Comparative Analysis of Molecular Interaction Networks

Jayesh Pandey[1], Mehmet Koyuturk[2], Shankar Subramaniam[3], Ananth Grama[1]

[1]Purdue University, [2]Case Western Reserve University
[3]University of California, San Diego

This work is supported by National Institutes of Health, National Science Foundation, and Intel.

Outline

- Why is comparative interactomics computationally challenging?
- Some results in conservation, alignment, and modularity.
- Statistical significance as an optimization metric for algorithms.
- Some open problems in computational interactomics.

Challenges in Computational Analysis

- Isomorphism Hurts!
  - Given two networks (unlabeled or labeled with potentially repeating node labels), are they identical? (complexity unknown)
  - Given two networks (unlabeled or labeled with potentially repeating node labels), what is the largest common component (NP Hard!).
- Must rely on nature of model and network – emphasizes analysis!
- Analytical modeling of specific network structures is in relative infancy. (e.g., what is the expected size of a clique in a scale-free graph?)
- Quantification of significance (e.g., p-values) is hard!

Challenges in Computational Analysis: Thesis

- Use a mix of modeling and measures to render problems easier.
- Use nature of underlying networks to develop effective algorithms.
- Use analysis to support claims of algorithmic effectiveness and efficiency.

Conservation in Interaction Networks

- "Evolution thinks modular" (Vespignani, *Nature Gen.*, 2003)
- Cooperative tasks require all participating units
  - Selective pressure on preserving interactions & interacting proteins
  - Interacting proteins follow similar evolutionary trajectories (Pellegrini et al., *PNAS*, 1999)
- Orthologs of interacting proteins are likely to interact (Wagner, *Mol. Bio. Evol.*, 2001)
  - Conservation of interactions may provide clues relating to conservation of function
- Modular conservation and alignment hold the key to critical structural, functional, and evolutionary concepts in systems biology

Conserved Interaction Patterns

- Given a collection of interaction networks (belonging to different species), find sub-networks that are common to an interesting subset of these networks.
  - A sub-network is a group of interactions inducing a single network (connected)
  - Frequency: The number of networks that contain a sub-network, is a coarse measure of statistical significance
- Computational challenges
  - How to relate molecules in different contexts/ organisms?
  - Requires solution of the intractable subgraph isomorphism problem
  - Must be scalable to potentially large number of networks
  - Networks are large (in the range of $10K$ edges)

Relating Proteins in Different Species

- Ortholog Databases
  - PPI networks: COG, Homologene, Pfam, ADDA
  - Metabolic pathways: Enzyme nomenclature
  - Reliable, but conservative
  - Domain families rely on domain information, but the underlying domains for most interactions are unknown $\Rightarrow$ Multiple node labels
- Sequence Clustering
  - Cluster protein sequences and label proteins according to this clustering
  - Flexible, but expensive and noisy
- Labels may span a large range of functional relationships, from protein families to ortholog groups
  - Without loss of generality, we call identically labeled proteins as orthologs

Problem Statement

- Given a set of proteins V, a set of interactions E, and a many-to-many mapping from V to a set of ortholog groups $\mathcal{L} = \{l_1, l_2, ..., l_n\}$, the corresponding interaction network is a labeled graph $G = (V, E, \mathcal{L})$.
  - $v \in V(G)$ is associated with a set of ortholog groups $L(v) \subseteq \mathcal{L}$.
  - $uv \in E(G)$ represents an interaction between u and v.

- S is a sub-network of G, *i.e.*, $S \sqsubseteq G$ if there is an injective mapping $\phi : V(S) \rightarrow V(G)$ such that for all $v \in V(S)$, $L(v) \subseteq L(\phi(v))$ and for all $uv \in E(S)$, $\phi(u)\phi(v) \in E(G)$.

Computational Problem

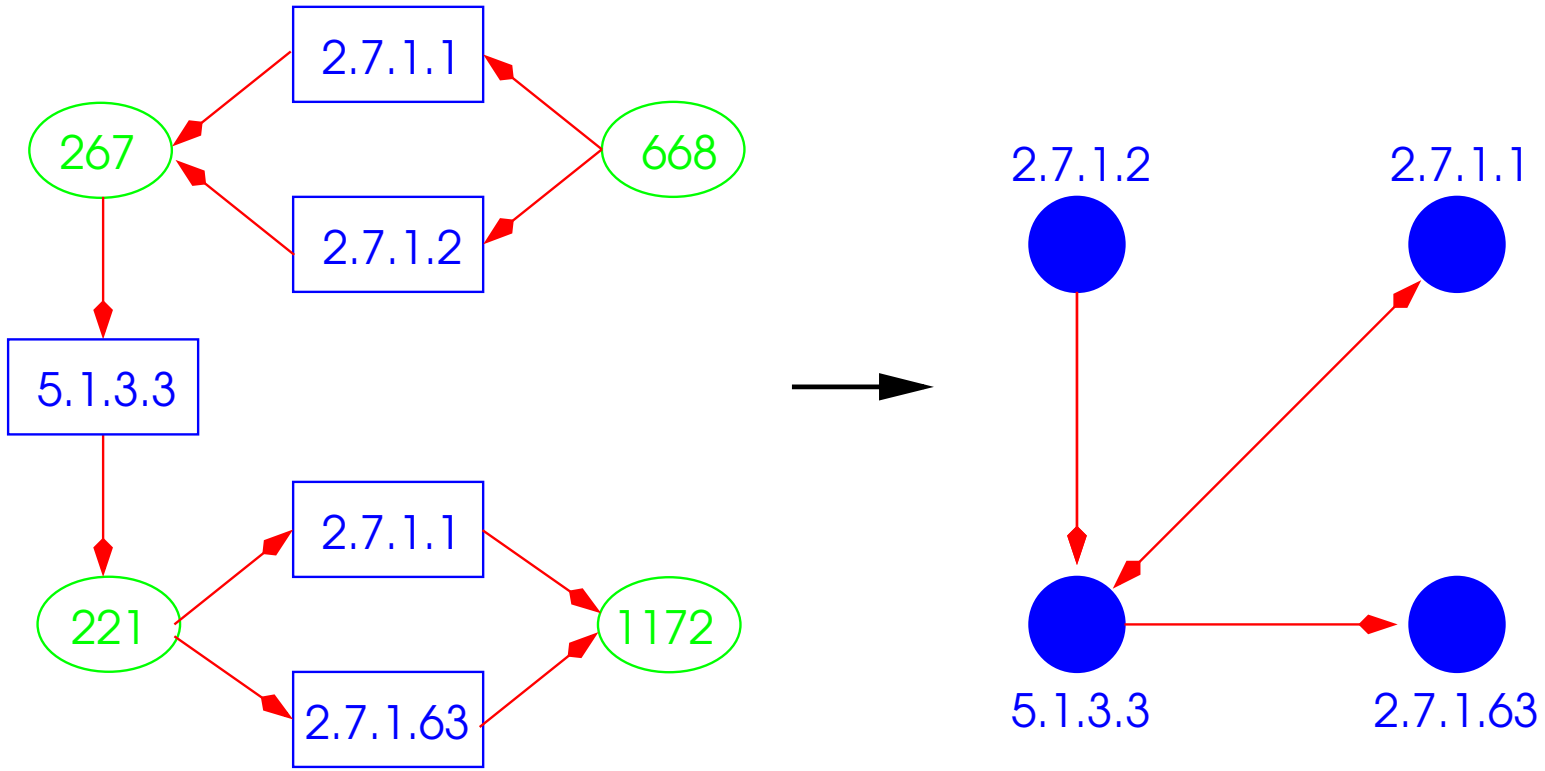
- Conserved sub-network discovery
  - Instance: A set of interaction networks $\mathcal{G} = \{G_1 = (V_1, E_1, \mathcal{L}), G_2 = (V_2, E_2, \mathcal{L}), ..., G_m = (V_m, E_m, \mathcal{L})\}$, each belonging to a different organism, and a frequency threshold σ^*.
  - Problem: Let $H(S) = \{G_i : S \sqsubseteq G_i\}$ be the occurrence set of graph S. Find all connected subgraphs S such that $|H(S)| \geq \sigma^*$, *i.e.*, S is a frequent subgraph in $\mathcal{G}$ and for all $S' \sqsupset S$, $H(S) \neq H(S')$, *i.e.*, S is maximal.

