Functional Coherence in Molecular Interaction Networks

Mehmet Koyutürk

Case Western Reserve University
Department of Electrical Engineering & Computer Science

Computer Science Colloquim
Kent State University
October 1, 2008

Outline

1. Background & Motivation

2. Annotation of Regulatory Pathways

3. Functional Coherence & Network Proximity

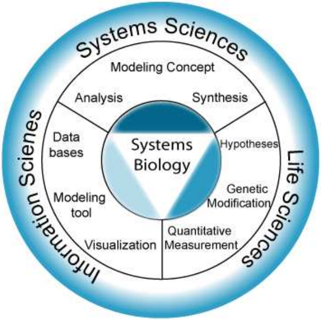
4. Acknowledgments

Outline

1. Background & Motivation
2. Annotation of Regulatory Pathways
3. Functional Coherence & Network Proximity
4. Acknowledgments

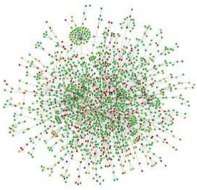
Systems Biology

- "To understand biology at the system level, we must examine the structure and dynamics of cellular and organismal function, rather than the characteristics of isolated parts of a cell or organism." (Kitano, *Science*, 2002)
  - Cell is not just an assembly of genes and proteins
  - Systems biology complements molecular biology

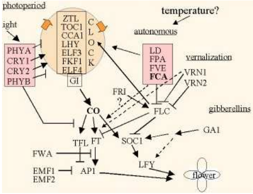


Modeling Cellular Organization: Networks

- Metabolism, genetic regulation, cellular signaling
- Nodes represent cellular components
    - Protein, gene, enzyme, metabolite
- Edges represent interactions
    - Binding, regulation, modification, complex membership, substrate-product relationship



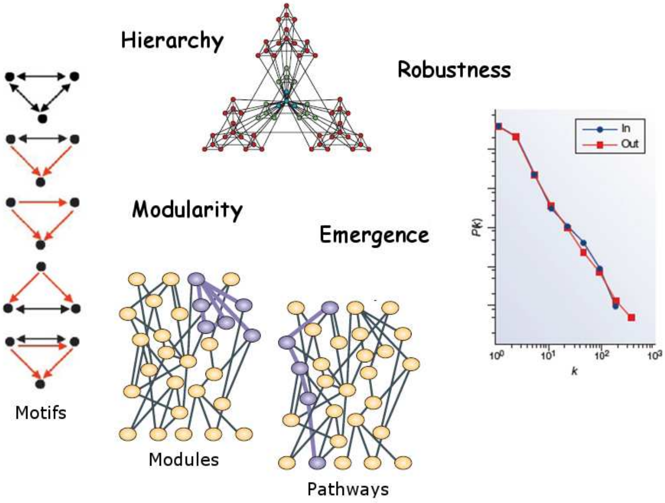
S.cerevisiae
PPI network



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

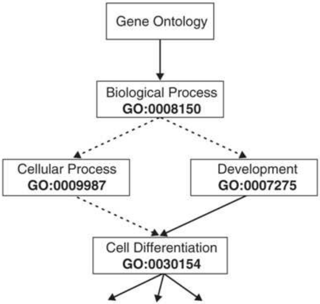
Function & Topology in Molecular Networks

How does function relate to network topology?



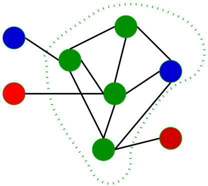
Characterizing Biological Function

- Significant progress on standardizing knowledge on biological function at the molecular level
    - Protein/domain families (COG, PFAM, ADDA)
    - Gene Ontology: Hierarchical classification of molecular functions, biological processes, and cellular components

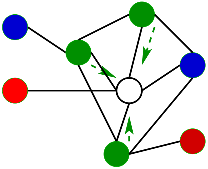


Functional Coherence

- Modularity manifests itself in terms of high connectivity in the network
  - Identification of modular subgraphs
  - Functional annotation of a group of molecules



