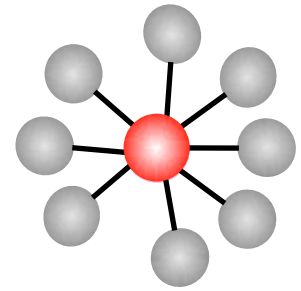
Different Types of Molecular Networks



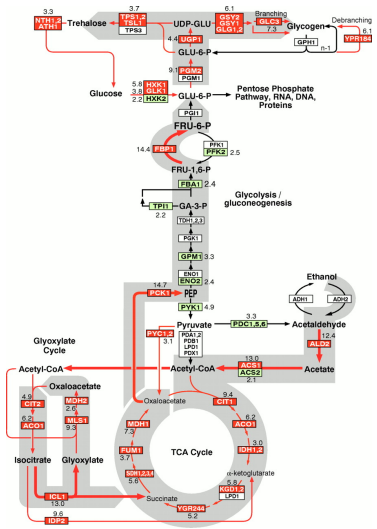
Protein-protein Interaction networks



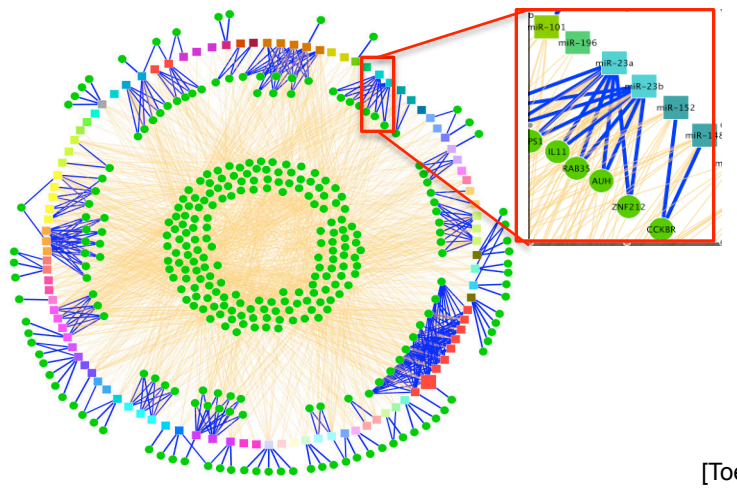
TF-target-gene Regulatory networks



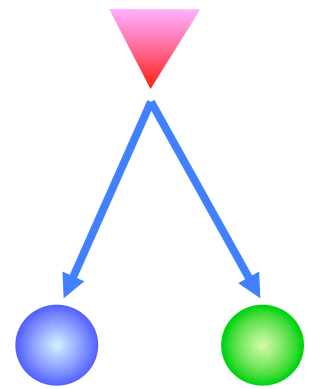
Undirected



Metabolic pathway networks



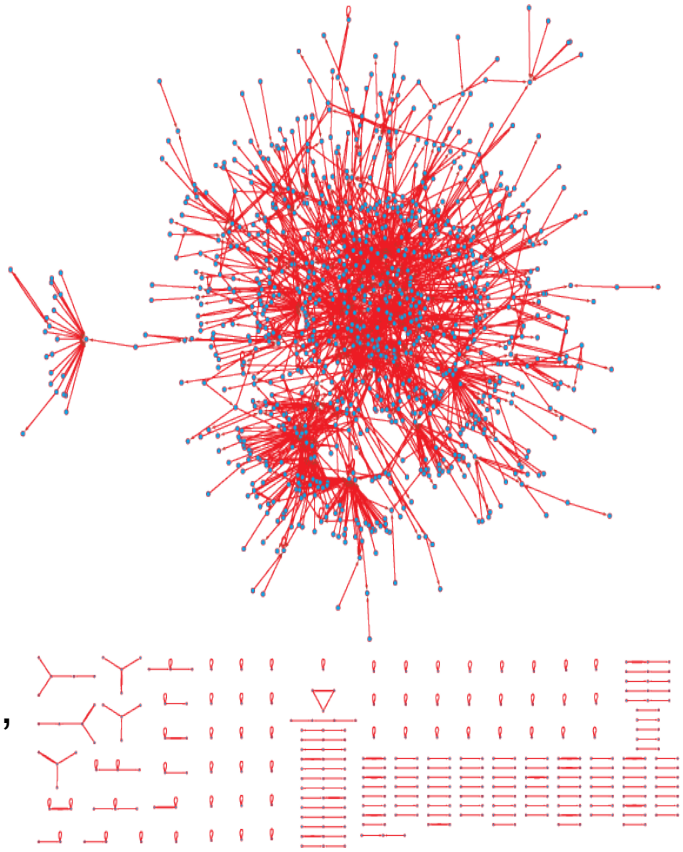
miRNA-target networks



Directed

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

Example: yeast PPI network



Actual size:

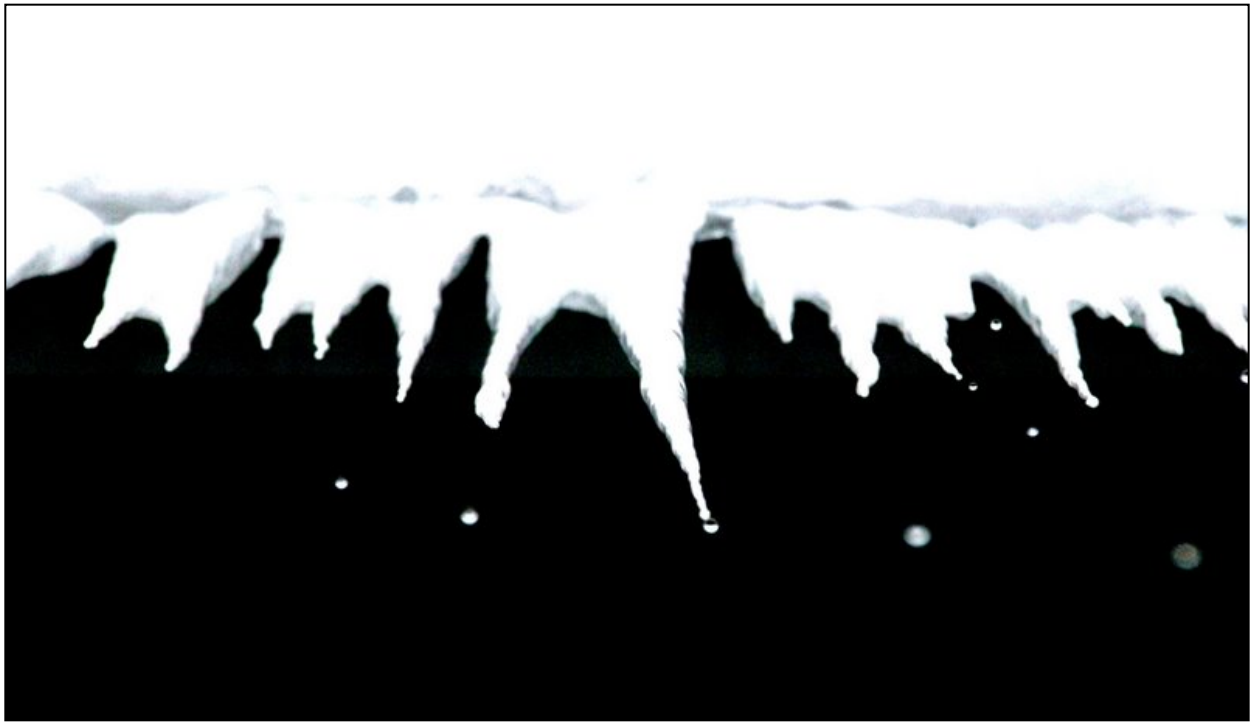
- ~6,000 nodes
  → Computational cost: ~18M pairs
- Estimated ~15,000 edges
  → Sparseness: 0.08% of all pairs (Yu et al., 2008)

Known interactions:

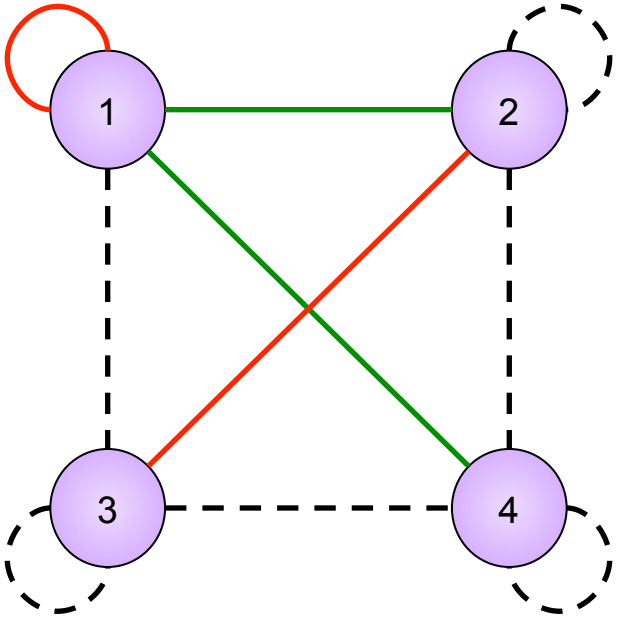
- Small-scale experiments: accurate but few
  → Overfitting: ~5,000 in BioGRID, involving ~2,300 proteins
- Large-scale experiments: abundant but noisy
  → Noise: false +ve/-ve for yeast two-hybrid data up to
  45% and 90% (Huang et al., 2007)

Predicting Networks

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



Training sets



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

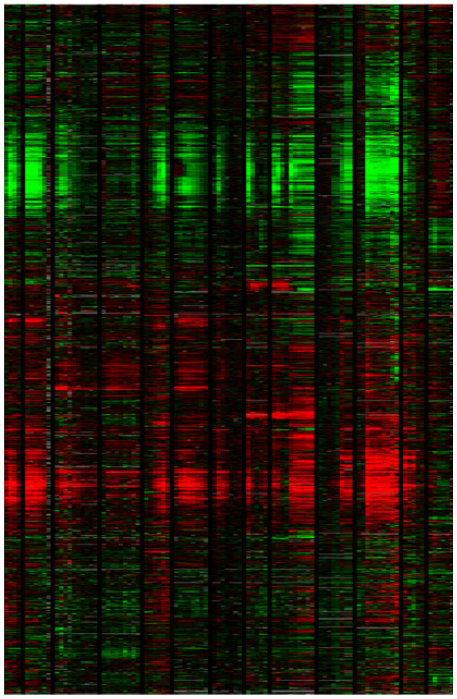
Known interactions

Known non-interactions

Unknown

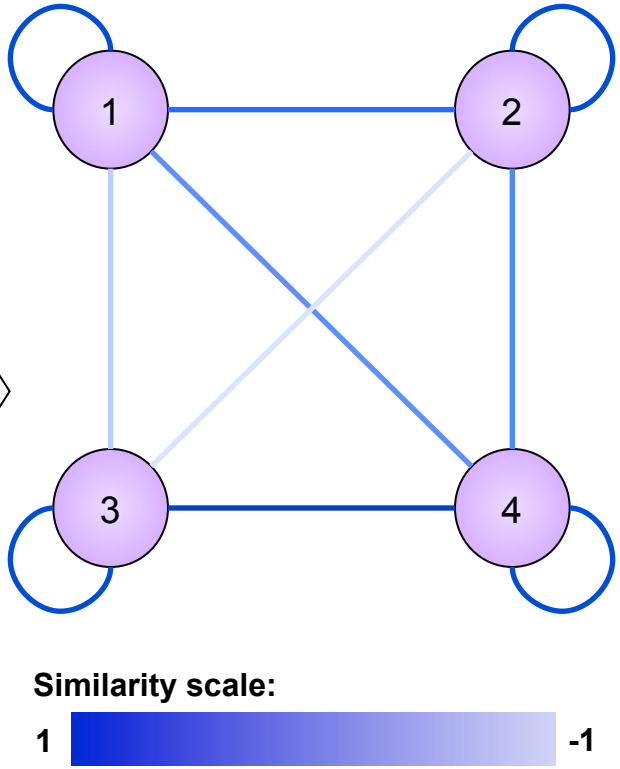
Network prediction: features

- Example 1: gene expression



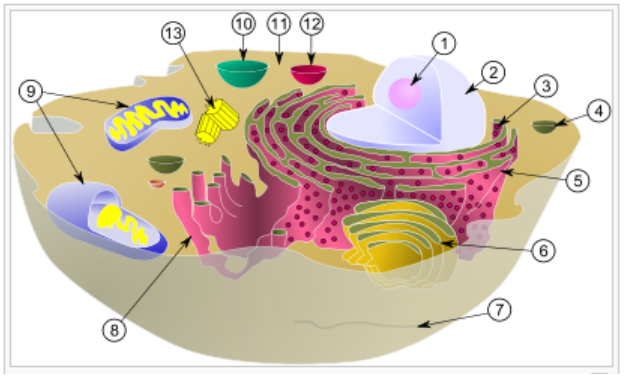
Gasch et al., 2000

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



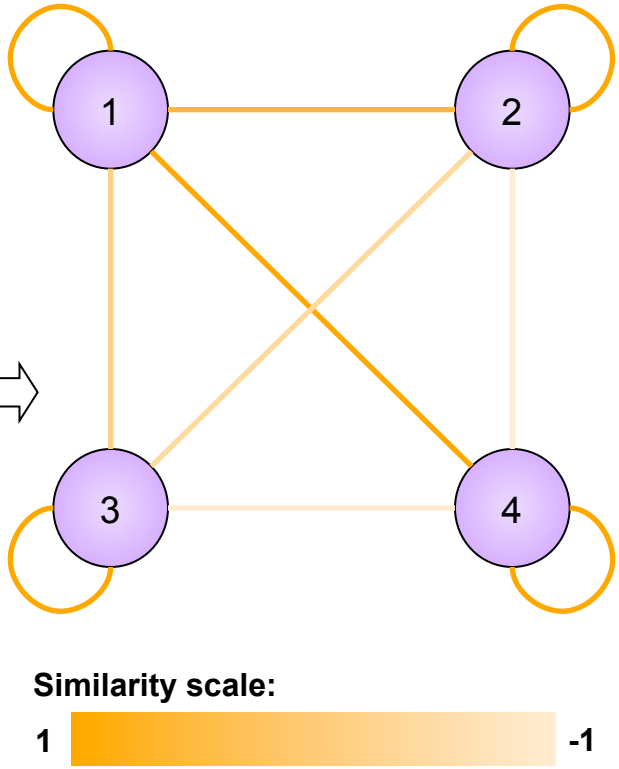
Network prediction: features

- Example 2: sub-cellular localization

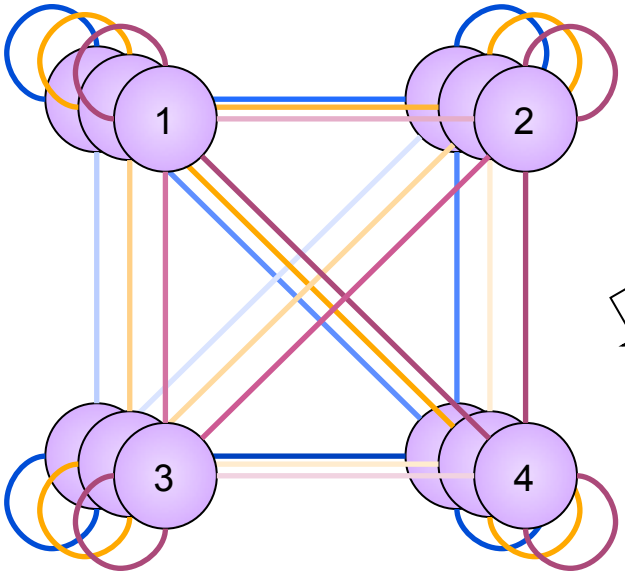


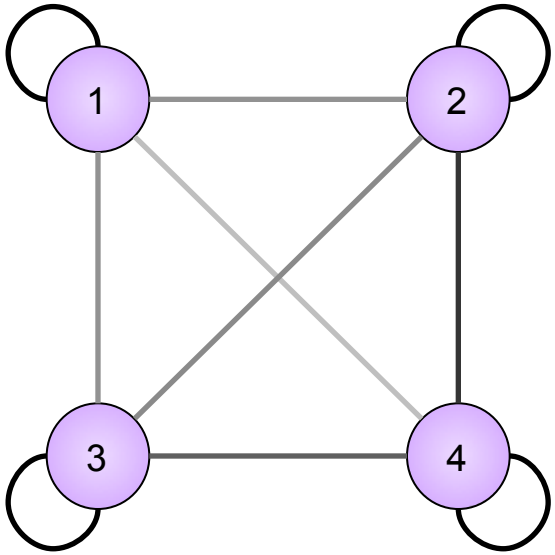
http://www.scq.ubc.ca/wp-content/yeasttwohybridtranscript.gif

