Discovering Commonalities in Interaction Networks

From Alignment to Canonical Modules

Mehmet Koyutürk

Case Western Reserve University
Department of Electrical Engineering & Computer Science

December 17, 2007

Outline

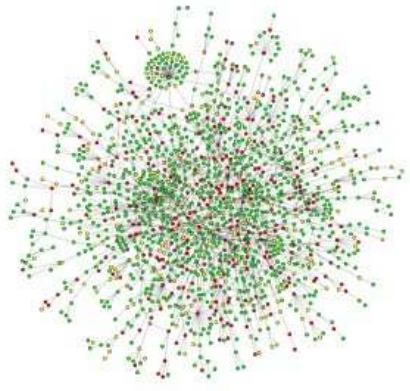
Outline

Why Systems Biology?

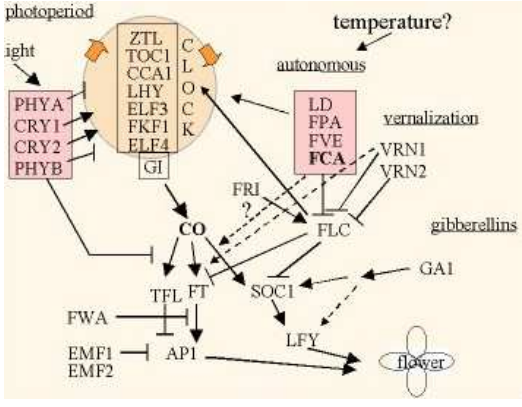
- Life is an emergent property
  - Emergent properties are those that are not demonstrated by individual parts and cannot be predicted even with full understanding of the parts alone
  - Cell is not just an assembly of genes and proteins
- Systems biology focuses on understanding the organization and dynamics of cellular and organismal function
  - Complements molecular biology
  - Organization: How are multiple components organized together?
  - Dynamics: How do multiple components behave together under different circumstances (time and space)?

Understanding Cellular Organization: Networks

- Nodes represent cellular components
  - Protein, gene, enzyme, metabolite
- Edges represent interactions
  - Binding, regulation, modification, complex membership
  - Weighted, directed, signed, hyper
- There are several interrelated models for different aspects of cellular organization and signaling

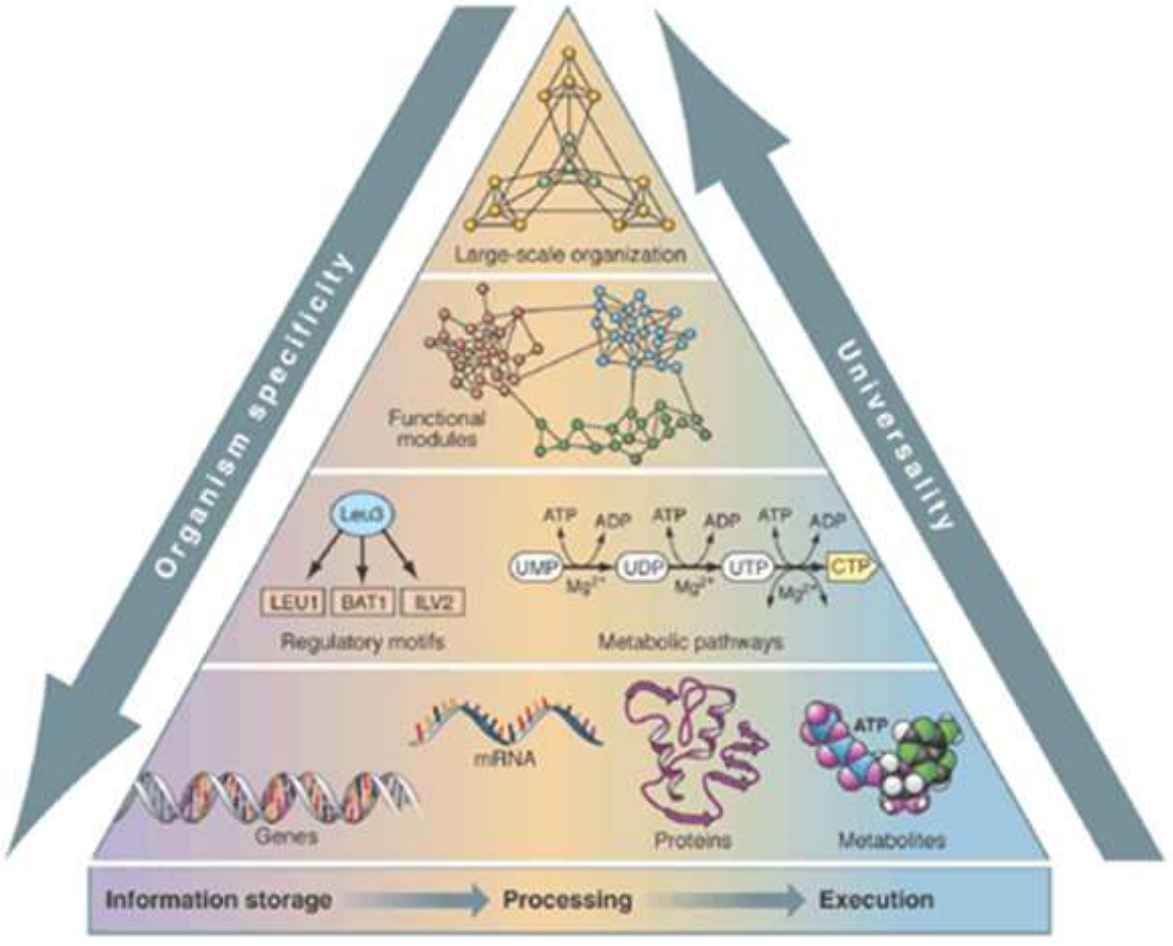


S.cerevisiae
PPI network



Genetic network that controls
flowering time in *A. Thaliana*

Life's Complexity Pyramid



Oltvai and Barabási, *Science*, 2002

Periodic Table of Systems Biology

"Once both the network structure and its functional properties are understood for a large number of regulatory circuits, studies on classifications and comparison of circuits will provide further insights into the richness of design patterns used and how design patterns of regulatory circuits have been modified or conserved through evolution. The hope is that intensive investigation will reveal a possible evolutionary family of circuits as well as a "periodic table" for functional regulatory circuits." –