- Functional association (similarity) is correlated with network proximity
  - Network based functional annotation
  - Identification of multiple disease markers



In This Talk

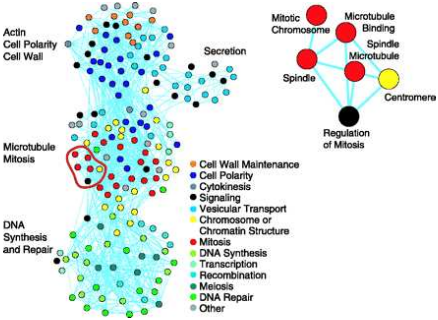
1. Recurrent functional interaction patterns
   - Crosstalk between different processes
   - "Periodic table of systems biology"
2. Functional coherence with respect to different types of interaction
   - What does proximity mean in domain-domain interaction networks?
   - Assessing functional similarity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

Outline

1. Background & Motivation
2. Annotation of Regulatory Pathways
3. Functional Coherence & Network Proximity
4. Acknowledgments

Functional Annotation: From Molecules to Systems

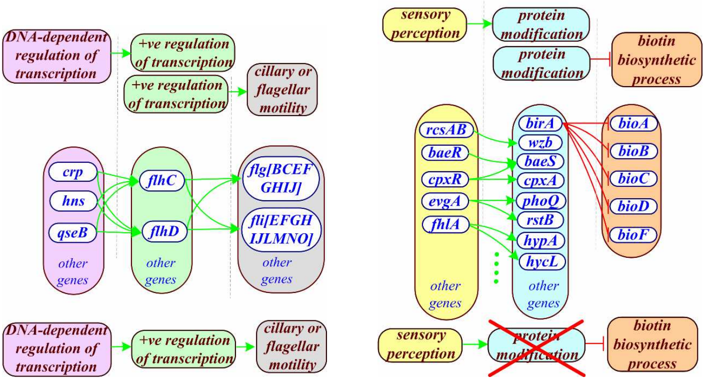
- Networks are species-specific
- Functional ontologies are described at the molecular level
- Can we map networks from gene space to an abstract (and unified) function space?



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

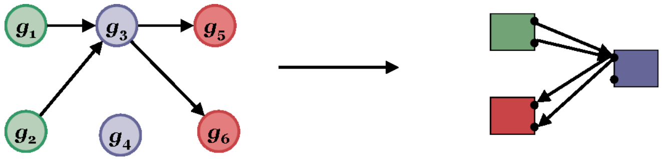
Gene Regulatory Networks: Indirect Regulation

- Assessment of pairwise interactions is simple, but not adequate


DNA-dependent regulation of transcription
+ve regulation of transcription
+ve regulation of transcription
cillary or flagellar motility
crp
hns
qseB
other genes
flhC
flhD
other genes
flg[BCEF GHIJ]
fli[EFGH IJLMNO]
other genes
DNA-dependent regulation of transcription
+ve regulation of transcription
cillary or flagellar motility
sensory perception
protein modification
protein modification
biotin biosynthetic process
rcsAB
baeR
cpxR
evgA
fhlA
other genes
birA
wzb
baeS
cpxA
phoQ
rstB
hypA
hycL
other genes
bioA
bioB
bioC
bioD
bioF
sensory perception
protein modification
biotin biosynthetic process

Functional Attribute Networks

- 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

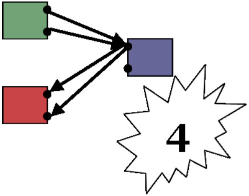


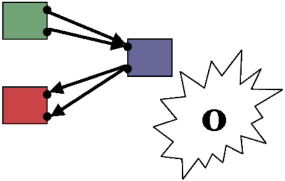
Gene Network

Functional Attribute Network

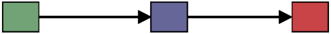
Frequency of a Multipath

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





Frequency of Multipath



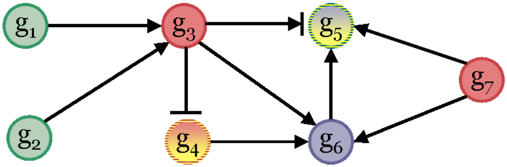
?