Algorithmic Insight: Ortholog Contraction

- Contract orthologous nodes into a single node
- No subgraph isomorphism
  - Graphs are uniquely identified by their edge sets
- Key observation: Frequent sub-networks are preserved $\Rightarrow$ No information loss
  - Sub-networks that are frequent in general graphs are also frequent in their ortholog-contracted representation
  - Ortholog contraction is a powerful pruning heuristic
- Discovered frequent sub-networks are still biologically interpretable!
  - Interaction between proteins becomes interaction between ortholog groups

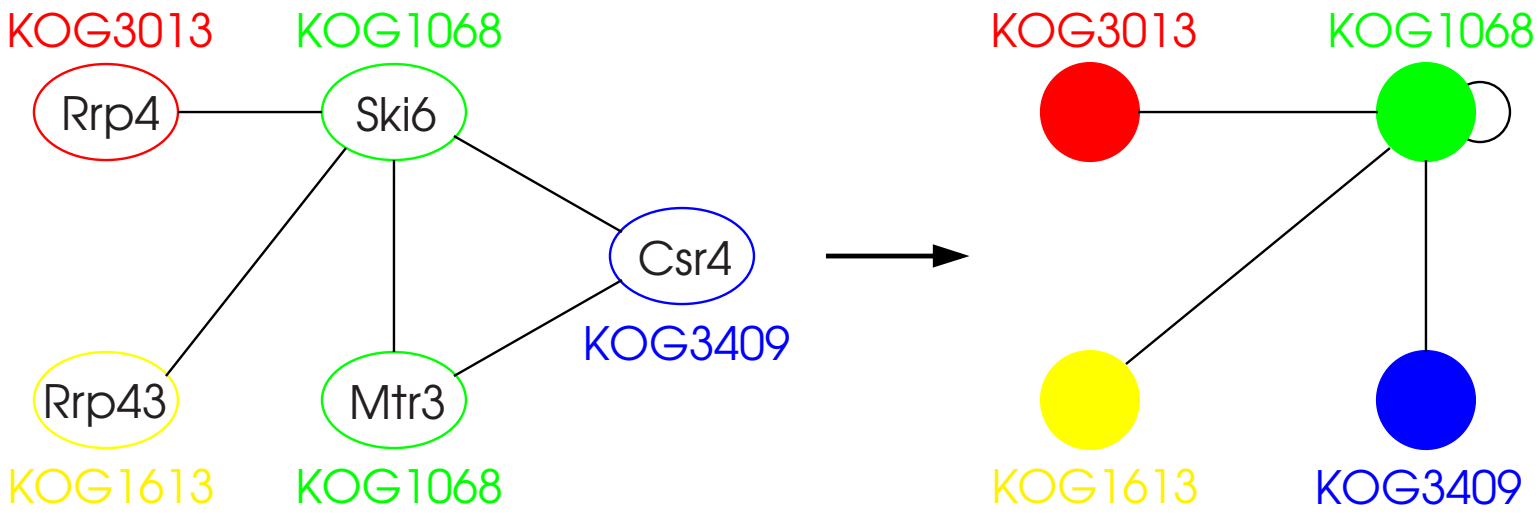
Ortholog Contraction in Metabolic Pathways

- Directed hypergraph → uniquely-labeled directed graph
  - Nodes represent enzymes
  - Global labeling by enzyme nomenclature (EC numbers)
  - A directed edge from one enzyme to the other implies that the second consumes a product of the first



Ortholog Contraction in PPI Networks

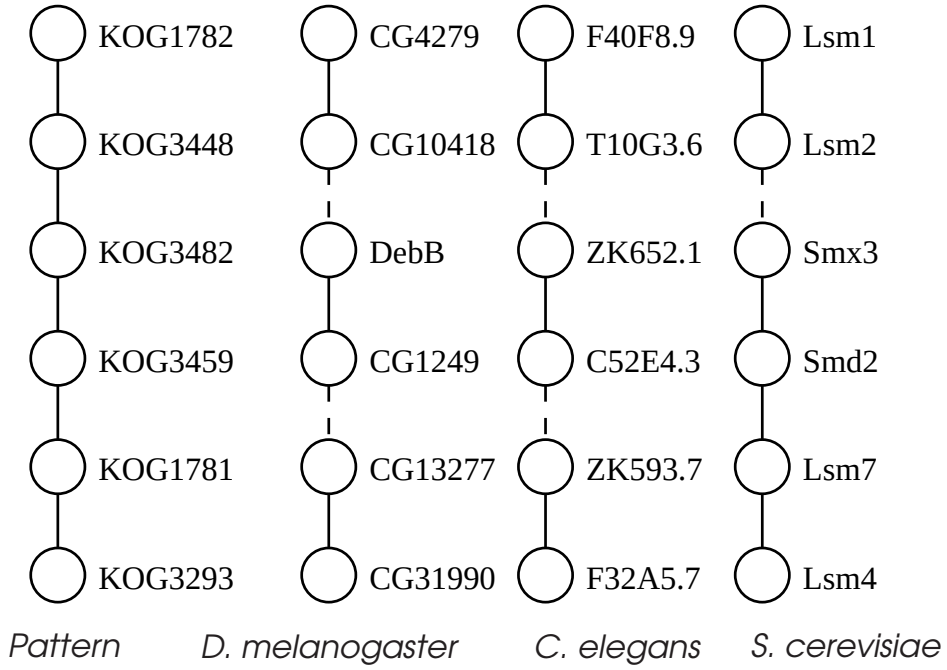
- Interaction between proteins → Interaction between ortholog groups or protein families



Results: Analyzing PPI Networks

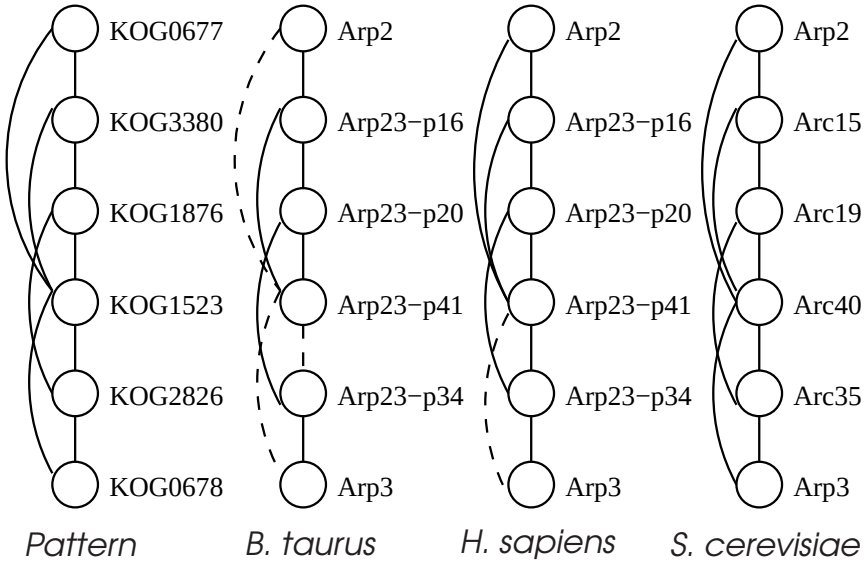
- PPI networks for 9 eukaryotic organisms derived from BIND and DIP
  - *A. thaliania, O. sativa, S. cerevisiae, C. elegans, D. melanogaster, H. sapiens, B. taurus, M. musculus, R. norvegicus*
  - # of proteins ranges from 288 (*Arabidopsis*) to 8577 (*fruit fly*)
  - # of interactions ranges from 340 (*rice*) to 28829 (*fruit fly*)
- Ortholog contraction
  - Group proteins according to existing COG ortholog clusters
  - Merge Homologene groups into COG clusters
  - Cluster remaining proteins via BLASTCLUST
  - Ortholog-contracted *fruit fly* network contains 11088 interactions between 2849 ortholog groups
- MULE is available at `http://www.cs.purdue.edu/pdsl/`

