Functional Characterization of Biological Networks

Jayesh Pandey

Department of Computer Science

December 3, 2009

Outline

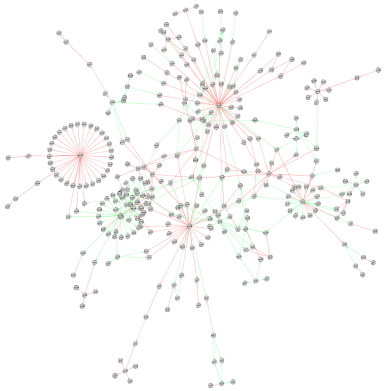
Outline

Molecular Interactions

- Regulation of molecular activity
  - Transcriptional regulation: Which genes will be expressed?
  - Post-transcriptional regulation & signaling: Phosphorylation, degradation, transport...
- Protein-protein interactions
  - High-throughput screening: Yeast Two-Hybrid, Affinity Purification
  - Noisy & incomplete
  - Nature, context, direction not known at a large scale
  - Small scale experiments are more reliable and informative

Modeling Molecular Interactions: Networks

- High level description of cellular organization
- Nodes represent cellular components
  - Protein, gene, enzyme, metabolite
- Edges represent interactions
  - Binding, regulation, modification, complex membership, substrate-product relationship

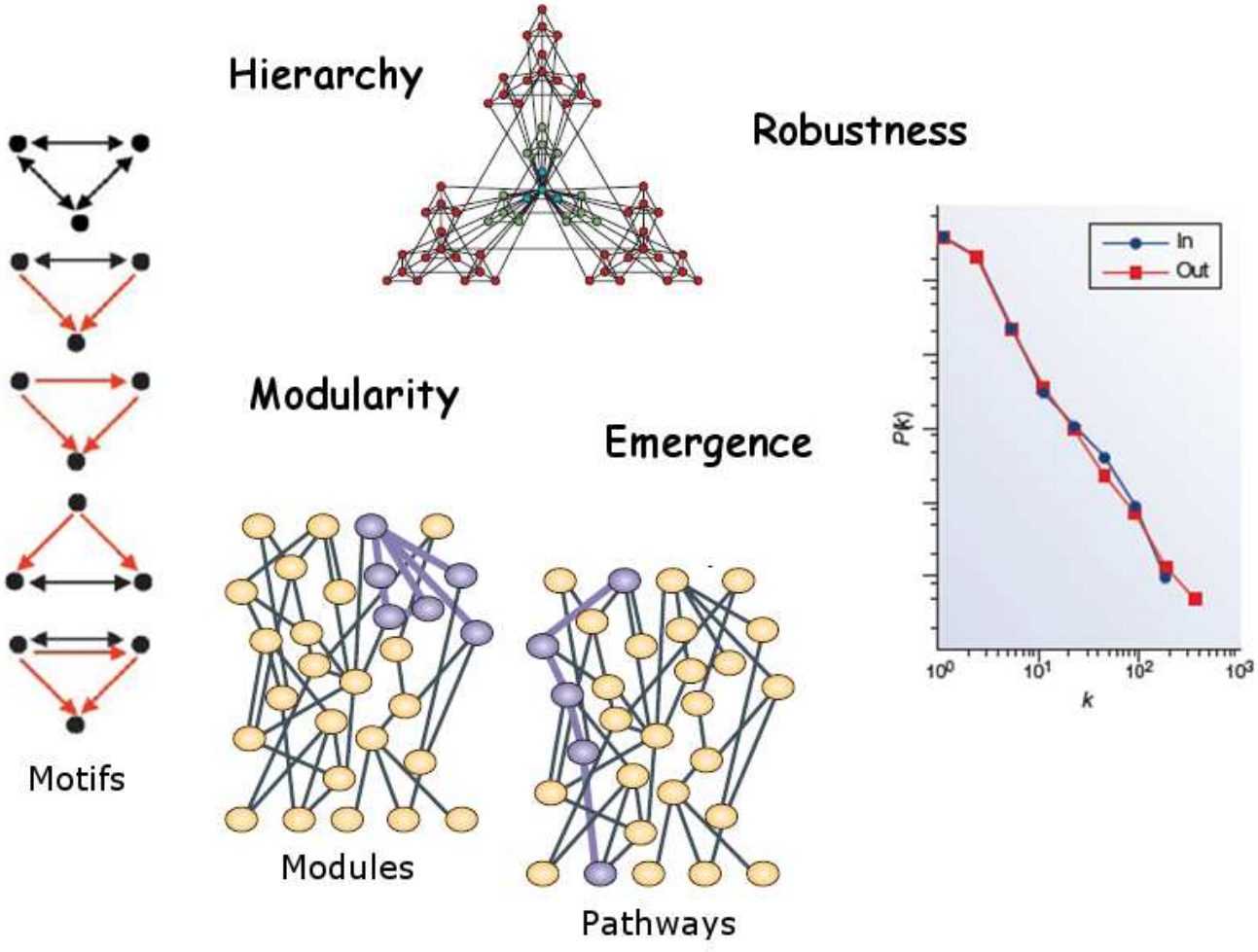


B.subtilis GRN

S.pombe PPI

Function & Topology in Molecular Networks

How does function relate to network topology?



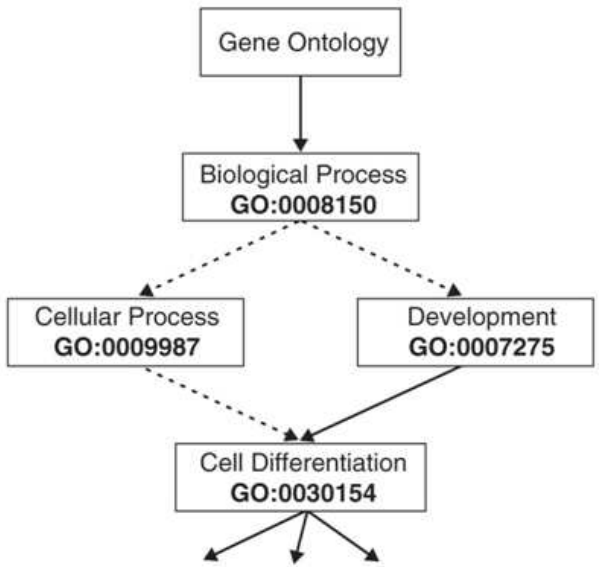
In this Talk

- Recurrent functional interaction patterns
  - Crosstalk between different processes
  - "Periodic table of systems biology"
- Functional coherence with respect to different types of interaction
  - What does proximity mean in domain-domain interaction networks?
  - Assessing functional similarity between two molecules
  - Development of incompleteness-aware approaches

Outline

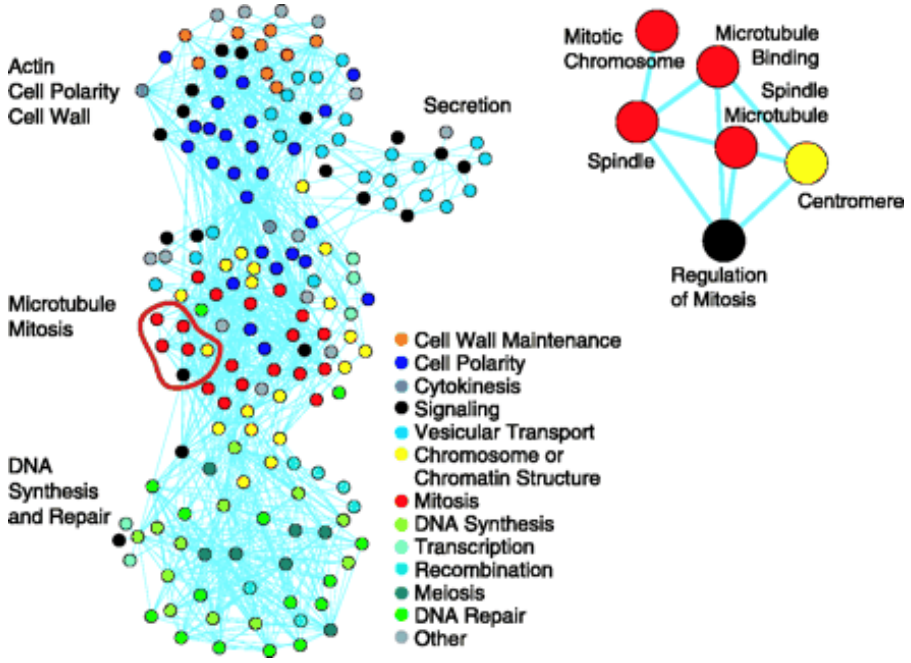
Molecular Annotation

- Significant progress on standardizing knowledge on biological function at the molecular level
  - Protein/domain families (COG, PFAM, ADDA)
- Molecular annotation provides a unified understanding of the underlying principles
- Gene Ontology
  - A controlled vocabulary of molecular functions, biological processes, and cellular components



From Molecules to Systems

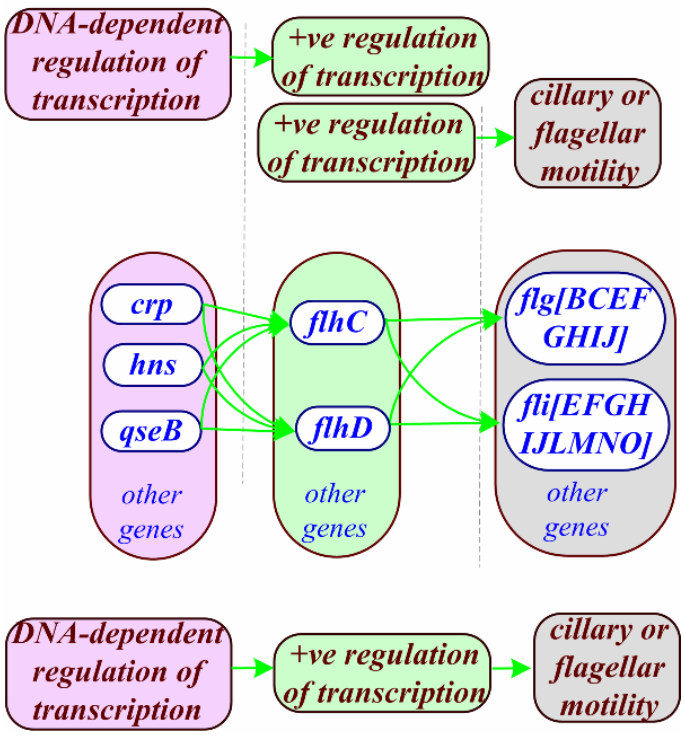
- Networks are species-specific
- Annotations are described at the molecular level
- Map networks from gene space to an abstract function space