Frequency *vs.* Statistical Significance

- We want to identify overrepresented pathways
  - 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 gene network is highly skewed
  - Pathways that contain common functional attributes have high frequency, but they are not necessarily interesting

Statistical Significance of a Pathway

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



φ(🟩 → 🟥 → 🟦) = φ(🟩 → 🟥 ⊣ 🟨) = 4

φ(🟩 → 🟥) = φ(🟥 ⊣ 🟨) = 2 φ(🟥 → 🟦) = 5

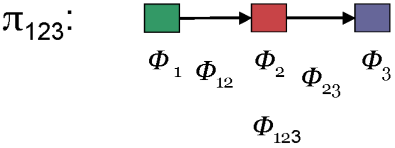
P(🟩 → 🟥 ⊣ 🟨) < P(🟩 → 🟥 → 🟦)

Significance of Pairwise Interactions

- A single regulatory interaction is the shortest pathway
  - Arbitrary degree distribution: The number of edges leaving and entering each functional attribute is specified
  - Edges are assumed to be independent
- The frequency of a regulatory interaction is a hypergeometric random variable
- $p_{ij} = P(\Phi_{ij} \geq \phi_{ij} | \mathcal{B}) = \sum_{\ell=\phi_{ij}}^{\min\{\beta_i \delta_j, n\}} \frac{\binom{\beta_i \delta_j}{\ell}\binom{m-\beta_i \delta_j}{n-\ell}}{\binom{m}{n}}.$
  - β_i = in-degree and δ_i = out-degree
  - m = pool of potential edges, n = number of edges in network

Significance of a Pathway

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



- Significance of pathway 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

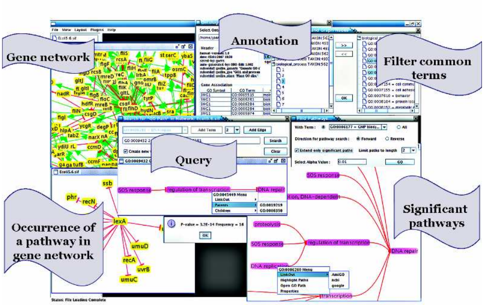
- Assume that interactions are independent
  - There are $\varphi_{12}\varphi_{23}$ possible pairs of π_{12} and π_{23} edges
  - The probability that a pair of π_{12} and π_{23} edges go through the same gene (corresponds to an occurrence of π_{123}) is $1/\varphi_2$
- The probability that at least 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 pathways
  - 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.*, *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



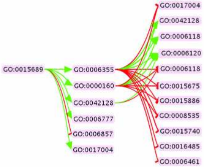
Significant Regulatory Pathways in Bacteria

- We use NARADA to identify significant pathways in the transcriptional networks of two bacterial species
  - *E. coli*: 1364 genes, 3159 regulatory interactions (RegulonDB)
  - *B. subtilis*: 562 genes, 604 regulatory interactions (DBTBS)

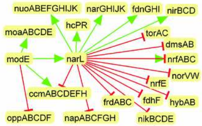
Strongly significant pathways ($p < 0.01$)

Pathway length	2	3	4
E. coli	143	753	1328
B. subtilis	22	78	202
Common	10	54	157

An Example: Molybdate Ion Transport



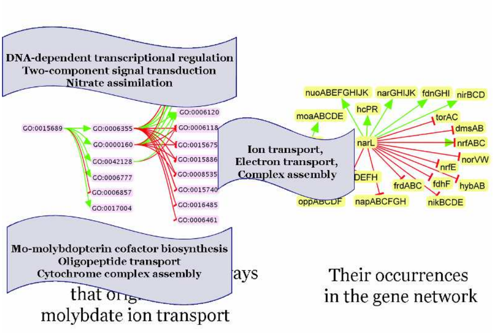
Significant regulatory pathways
that originate at
molybdate ion transport