Conserved Protein Interaction Patterns



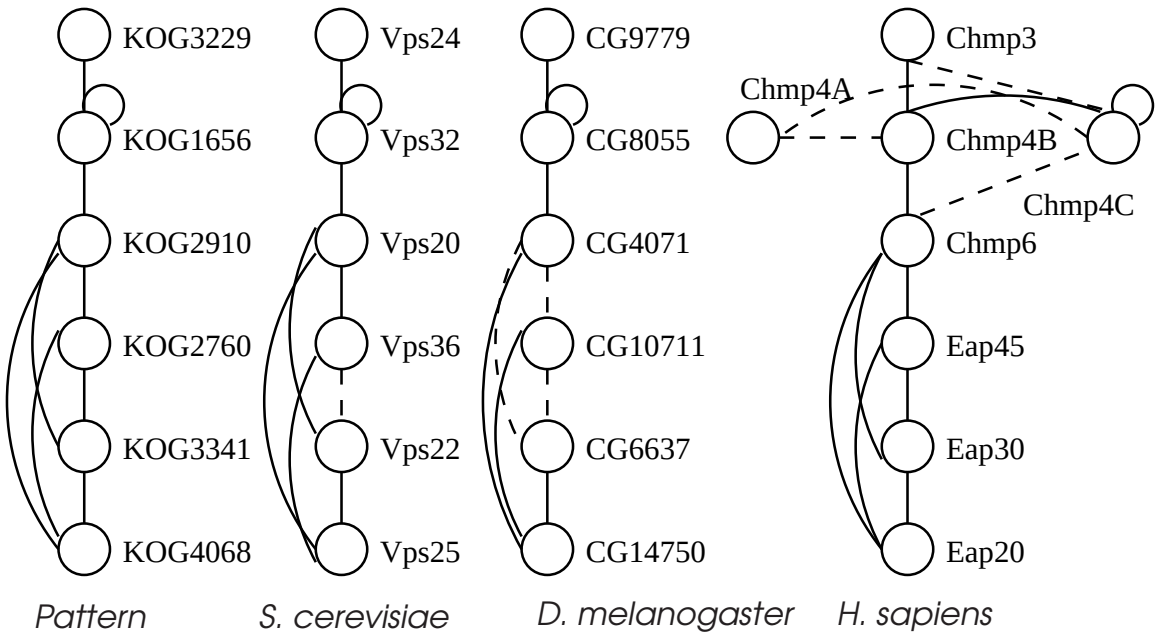
Small nuclear ribonucleoprotein complex ($p < 2e - 43$)

Conserved Protein Interaction Patterns



Actin-related protein Arp2/3 complex ($p < 9e - 11$)

Conserved Protein Interaction Patterns



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

Discussion

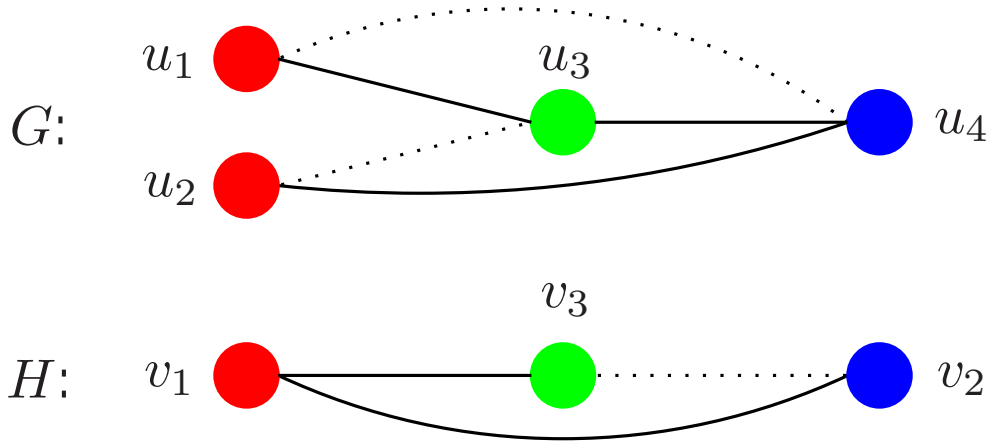
- Ortholog contraction is fast & scalable
  - Graph cartesian product based methods (Sharan et al., *PNAS*, 2004), (Koyutürk et al., *RECOMB*, 2005) create m^n product nodes for an ortholog group that has m proteins in each of n organisms
  - Ortholog contraction represents the same group with only n contracted nodes
  - Isomorphism-based graph analysis algorithms do not scale to large networks
- Ortholog contraction implicitly accounts for noise by eliminating false positives by thresholding frequency, and false negatives by contraction
- Key Open Problems: (i) Frequency is not significance (ii) How do we compute optimal ortholog groups?

Alignment of PPI Networks

- Given two PPI networks that belong to two different organisms, identify sub-networks that are similar to each other
  - Biological implications
  - Mathematical modeling
- Existing algorithms
  - PathBLAST aligns pathways (linear chains) to simplify the problem while maintaining biological meaning (Kelley et al., *PNAS*, 2004)
  - NetworkBLAST compares conserved complex model with null model to identify significantly conserved subnets (Sharan et al., *J. Comp. Biol.*, 2005)
- Our approach:
  - Guided by models of evolution
  - Scores evolutionary events
  - Identifies sets of proteins that induce high-scoring sub-network pairs

Match, Mismatch, and Duplication

- Evolutionary events as graph-theoretic concepts
  - A match $\in \mathcal{M}$ corresponds to two pairs of homolog proteins from each organism such that both pairs interact in both PPIs. A match is associated with score μ.
  - A mismatch $\in \mathcal{N}$ corresponds to two pairs of homolog proteins from each organism such that only one pair is interacting. A mismatch is associated with penalty ν.
  - A duplication $\in D$ corresponds to a pair of homolog proteins that are in the same organism. A duplication is associated with score δ.



Scoring Matches, Mismatches and Duplications

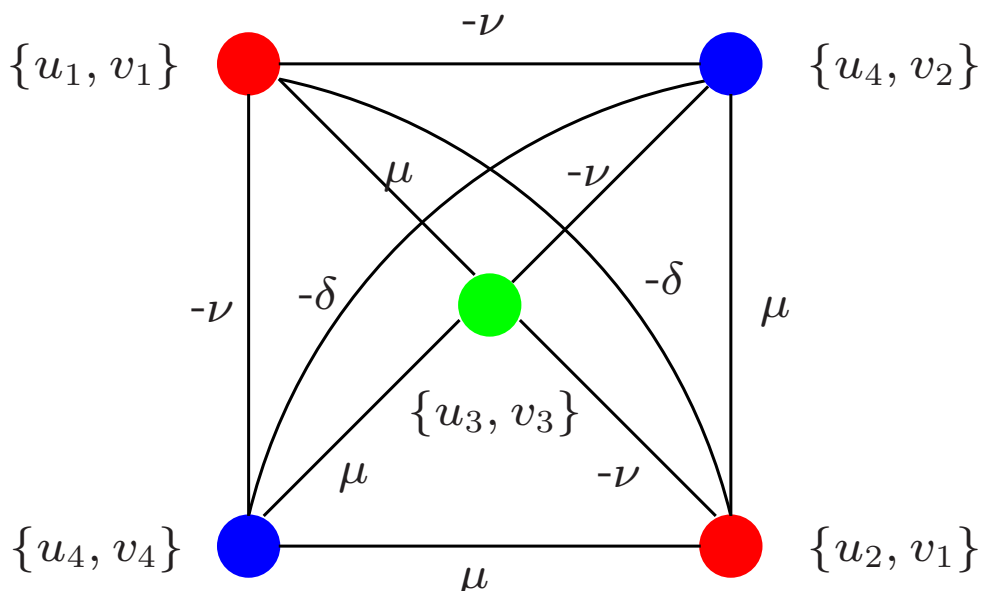
- Quantifying similarity between two proteins
  - Confidence in two proteins being orthologous
  - BLAST E-value: $S(u,v) = log_{10}\frac{p(u,v)}{p_{random}}$
  - Ortholog clustering: $S(u,v) = c(u)c(v)$
- Match score
  - $\mu(uu', vv') = \bar{\mu}\min\{S(u,v), S(u',v')\}$
- Mismatch penalty
  - $\nu(uu', vv') = \bar{\nu}\min\{S(u,v), S(u',v')\}$
- Duplication score
  - $\delta(u,u') = \bar{\delta}(\hat{\delta} - S(u,u'))$
  - $\hat{\delta}$ specifies threshold for sequence similarity to be considered functionally conserved