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



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)

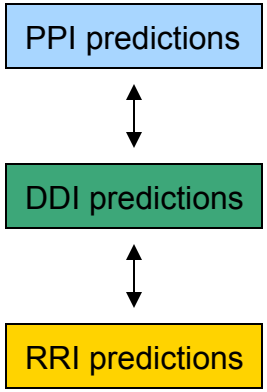
Our work: efficiently propagating known information

Training set expansion

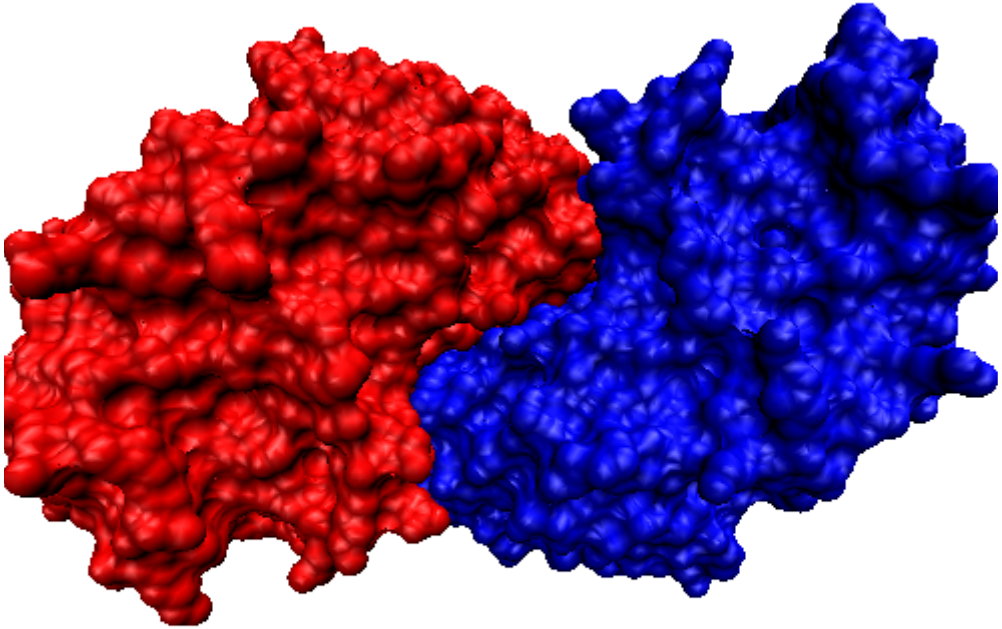
- Motivation: lack of training examples
- Expand training sets horizontally

Multi-level learning

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



Protein interaction

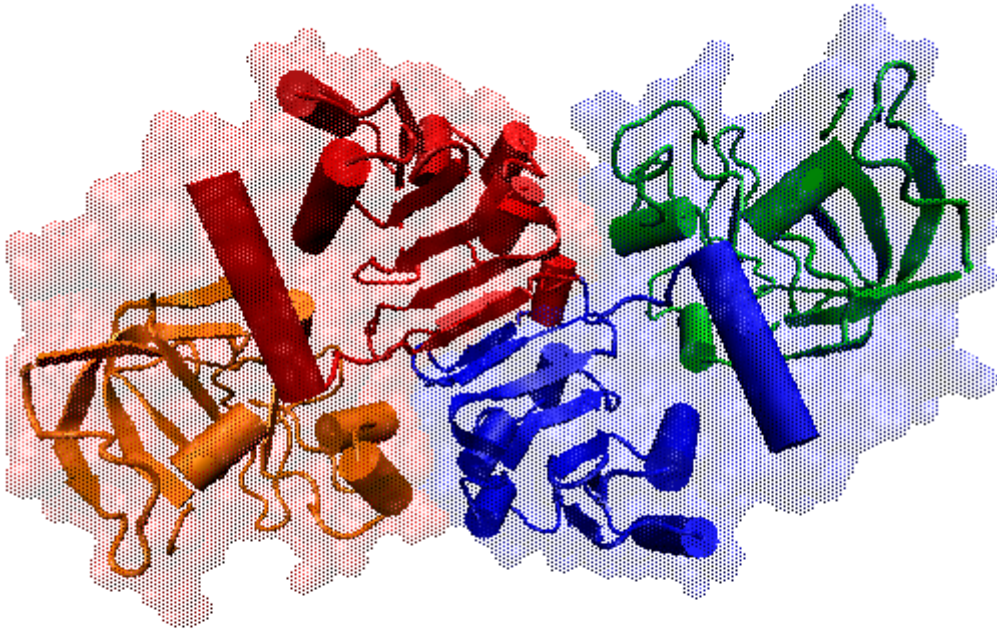


Yeast NADP-dependent alcohol dehydrogenase 6 (PDB: 1piw)

Protein-level features for interaction prediction: functional genomic information

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

Domain interaction

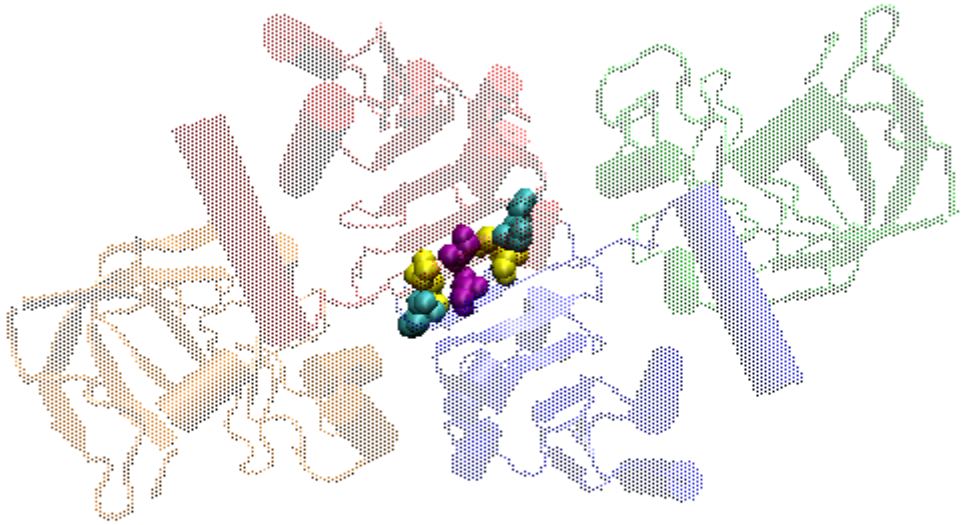


Pfam domains: PF00107 (inner) and PF08240 (outer)

Domain-level features for interaction prediction: evolutionary information

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

Residue interaction

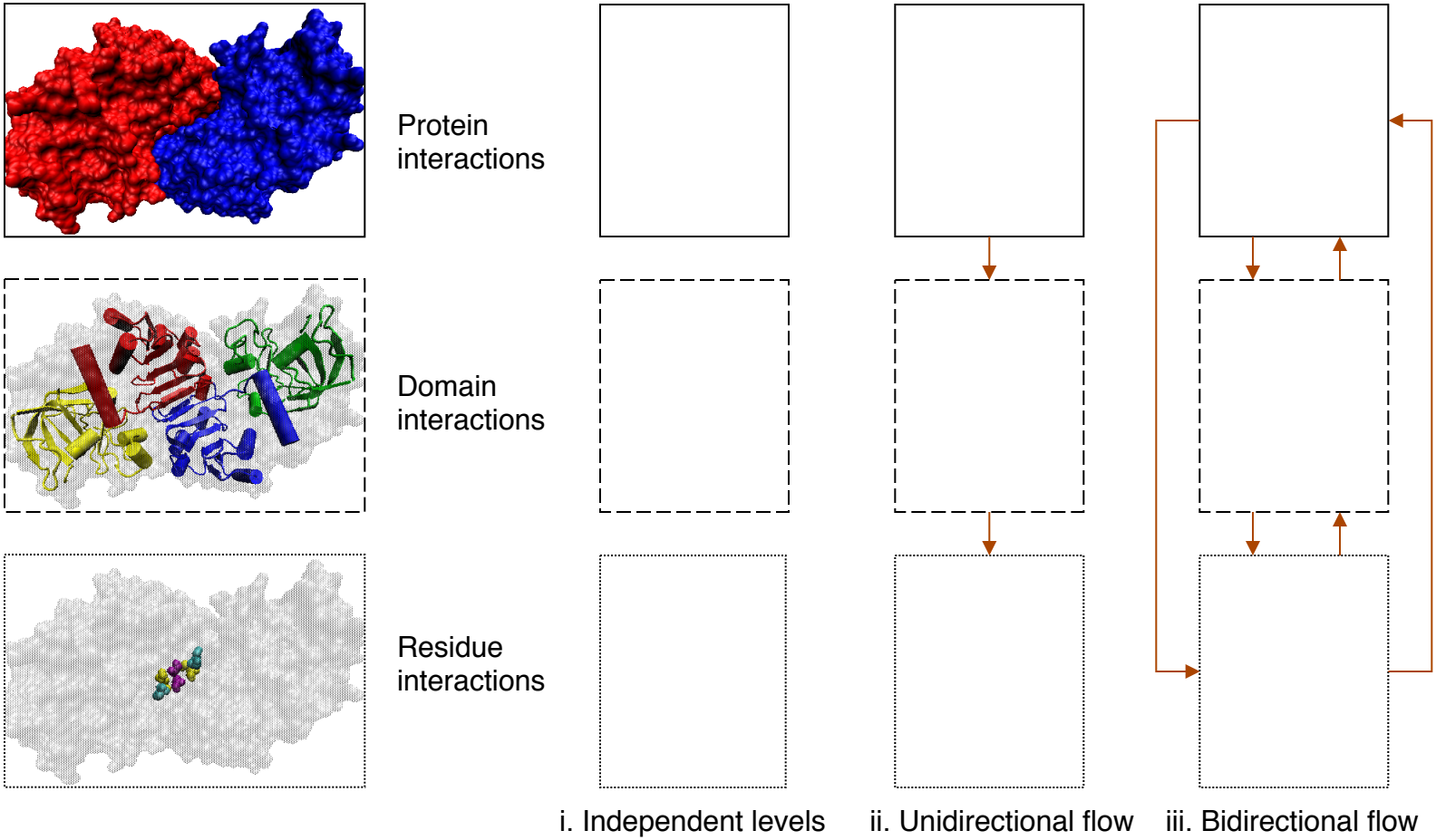


Interacting residues: 283 (yellow) with 287 (cyan), and 285 (purple) with 285

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

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

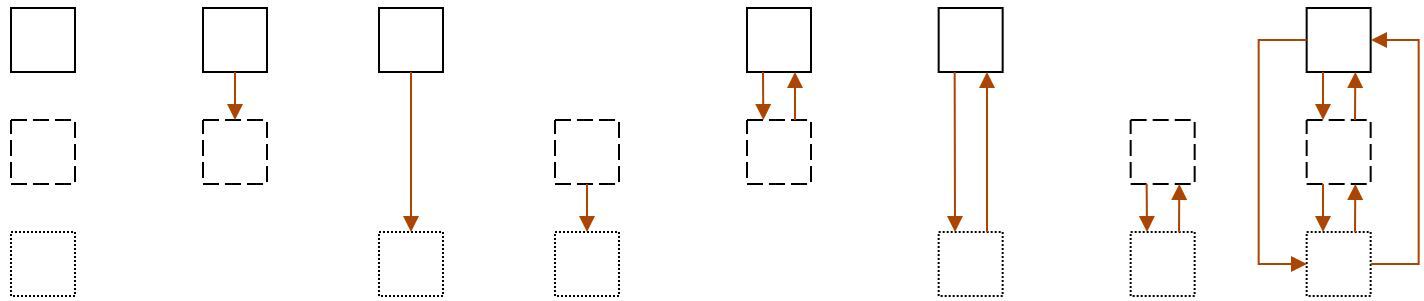
Combining the three problems



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

Empirical results (AUCs)

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



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

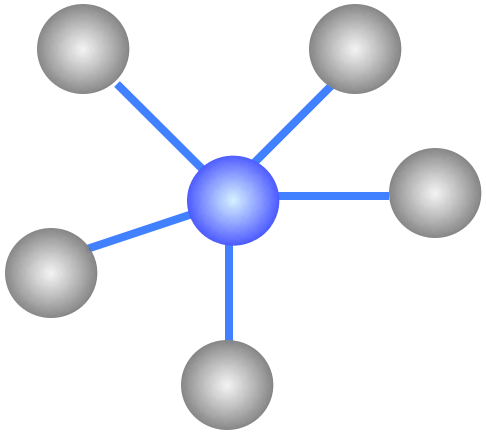
[Yip and Gerstein, BMC Bioinfo. ('09, press)]

Finding Central Points in Networks: Hubs & Bottlenecks

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

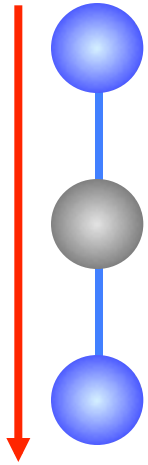
Global topological measures

Indicate the gross topological structure of the network



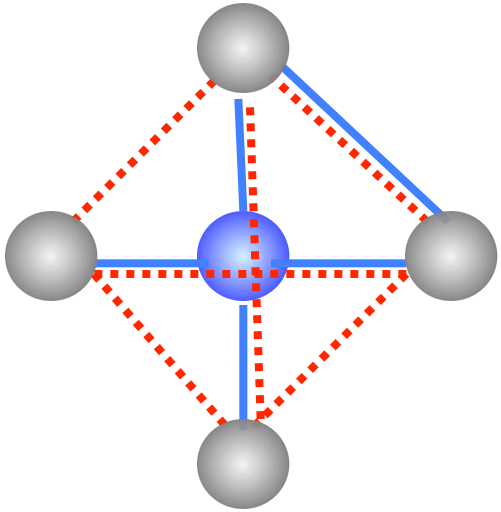
Degree (K)

5



Path length (L)

2



Clustering coefficient (C)

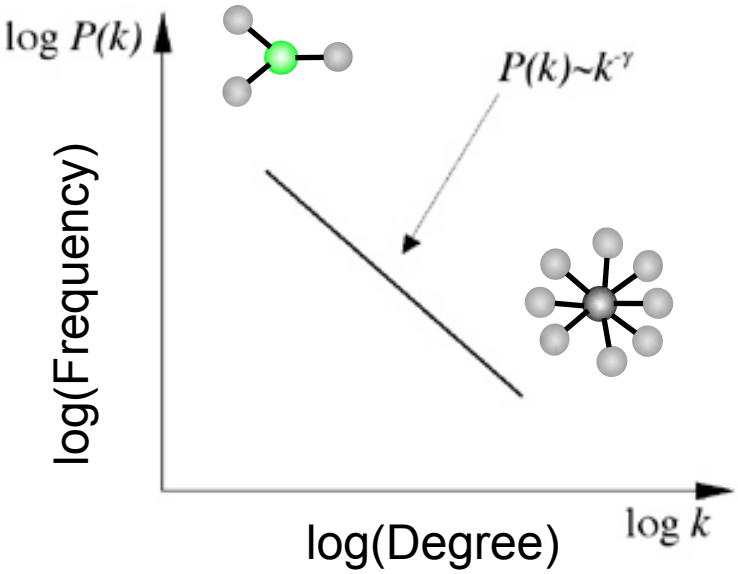
1/6

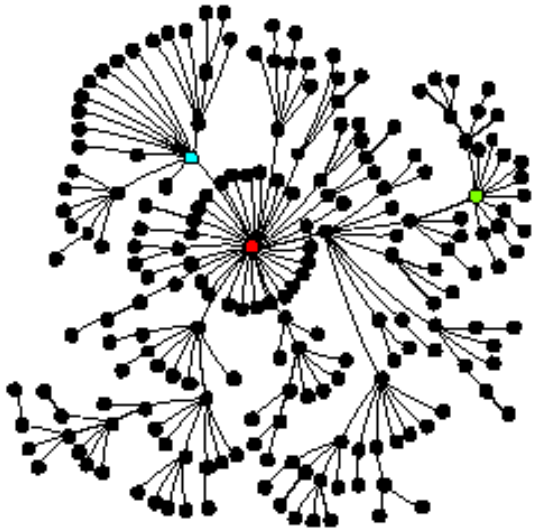
Interaction and expression networks are ***undirected***