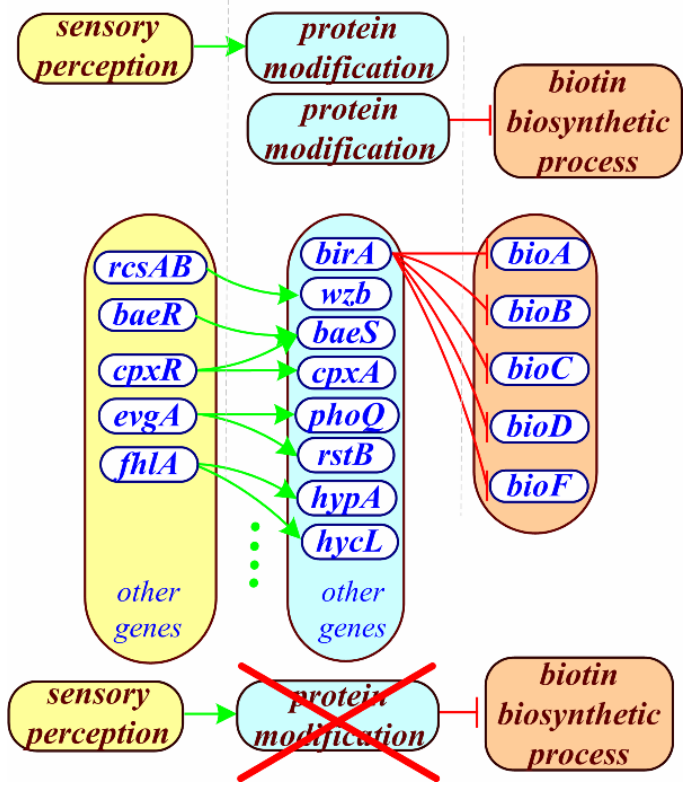


Network of GO terms based on significance of pairwise interactions in S. cerevisiae Synthetic Gene Array (SGA) network (Tong et al., Science, 2004)

Indirect Regulation

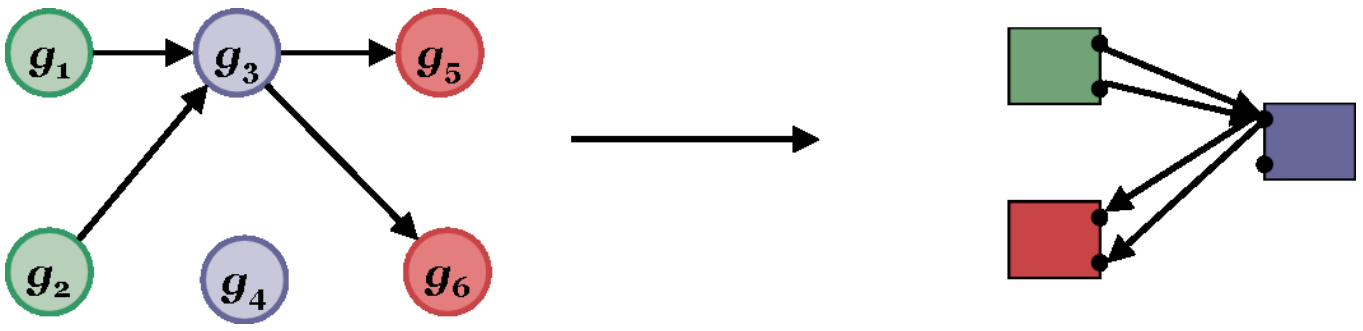
- Assessment of pairwise interactions is simple, but not adequate





Functional Attribute Network

- Multigraph model
  - A gene is associated with multiple functional attributes
  - A functional attribute is associated with multiple genes
  - Functional attributes are represented by nodes
  - Genes are represented by ports, reflecting context

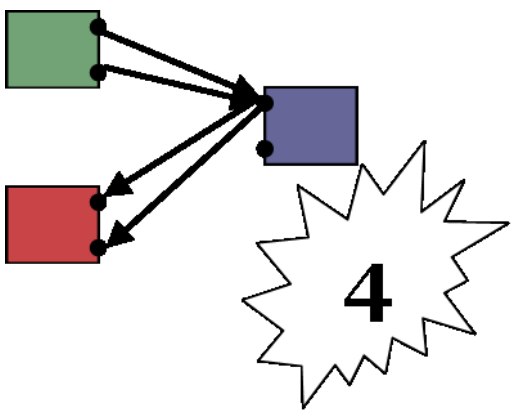


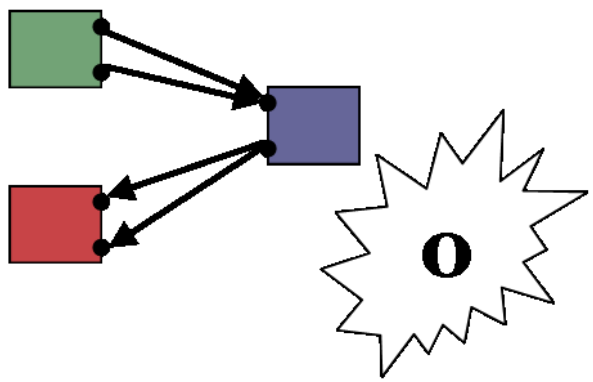
Gene Network

Functional Attribute Network

Frequency of a Multipath

- A pathway of functional attributes occurs in various contexts in the gene network
  - Multipath in the functional attribute network





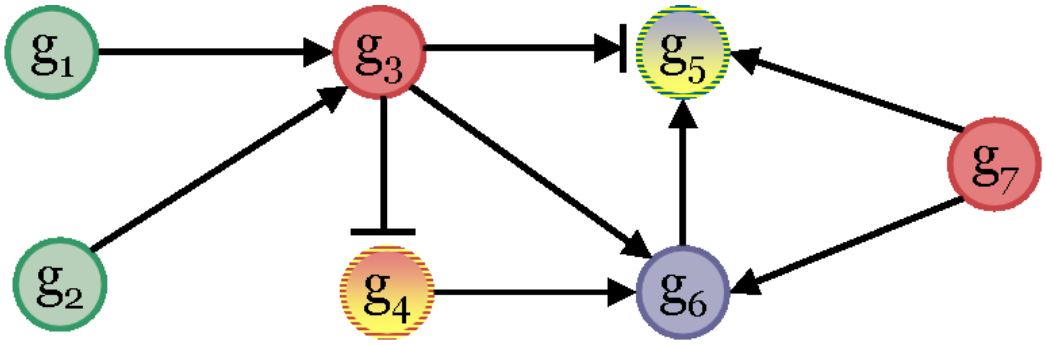
Frequency of Multipath  ?

Frequency *vs.* Statistical Significance

- We want to identify patterns with unusual frequency
  - These might correspond to modular pathways
- Frequency alone is not a good measure of statistical significance
  - The distribution of functional attributes among genes is not uniform
  - The degree distribution in the network is highly skewed
  - Pathways that contain common functional attributes have high frequency, but they are not necessarily interesting

Statistical Significance of a Pattern

- Emphasize modularity of pattern (Pandey *et al.*, *ISMB*, 2007)
  - Condition on frequency of building blocks
  - Evaluate the significance of the coupling of building blocks



$\varphi(\rightarrow\rightarrow) = \varphi(\rightarrow\rightarrow\dashv) = 4$

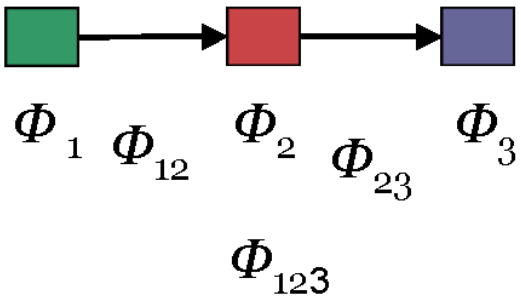
$\varphi(\rightarrow) = \varphi(\rightarrow\dashv) = 2 \qquad \varphi(\rightarrow) = 5$

$P(\rightarrow\rightarrow\dashv) < P(\rightarrow\rightarrow\rightarrow)$

Significance of a Pattern

- We denote each frequency random variable by ϕ, their observed value by φ

π_{123}:



- Significance of pattern π_{123} (p_{123}) is defined as $P(\phi_{123} \geq \varphi_{123} | \phi_{12} = \varphi_{12}, \phi_{23} = \varphi_{23}, \phi_1 = \varphi_1, \phi_2 = \varphi_2, \phi_3 = \varphi_3)$