Pairwise Alignment of PPIs as an Optimization Problem

- Alignment score:
  $\sigma(\mathcal{A}(P)) = \sum_{M \in \mathcal{M}} \mu(M) - \sum_{N \in \mathcal{N}} \nu(N) + \sum_{D \in \mathcal{D}} \delta(D)$
  - Matches are rewarded for conservation of interactions
  - Duplications are rewarded/penalized for functional conservation/differentiation after split
  - Mismatches are penalized for functional divergence (what about experimental error?)
- Scores are functions of similarity between associated proteins
- Problem: Find all protein subset pairs with significant alignment score
  - High scoring protein subsets are likely to correspond to conserved modules
- A graph equivalent to BLAST

Weighted Alignment Graph

- $\mathbf{G(V, E)}$: $\mathbf{V}$ 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 $vv' \in V$, with weight $w(\mathbf{vv'}) = \mu(uv, u'v')$
  - mismatch edge if $uu' \in E$ and $vv' \notin V$ or vice versa, with weight $w(\mathbf{vv'}) = -\nu(uv, u'v')$
  - duplication edge if $S(u, u') > 0$ or $S(v, v') > 0$, with weight $w(\mathbf{vv'}) = \delta(u, u')$ or $w(\mathbf{vv'}) = \delta(v, v')$



Maximum Weight Induced Subgraph Problem

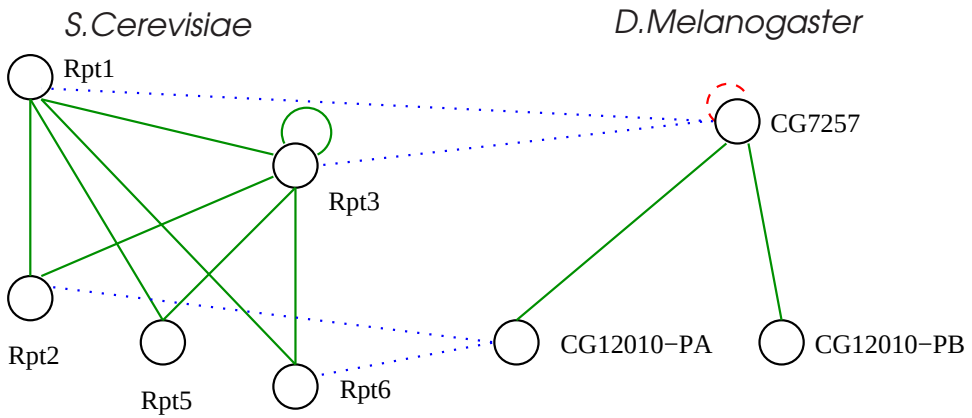
- Definition: (MAWISH)
  - Given graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and a constant ϵ, find $\tilde{\mathcal{V}} \in \mathcal{V}$ such that $\sum_{\mathbf{v},\mathbf{u}\in\tilde{\mathcal{V}}} w(\mathbf{vu}) \geq \epsilon$.
  - NP-complete by reduction from Maximum-Clique
- Theorem: (MAWISH $\equiv$ Pairwise alignment)
  - If $\tilde{\mathcal{V}}$ is a solution for the MAWISH problem on $\mathcal{G}(\mathcal{V}, \mathcal{E})$, then $P = \{\tilde{U}, \tilde{V}\}$ induces an alignment $\mathcal{A}(P)$ with $\sigma(\mathcal{A}) \geq \epsilon$, where $\tilde{\mathcal{V}} = \tilde{U} \times \tilde{V}$.
- Solution: Local graph expansion
  - Greedy graph growing + iterative refinement
  - Linear-time heuristic
- Source code available at `http://www.cs.purdue.edu/pdsl/`

Alignment of Yeast and Fruit Fly PPI Networks

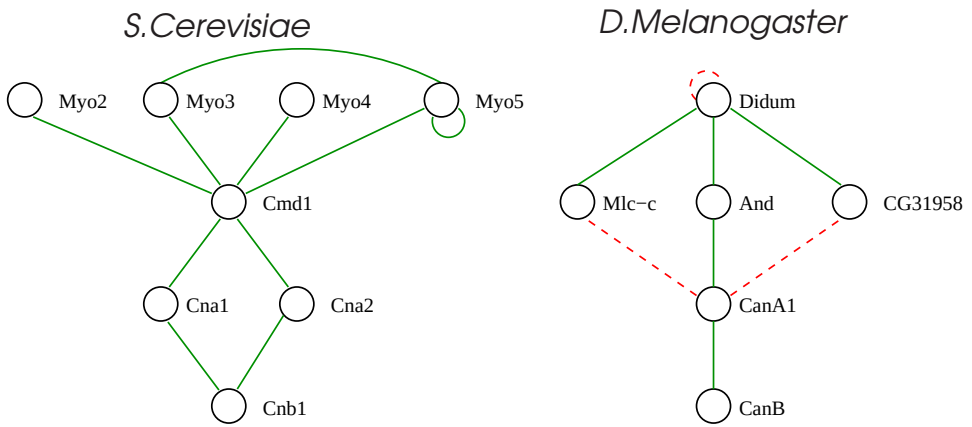
Rank	Score	z-score	# Proteins	# Matches	# Mismatches	# Dups.
1	15.97	6.6	18 (16, 5)	28	6	(4, 0)
	protein amino acid phosphorylation (69%)					
	JAK-STAT cascade (40%)					
2	13.93	3.7	13 (8, 7)	25	7	(3, 1)
	endocytosis (50%) / calcium-mediated signaling (50%)					
5	8.22	13.5	9 (5, 3)	19	11	(1, 0)
	invasive growth (sensu Saccharomyces) (100%)					
	oxygen and reactive oxygen species metabolism (33%)					
6	8.05	7.6	8 (5, 3)	12	2	(0, 1)
	ubiquitin-dependent protein catabolism (100%)					
	mitosis (67%)					
21	4.36	6.2	9 (5, 4)	18	13	(0, 5)
	cytokinesis (100%, 50%)					
30	3.76	39.6	6 (3, 5)	5	1	(0, 6)
	DNA replication initiation (100%, 80%)					

Subnets Conserved in Yeast and Fruit Fly

Proteosome regulatory particle subnet



Calcium-dependent stress-activated signaling pathway



Discussion

- Comparison to other approaches: NetworkBlast (Sharan et al., *PNAS*, 2005), NUKE (Novak et al., *Genome Informatics*, 2005)
  - Faster than NetworkBLAST, but provides less coverage
  - Comparable to NUKE depending on speed vs coverage trade-off
- Scores evolutionary events
  - Flexible, allows incorporation of different evolutionary models, experimental bases, target structures
  - Somewhat ad-hoc, what is a good weighting of scores?

Analytical Assessment of Statistical Significance

- What is the significance of a dense component in a network?
- What is the significance of a conserved component in multiple networks?
- Existing techniques
  - Mostly computational (*e.g.*, Monte-Carlo simulations)
  - Compute probability that the pattern exists rather than a pattern with the property (*e.g.*, size, density) exists
  - Overestimation of significance

Random Graph Models

- Interaction networks generally exhibit power-law property (or exponential, geometric, etc.)
- Analysis simplified through independence assumption (Itzkovitz et al., *Physical Review*, 2003)
- Independence assumption may cause problems for networks with arbitrary degree distribution
- $P(uv \in E) = d_u d_v / |E|$, where d_u is expected degree of u, but generally $d_{\max}^2 > |E|$ for PPI networks
- Analytical techniques based on simplified models (Koyutürk, Grama, Szpankowski, *RECOMB*, 2006)
  - Rigorous analysis on $G(n, p)$ model
  - Extension to piecewise $G(n, p)$ to capture network characteristics more accurately

Significance of Dense Subgraphs

- A subnet of r proteins is said to be ρ-dense if $F(r) \geq \rho r^2$, where $F(r)$ is the number of interactions between these r proteins
- What is the expected size of the largest ρ-dense subgraph in a random graph?
  - Any ρ-dense subgraph with larger size is statistically significant!
- $G(n, p)$ model
  - n proteins, each interaction occurs with probability p
  - Simple enough to facilitate rigorous analysis
  - If we let $p = d_{\max}/n$, largest ρ-dense subgraph in $G(n, p)$ stochastically dominates that in a graph with arbitrary degree distribution
- Piecewise $G(n, p)$ model
  - Few proteins with many interacting partners, many proteins with few interacting partners
  - Captures the basic characteristics of PPI networks
  - Analysis of $G(n, p)$ model immediately generalized to this model

Largest Dense Subgraph

- Theorem: If G is a random graph with n nodes, where every edge exists with probability p, then

$$\lim_{n\to\infty} \frac{R_\rho}{\log n} = \frac{1}{\kappa(p,\rho)} \qquad (pr.), \tag{1}$$

where

$$\kappa(p,\rho) = \rho \log \frac{\rho}{p} + (1-\rho)\log \frac{1-\rho}{1-p}. \tag{2}$$

More precisely,

$$P(R_\rho \geq r_0) \leq O\left(\frac{\log n}{n^{1/\kappa(p,\rho)}}\right), \tag{3}$$

where

$$r_0 = \frac{\log n - \log\log n + \log \kappa(p,\rho)}{\kappa(p,\rho)} \tag{4}$$

for large n.