[Barabasi]

Scale-free networks

Power-law distribution





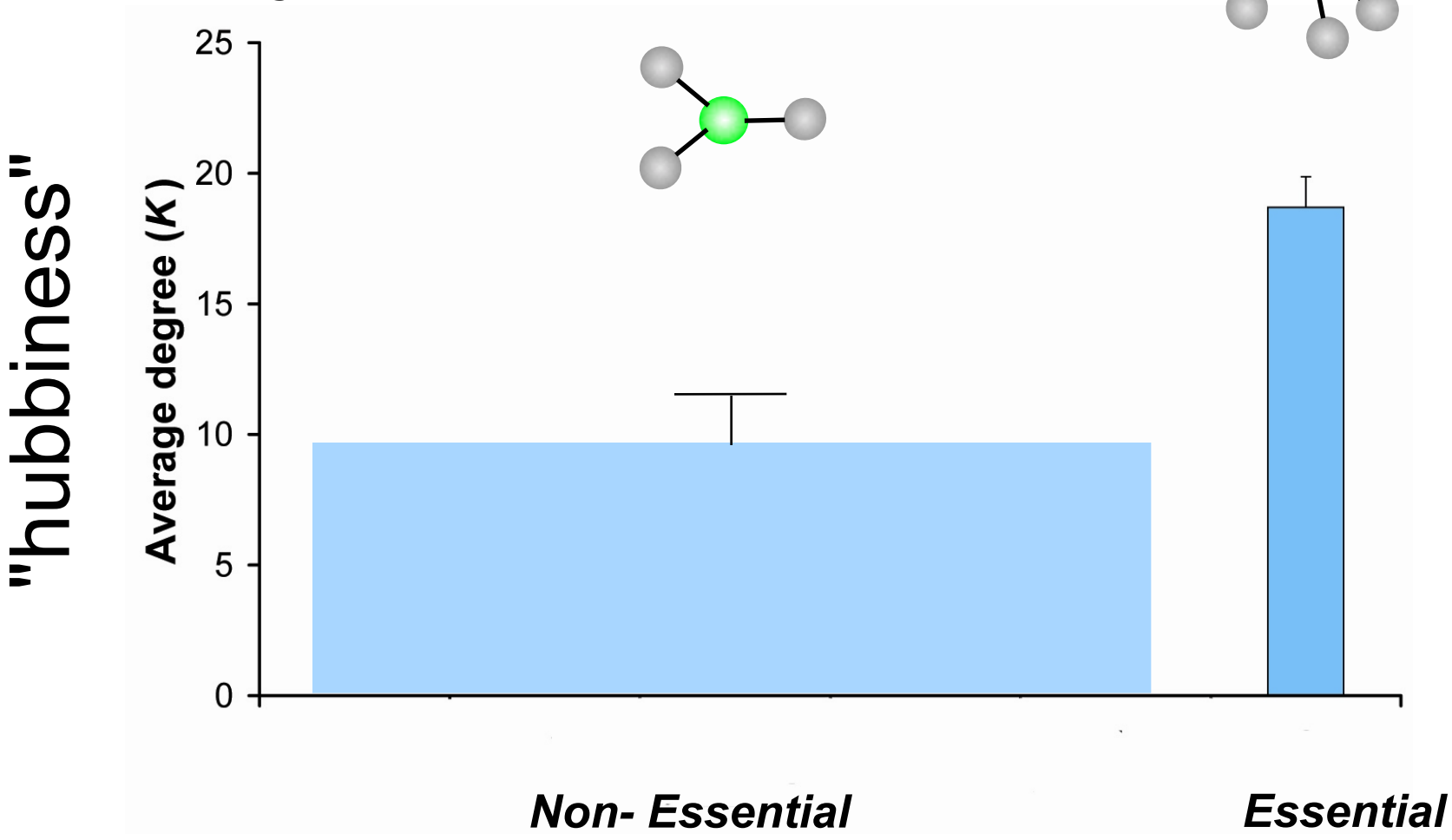
Hubs dictate the structure of the network

[Barabasi]

Hubs tend to be Essential

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

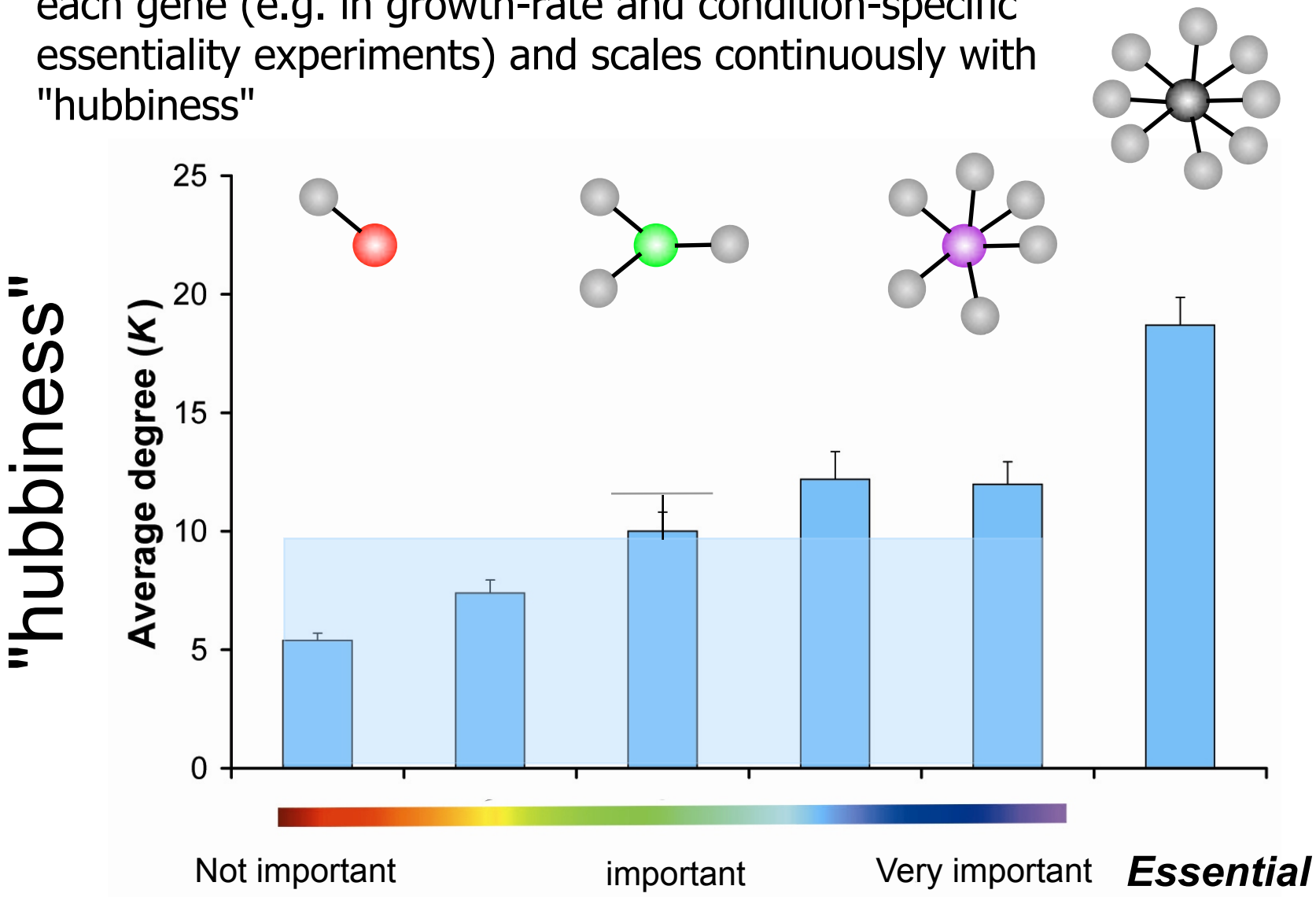
[Lauffenburger, Barabasi]



[Yu et al., 2003, TIG]

Relationships extends to "Marginal Essentiality"

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



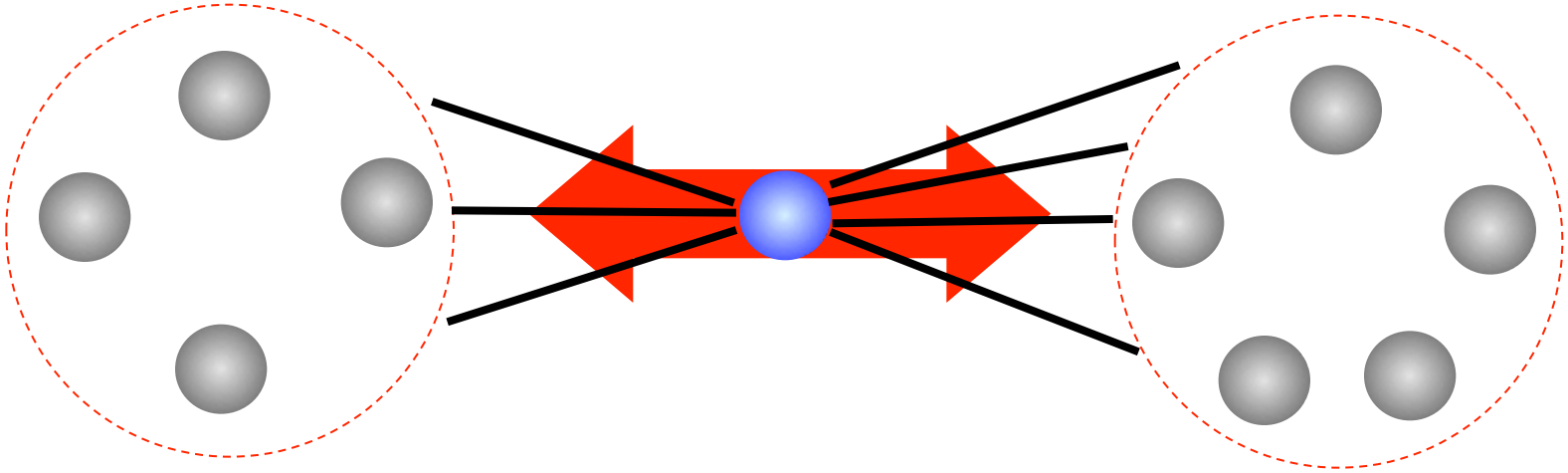
[Yu et al., 2003, TIG]

Another measure of Centrality: Betweenness centrality

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

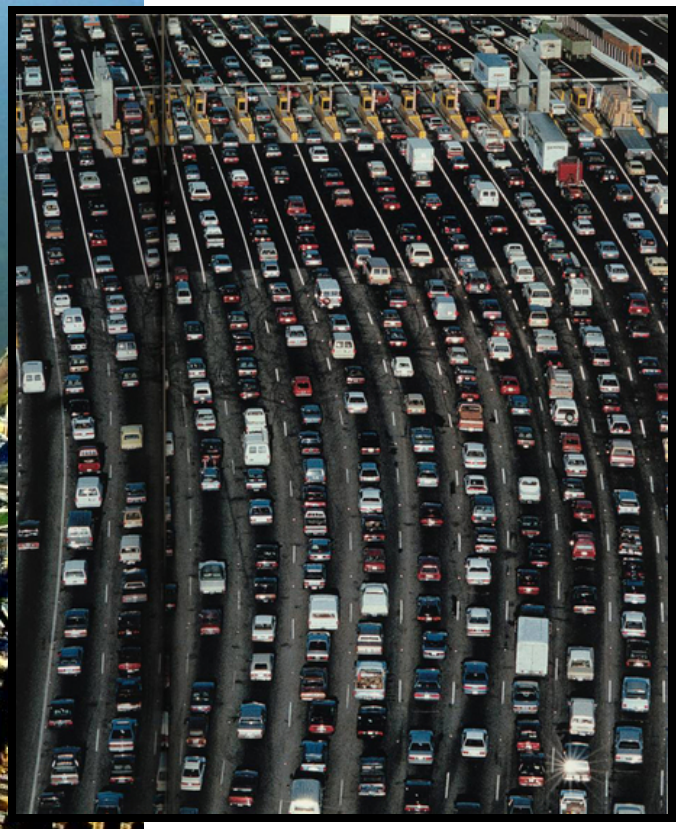
Freeman LC (1977) Set of measures of centrality based on betweenness. Sociometry 40: 35–41.

Girvan & Newman (2002) PNAS 99: 7821.