H. Kitano, *Science*, 2002.

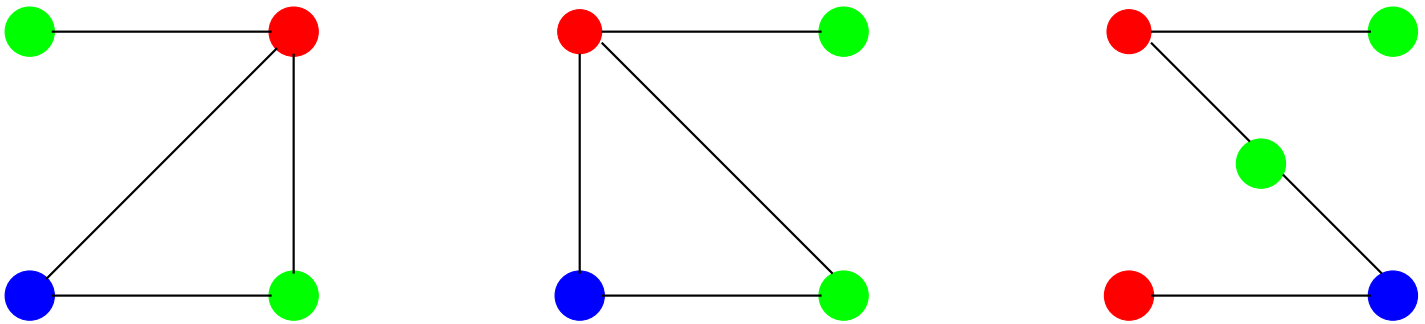
Conservation of Modularity

- Selective pressure on preserving collective function
  - Interacting proteins follow similar evolutionary trajectories (Pellegrini *et al.*, *PNAS*, 1999)
  - Orthologs of interacting proteins are likely to interact (Wagner, *Mol. Bio. Evol.*, 2001)
  - Proteins that are involved in coherent interaction patterns (*e.g.*, feedback loop) are more likely to be conserved (Wuchty *et al.*, *Nature Genetics*, 2003)
  - Sequence homology does not always imply functional orthology (Kelley *et al.*, *PNAS*, 2003)
- Goal: Uncovering evolutionary design principles
  - Interaction patterns that are conserved across (a subgroup of) species (or, more generally, that recur in various contexts) are likely to correspond to a modular component of cellular organization

Outline

Frequent Subgraph Discovery

- Computationally intractable: Existing algorithms are limited to very small graphs and subgraphs
- How can we color the nodes consistently across species?

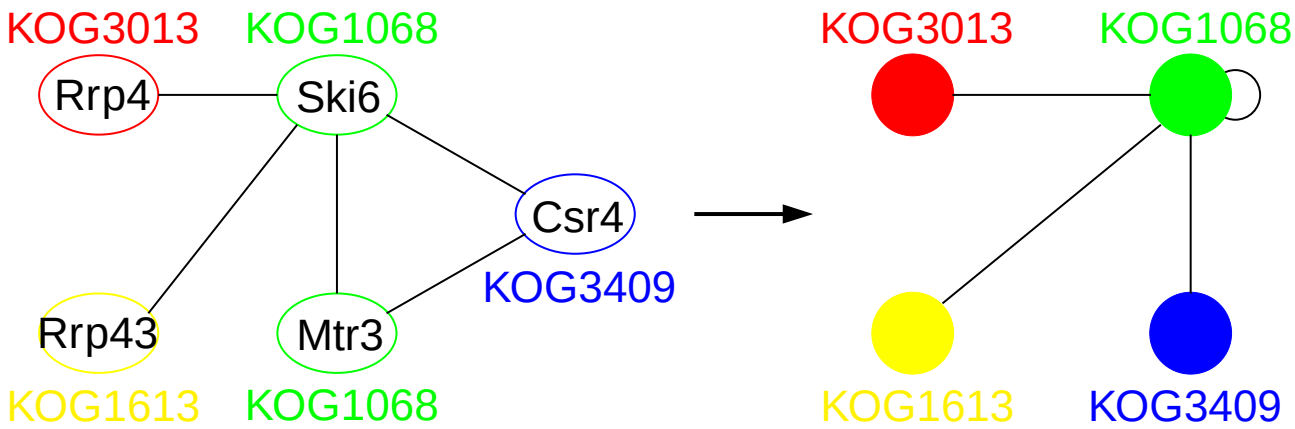


Network database

Interaction patterns that are common to all networks

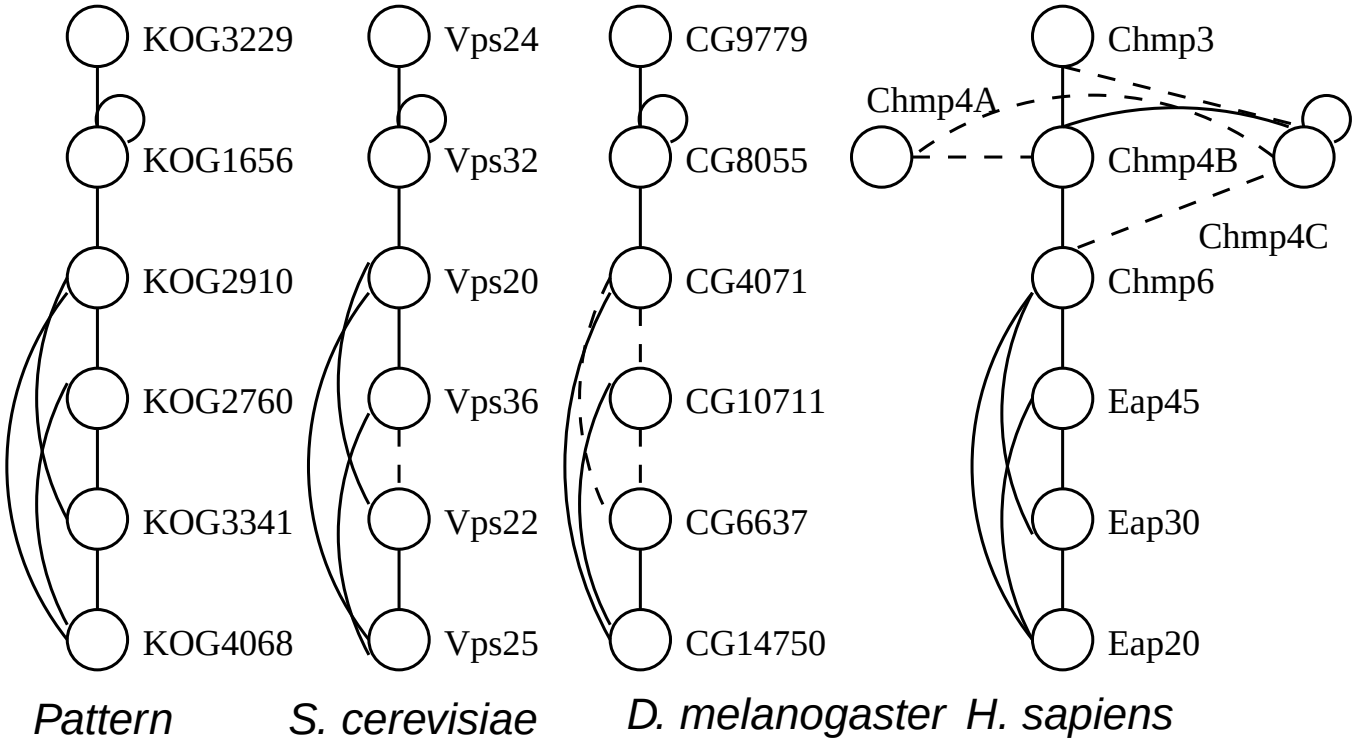
Ortholog Contraction

- Interaction between proteins → Interaction between ortholog groups or protein families (Koyutürk *et al.*, *ISMB*, 2004)