Computing Significance

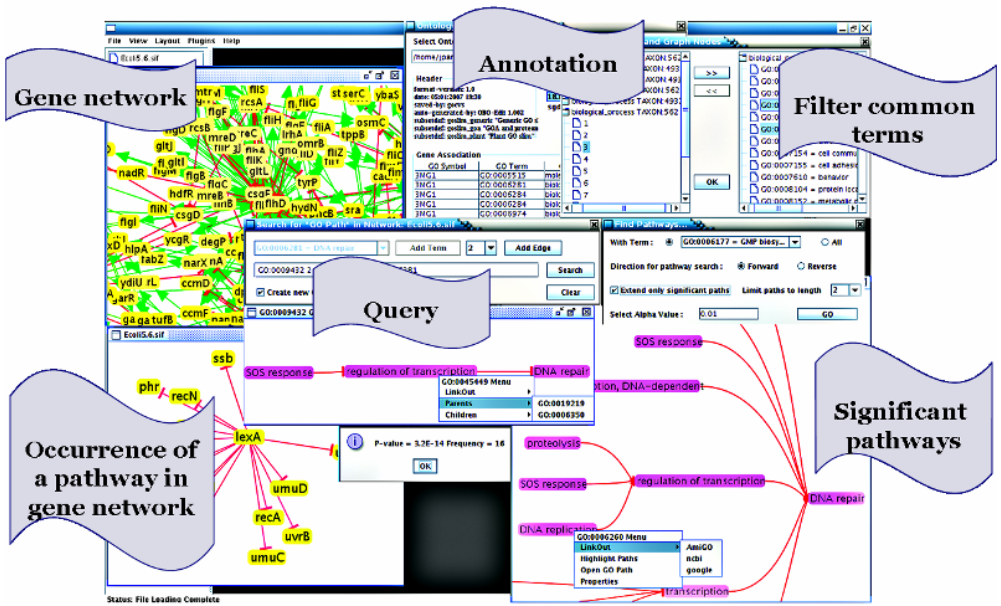
- Assume that interactions are independent
  - There are $\varphi_{12}\varphi_{23}$ possible pairs of π_{12} and π_{23} edges
  - The probability that a pair of π_{12} and π_{23} edges go through the same gene (corresponds to an occurrence of π_{123}) is $1/\varphi_2$
- The probability that at least φ_{123} of these pairs go through the same gene can be bounded by
  - $p_{123} \leq exp(\varphi_{12}\varphi_{23} H_q(t))$ where $q = 1/\varphi_2$ and $t = \varphi_{123}/\varphi_{12}\varphi_{23}$
  - $H_q(t) = t\log(q/t) + (1-t)\log((1-q)/(1-t))$ is divergence
  - Bonferroni-corrected for multiple testing (adjusted by $\prod_{j=1}^{k} |\cup_{g_\ell \in T_{i_j}} \mathcal{F}(g_\ell)|$)

Algorithmic issues

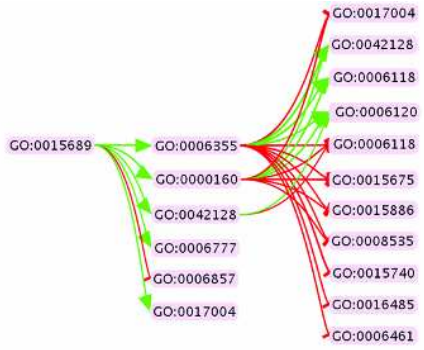
- Significance is not monotonic with respect to size
  - Need to enumerate all pathways?
- Strongly significant patterns
  - A pathway is strongly significant if all of its building blocks and their coupling are significant (defined recursively)
  - Allows pruning out the search space effectively
- Shortcircuiting common functional attributes
  - Transcription factors, DNA binding genes, etc. are responsible for mediating regulation
  - Shortcircuit these terms, consider regulatory effect of different processes on each other directly

NARADA

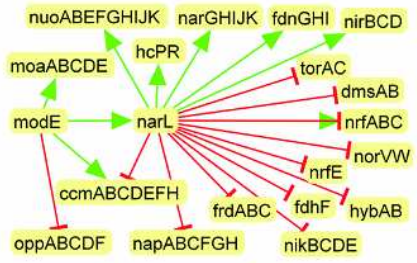
- A software for identification of significant pathways (Pandey *et al.*, *PSB*, 2008)
  - Given functional attribute *T*, find all significant pathways that originate (terminate) at *T*
  - User can explore back and forth between the gene network and the functional attribute network



An Example: Molybdate Ion Transport



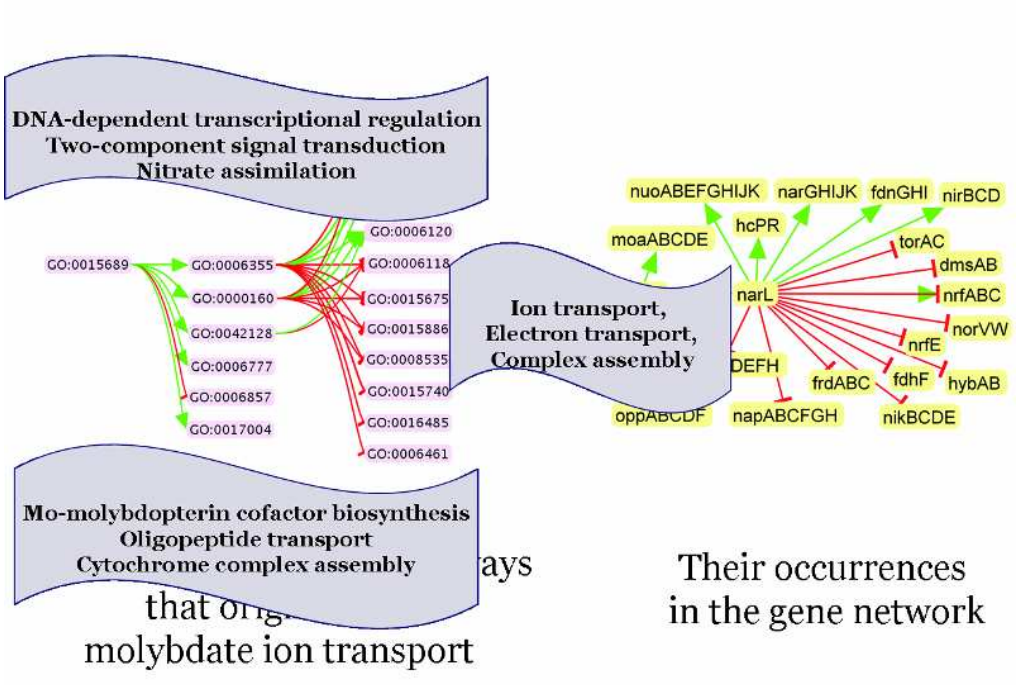
Significant regulatory pathways that originate at molybdate ion transport



Their occurrences in the gene network

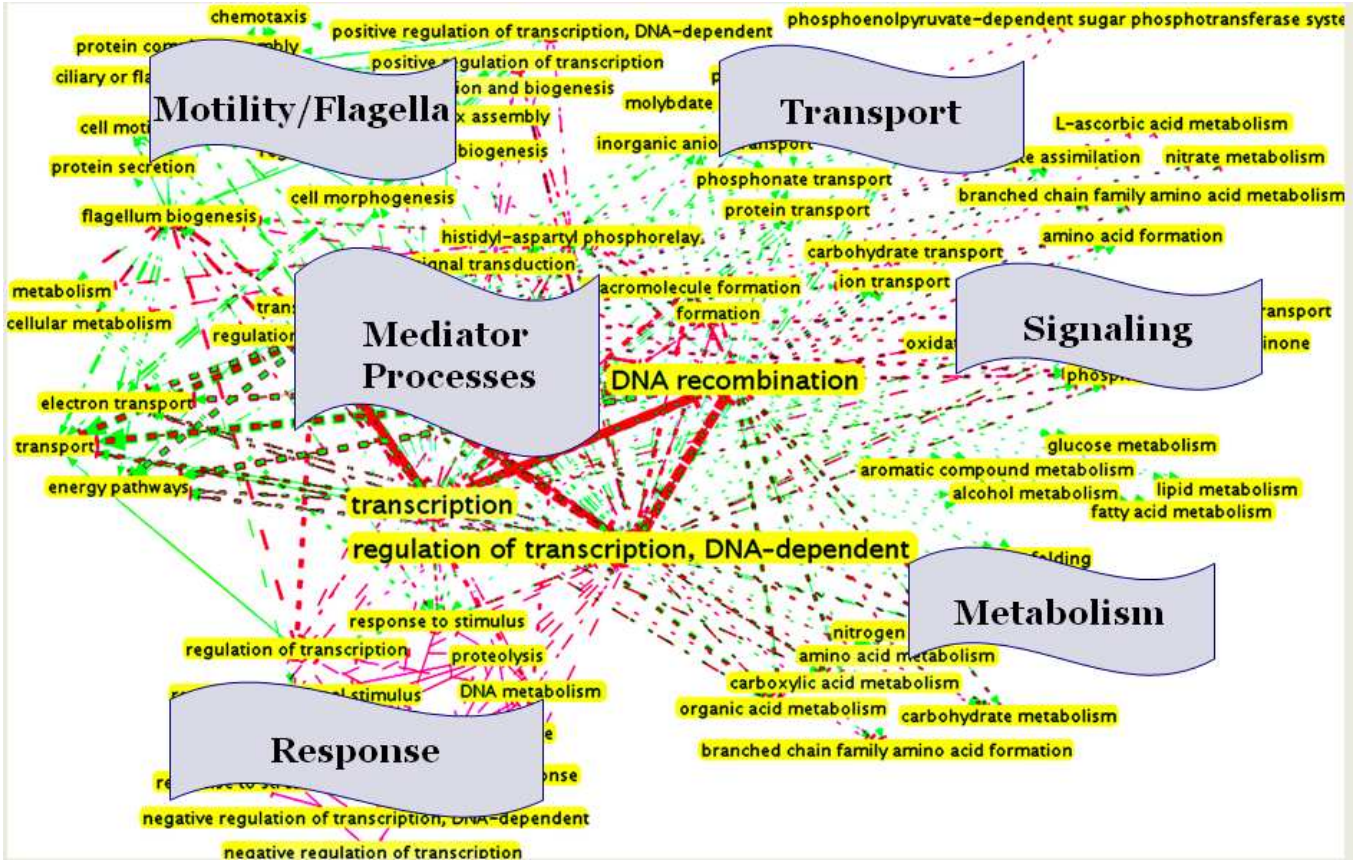
- modE regulates various processes directly
- It regulates various other processes indirectly
  - Regulation of these mediator processes is not significant on itself
  - NARADA captures modularity of indirect regulation!