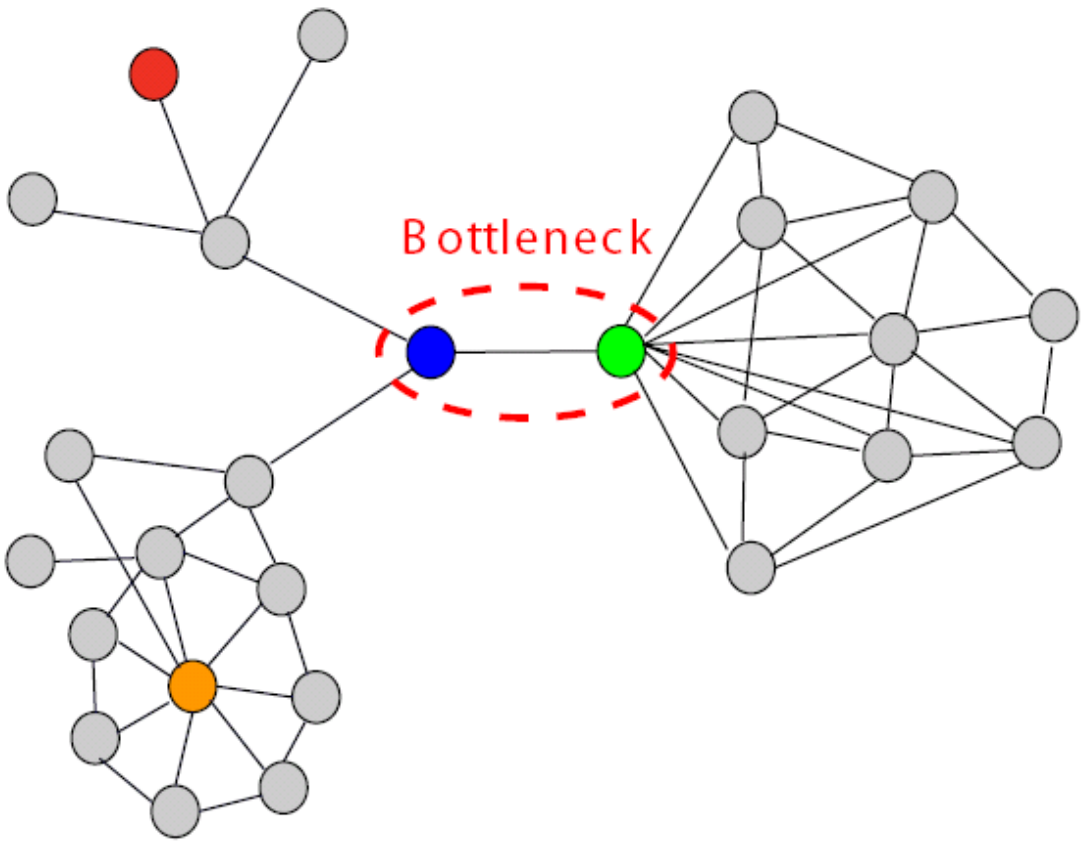


Betweenness centrality -- Bottlenecks

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



Bottlenecks & Hubs



Hub-bottleneck **node**

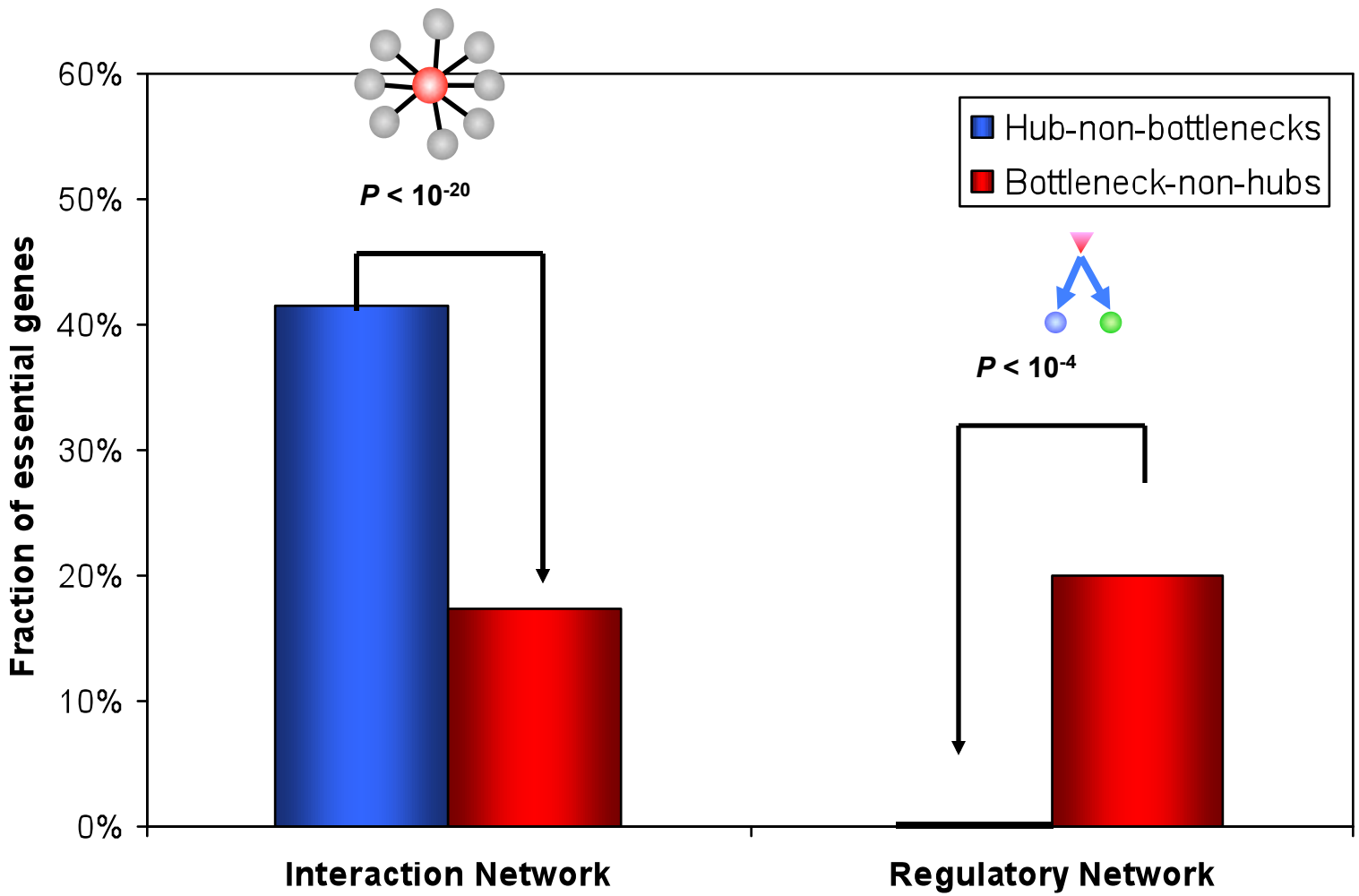
Non-hub-bottleneck **node**

Hub-non-bottleneck **node**

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

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

Bottlenecks are what matters in regulatory networks



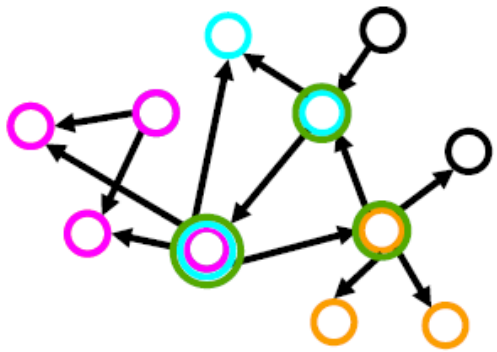
[Yu *et al.*, *PLoS Comput Biol* (2007)]

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

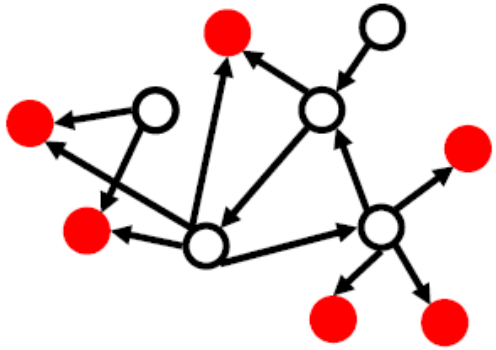
Where are key points networks ? How do we locate them ?

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

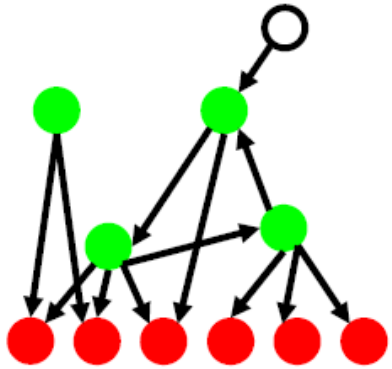
I. Example network with all 4 motifs



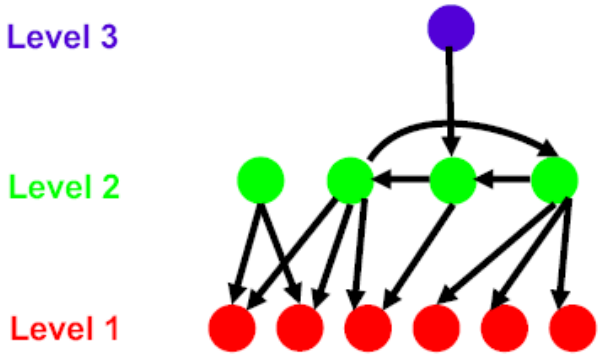
II. Finding terminal nodes (Red)



III. Finding mid-level nodes (Green)

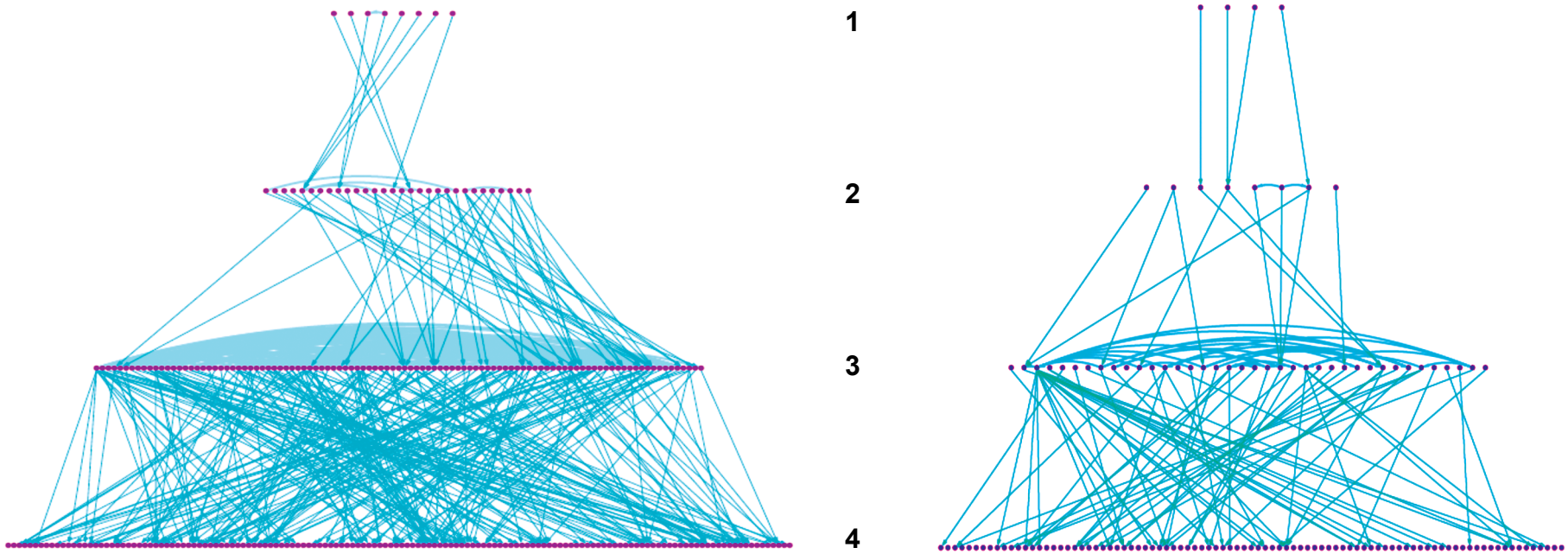


IV. Finding top-most nodes (Blue)



[Yu et al., PNAS (2006)]