Piecewise $G(n,p)$ model

- The size of largest dense subgraph is still proportional to $\log n/\kappa$ with a constant factor depending on number of hubs

- Model:

$$P(uv \in E(G)) = \begin{cases} p_h & \text{if } u,v \in V_h \\ p_l & \text{if } u,v \in V_l \\ p_b & \text{if } u \in V_h, v \in V_l \text{ or } u \in V_l, v \in V_h \end{cases}$$

- Result:
  Let $n_h = |V_h|$. If $n_h = O(1)$, then $P(R_n(\rho) \geq r_1) \leq O\left(\frac{\log n}{n^{1/\kappa(p_l,\rho)}}\right)$, where

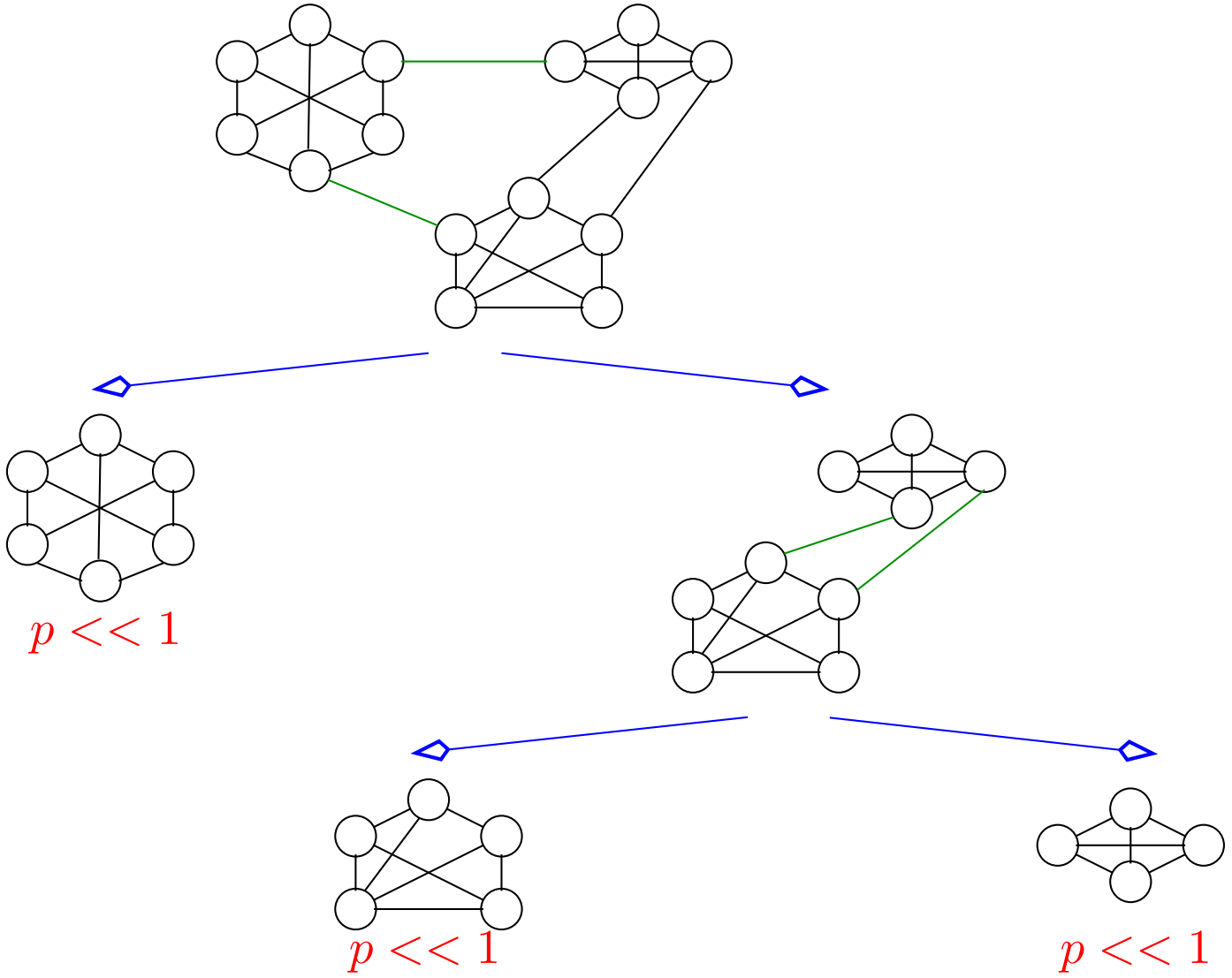
$$r_1 = \frac{\log n - \log\log n + 2n_h \log B + \log \kappa(p_l,\rho) - \log e + 1}{\kappa(p_l,\rho)}$$

  and $B = \frac{p_b q_l}{p_l} + q_b$, where $q_b = 1 - p_b$ and $q_l = 1 - p_l$.

Algorithms Based on Statistical Significance

- Identification of topological modules
- Use statistical significance as a stopping criterion for graph clustering heuristics
- HCS Algorithm (Hartuv & Shamir, *Inf. Proc. Let.*, 2000)
  - Find a minimum-cut bipartitioning of the network
  - If any of the parts is dense enough, record it as a dense cluster of proteins
  - Else, further partition them recursively
- SIDES: Use statistical significance to determine whether a subgraph is sufficiently dense
  - For given number of proteins and interactions between them, we can determine whether those proteins induce a significantly dense subnet