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



DNA-dependent transcriptional regulation
Two-component signal transduction
Nitrate assimilation

Ion transport,
Electron transport,
Complex assembly

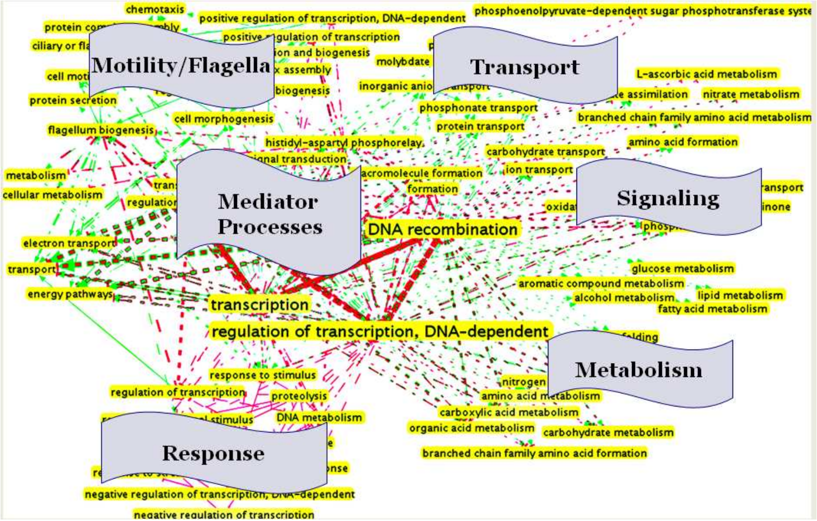
Mo-molybdopterin cofactor biosynthesis
Oligopeptide transport
Cytochrome complex assembly

ays
that or
molybdate ion transport

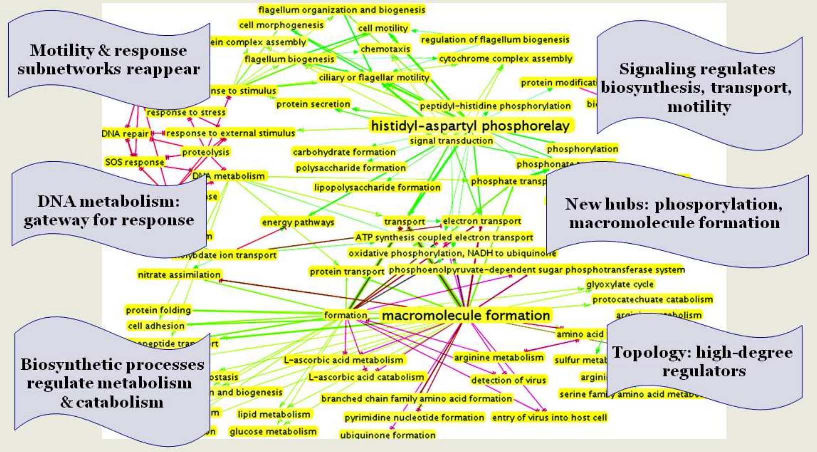
Their occurrences
in the gene network

- 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!

Functional View of *E. coli* Regulatory Network



Short-Circuiting Mediator Processes



Motility & response subnetworks reappear

Signaling regulates biosynthesis, transport, motility

DNA metabolism: gateway for response

New hubs: phosporylation, macromolecule formation

Biosynthetic processes regulate metabolism & catabolism

Topology: high-degree regulators

Applications

- Projecting from functional space back to molecular space
    - Pattern-based functional annotation (Kirac *et al.*, *RECOMB*, 2008)
    - Pathway identification through cross-species projection (Cakmak *et al.*, *Bioinformatics*, 2008)
- Ongoing work: Interaction prediction
    - Identify significant functional pathways in *E. coli* transcriptional network
    - Find (partial) occurrences of these pathways in the *B.subtilis* transcriptional network
    - "Interpolate" these pathways to predict novel interactions