- Ortholog contraction preserves frequent subgraphs (but it may add non-existing patterns)
  - Can be used as a filtering mechanism

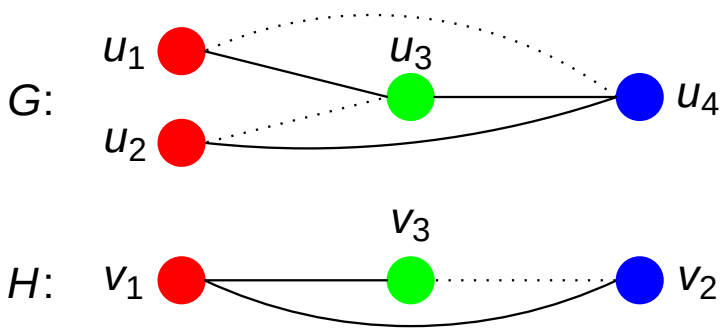
Frequent Protein Interaction Patterns



Endosomal sorting ($p < 1e - 78$)

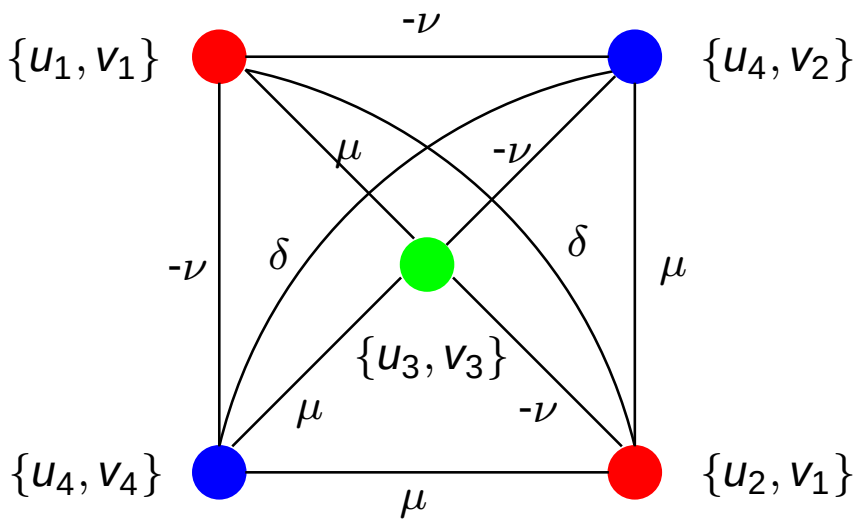
Aligning PPI Networks

- Given two PPI networks that belong to two different organisms, identify sub-networks that are similar to each other
  - Biological meaning, mathematical modeling
- MAWISH (Koyutürk *et al.*, *J. Comp. Biol.*, 2006)
  - Each evolutionary event (match, mismatch, duplication) is associated with a score
  - The score of a pair of subgraphs, each from one network, is a linear combination of scores of individual evolutionary events



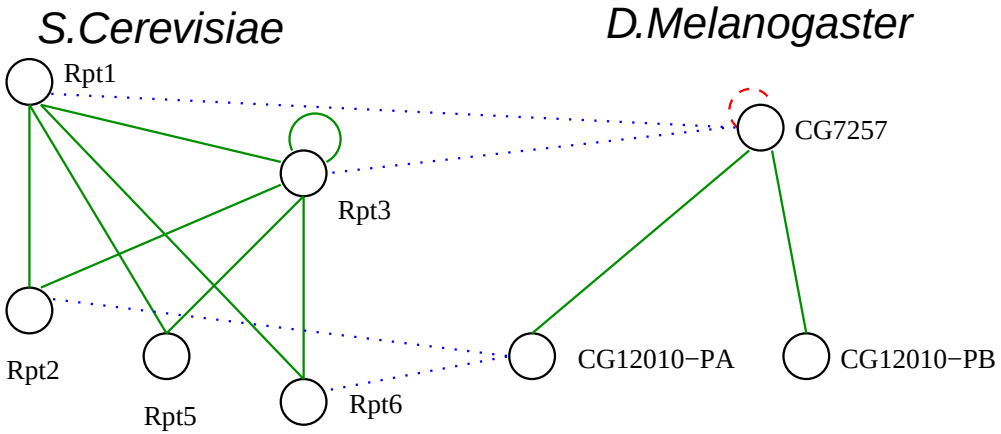
Alignment Graph

- $\mathbf{G}(\mathbf{V}, \mathbf{E})$: $\mathbf{V}$ consists of all pairs of homolog proteins $\mathbf{v} = \{u \in U, v \in V\}$
- An edge $\mathbf{vv}' = \{uv\}\{u'v'\} \in \mathbf{E}$ is a
  - match edge if $uu' \in E$ and $vv' \in V$
  - mismatch edge if $uu' \in E$ and $vv' \notin V$ or vice versa
  - duplication edge if $S(u, u') > 0$ or $S(v, v') > 0$
- The edges are weighted accordingly

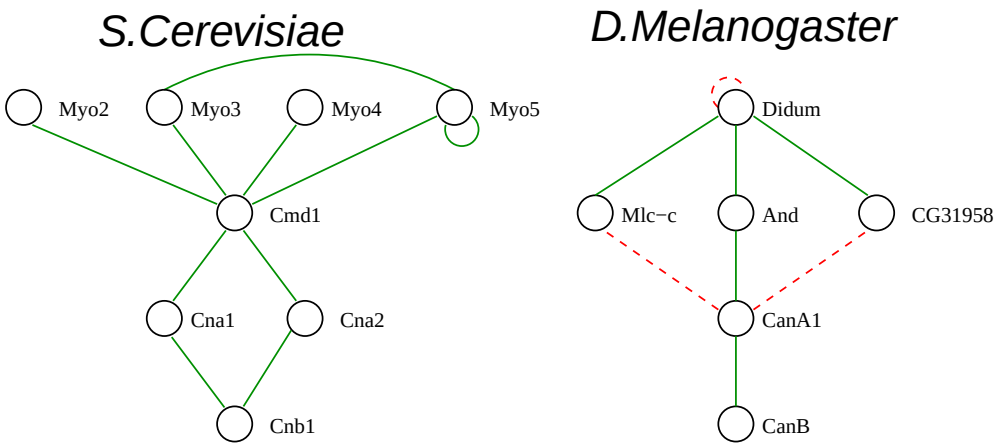


Subnets Conserved in Yeast and Fruit Fly

Proteosome regulatory particle



Calcium-dependent stress-activated signaling pathway



Outline

Modularity and Connectivity

- A functional module is generally defined as a group of molecules that perform a distinct function together
  - Functional modules often manifest themselves as dense (highly interconnected) subgraphs in PPI networks
- There exist many algorithms for "graph clustering", *i.e.*, finding proteins in the network that induce a dense subgraph
- Noise, ambiguity in definition of modules: No unique way to decide what is "dense" in a biologically sound manner
- Our approach: Statistical significance – Find subgraphs that are "unusually" dense (Koyuturk *et al.*, *J. Comp. Bio.*, 2007)