SIDES **Algorithm**



SIDES is available at `http://www.cs.purdue.edu/pdsl`

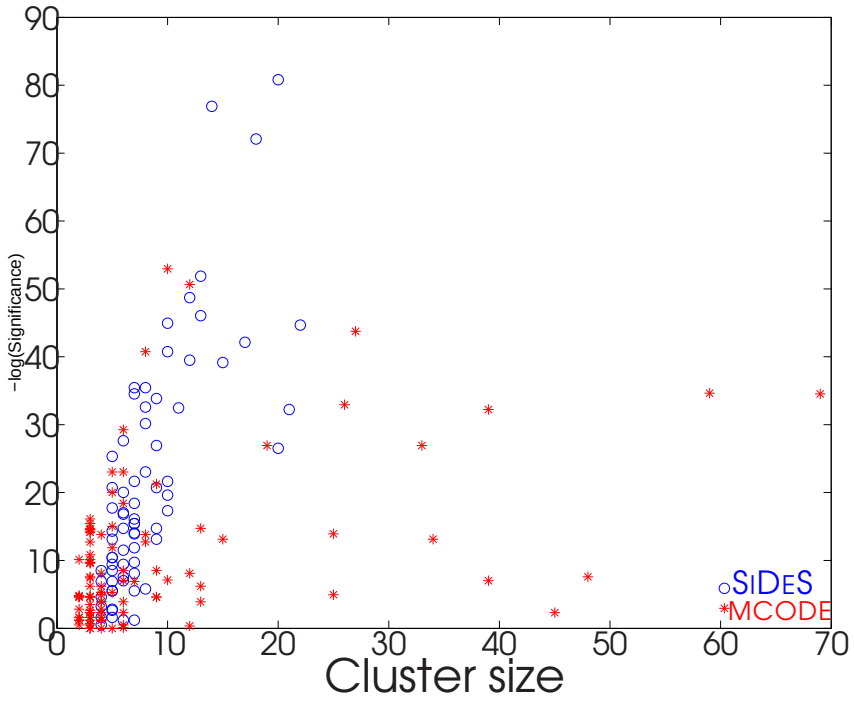
Performance of SIDES

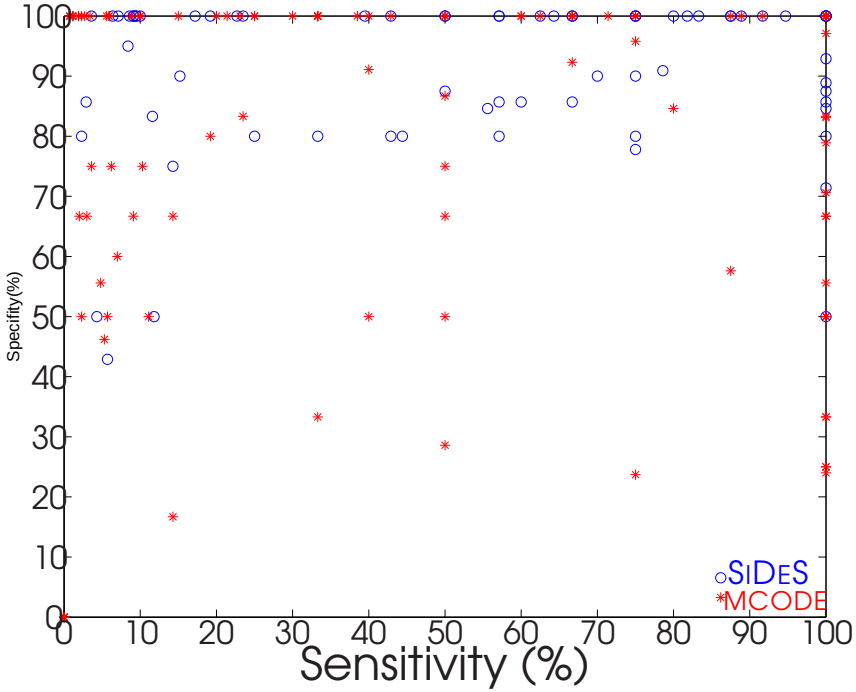
- Biological relevance of identified clusters is assessed with respect to Gene Ontology (GO)
  - Estimate the statistical significance of the enrichment of each GO term in the cluster
- Quality of the clusters with respect to GO annotations
  - Assume cluster C containing n_C genes is associated with term T that is attached to n_T genes and n_{CT} of genes in C are attached to T
  - specificity = $100 \times n_{CT}/n_C$
  - sensitivity = $100 \times n_{CT}/n_T$

	SIDES			MCODE		
	Min.	Max.	Avg.	Min.	Max.	Avg.
Specificity (%)	43.0	100.0	91.2	0.0	100.0	77.8
Sensitivity (%)	2.0	100.0	55.8	0.0	100.0	47.6

Comparison of SIDES with MCODE (Bader & Hogue, *BMC Bioinformatics*, 2003)

Performance of SIDES





Size vs Significance

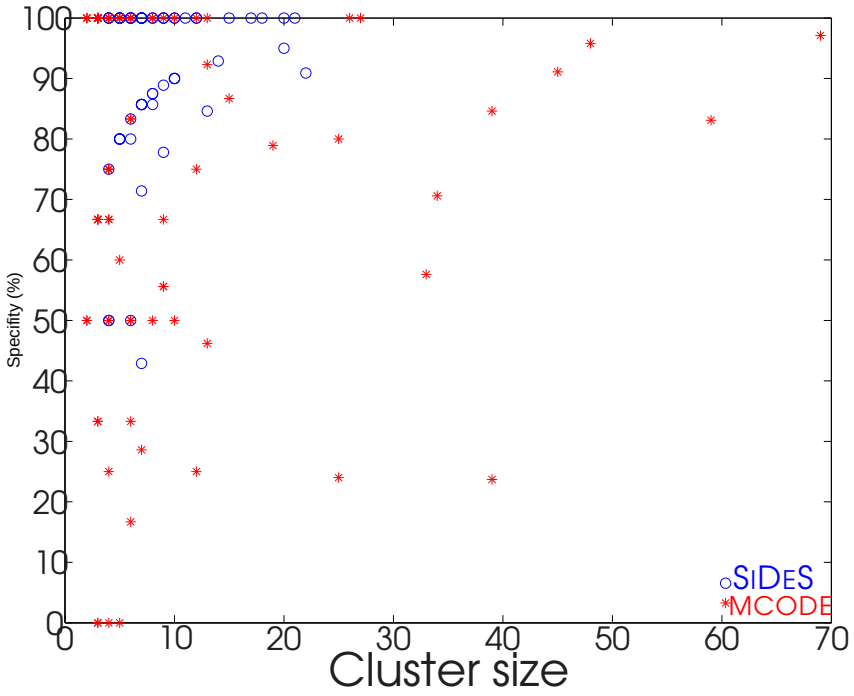
Sensitivity vs Specifity

Correlation

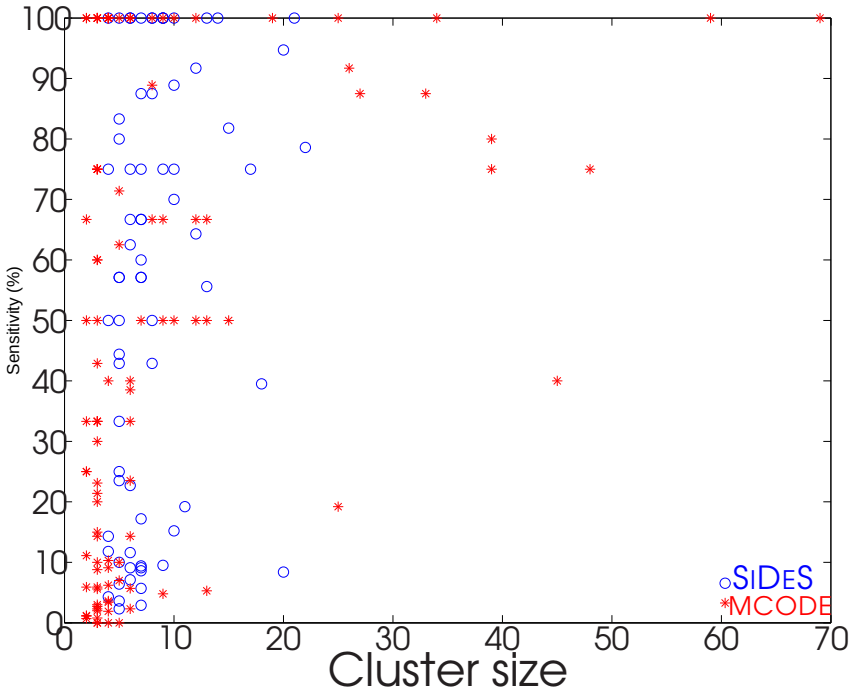
SIDES: 0.76

MCODE: 0.43

Performance of SIDES



Size vs Specificity



Size vs Sensitivity

Correlation

SIDES: 0.22
MCODE: -0.02

SIDES: 0.27
MCODE: 0.36

Statistical Significance as an Optimization Criteria

- Most algorithms satisfy queries and quantify the significance of the answer.
- Can we pose this as an optimization problem on the signficance – give me the most significant result for the query?
- We address this problem in the simple context of finding significant pathways.

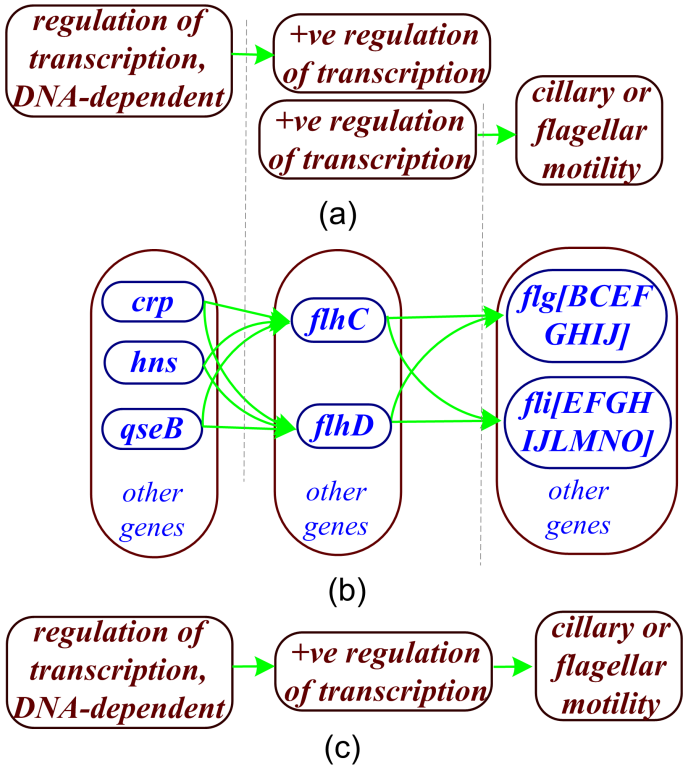
Finding Statistically Significant Functional Pathways

- Stack functional annotation of a molecule (gene) from an ontology on to the network.
- Generate a null-hypothesis for node functional annotation.
- Find all pathways with p-values lower than a threshold.