Outline

1 Background & Motivation

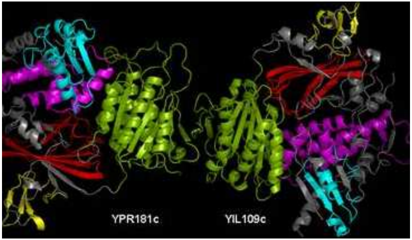
2 Annotation of Regulatory Pathways

3 Functional Coherence & Network Proximity

4 Acknowledgments

Domain-Domain Interactions

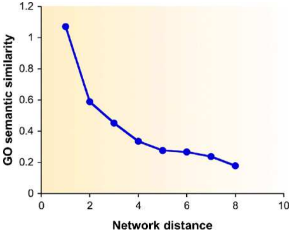
- Most proteins are composed of multiple domains
- Many domains are reused in several (evolutionarily/functionally related) proteins
- Interactions between domains underlie observed protein-protein interactions
- Many algorithms exist to infer domain-domain interactions



Jothi *et al.*, *JMB*, 2006

PPI Networks *vs.* DDI Networks

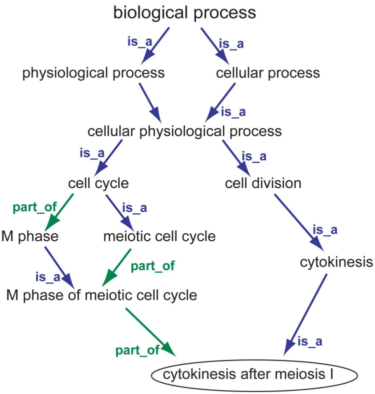
- Protein-protein interaction (PPI) networks are used extensively for functional inference
    - Network-based functional annotation
    - Identification of functional modules
- In PPI networks, functional coherence manifests itself in terms of network proximity
    - How about DDI "networks"?



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

Assessing Functional Similarity

- Gene Ontology (GO) provides a hierarchical taxonomy of biological function
- 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
    - $\lambda(c_i, c_j) = \text{argmax}_{c \in A_i \cap A_j} I(c)$: Minimum common ancestor of c_i and c_j

Functional Similarity of Molecules

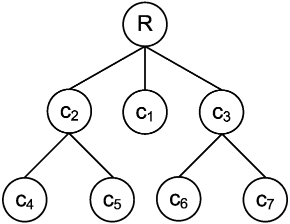
- Each molecule (protein or domain) is associated with multiple GO terms
- Available annotations are incomplete
- Domain annotations are often derived from protein annotations
  - A domain is associated with terms at the intersection of proteins that contain the domain
- Is it possible to compare functional 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

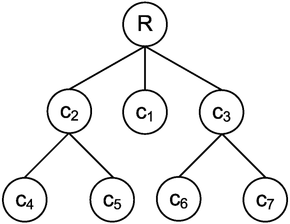
- Generalize the concept of minimum common ancestor to sets of terms (Pandey *et al.*, *ECCB*, 2008)

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

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

  - $G_{\Lambda(S_i,S_j)} = \bigcap_{c_k \in \Lambda(S_i,S_j)} G_{c_k}$ 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 $c_k \not\sim c_n$ for all $c_n \in S_i \cup S_j$, we have

   $$\Lambda(S_i \cup c_k, S_j \cup c_k) = \Lambda(S_i, S_j) \sqcup \Lambda(S_i \sqcup S_j, \{c_k\}) \sqcup \{c_k\} \supseteq \Lambda(S_i, S_j) \cup \{c_k\},$$