Statistical Significance of Subgraph Connectivity

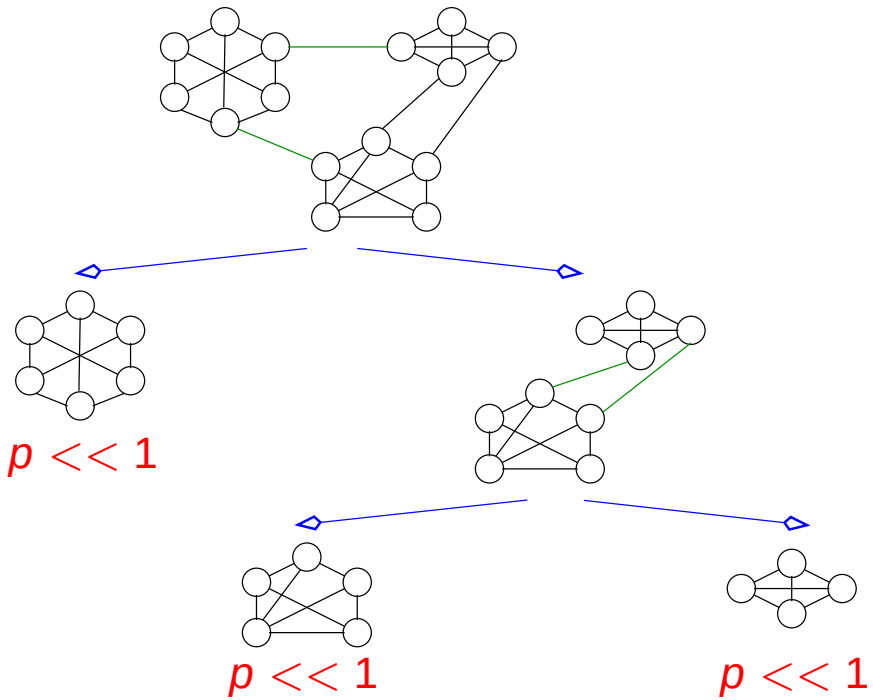
- For a given reference network generation model and fixed density threshold ρ, we are interested in the largest subgraph with density at least ρ
  - If a larger subgraph with density ρ exists in the observed network, it is likely to be generated by factors other than randomness (with respect to the reference model)
  - By looking at the largest dense subgraph, we account for multiple hypothesis testing
- In network analysis, statistical significance is generally evaluated using Monte-Carlo simulations
  - Analytical models are computationally less expensive, and provide information on the distribution of the pattern before pattern is found
  - For given density, we can decide on the range of subgraph size that is interesting, and vice versa

Largest Dense Subgraph

- Let $r_{\mathcal{M}}$ denote the expected size of the largest ρ-dense subgraph with respect to reference model $\mathcal{M}$
- $G(n,p)$ model
  - $r_0 \approx \frac{\log n}{\kappa(p,\rho)}$, where $\kappa(p,\rho) = \rho \log \frac{\rho}{p} + (1-\rho) \log \frac{1-\rho}{1-p}$ denotes divergence, n denotes number of proteins, p denotes interaction probability
- Piecewise $G(n,p)$ model
  - The number of hub nodes $n_h << n$ contributes a constant factor to the size of the largest dense subgraph, *i.e.*, $r_1 \approx \frac{\log n + 2n_h \log B}{\kappa(p_l,\rho)}$, where $B = \frac{p_h + p - 2p_h p}{p}$ and $p_h >> p$ denotes the probability of interaction with a hub protein
- How about power-law graphs?

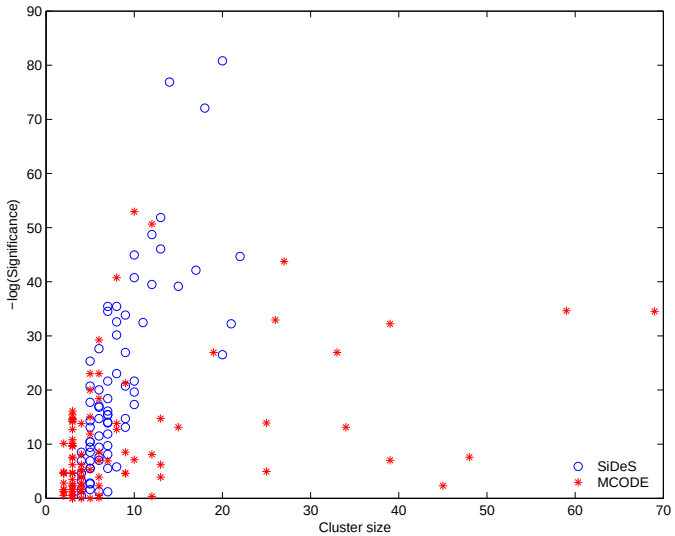
Identifying Significantly Dense Subgraphs

- SIDES algorithm
  - Recursive min-cut until density of subgraph is statistically significant
  - Based on HCS (Hartuv & Shamir, *Information Processing Letters*, 2000)

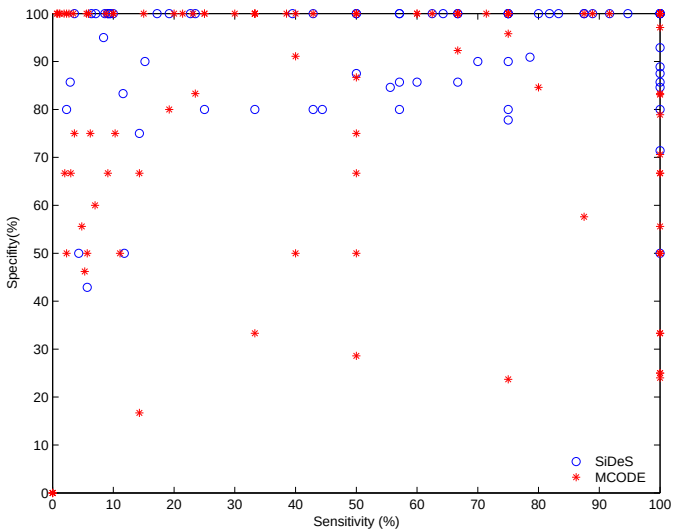


Performance of SIDES

- Biological relevance of identified subgraphs is evaluated with respect to Gene Ontology (GO)
  - If there is a GO term that is significantly enriched in the subgraph, then it is likely to correspond to a functional module