Finding Statistically Significant Functional Pathways

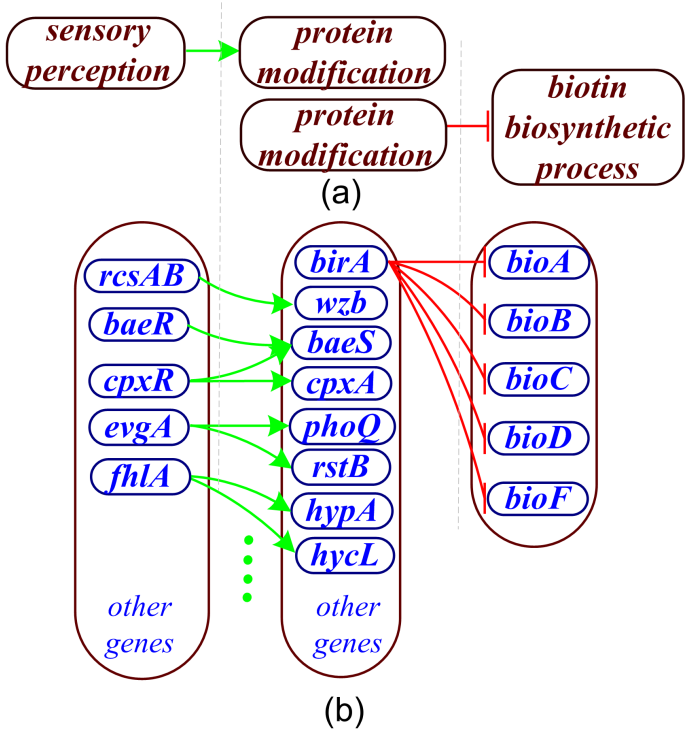
- Statistical significance is not monotonic in pathway space (cannot build longer pathways from known significant pathways)
- Statistical significance is not monotonic in ontology space (cannot build pathways at coarse levels in ontology and refine)
- The above statements are true for a broad class of measures of statistical significance we refer to as statistically interpretable.

Finding Statistically Significant Functional Pathways



From interactions between functional attributes to pathways of functional attributes: (a) statistically significant regulatory interactions between *DNA-dependent regulation of transcription*, *positive regulation of transcription*, and *cillary and flagellar motility*, (b) the two regulatory interactions are connected in the gene network as well, (c) these separate interactions may be combined into a pathway of functional attributes.

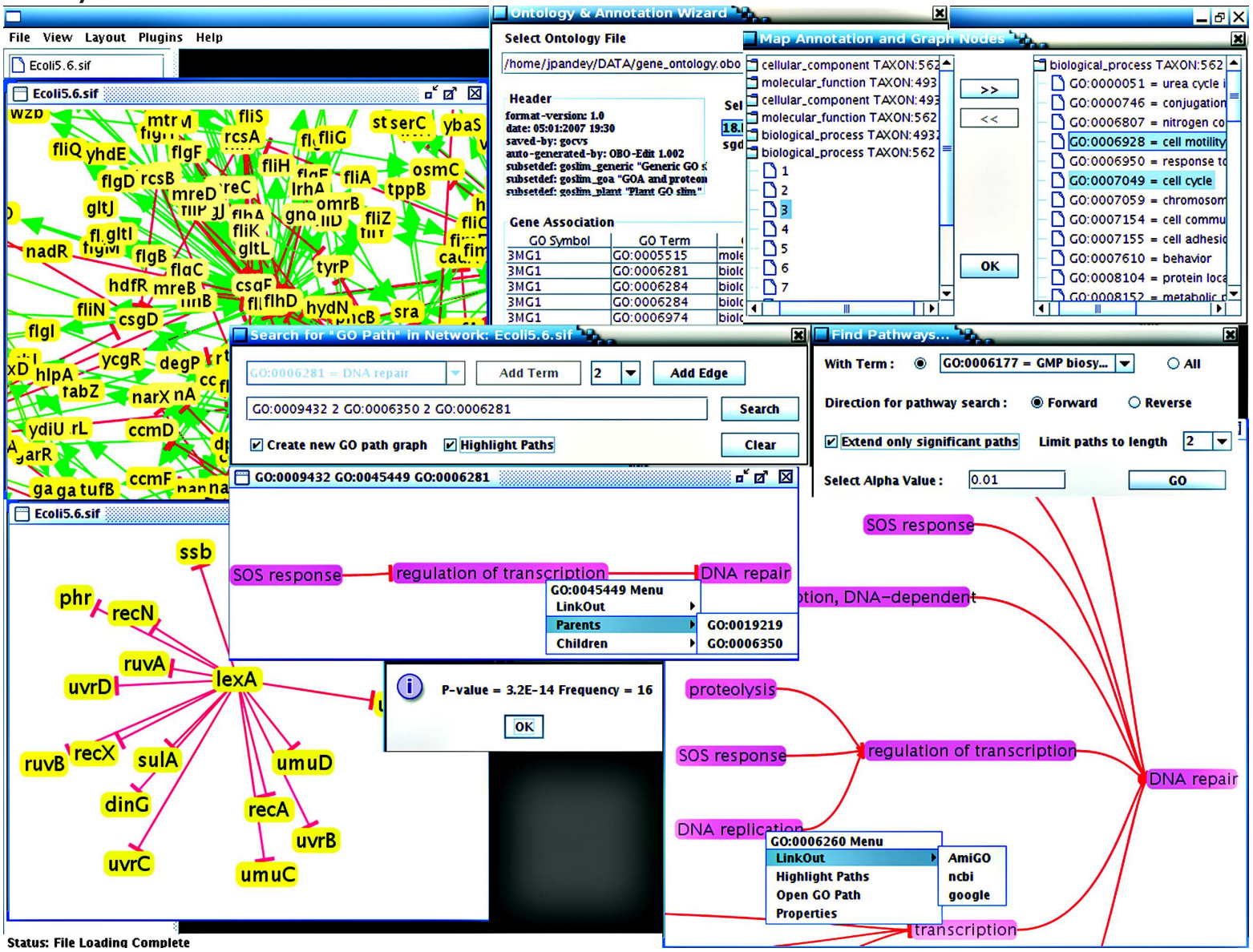
Finding Statistically Significant Functional Pathways