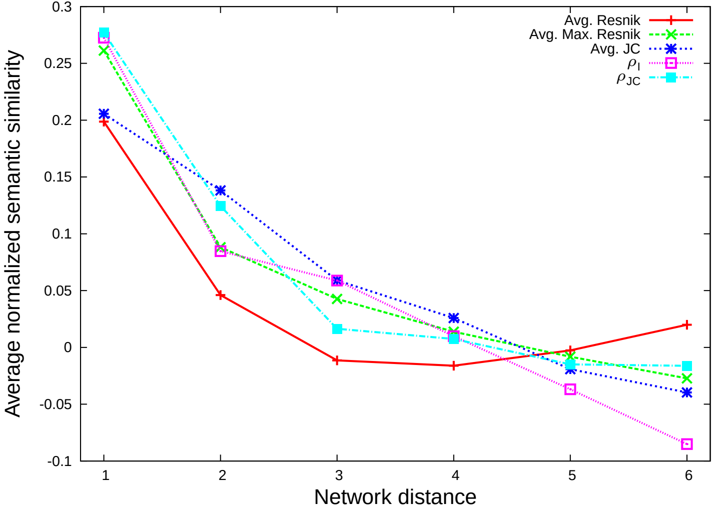
   leading to $G_{\Lambda(S_i \cup c_k, S_j \cup c_k)} \subseteq G_{\Lambda(S_i,S_j)}$ and $|G_{\Lambda(S_i \cup c_k, S_j \cup c_k)}| \leq |G_{\Lambda(S_i,S_j)}|$.

   Consequently, $\rho_I(S_i \cup c_k, S_j \cup c_k) \geq \rho_I(S_i, S_j)$.
4. Generality:

   $\Lambda(S_i, S_j \cup S_k) = \Lambda(S_i, S_j) \sqcup \Lambda(S_i, S_k) \sqsupseteq \Lambda(S_i, S_j)$.

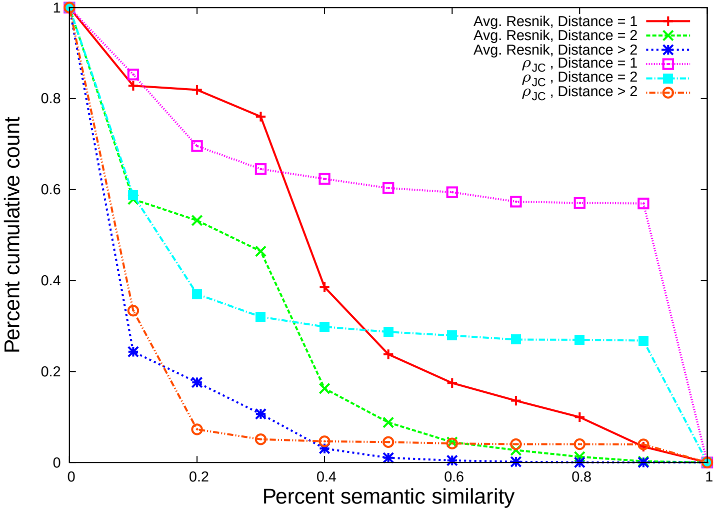
   Therefore, $G_{\Lambda(S_i, S_j \cup S_k)} \subseteq G_{\Lambda(S_i,S_j)}$, leading to $\rho_I(S_i, S_j \cup S_k) \geq \rho_I(S_i, S_j)$.

Comparison of Similarity Measures



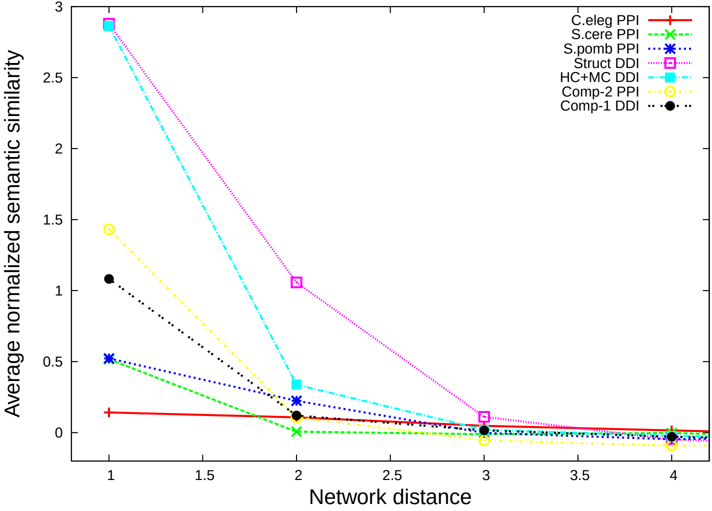
Network distance *vs.* functional similarity on *C. elegans* PPI network