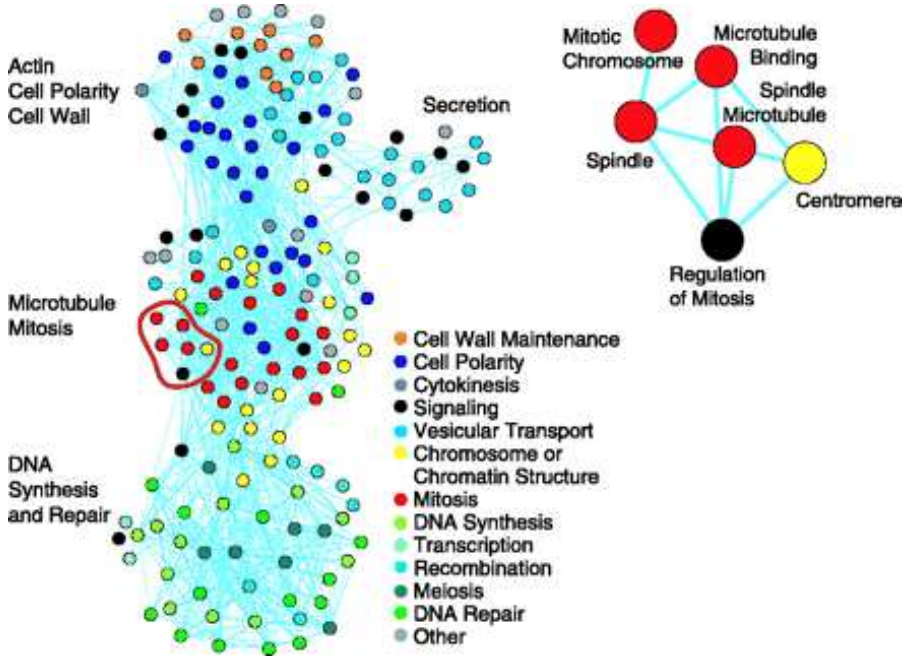
Size vs Biological Significance



Sensitivity vs Specificity

Annotation: From Individual Molecules to Systems

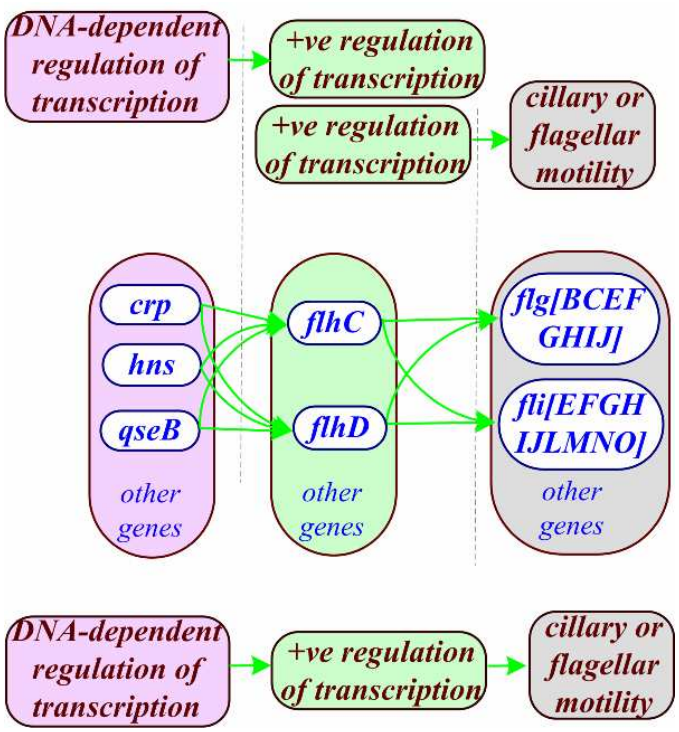
- Networks are species-specific
- Annotation is at the molecular level
- Map networks from gene space to an abstract (and unified) function space

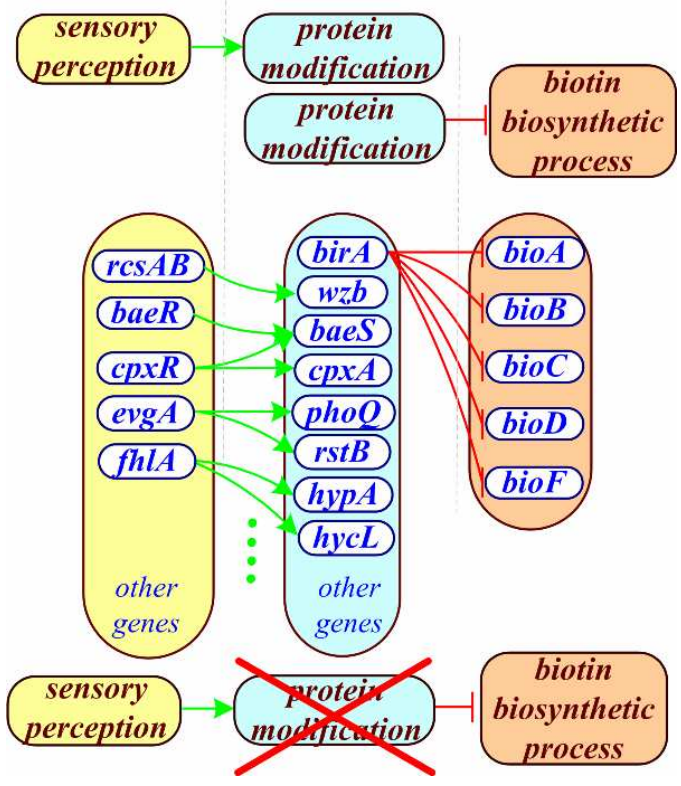


Network of GO terms based on significance of pairwise interactions in *S. cerevisiae* Synthetic Gene Array (SGA) network (Tong *et al.*, *Science*, 2004)

Gene Regulatory Networks: Indirect Regulation

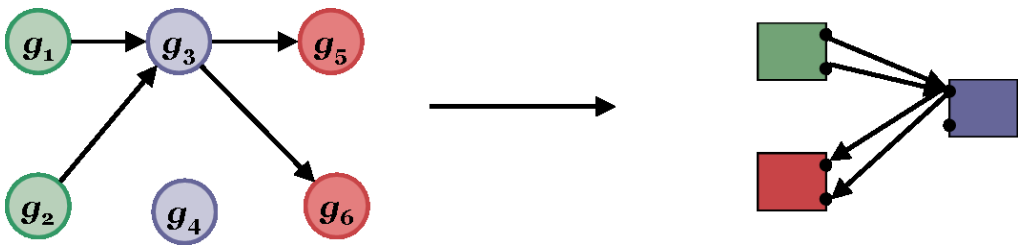
- Assessment of pairwise interactions is simple, but not adequate





Pathways of Functional Attributes

- A pathway of functional attributes maps to multiple pathways of genes in different contexts
  - We want to identify functional pathways that are overrepresented in the gene network
  - These might help build a "periodic table of systems biology"
- Frequency alone is not a good measure of statistical significance
  - The distribution of functional attributes and degree distribution of genes are not uniform