Regulatory Networks have similar hierarchical structures



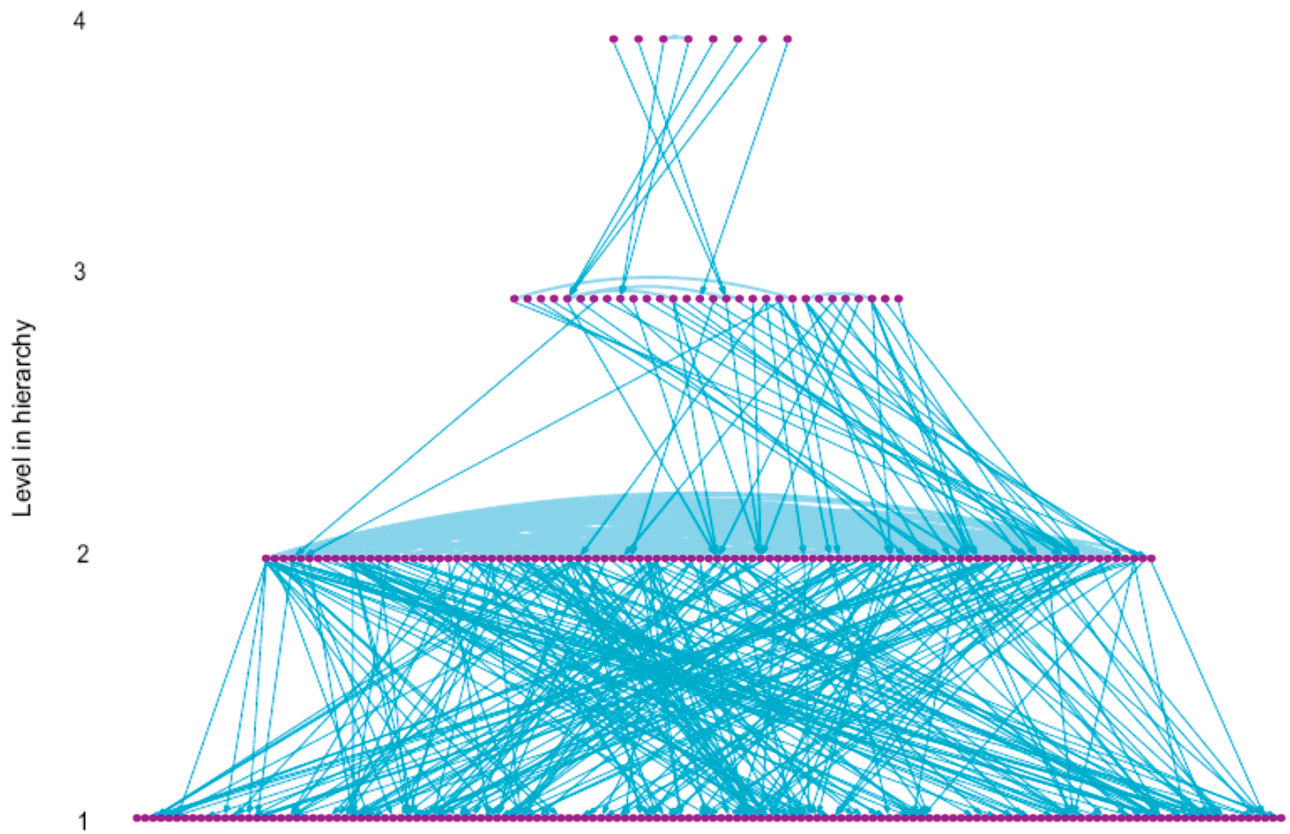
S. cerevisiae

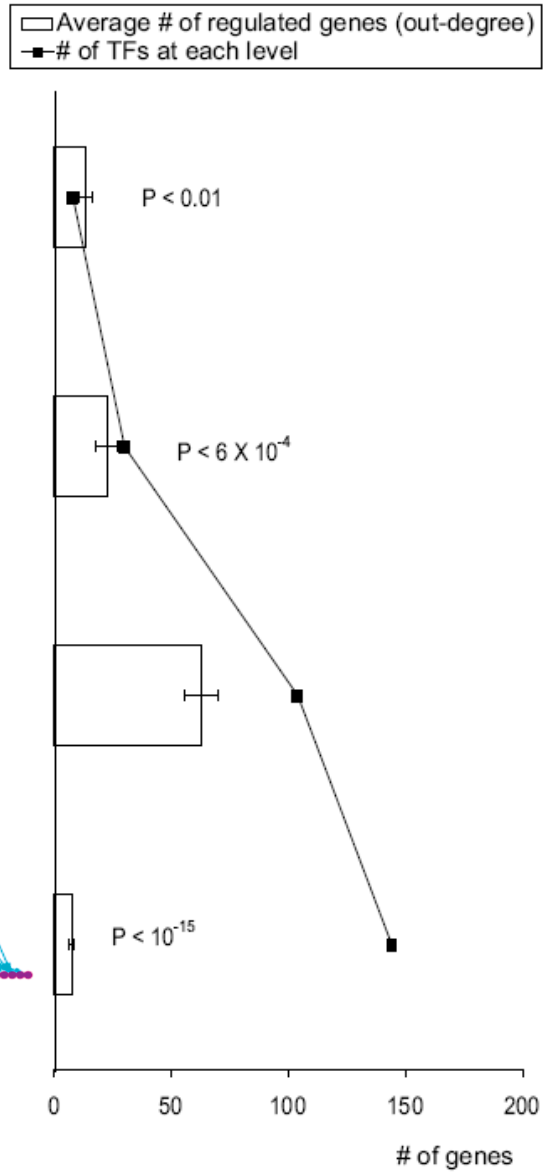
E. coli

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

Yeast Regulatory Hierarchy: the Middle-managers Rule

A. Regulatory hierarchy in *S. cerevisiae*

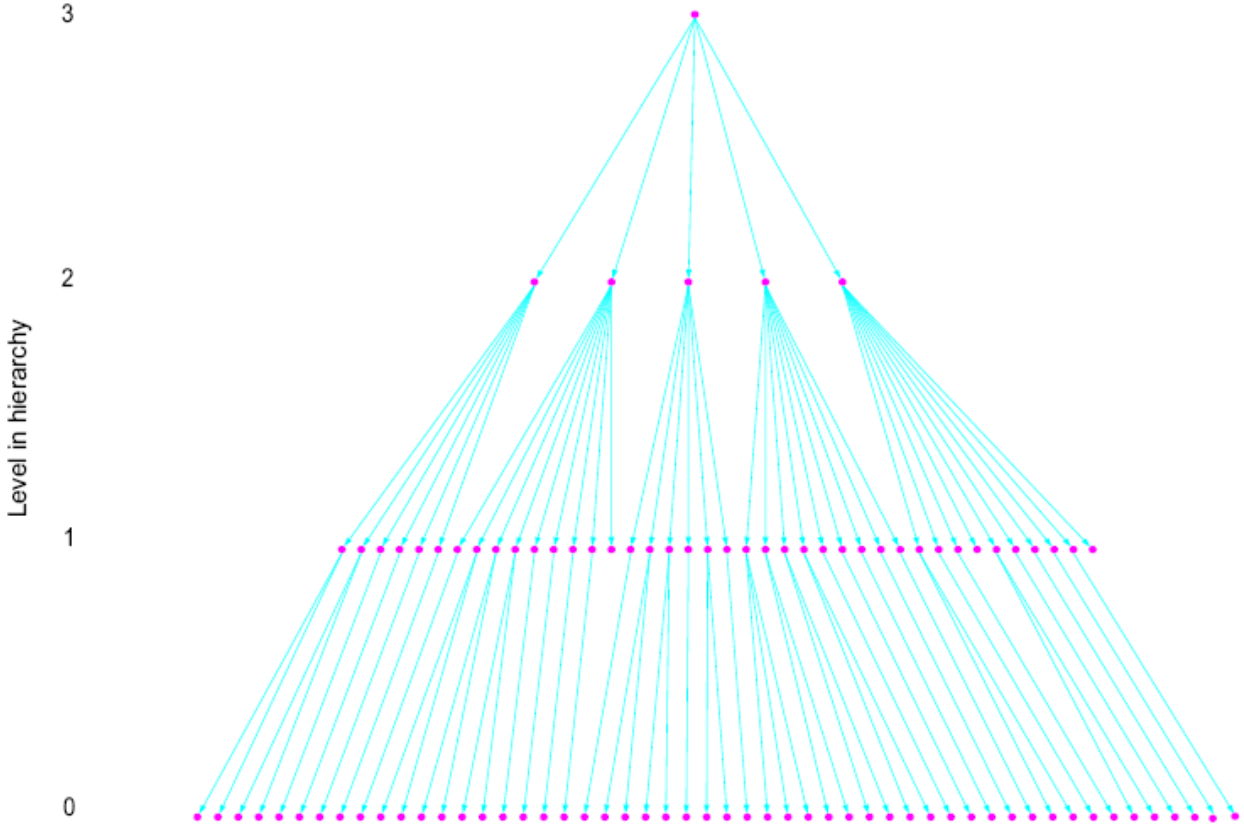


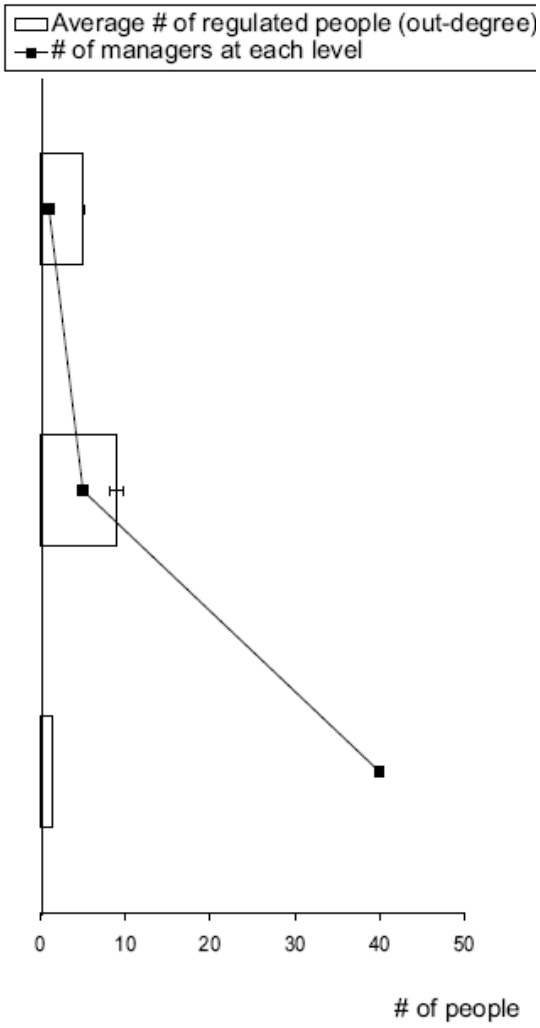


[Yu et al., PNAS (2006)]

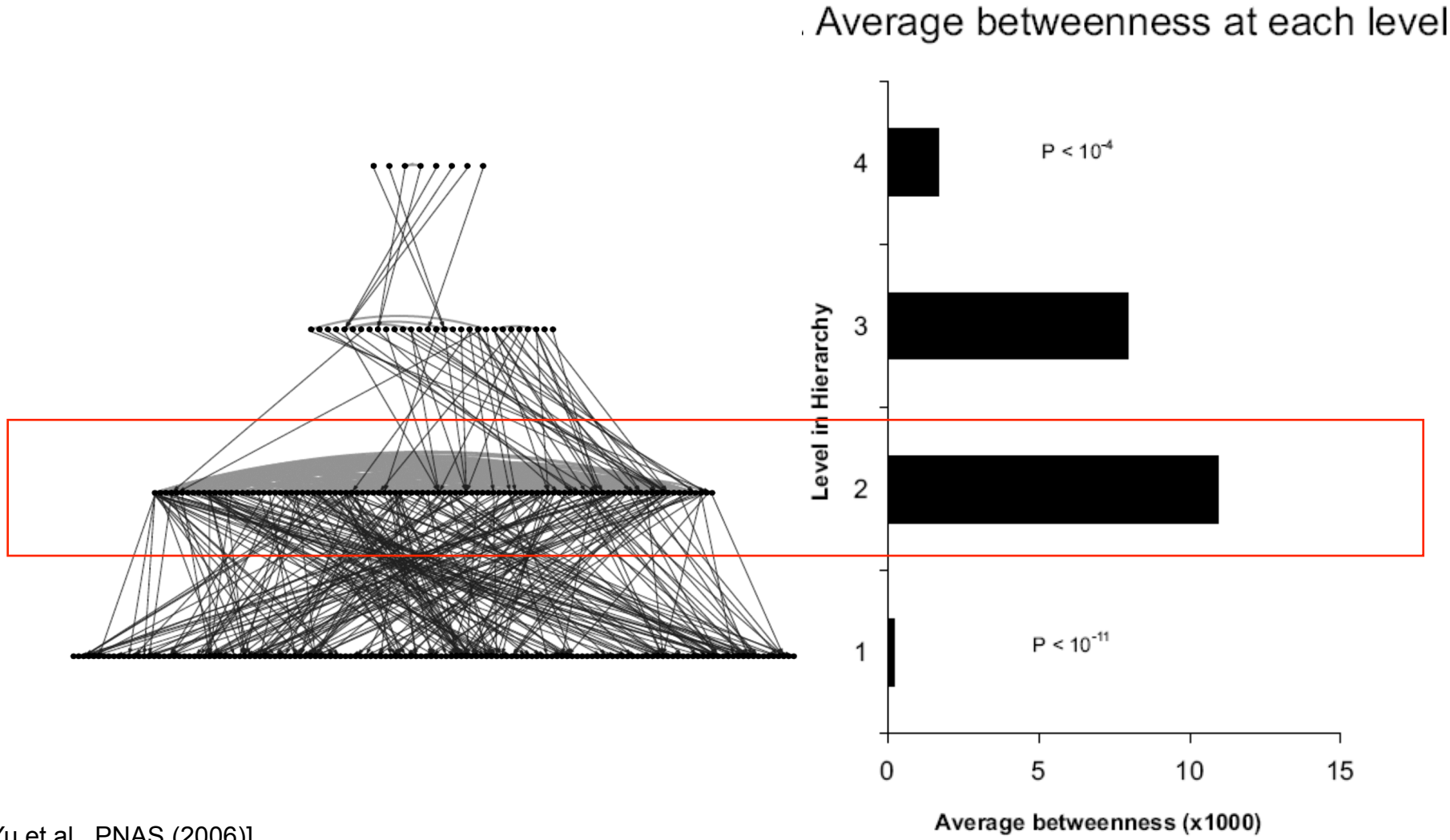
Yeast Network Similar in Structure to Government Hierarchy with Respect to Middle-managers

B. Governmental hierarchy of a representive city (Macao)



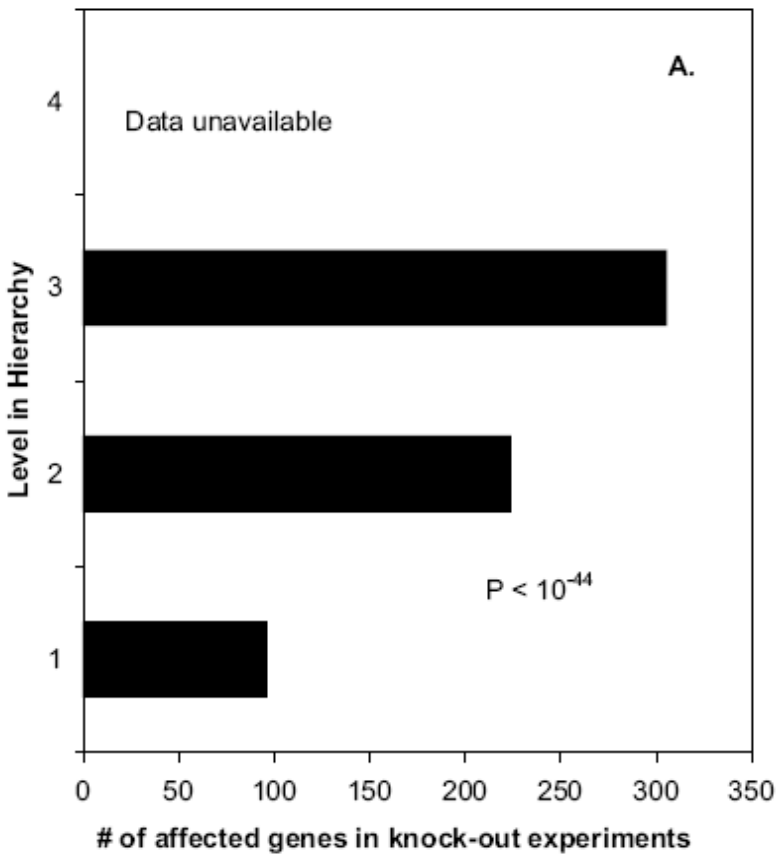


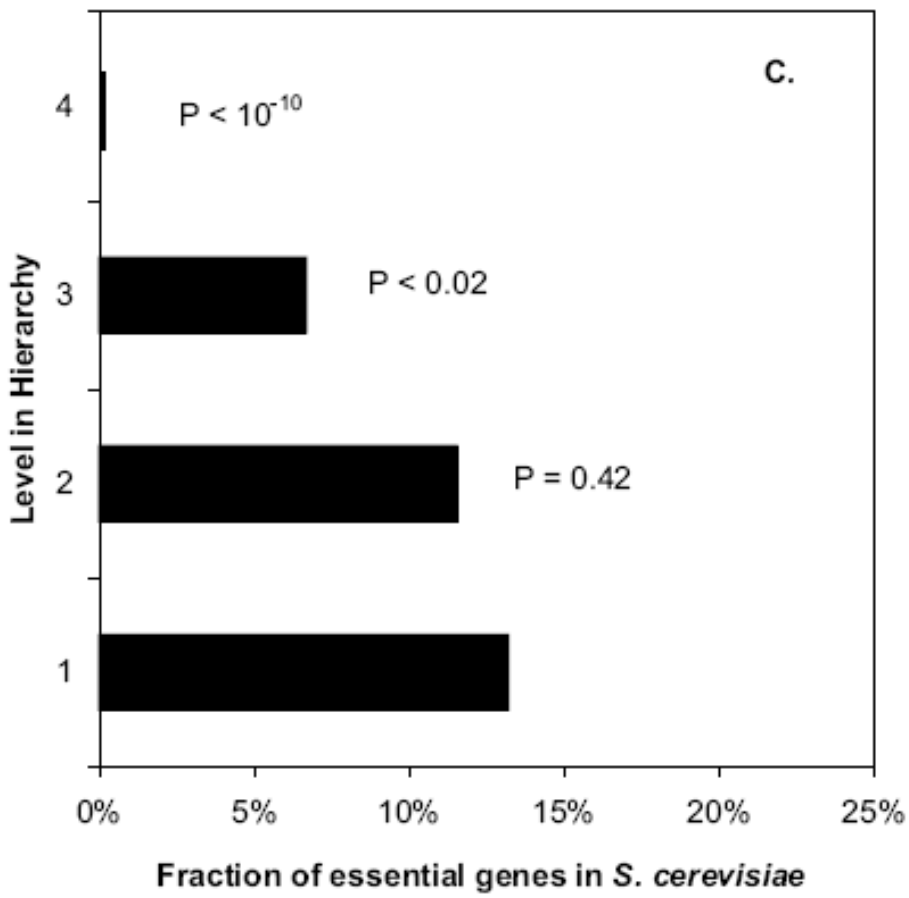
Characteristics of Regulatory Hierarchy: Middle Managers are Information Flow Bottlenecks



[Yu et al., PNAS (2006)]

Characteristics of Regulatory Hierarchy: The Paradox of Influence and Essentiality



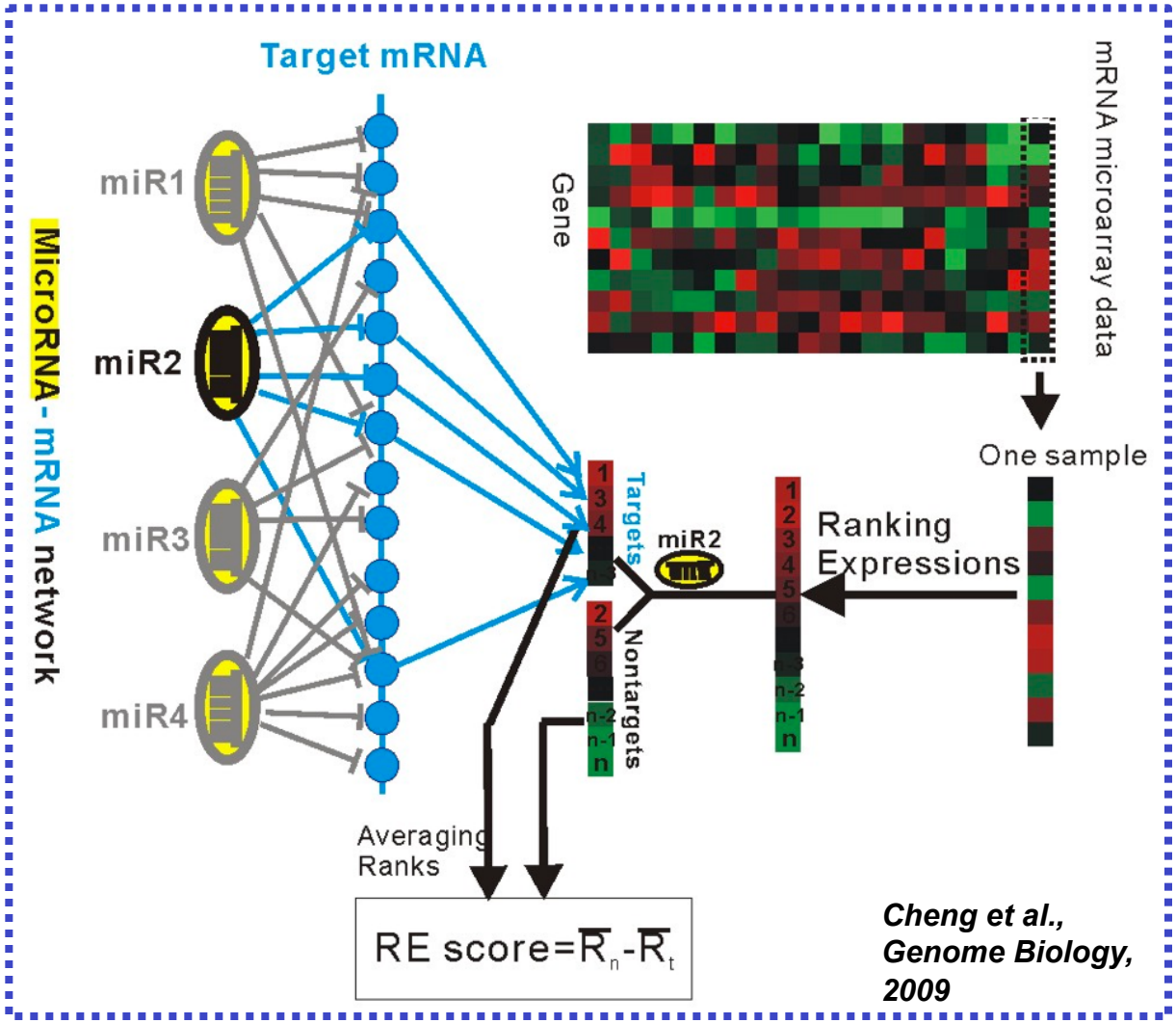


[Yu et al., PNAS (2006)]