Comparison of Similarity Measures



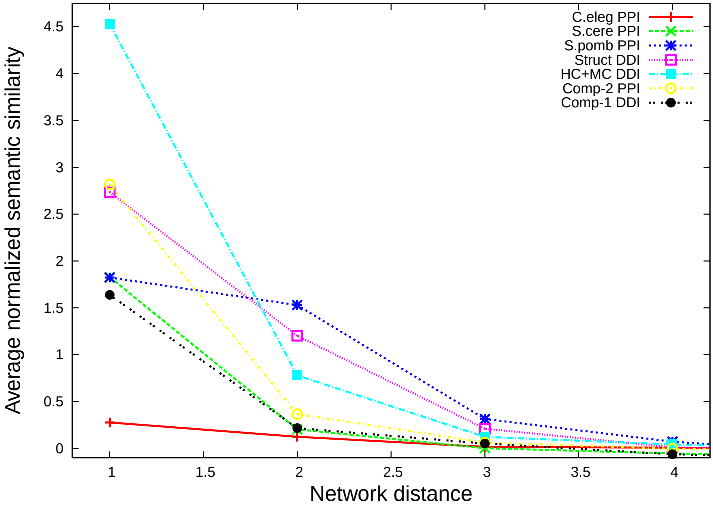
Distribution of functional similarity scores for structurally inferred DDIs

Comparison of PPI and DDI Networks



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

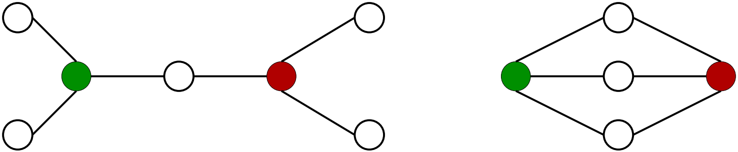
Comparison of PPI and DDI Networks



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

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)



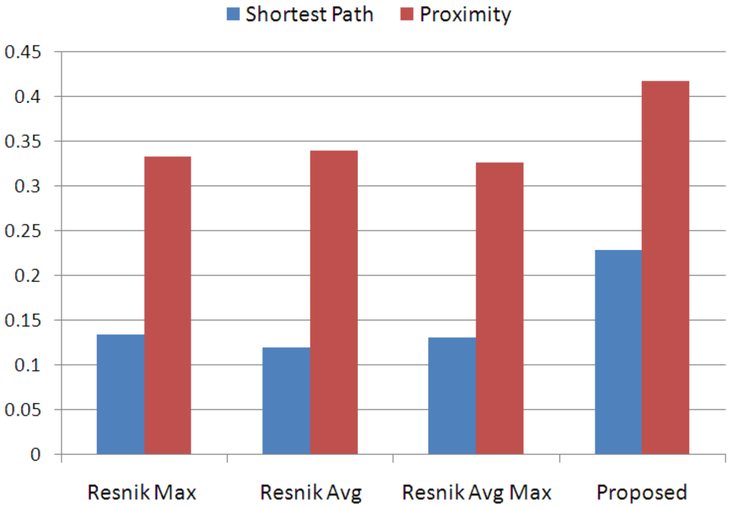
Proximity Based On Random Walks

- 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*

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

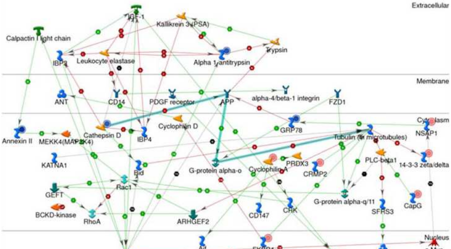
- $\Phi(i,j)$: Network proximity between protein *i* and protein *j*
- *A*: Stochastic matrix derived from the adjacency matrix of the network
- *I*: Identity matrix
- ρ: Restart probability