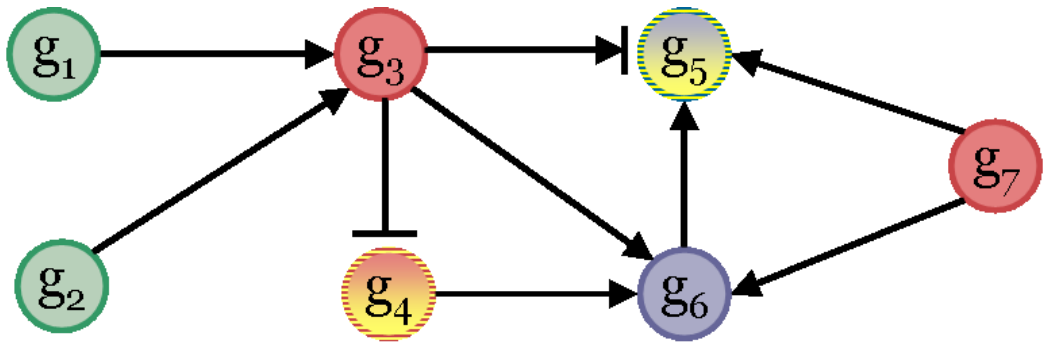


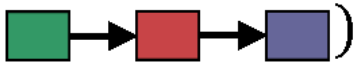
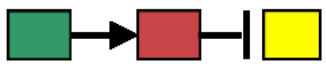
Gene Network

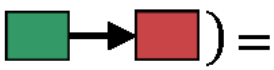
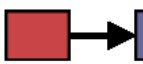
Functional Attribute Network

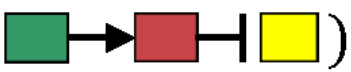
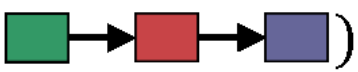
Statistical Significance of a Pathway

- Emphasize modularity of pathways
  - Condition on frequency of building blocks
  - Evaluate the significance of the coupling of building blocks



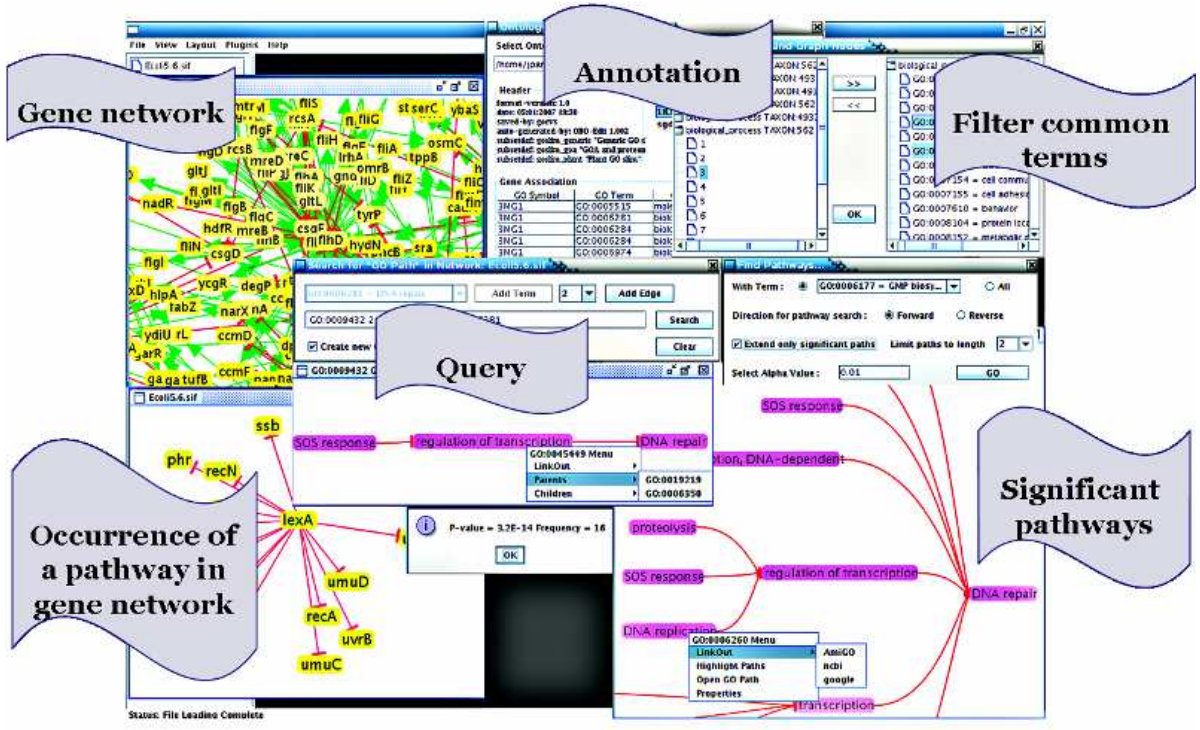
$\varphi($  $) = \varphi($  $) = 4$

$\varphi($  $) = \varphi($  $) = 2$ $\varphi($  $) = 5$

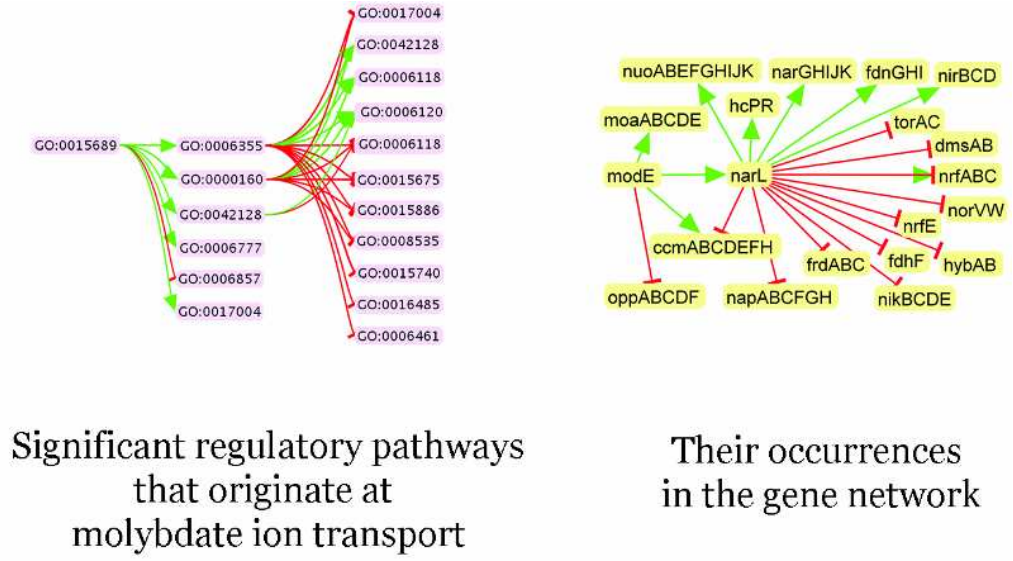
P() < P()