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

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

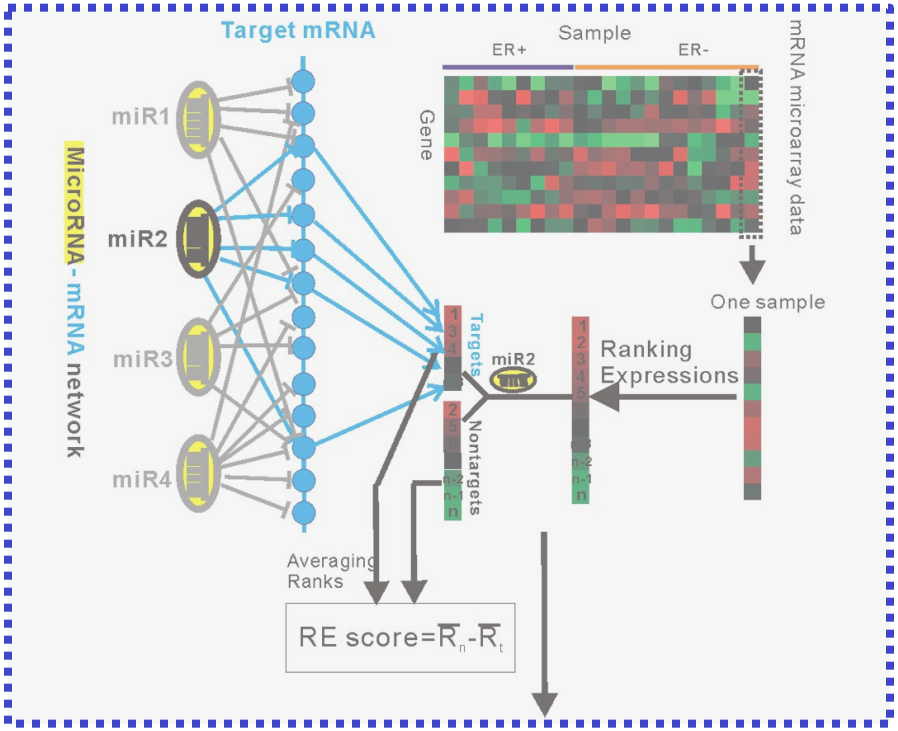
RE-score: Another way to identify "important" network nodes



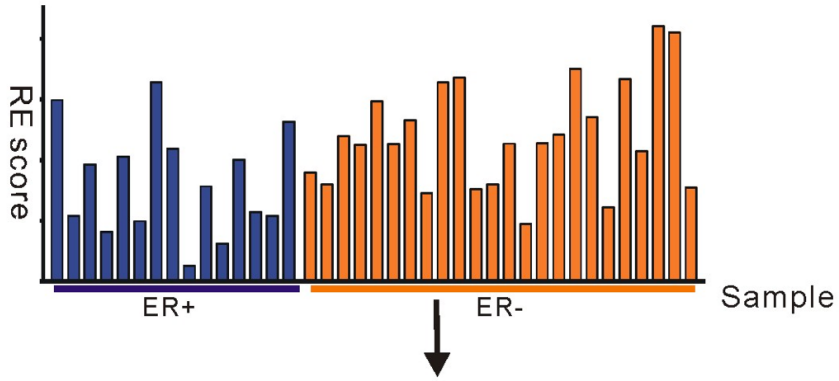
Cheng et al., Genome Biology, 2009

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

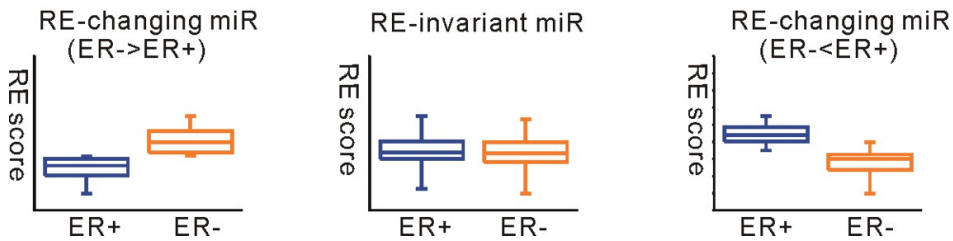
(ER- and ER+ breast cancer)



Calculating RE scores of a miRNA in each sample



Comparing the RE scores between ER+ and ER-

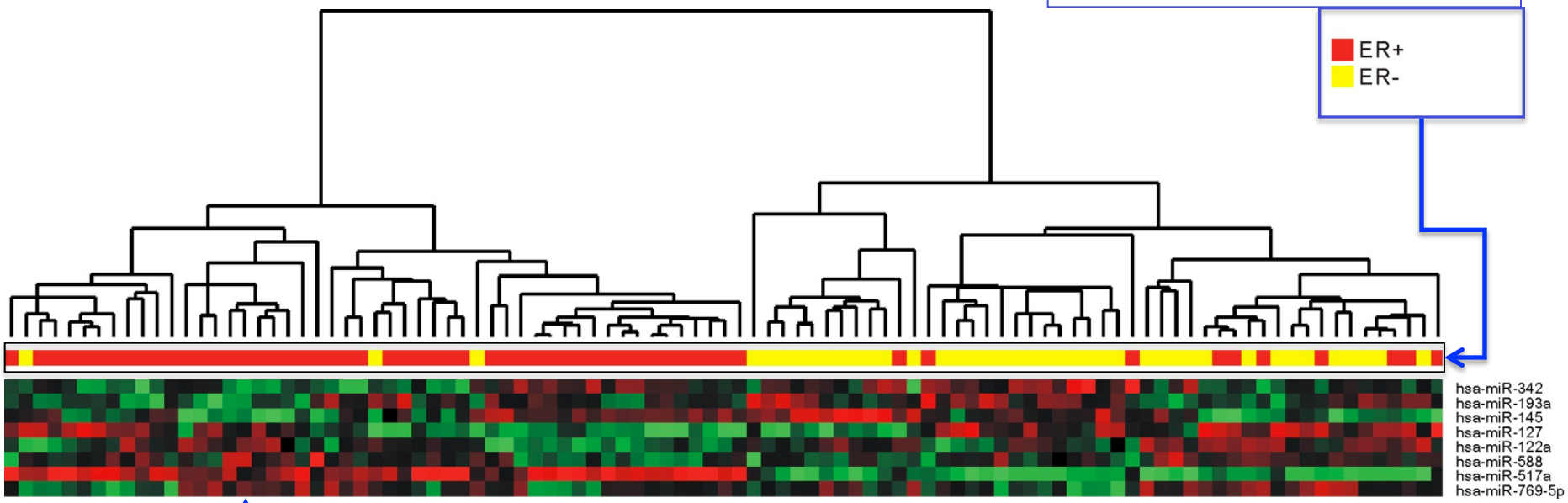


Cheng et al., Genome Biology, 2009

RE-score can be used to classify cancers

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

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

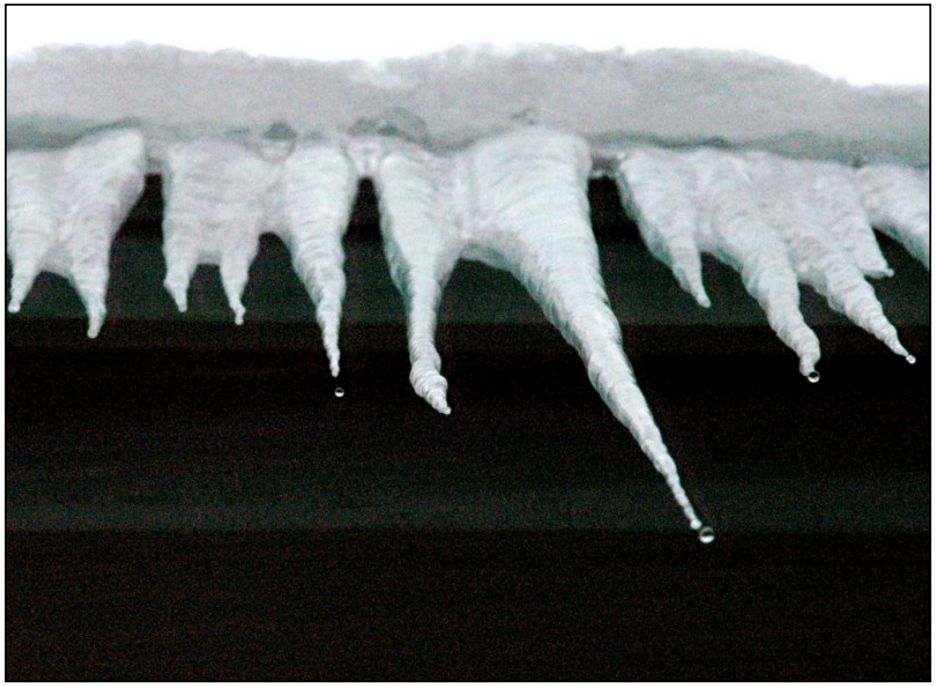


(1) RE-score profile for diff. miRNA in 1 cancer sample.
(2) Tabulate over many different breast cancer samples

Cheng et al., Genome Biology, 2009

Network Dynamics: Environments

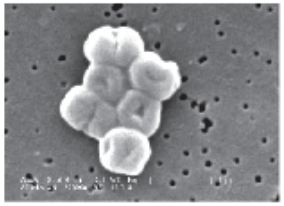
How do molecular networks change across environments?
What pathways are used more ?
Used as a biosensor ?



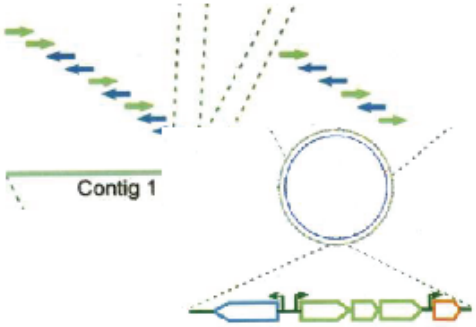
What is metagenomics?

Genomics Approach

Culture Microbes → Extract DNA → Sequence → Assemble and Annotate



ATCGTATA
CGCGAAG
ACGTCTGA
AGTGCTGCT

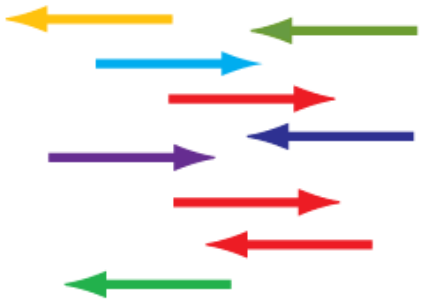


PROBLEM: Estimated that less than 1% can be cultured in the lab

Metagenomics Approach

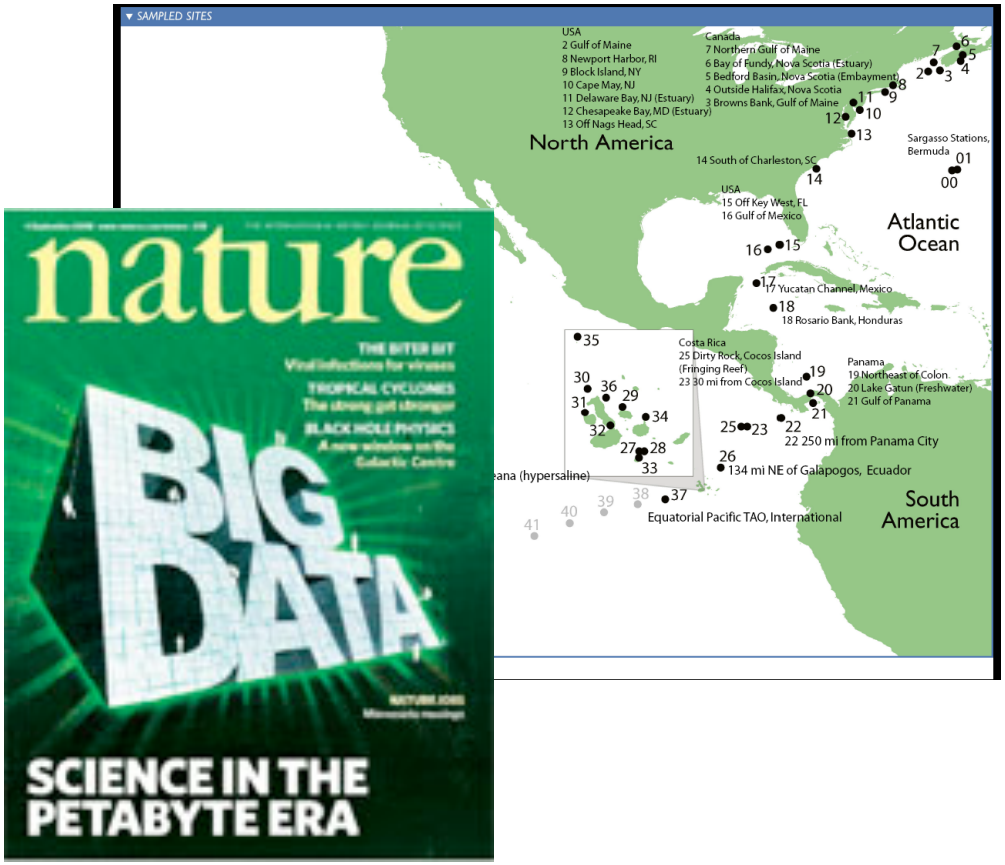
Collect Sample → Extract DNA → Sequence → Partially Assemble and Annotate

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



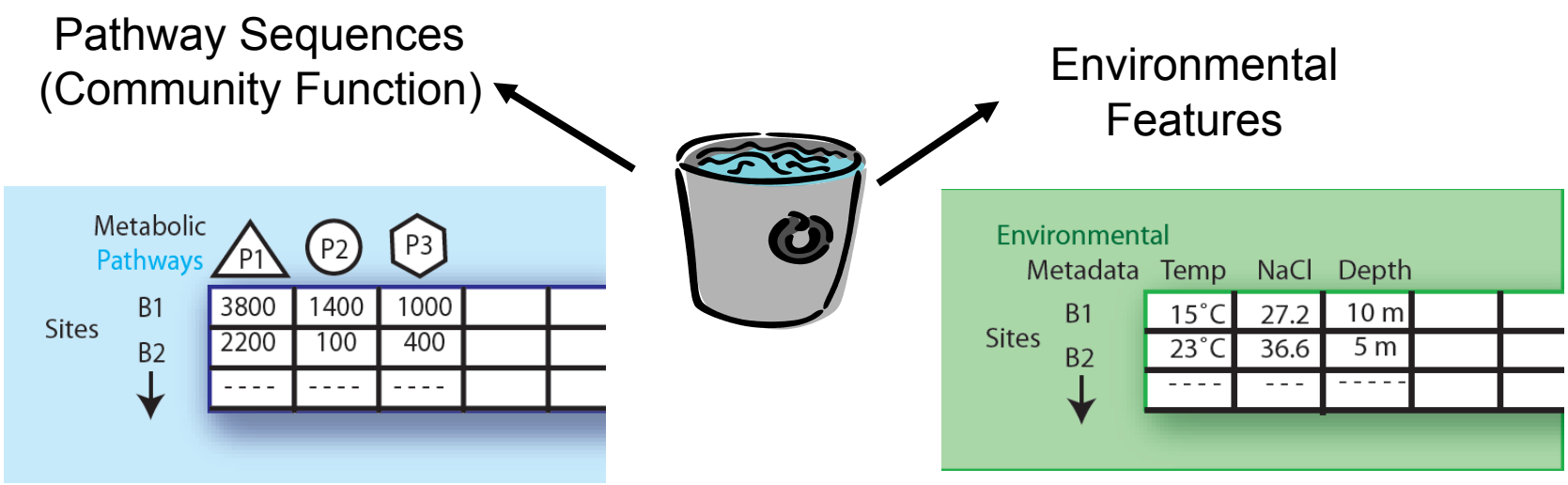
PROBLEM: Lose information about which gene belongs to which microbe.