An Example: Molybdate Ion Transport

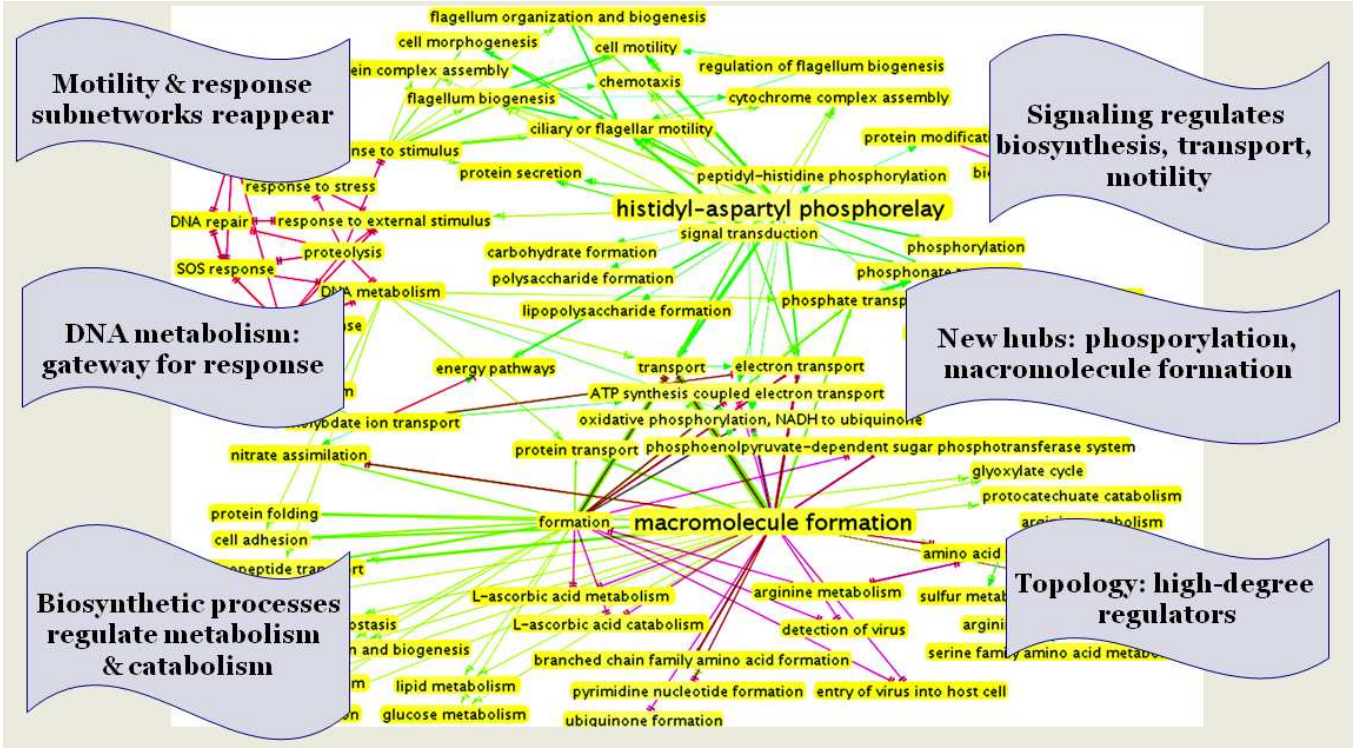


- modE regulates various processes directly
- It regulates various other processes indirectly
  - Regulation of these mediator processes is not significant on itself
  - NARADA captures modularity of indirect regulation!

Functional View of *E. coli* Regulatory Network



Short-Circuiting Mediator Processes



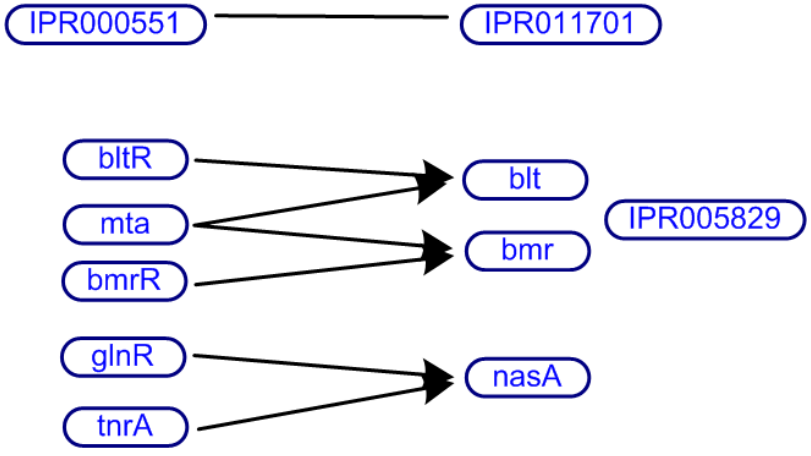
Significant Patterns in Bacteria

- We use NARADA to identify significant patterns in the transcriptional networks of two bacterial species
  - *E. coli*: 1502 genes, 3586 regulatory interactions (RegulonDB)
  - *B. subtilis*: 996 genes, 1381 regulatory interactions (DBTBS)

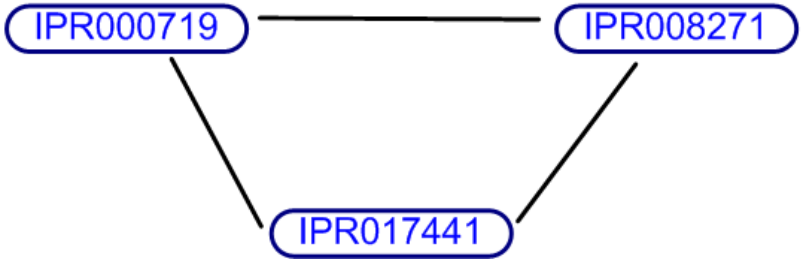
Significant patterns ($p < 0.01$)

Patterns	B. subtilis	E. coli	BS in EC	EC in BS
linear path	34	308	0	0
feedback	27	114	25	25
feedforward	77	659	77	86
sink hub	18	344	18	18
source hub	4907	8331	4270	4815

Domain Annotation Patterns



B. subtilis GRN

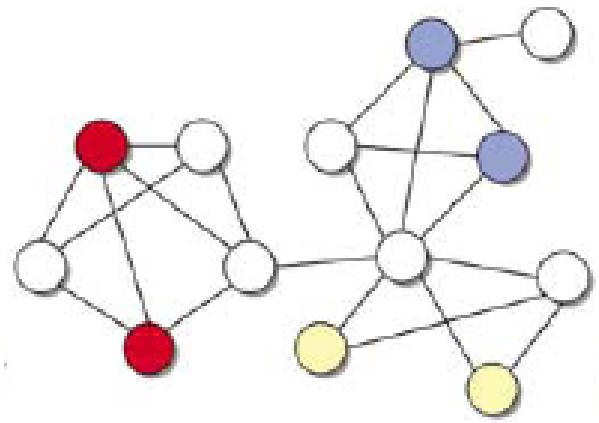


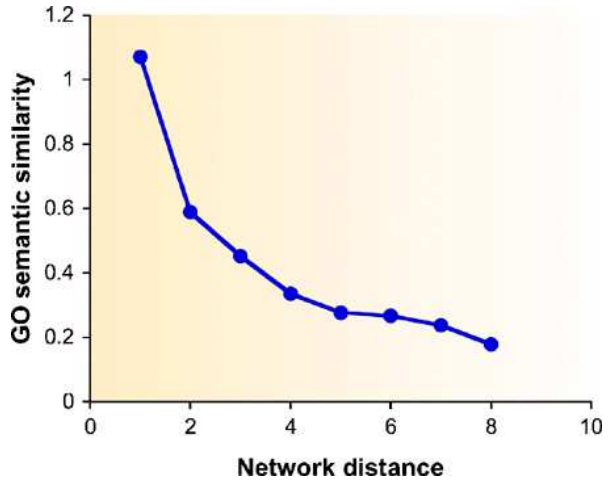
D. mela. PPI & S. cere. PPI

Outline

Functional Coherence in Networks

- Modularity manifests itself in terms of high connectivity in the network
- Functional association (similarity) is correlated with network proximity
- Protein-protein interaction (PPI) networks are used extensively for functional inference
- In PPI networks, functional coherence manifests itself in terms of network proximity
  - How about DDI "networks"?

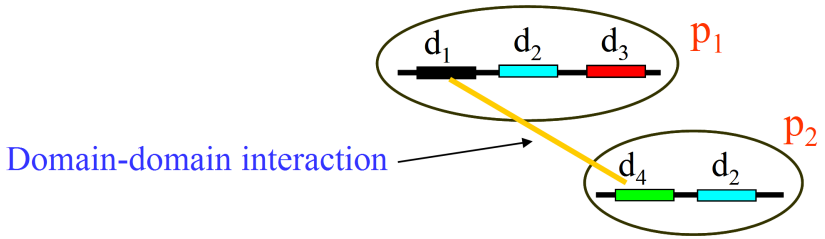




Sharan *et al.*, *MSB*, 2007

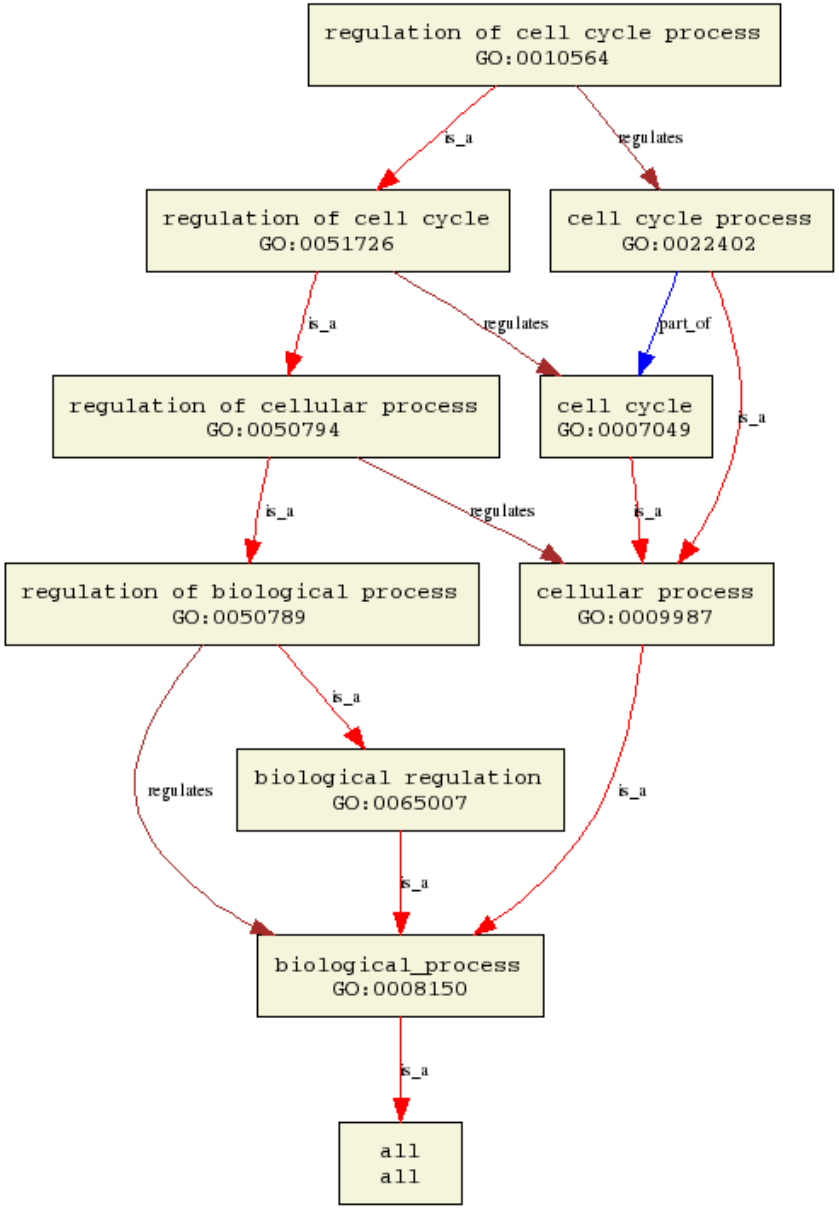
Domain-Domain Interactions

- Most proteins are composed of multiple domains
- Many domains are independent units reused in several related proteins
- Interactions between domains underlie observed protein-protein interactions
- Inferred by experimental and computational techniques