NARADA

- A software for identification of significant pathways (Pandey *et al.*, *ISMB*, 2007)
  - Given functional attribute T, find all significant pathways that originate (terminate) at T
  - User can explore back and forth between the gene network and the functional attribute network



An Example: Molybdate ion transport

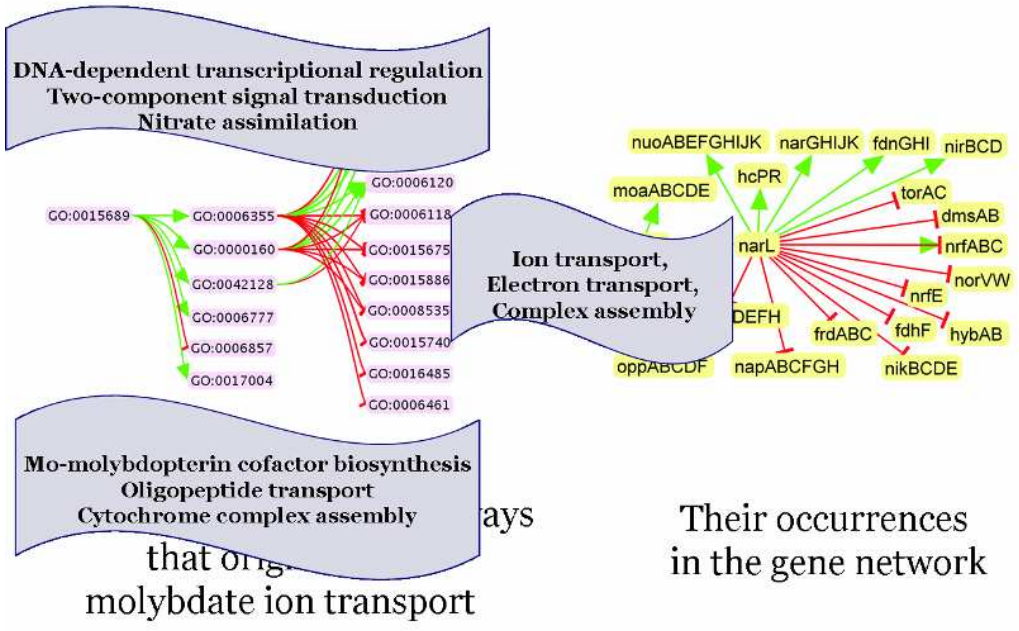


Significant regulatory pathways that originate at molybdate ion transport

Their occurrences in the gene network

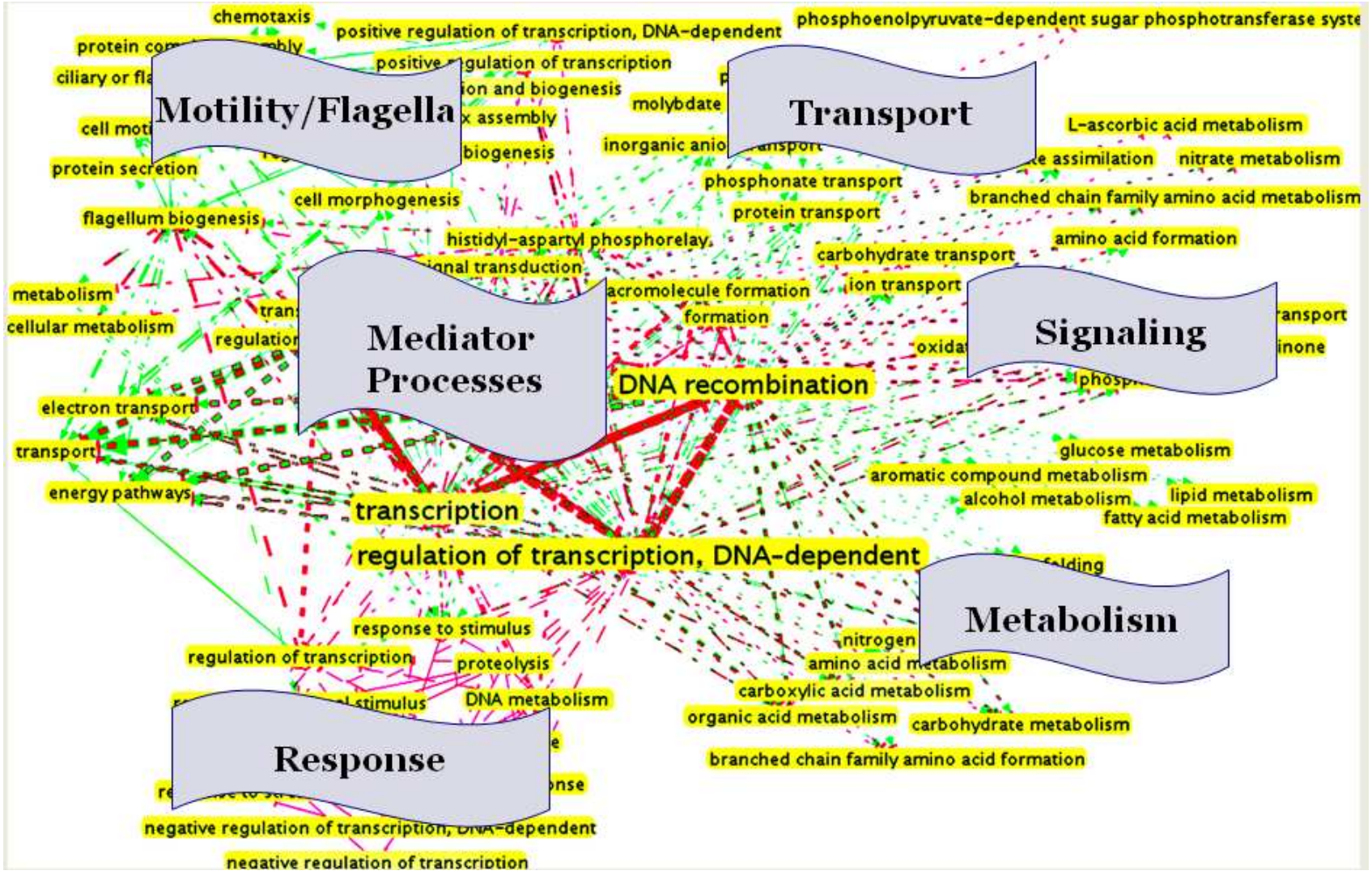
- modE regulates various processes directly
- It regulates various other processes indirectly
  - Regulation of these mediator processes is not significant on itself
  - NARADA captures modularity of indirect regulation!