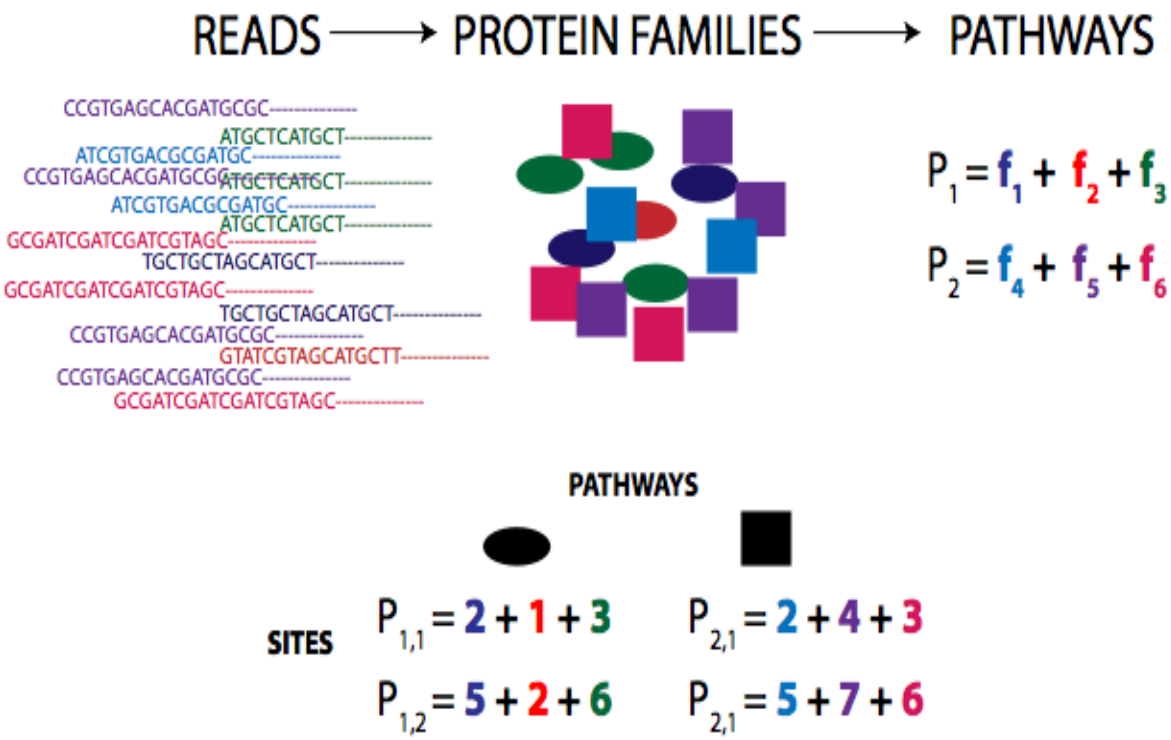
Global Ocean Survey Statistics (GOS)



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

Rusch, et al., PLOS Biology 2007

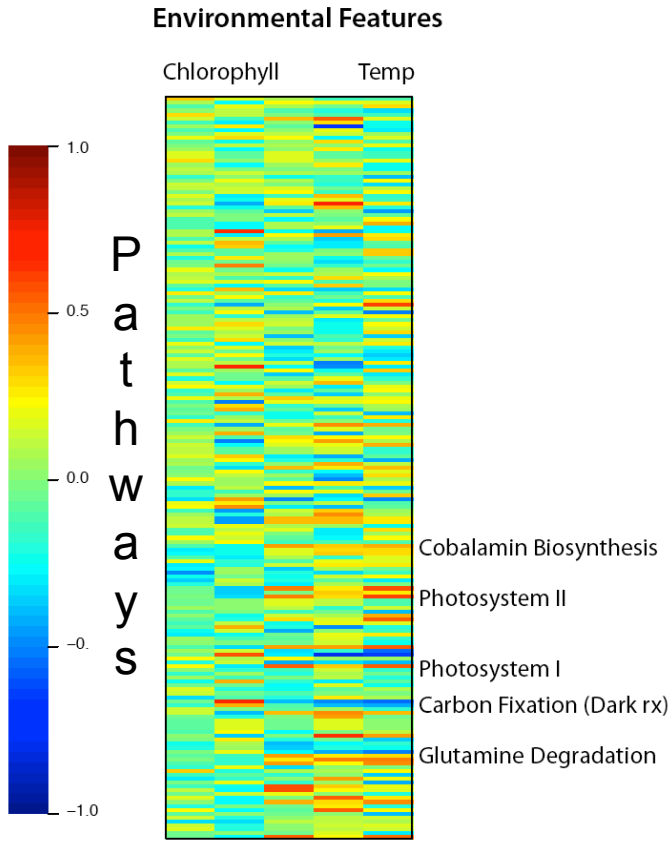


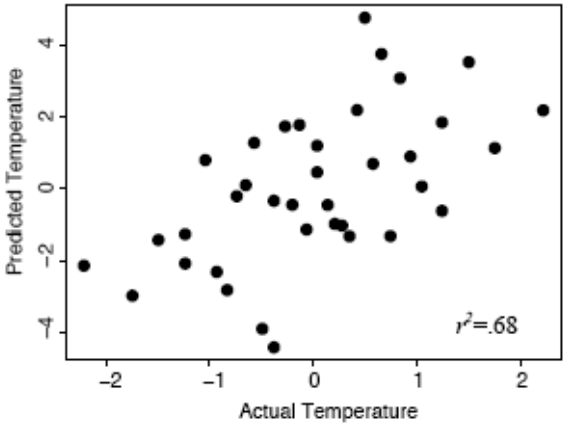


Expressing data as matrices indexed by site, env. var., and pathway usage

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

Simple Relationships: Pairwise Correlations

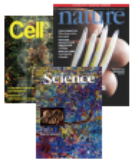




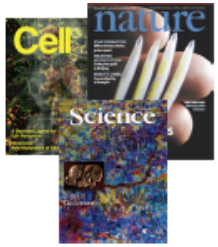
[Gianoulis et al., PNAS (in press, 2009)]

Canonical Correlation Analysis: Simultaneous weighting

Score	# of papers published
GRE	

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

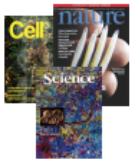
UPI = a GRE + b  + c GPA

GPI = a'  + b'  + c' 

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

Canonical Correlation Analysis: Simultaneous weighting

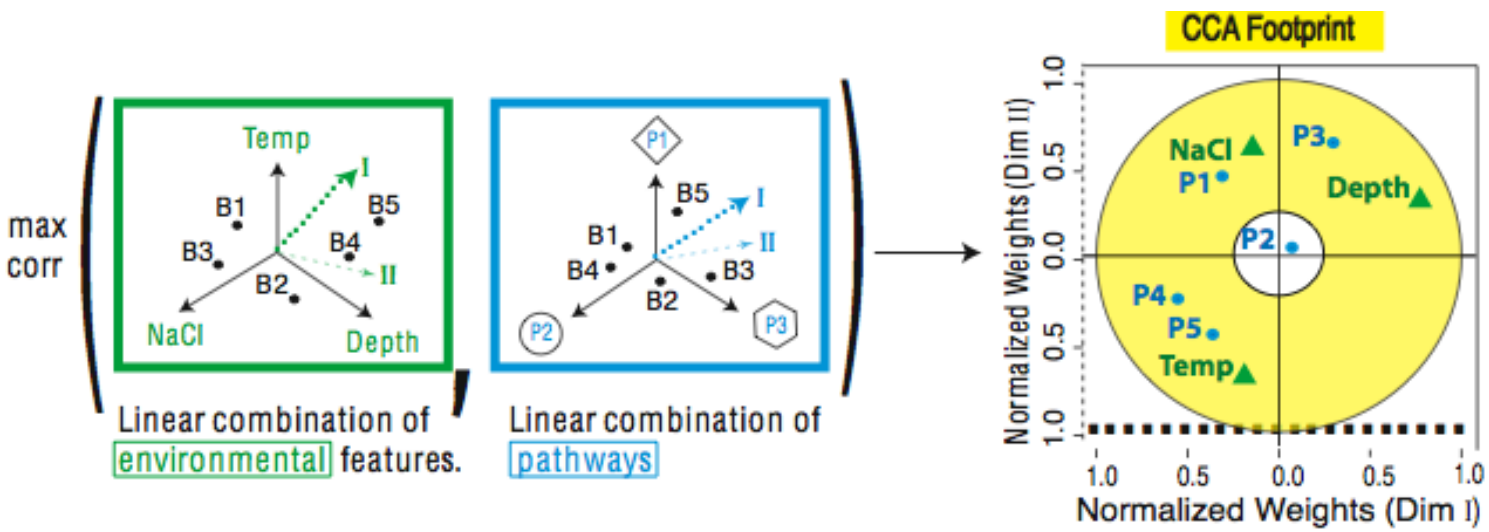
Score	# of papers published
GRE	

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

Environmental Features	Metabolic Pathways
Temp etc	Photosynthesis etc
Chlorophyll	Lipid Metabolism

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

Environmental-Metabolic Space



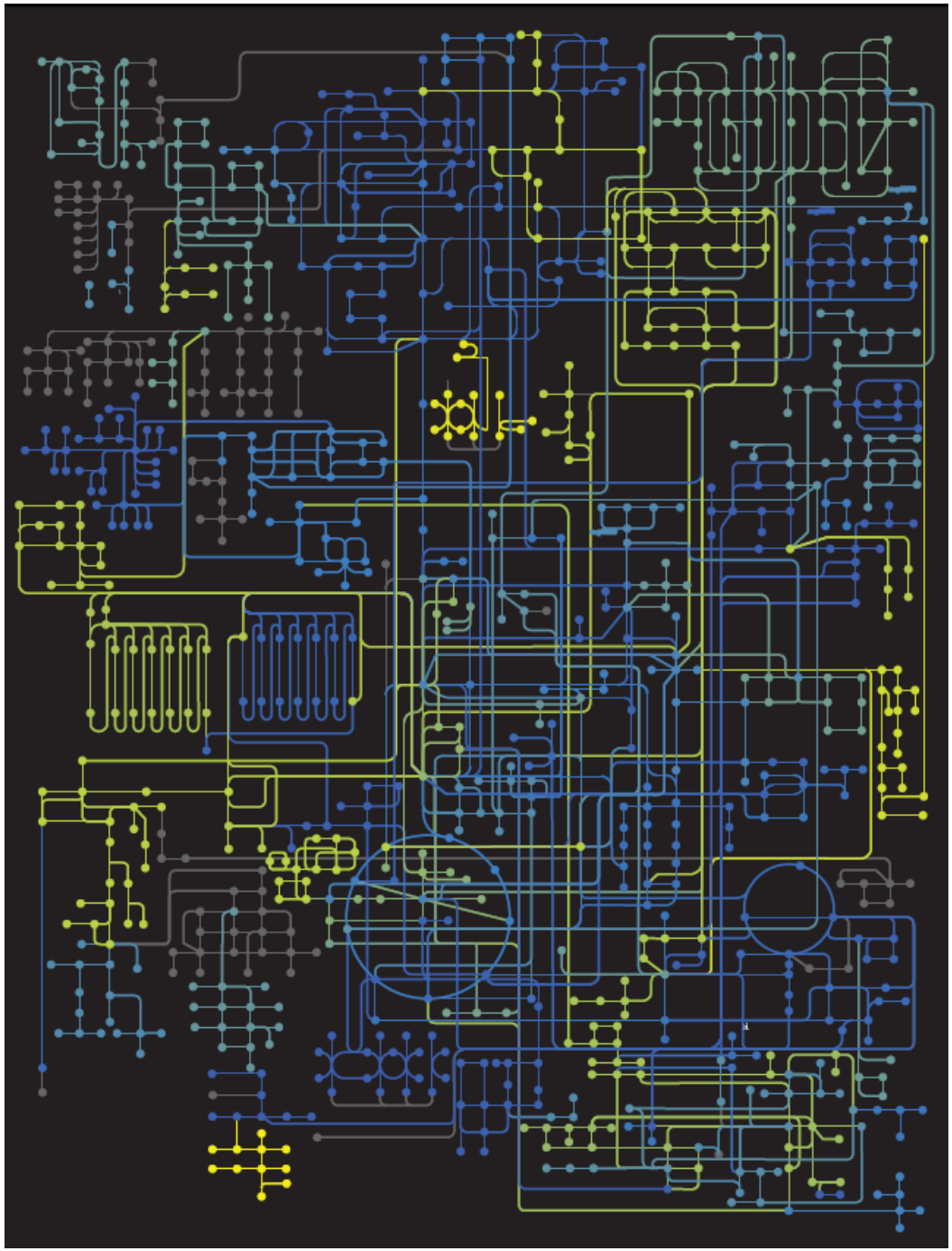
The goal of this technique is to interpret cross-variance matrices
We do this by defining a change of basis.

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

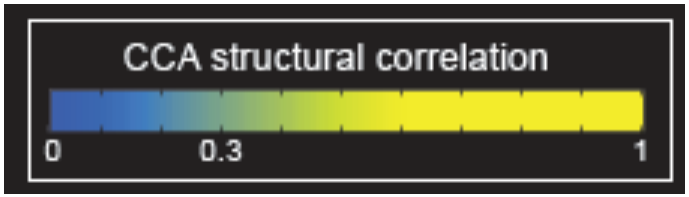
$$C = \begin{matrix} \Sigma_X & \Sigma_{X,Y} \\ \Sigma_Y & \Sigma_{Y,X} \end{matrix}$$