Assessing Functional Similarity

- Gene Ontology (GO) provides a hierarchical taxonomy of biological process, molecular function and cellular component
- Assessment of semantic similarity between concepts in a hierarchical taxonomy is well studied (Resnik, *IJCAI*, 1995)



Semantic Similarity of GO Terms

- Resnik's measure based on information content:

$$I(c) = -\log_2(|G_c|/|G_r|)$$

$$\delta_I(c_i, c_j) = \max_{c \in A_i \cap A_j} I(c)$$

  - G_c: Set of molecules that are associated with term c
  - r: Root term
  - A_i: Ancestors of term c_i in the hierarchy
  - $\lambda(c_i, c_j) = \text{argmax}_{c \in A_i \cap A_j} I(c)$: Lowest common ancestor of c_i and c_j

Functional Similarity of Molecules

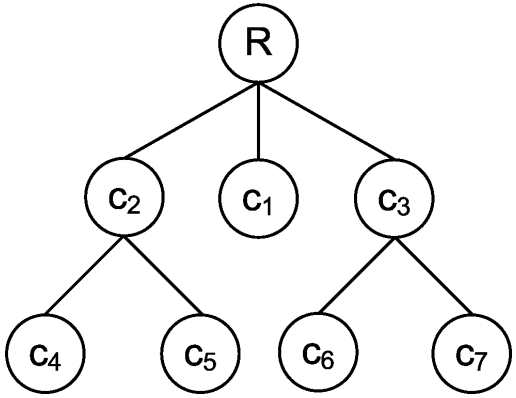
- Each molecule (protein or domain) is associated with multiple GO terms
- Available annotations are incomplete
- Domain annotations are often derived from protein annotations
- Is it possible to compare functional similarity between domains and functional similarity between proteins at all?

Properties of Admissible Measures

What are the basic required properties of an admissible measure of similarity between two sets?

1. Symmetry: $\rho(S_i, S_j) = \rho(S_j, S_i)$ for all S_i, S_j
2. Consistency: $\rho(S_i, S_j) \leq \rho(S_j, S_j)$ for all S_i, S_j
3. Monotonicity: $\rho(S_i, S_j) \leq \rho(S_i \cup c_k, S_j \cup c_k)$
4. Generality: $\rho(S_i, S_j) \leq \rho(S_i, S_j \cup S_k)$ for all S_i, S_j, S_k
   - Incompleteness-aware measures: No conclusions based on negative evidence!

Illustration of Properties



$$S_1 = \{c_4\},\ S_2 = \{c_7\}$$
$$S_3 = \{c_6\},\ S_4 = \{c_4, c_6\}$$
$$S_5 = \{c_6, c_7\}$$

- Monotonicity:
  $\rho(S_1, S_2) \leq \rho(S_4, S_5)$
- Generality:
  $\rho(S_2, S_3) \leq \rho(S_2, S_4)$

Existing Measures are not Admissible

- Average (Lord *et al.*, *Bioinformatics*, 2003)

$$\rho_A(S_i, S_j) = \frac{1}{|S_i||S_j|} \sum_{c_k \in S_i} \sum_{c_l \in S_j} \delta(c_k, c_l)$$

  - Fails consistency, monotonicity, generality
- Maximum (Sevilla *et al.*, *IEEE TCBB*, 2005)

$$\rho_M(S_i, S_j) = \max_{c_k \in S_i, c_l \in S_j} \delta(c_k, c_l)$$

  - Principle: Similarity in a single pair of terms is sufficient
  - Fails monotonicity

Existing Measures are not Admissible

- Average of Maxima (Schlicker *et al.*, *Bioinformatics*, 2007)

$$\rho_H(S_i, S_j) = \max \left\{ \frac{1}{|S_i|} \sum_{c_k \in S_i} \max_{c_l \in S_j} \delta(c_k, c_l), \frac{1}{|S_j|} \sum_{c_l \in S_j} \max_{c_k \in S_i} \delta(c_k, c_l) \right\}$$

  - Principle: Similarity with a single term is sufficient for each term
  - Fails consistency, monotonicity, generality

Information Content Based Set Similarity

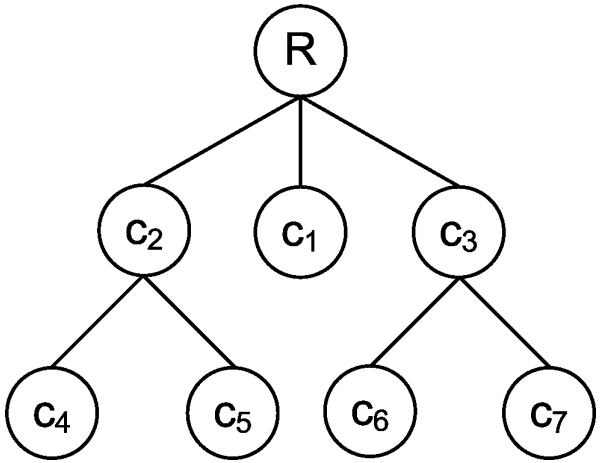
- Generalize the concept of lowest common ancestor to sets of terms (Pandey *et al.*, *ECCB*, 2008)

$$\Lambda(S_i, S_j) = \bigsqcup_{c_k \in S_i, c_l \in S_j} \lambda(c_k, c_l)$$

$$\rho_I(S_i, S_j) = I(\Lambda(S_i, S_j)) = -\log_2 \left(\frac{|G_{\Lambda(S_i,S_j)}|}{|G_r|} \right)$$

  - $G_{\Lambda(S_i,S_j)} = \bigcap_{c_k \in \Lambda(S_i,S_j)} G_{c_k}$ is the set of molecules that are associated with all terms in the MCA set

Illustration of Information Content Based Measure



$$S_1 = \{c_4, c_6, c_7\}$$
$$S_2 = \{c_4\}$$
$$S_3 = \{c_4, c_6\}$$
$$S_4 = \{c_6, c_7\}$$
$$S_5 = \{c_4, c_3\}$$

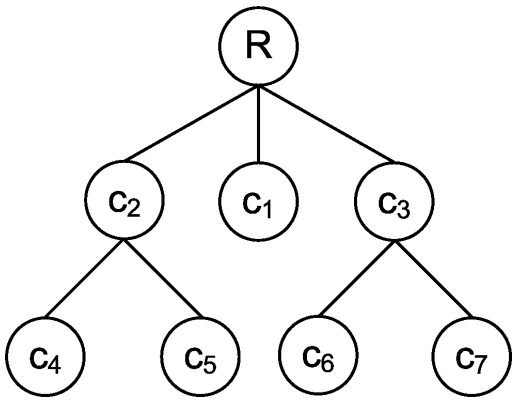
- $\lambda(c_4, c_4) = c_4$,
  $\lambda(c_6, c_4) = \lambda(c_7, c_4) = R$
- $\Lambda(S_1, S_2) = \{c_4\} \Rightarrow$
  $\rho_I(S_1, S_2) =$
  $-\log_2(|G_{c_4}|/|G_R|) =$
  $\log_2(5/4)$
- $\Lambda(S_1, S_3) = \{c_4, c_6\} \Rightarrow$
  $\rho_I(S_1, S_3) = \log_2(5/2)$

Information Content Based Measure Is Admissible

1. Symmetry: Trivially, $\rho_I(S_i, S_j) = \rho_I(S_j, S_i)$ for all S_i, S_j.
2. Consistency: Clearly, $c_k \preceq \lambda(c_k, c_l)$ for any c_k, c_l. Now consider any $c_m \in \Lambda(S_i, S_j)$. Since $c_m = \lambda(c_k, c_l)$ for some $c_k \in S_i$ and $c_l \in S_j$, there always exists $c_n \in \Lambda(S_i, S_i)$ such that $c_n \preceq c_k \preceq c_m$. Consequently, we must have $G_{\Lambda(S_i,S_i)} \subseteq G_{\Lambda(S_i,S_j)}$, leading to $\rho_I(S_i, S_j) \leq \rho_I(S_i, S_i)$.
3. Monotonicity: Since $c_k \nsim c_n$ for all $c_n \in S_i \cup S_j$, we have $\Lambda(S_i \cup c_k, S_j \cup c_k) = \Lambda(S_i, S_j) \sqcup \Lambda(S_i \sqcup S_j, \{c_k\}) \sqcup \{c_k\} \supseteq \Lambda(S_i, S_j) \cup \{c_k\}$, leading to $G_{\Lambda(S_i \cup c_k, S_j \cup c_k)} \subseteq G_{\Lambda(S_i,S_j)}$ and $|G_{\Lambda(S_i \cup c_k, S_j \cup c_k)}| \leq |G_{\Lambda(S_i,S_j)}|$. Consequently, $\rho_I(S_i \cup c_k, S_j \cup c_k) \geq \rho_I(S_i, S_j)$.
4. Generality: $\Lambda(S_i, S_j \cup S_k) = \Lambda(S_i, S_j) \sqcup \Lambda(S_i, S_k) \sqsupseteq \Lambda(S_i, S_j)$. Therefore, $G_{\Lambda(S_i, S_j \cup S_k)} \subseteq G_{\Lambda(S_i,S_j)}$, leading to $\rho_I(S_i, S_j \cup S_k) \geq \rho_I(S_i, S_j)$.