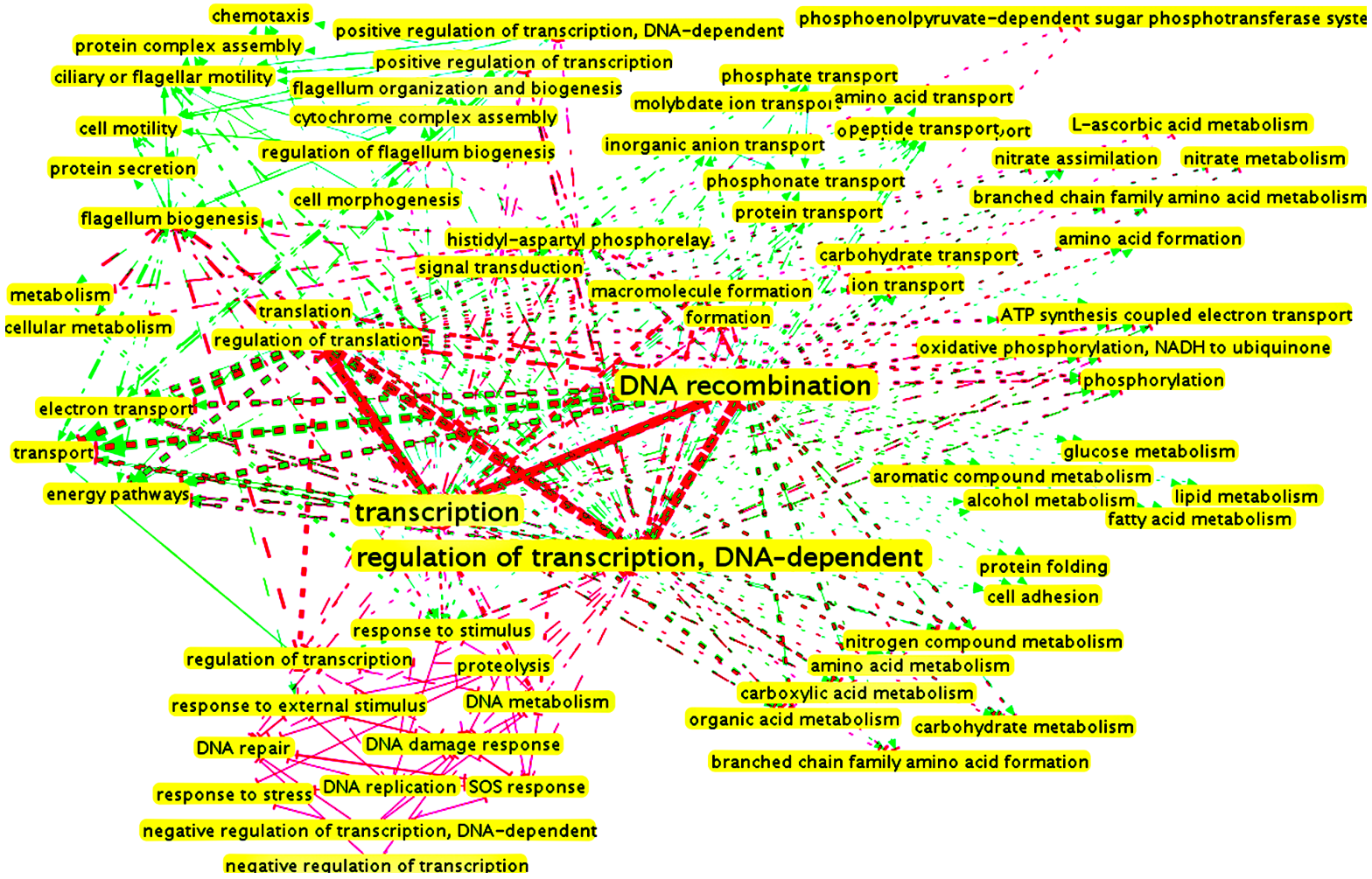
Pairwise assessment of interactions between functional attributes does not necessarily imply indirect paths: (a) regulation of *protein modification* by *sensory perception* is significantly overrepresented, as well as the regulation of *biotin biosynthetic process* by *protein modification*; (b) the two frequent interactions never go through the same gene.

Finding Statistically Significant Regulatory Pathways

Greedy Enumeration to the Rescue!

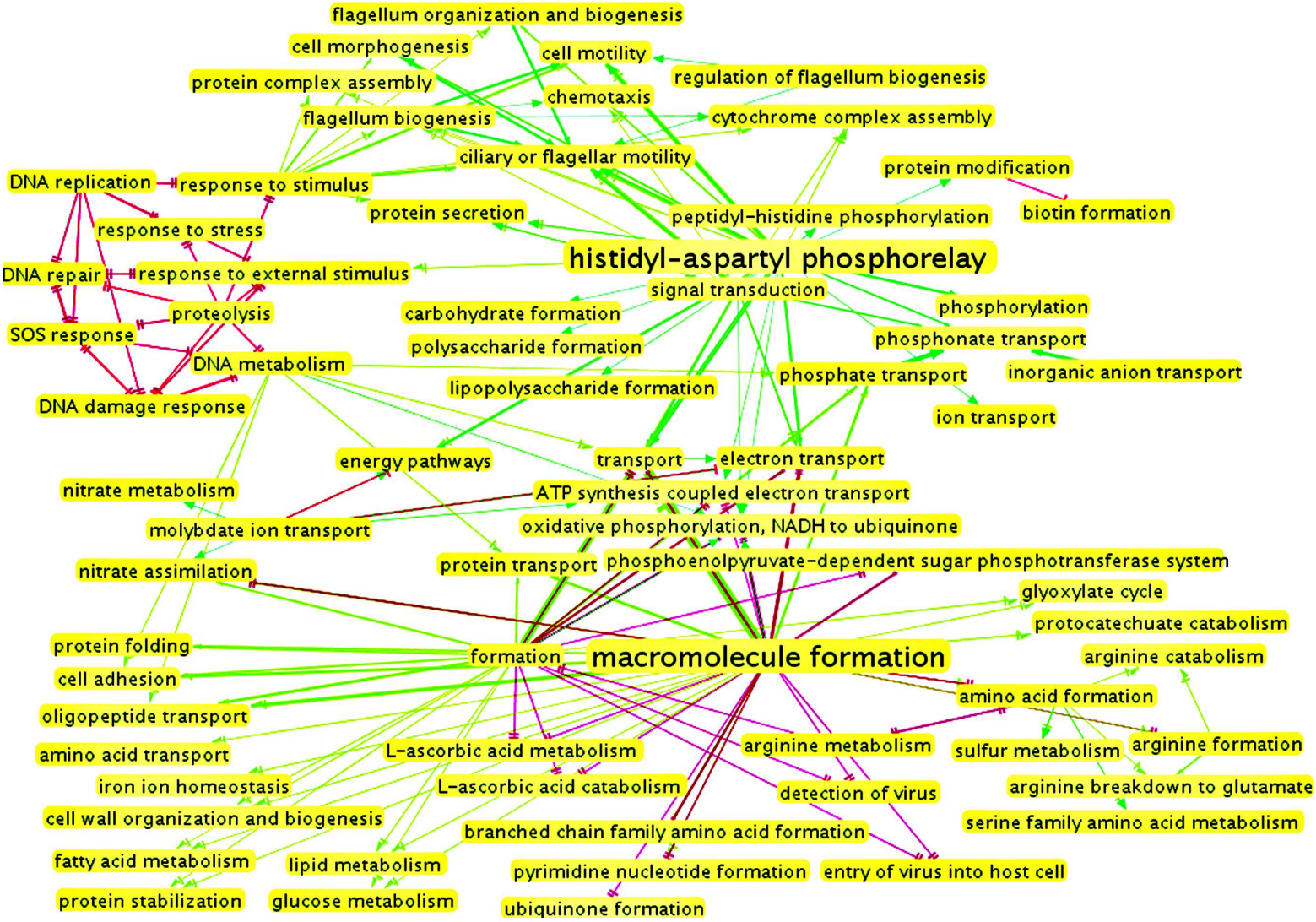


Finding Statistically Significant Regulatory Pathways



A global view of *E. coli* transcriptional network mapped to cellular processes described by GO.

Finding Statistically Significant Regulatory Pathways



A global view of *E. coli* transcriptional network obtained after short-circuting common mediator processes related to transcription and translation.

Finding Statistically Significant Regulatory Pathways

Frequency	p-value	Pathway
276	5E-94	metabolic process ⊣ flagellum biogenesis ⟶ transport
136	3.1E-71	regulation of translation ⊣ DNA recombination ⟶ transport
38	4.9E-47	response to stimulus ⊣ transcription ⟶ cell motility
36	6.6E-35	flagellum biogenesis ⟶ ciliary or flagellar motility
56	1.4E-24	regulation of translation ⊣ transcription ⟶ carboxylic acid metabolism
178	8.3E-21	signal transduction ⊣ transcription ⟶ transport
14	8.6E-20	phosphate transport ⟶ transcription ⟶ phosphonate transport
16	2E-16	SOS response ⊣ regulation of transcription ⊣ DNA repair
501	1.2E-13	regulation of transcription, DNA-dependent ⟶ transport
12	3.6E-10	proteolysis ⊣ regulation of transcription ⊣ response to external stimulus
15	3.8E-7	nitrate assimilation ⊣ cytochrome complex assembly
10	1.4E-6	cell morphogenesis ⊣ protein secretion
178	3.8E-4	transcription ⟶ carbohydrate metabolic process

Narada available at:

`http://www.cs.purdue.edu/homes/jpandey/narada/`

Outstanding Challenges

- Models, measures, algorithms and analysis!
- Data and data quality.
- Discriminative analysis.