Shortest Path *vs.* Proximity



Application: Identifying Indirectly Implicated Genes

- Premise: Small changes in mRNA expression may lead to significant changes in post-transcriptional activity
  - Human colorectal cancer: Identify proteins with significant fold change (between metastatic and control samples) using 2D-PAGE
  - Map these "seed proteins" on the PPI network to extract "implicated subnets"
  - Refine these subnets using gene expression data



"Regulation of developmental proteins" subnet, differentially expressed in metastatic stages of human colorectal cancer

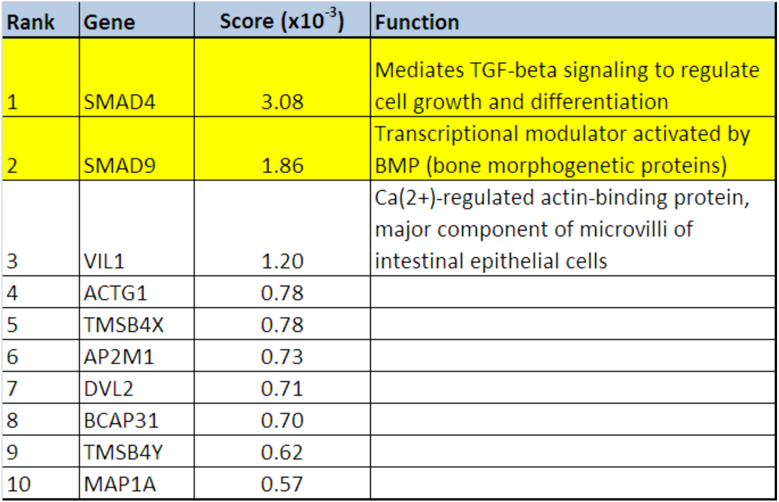
Using Network Proximity to Find Implicated Genes

- Generalize random walk with restarts
  - Restart at any of the seed proteins!

$$\phi(0) = r,\ \phi(t+1) = (1-\rho)A\phi(t) + \rho r,\ \phi = \lim_{t\to\infty} \phi(t)$$

- $\phi(j)$: Proximity of protein j to seed proteins
- r : Restart vector, $||r||_1 = 1$
- $r(i) = |z_i|$ if fold change z_i of protein i is significant
- Prioritize all proteins in the network based on $\phi(j)$

Genes Implicated by Network Proximity



Rank	Gene	Score (x10^{-3})	Function
1	SMAD4	3.08	Mediates TGF-beta signaling to regulate cell growth and differentiation
2	SMAD9	1.86	Transcriptional modulator activated by BMP (bone morphogenetic proteins)
3	VIL1	1.20	Ca(2+)-regulated actin-binding protein, major component of microvilli of intestinal epithelial cells
4	ACTG1	0.78	
5	TMSB4X	0.78	
6	AP2M1	0.73	
7	DVL2	0.71	
8	BCAP31	0.70	
9	TMSB4Y	0.62	
10	MAP1A	0.57	

Outline

1. Background & Motivation
2. Annotation of Regulatory Pathways
3. Functional Coherence & Network Proximity
4. Acknowledgments

Thanks

- CWRU
  - Sinan Erten
- Case School of Medicine
  - Mark Chance, Rod Nibbe, Vishal Patel
- Purdue
  - Jayesh Pandey, Wojciech Szpankowski, Ananth Grama
- UC-San Diego
  - Yohan Kim, Shankar Subramaniam
- T. & D. Schroeder for endowed chair!