Functional Coherence of Module

Each module is associated with set of molecular entities, and each molecule associated with set of terms.



$$S_1 = \{c_4\},\ S_2 = \{c_4\},$$
$$S_3 = \{c_4, c_6\},$$
$$S_4 = \{c_1, c_6\},\ S_5 = \{c_1\},$$
$$S_6 = \{c_6\}$$

Sets:

- $\mathcal{R}_1 = \{S_1, S_2, S_3, S_4\}$
- $\mathcal{R}_2 = \{S_1, S_2, S_3\}$
- $\mathcal{R}_3 = \{S_3, S_4\}$

Existing Measure

- Average (Pu *et al.*, *Proteomics*, 2007)

$$\sigma_A(\mathcal{R}) = \frac{1}{n(n-1)/2} \sum_{1 \le i < j \le n} \rho(S_i, S_j).$$

- Example: $\sigma_A(S_1, S_2, S_3, S_4) =$

$$\frac{1}{6}(3 * \sigma_A(S_1, S_2, S_3) + \rho(S_3, S_4) + \rho(S_1, S_4) + \rho(S_2, S_4))$$

Generalized Information Content

Extend the notion of the minimum common ancestor of pairs of terms to tuples of terms (Pandey *et al.*, *APBC*, 2010)

$\lambda(c_{i_1}, \ldots, c_{i_n}) = \text{argmax}_{c \in \cap_{k=1}^{n} A_{i_k}} I(c)$

$$\sigma_I(\mathcal{R}) = I(\Lambda(S_1, \ldots, S_n)) = -\log_2 \left(\frac{|G_{\Lambda(S_i,\ldots,S_j)}|}{|G_r|} \right).$$

where

$$\Lambda(S_1, S_2, \ldots, S_n) = \bigsqcup_{c_{i_j} \in S_j, 1 \leq j \leq n} \lambda(c_{i_1}, c_{i_2}, \ldots, c_{i_n})$$

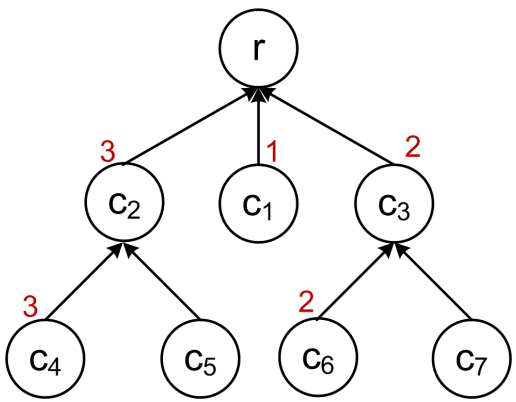
Example: $\sigma_I(S_1, S_2, S_3, S_4) = I(r) = 0$

Weighted Information Content

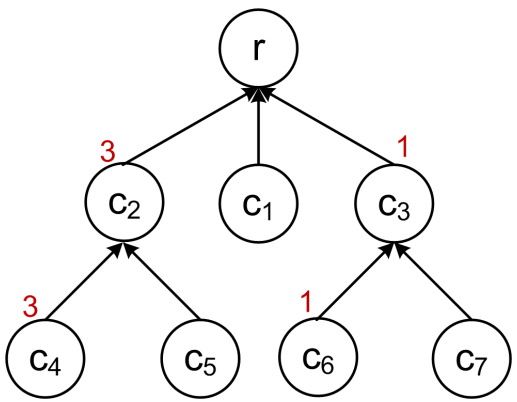
Weigh the information content of shared functionality by the number of molecules that contribute to the shared functionality

$$\sigma_W(\mathcal{R}) = 1 - \frac{\sum_{1 \le i \le n} \sum_{c \in \mathcal{A}'_i} I(c)}{\sum_{1 \le i \le n} \sum_{c \in \mathcal{A}_i} I(c)}$$

$\sigma_W(S_1, S_2, S_3, S_4) = 0.86$

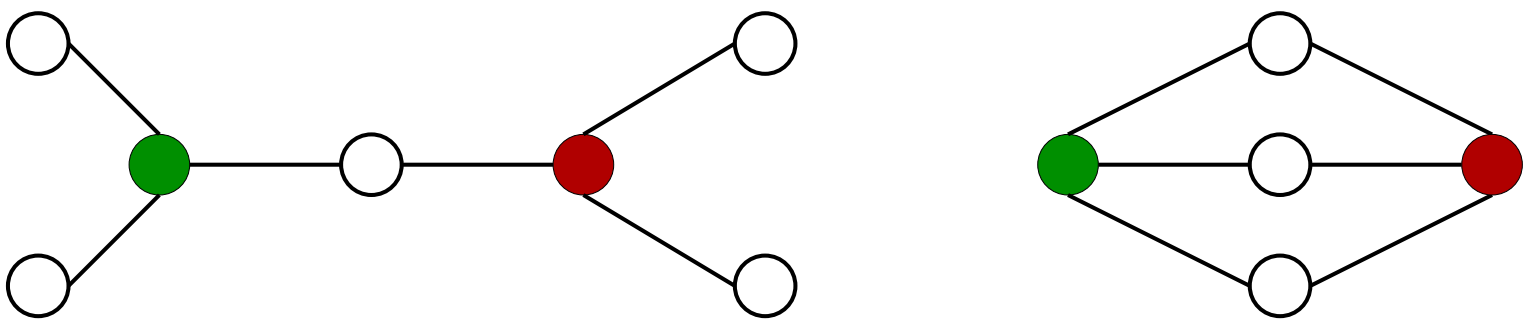


$\sigma_W(S_1, S_2, S_3) = 0.75$



Accounting for Multiple Paths

- Is "shortest path" a good measure of network proximity?
  - Multiple alternate paths might indicate stronger functional association
  - In well-studied pathways, redundancy is shown to play an important role in robustness & adaptation (*e.g.*, genetic buffering)



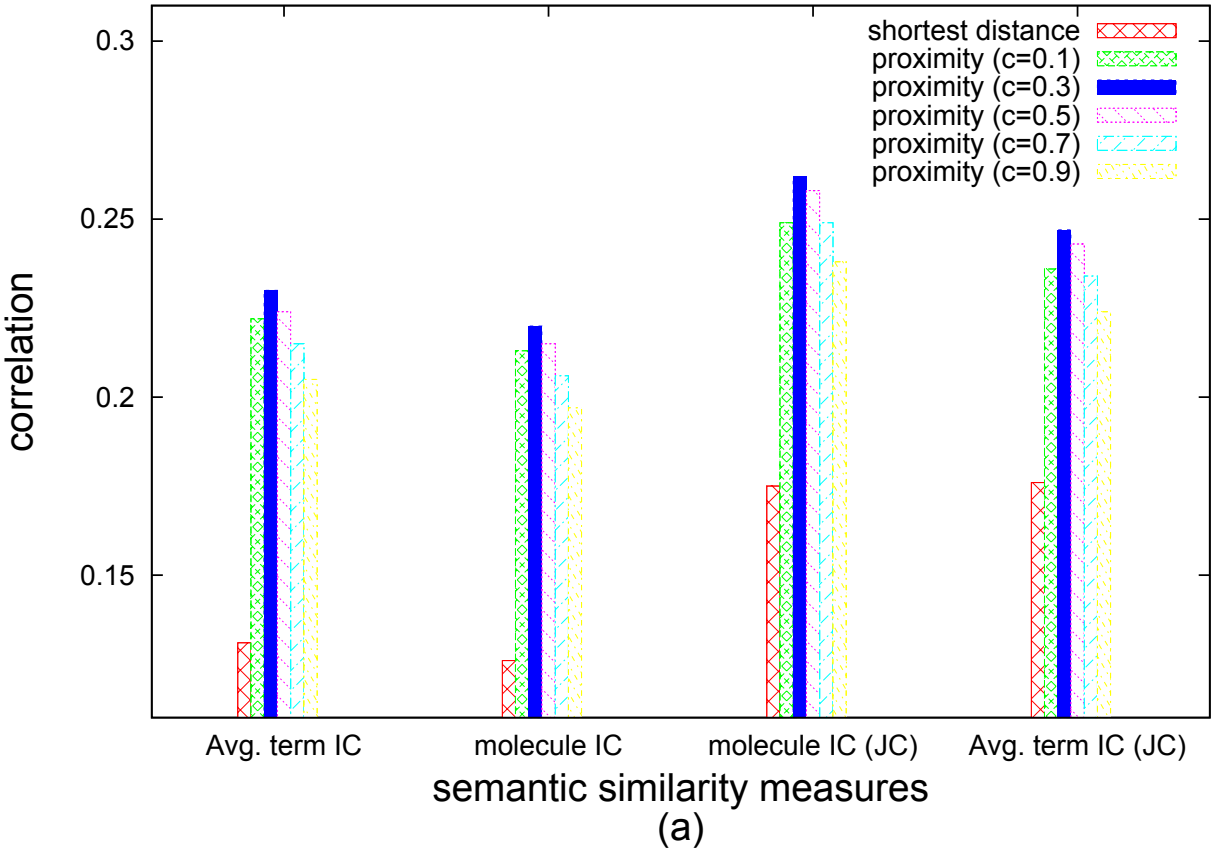
Proximity Based On Random Walks

- Simulate an infinite random walk with random restarts at protein i
- Proximity between proteins i and j is given by the relative amount of time spent at protein j

$$\Phi(0) = I,\ \Phi(t+1) = (1-c)A\Phi(t) + cI,\ \Phi = \lim_{t\to\infty} \Phi(t)$$