An Example: Molybdate ion transport

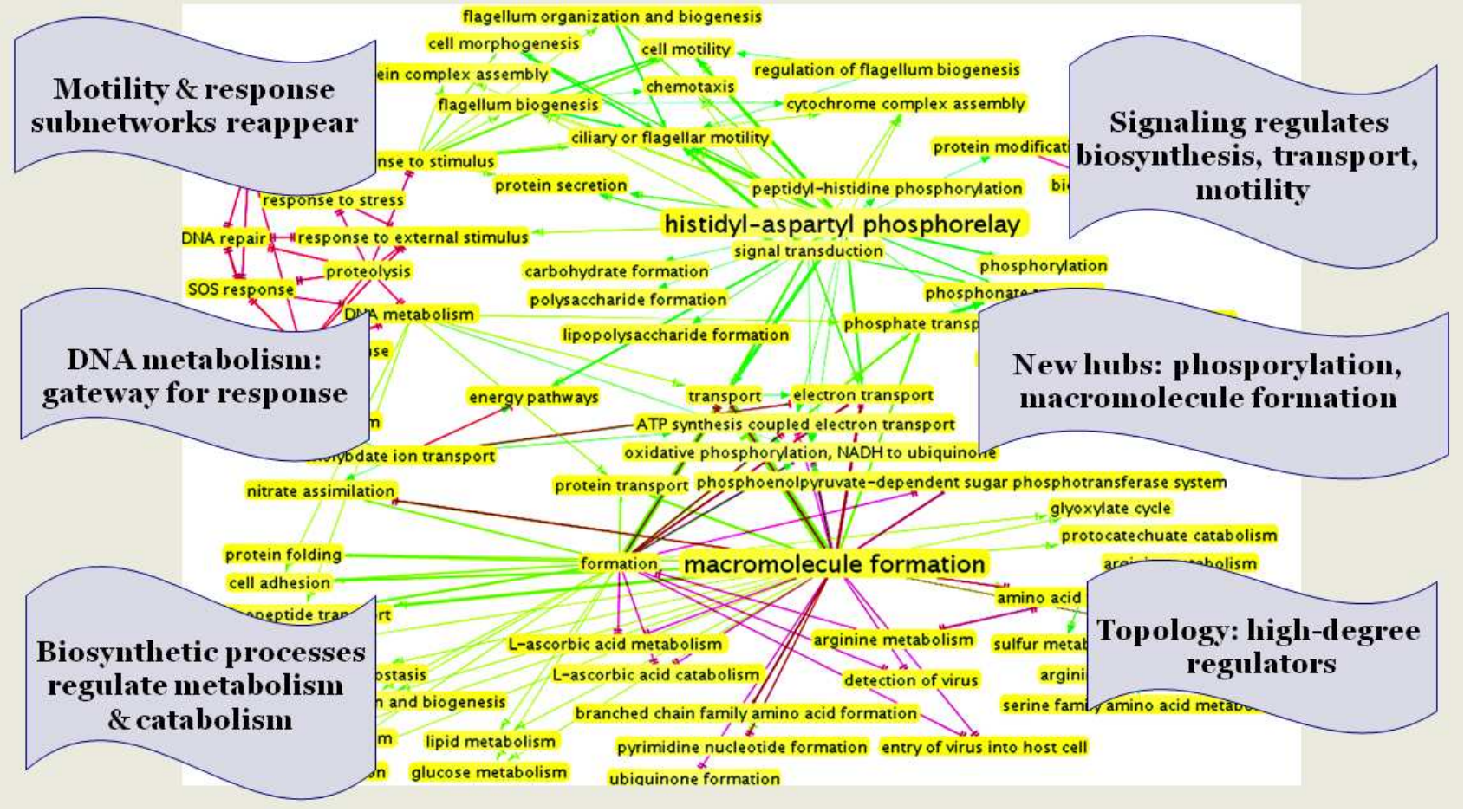


- modE regulates various processes directly
- It regulates various other processes indirectly
  - Regulation of these mediator processes is not significant on itself
  - NARADA captures modularity of indirect regulation!

Functional View of *E. coli* Regulatory Network

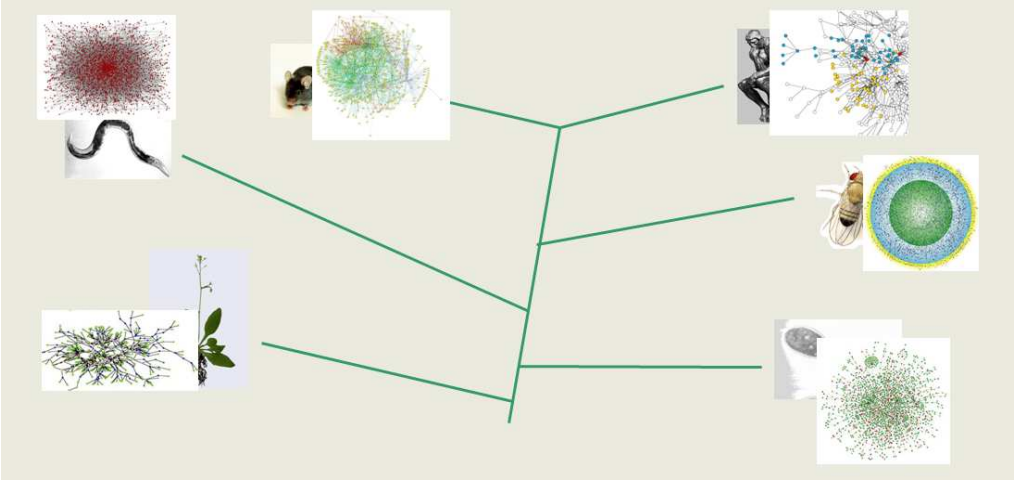


Short-Circuiting Mediator Processes



Outline

Phylogenetic Analysis of Cellular Organization



- Can we guide comparative network analysis based on available phylogenetic information?
- Can we reconstruct the phylogenetic tree based on network comparisons?
  - Existing approaches: Define a measure of (global) similarity between a pair of networks

Feature Based Network Analysis

- Consider each cell as a set of features that represent components and properties of cellular organization
  - Components: Modular subgraphs, pathways, topological motifs
  - Properties: Degree distribution, clustering coefficients
  - Facilitates integration of various network models (PPI, gene regulation, metabolic etc.)
  - Use features to distinguish cells, use cells to distinguish features
  - Enables multi-level analysis: The cells can be considered at species, individual, or tissue level
- We calibrate and verify methods using networks that are generated based on evolutionary models

From Functional Domain to Molecular Domain

- Recall that identified functional pathways are "abstract"
  - How can we test the conclusions derived from identified pathways?
- Project pathways that are identified in one species on another species
  - Identify significant functional pathways in *E. coli* transcriptional network
  - Then, find (partial) occurrences of these pathways in the *B.subtilis* transcriptional network
  - Score missing interactions based on the significance of partial pathways that contain them
- Preliminary results are promising

Outline

Thanks

- CWRU
  - Sinan Erten, Xin Li, Gurkan Bebek, Jing Li
- Purdue
  - Jayesh Pandey, Wojciech Szpankowski, Ananth Grama
- UC-San Diego
  - Yohan Kim, Shankar Subramaniam