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

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

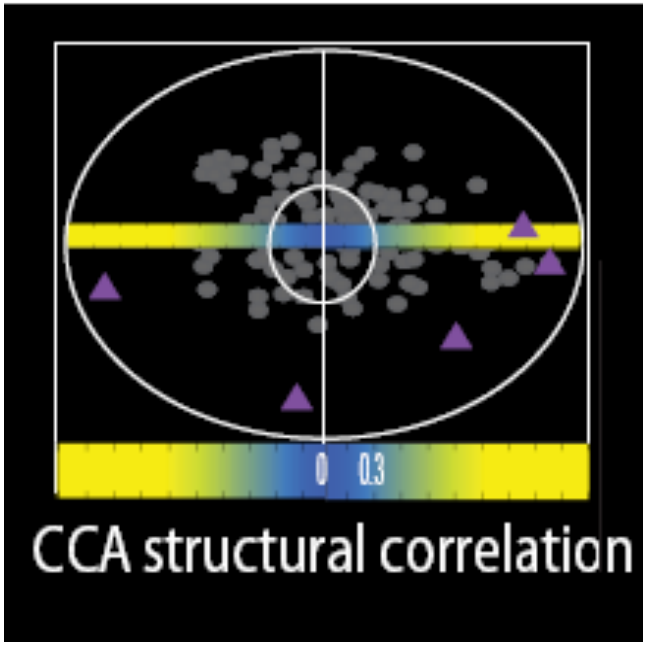


Strength of Pathway co-variation with environment



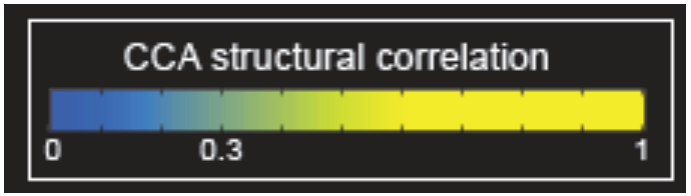
Environmentally invariant

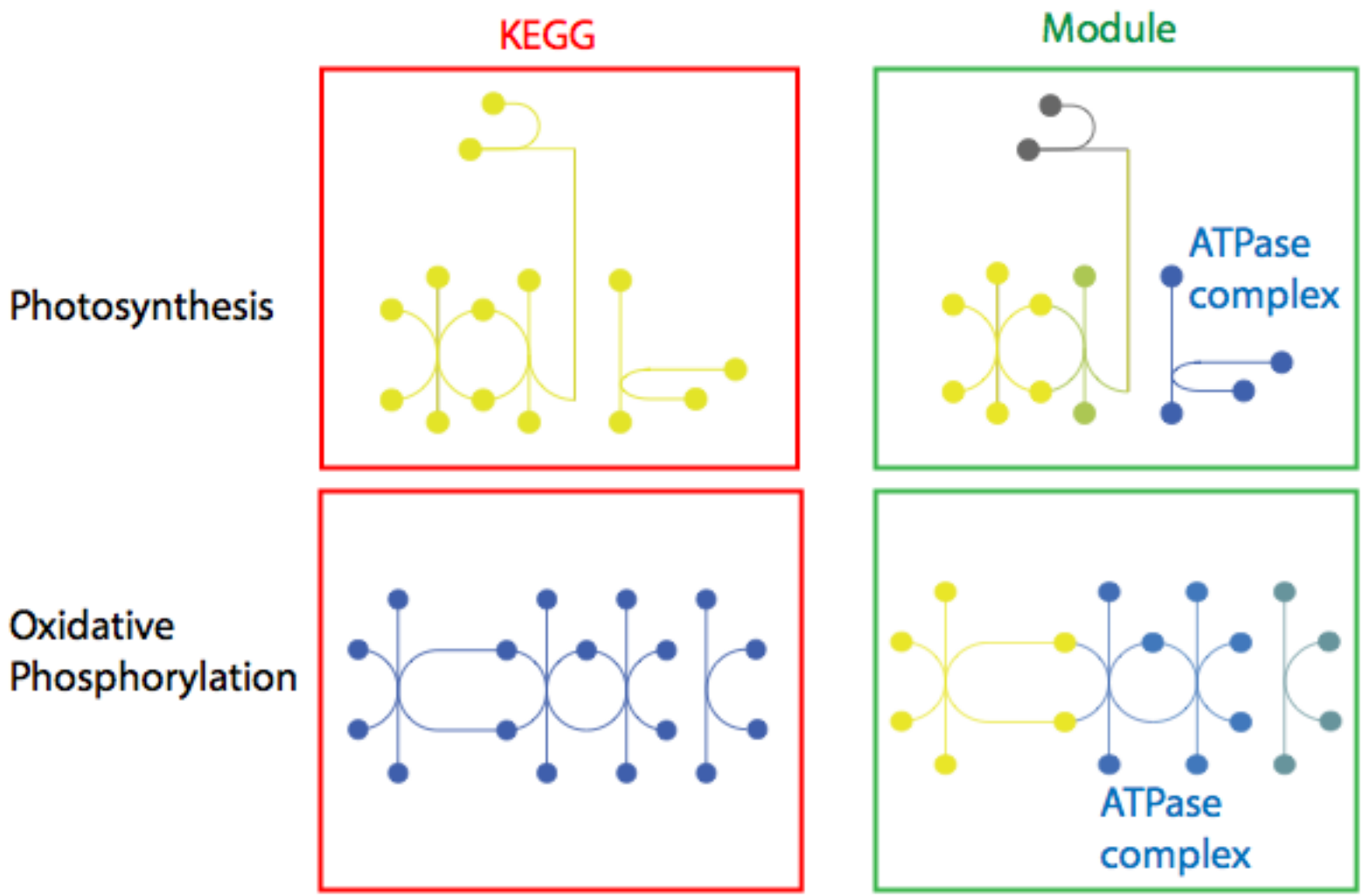
Environmentally variant



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

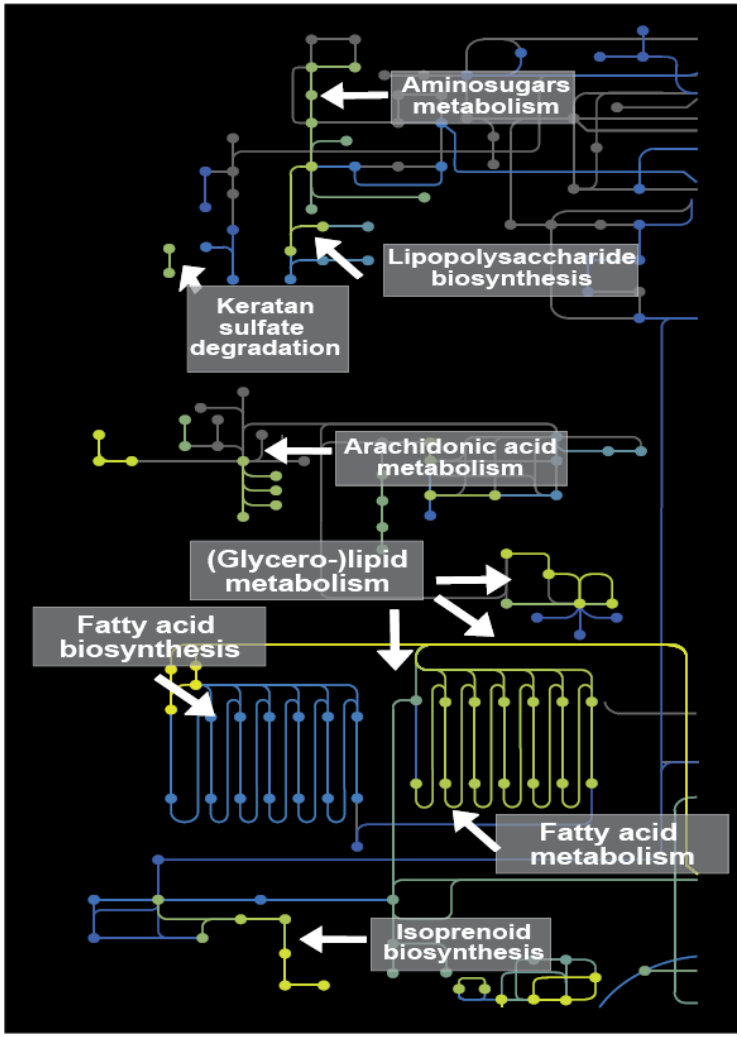
Conclusion #1: energy conversion strategy, temp and depth

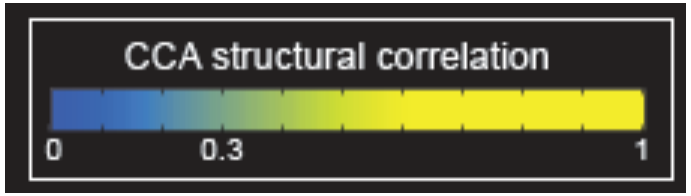




[Gianoulis et al., PNAS (in press, 2009)]

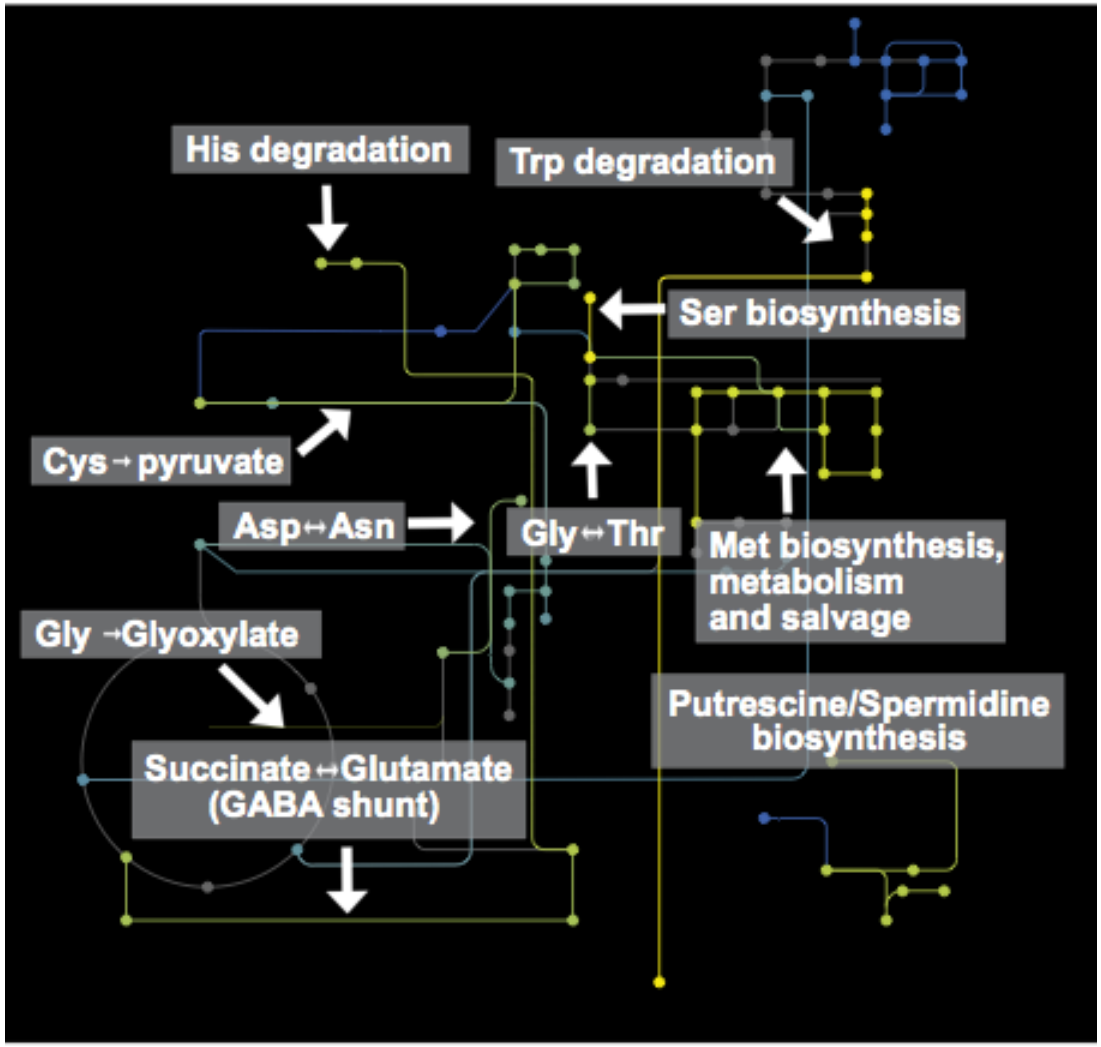
Conclusion #2: Outer Membrane components vary the environment





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

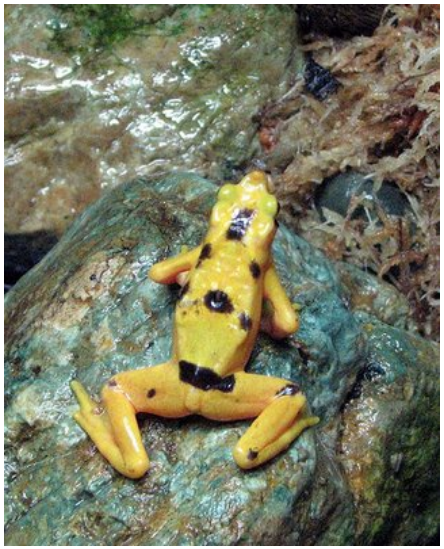
Conclusion #3: Covariation of AA biosynthesis and Import



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

Conclusions on Networks: Predictions

◊ Extrapolating from training sets

◊ Principled ways of using known information in the fullest possible fashion

- Prediction Propagation
- Multi-level learning

Conclusions:

Centrality Measures in Protein Networks

- Hubs & Bottlenecks
  - ◊ Importance of later in regulatory networks
- Regulatory Network Hierarchies
  - ◊ Middle managers dominate, sitting at info. flow bottlenecks
  - ◊ Paradox of influence and essentiality
  - ◊ Topmost proteins sit at center of interaction network
- RE-score
  - ◊ measures degree of (down) regulation of targets v. non-targets
  - ◊ Application to miRNA network
  - ◊ Different miRNA RE-scores in cancer classification

Conclusions: Networks Dynamics across Environments

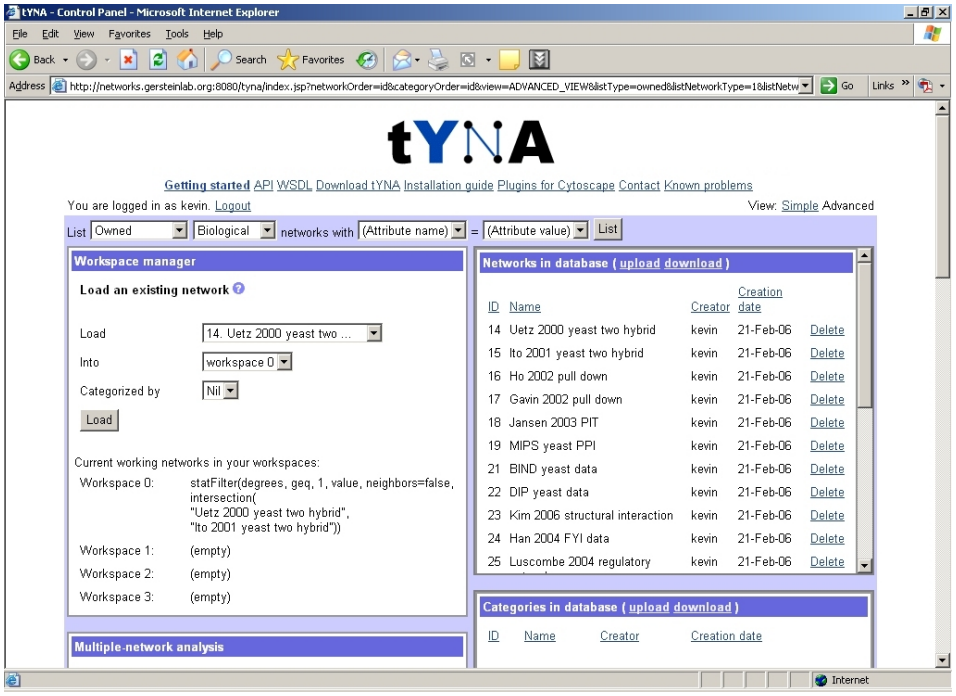
- Developed and adapted techniques (CCA) to connect quantitative features of environment to metabolism
- Identified footprints predictive of environment (potentially as a biosensor)
- clear relationship exists between a community's energy conversion strategies and its environmental parameters (e.g. temperature and chlorophyll)
- Suggest that limiting amounts of cofactor can (partially) explain increased import of amino acids in nutrient-limited conditions.

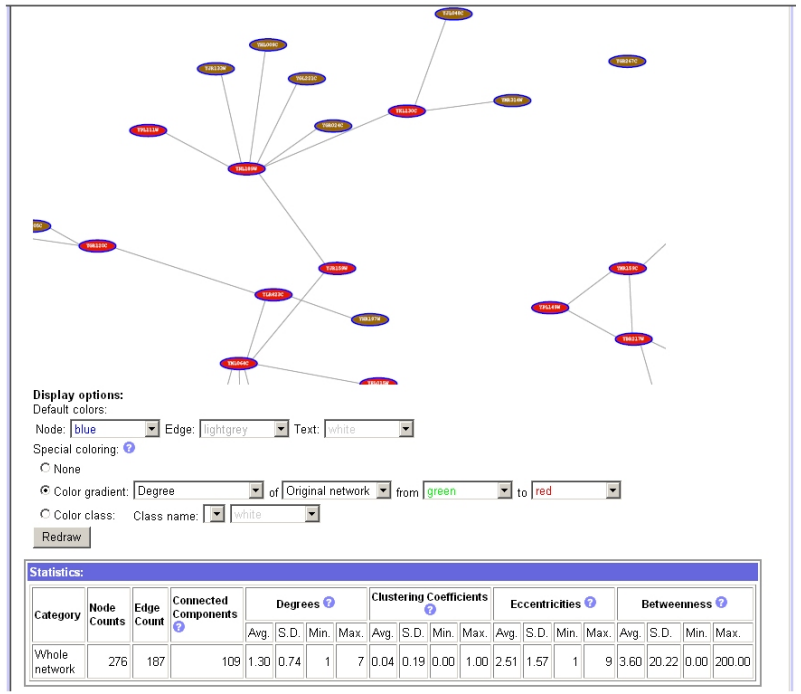
TopNet - an automated web tool

Topology of Networks

tYNA

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





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

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

Acknowledgements

H Yu
P Kim
K Yip
T Gianoulis
C Cheng

A Paccanaro
P Alves
M Seringhaus
Y Xia
A Sboner
P Patel
P Bork
J Raes
M Snyder
N Bhardwaj
R Alexander
L Jensen
T Yamada

Networks.GersteinLab.org

Job opportunities currently for postdocs & students

More Information on this Talk

SUBJECT: Networks

DESCRIPTION:
Functional Genomics & Systems Biology Workshop, Welcome Trust workshop, Cambridge, UK; 2009.11.30, 17:20-17:50; [I:**WTSYSBIO**] (Medium networks talk, shortened from [I:**MBINETS**].)

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

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

.

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