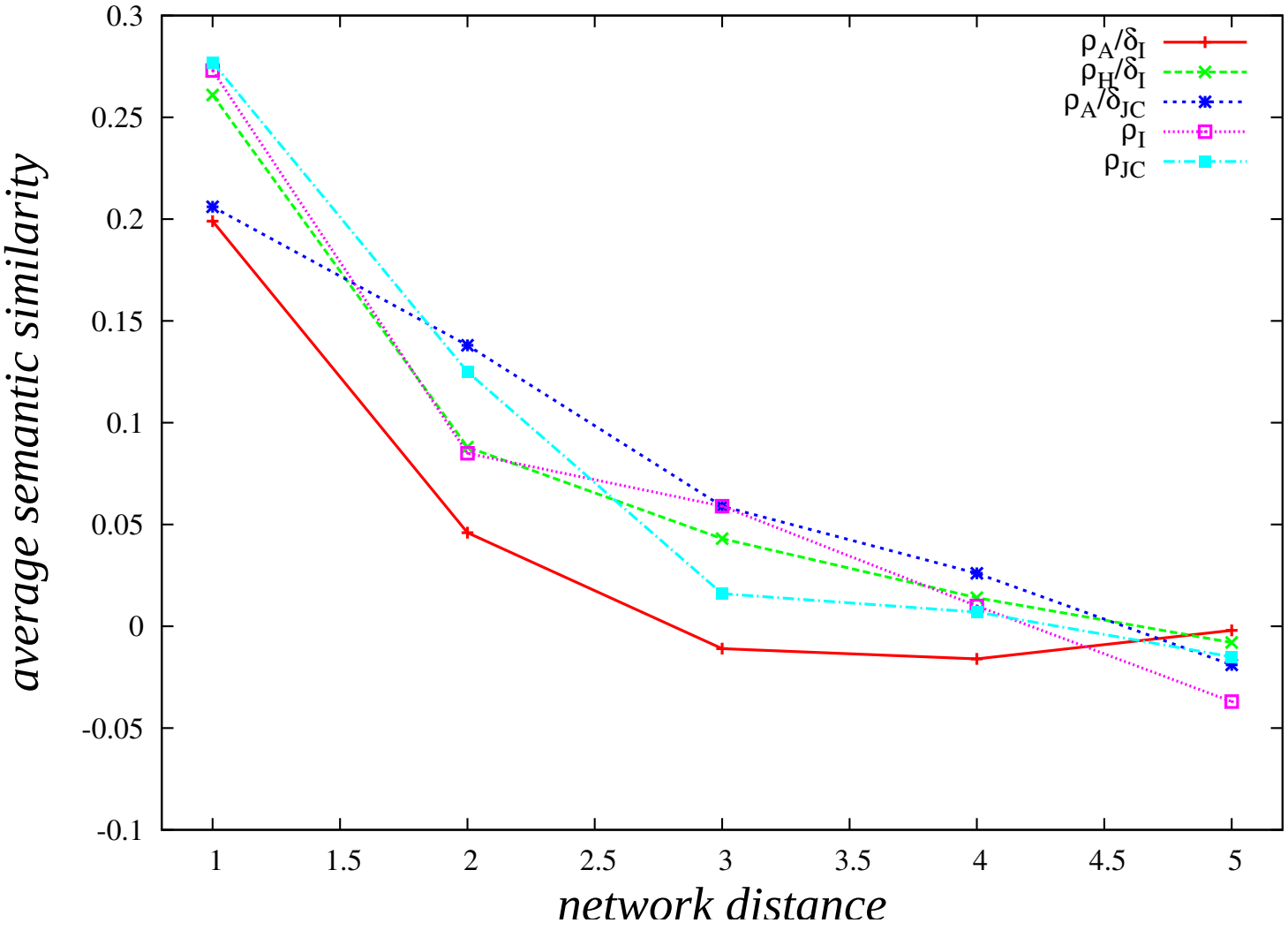
  - $\Phi(i,j)$: Network proximity between protein i and protein j
  - A: Stochastic matrix derived from the adjacency matrix of the network
  - I: Identity matrix
  - c: Restart probability

Network Proximity & Functional Similarity



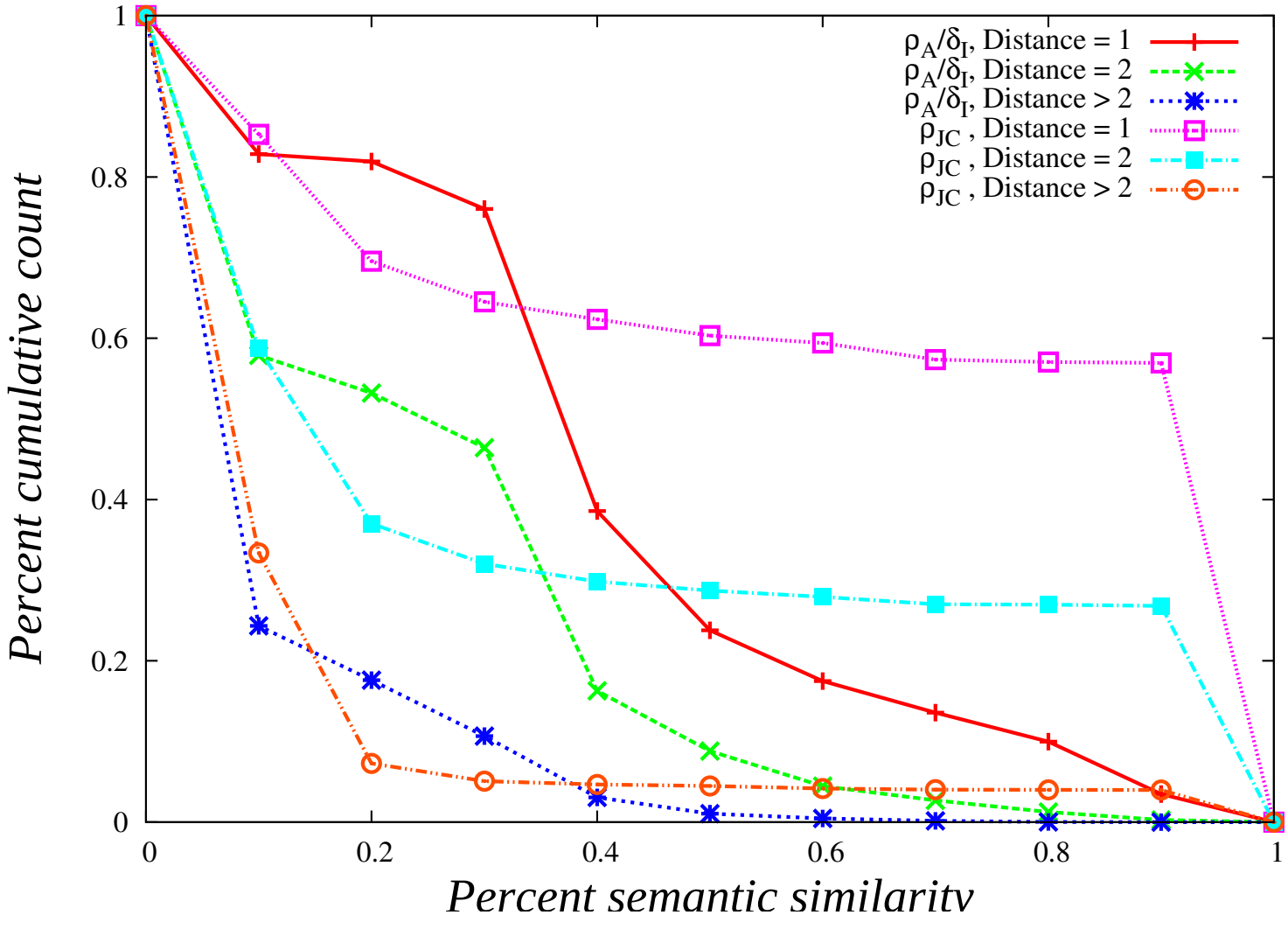
Correlation between functional similarity and network proximity

Comparison of Similarity Measures



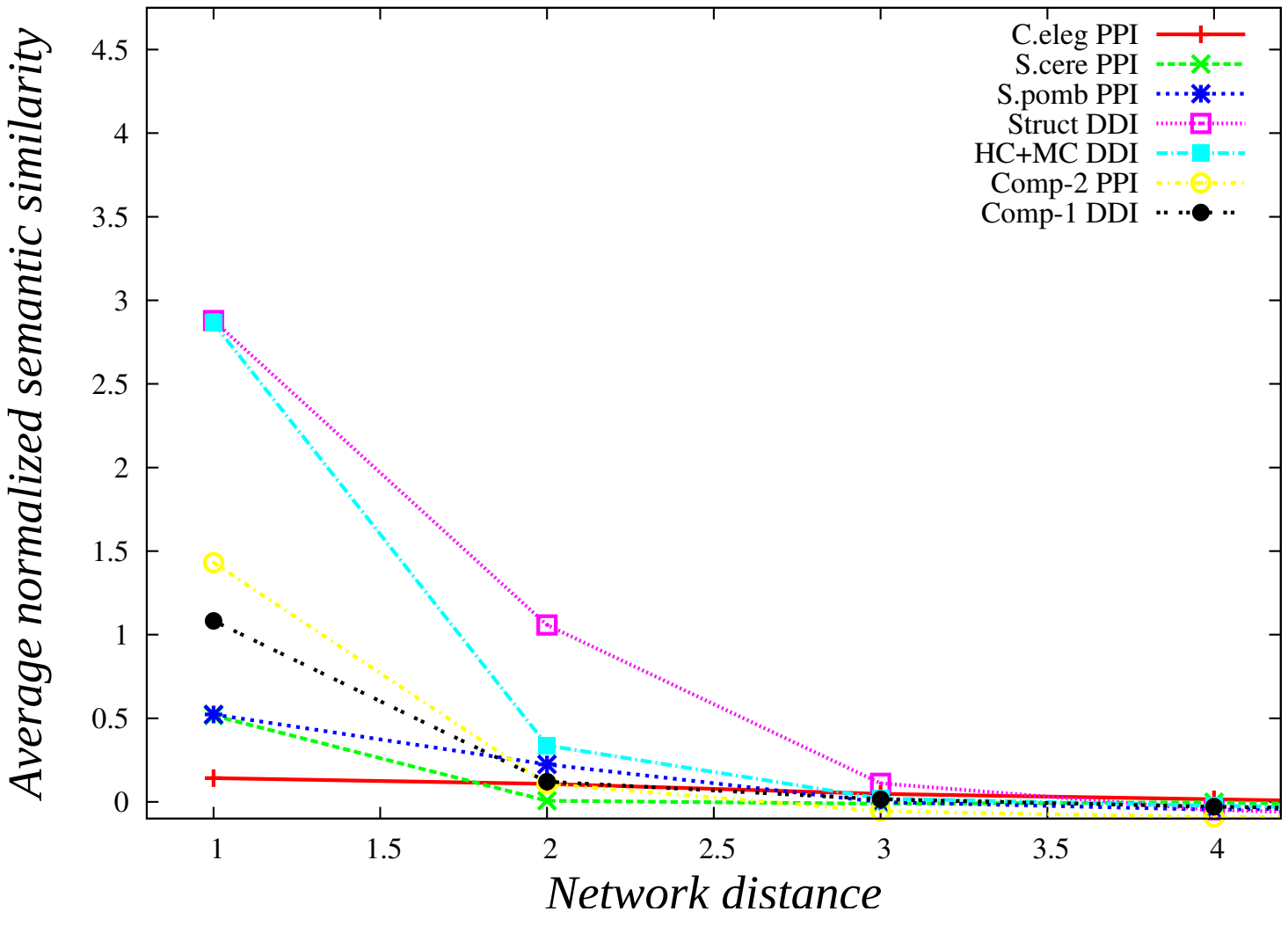
Network distance *vs.* functional similarity on *C. elegans* PPI network

Comparison of Similarity Measures



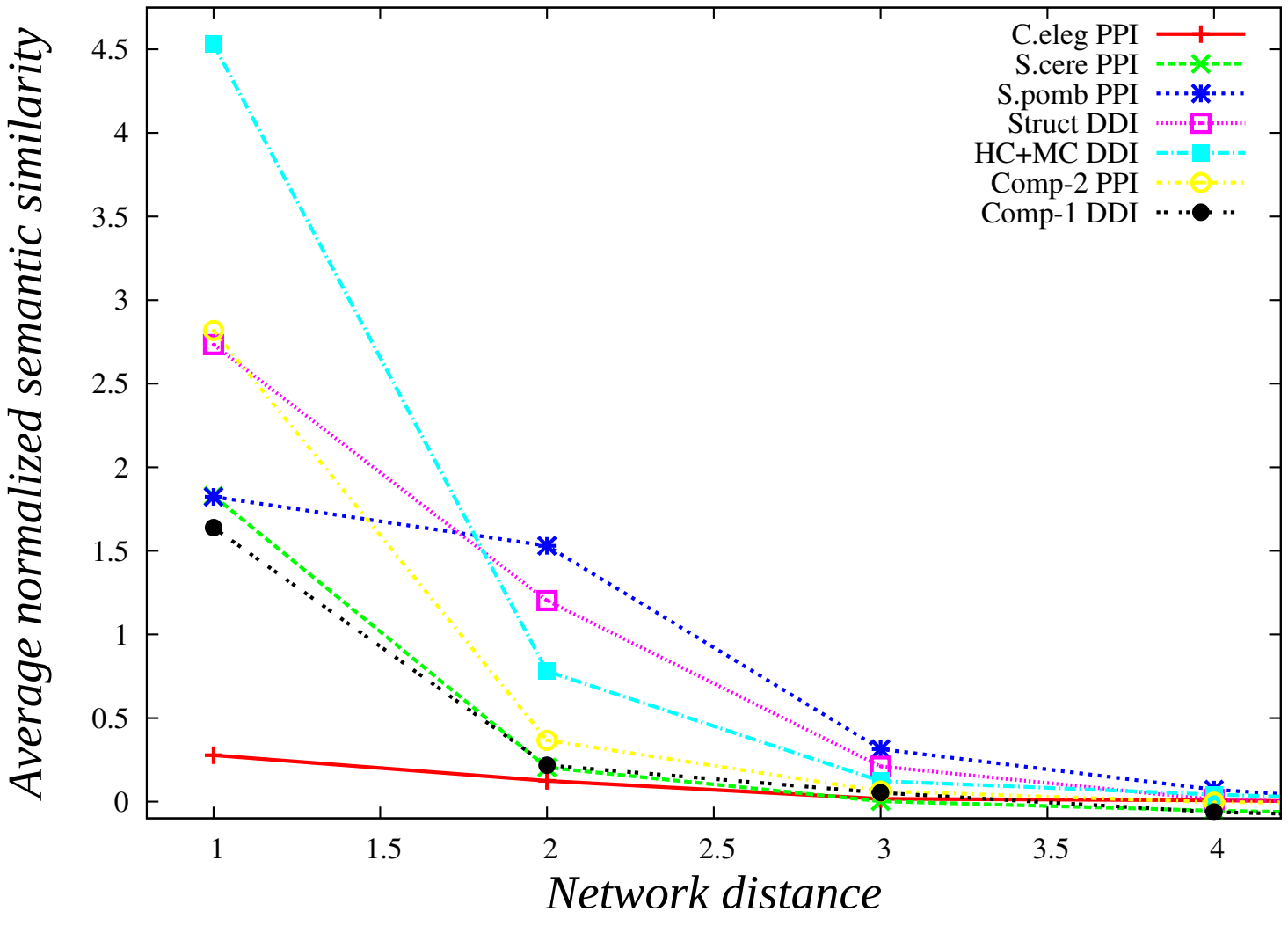
Distribution of functional similarity scores for structurally inferred DDIs

Comparison of PPI and DDI Networks



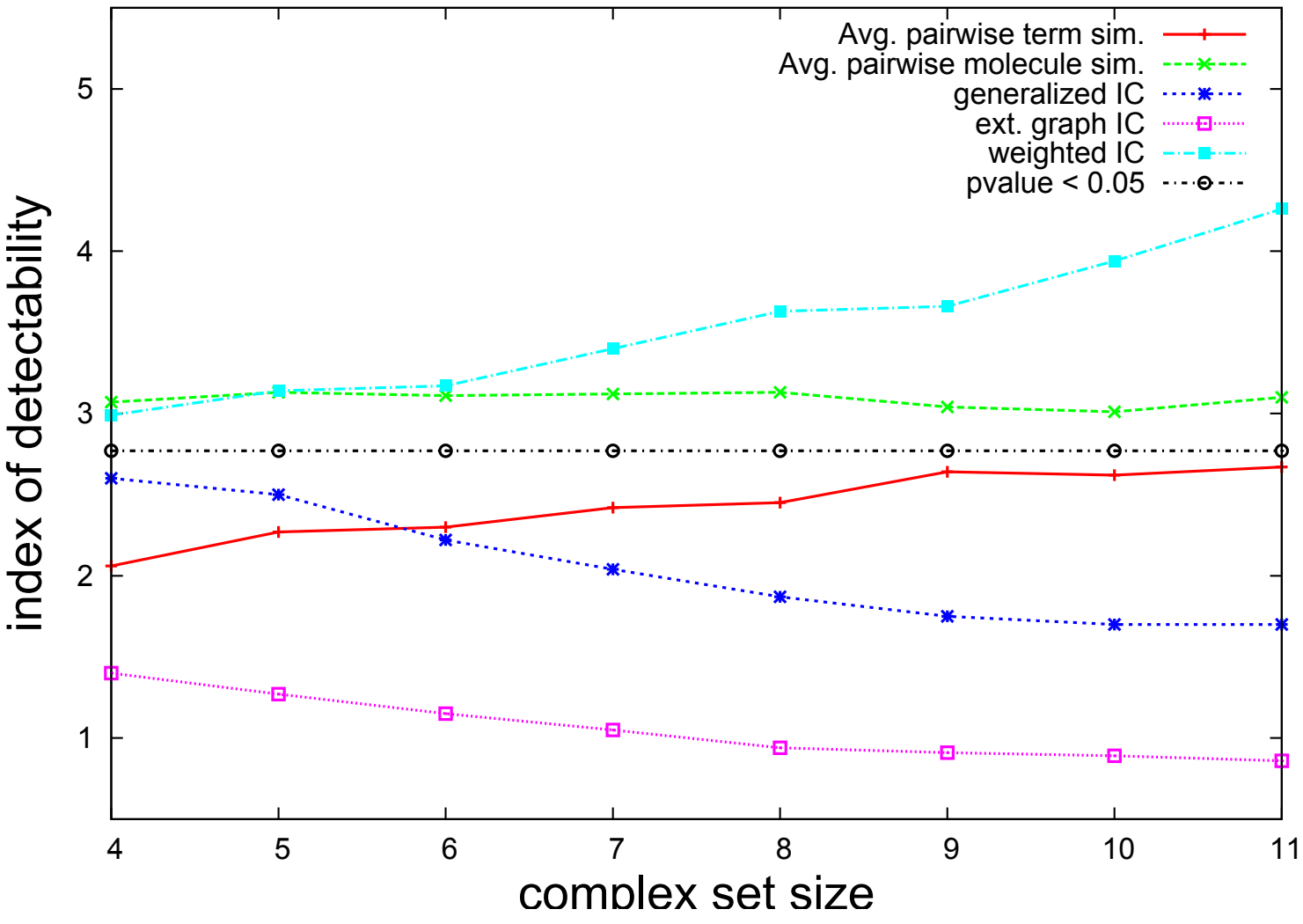
Network distance *vs.* functional similarity based on molecular functions

Comparison of PPI and DDI Networks



Network distance *vs.* functional similarity based on biological processes

Comparison of Coherence Meaures



Index of Dectectability vs. complex sizes

Outline

Conclusions & Avenues for Future Research

- Computational tools to analyze biological networks in context of functions of individual bio-molecules
- Conclusions
  - Patterns describe essential mechanisms in biological systems
  - Coherence and proximity measures suitable to work with noisy and incomplete data
- Avenues for Future Research
  - Pattern-based protein function prediction
  - Phylogenetic analysis of identified patterns
  - Using proximity measure to find disease implicated genes

Thanks...

- For their guidance and support
  - Ananth Grama
  - Mehmet Koyuturk (CWRU)
- For constructive feedback
  - Wojciech Szpankowski
  - Shankar Subramaniam of UCSD
  - Daisuke Kihara
  - Alex Pothen
- For productive and intriguing discussions
  - Members of Parallel & Distributed Systems Lab