Functional Characterization and Topological Modularity of Molecular Interaction Networks

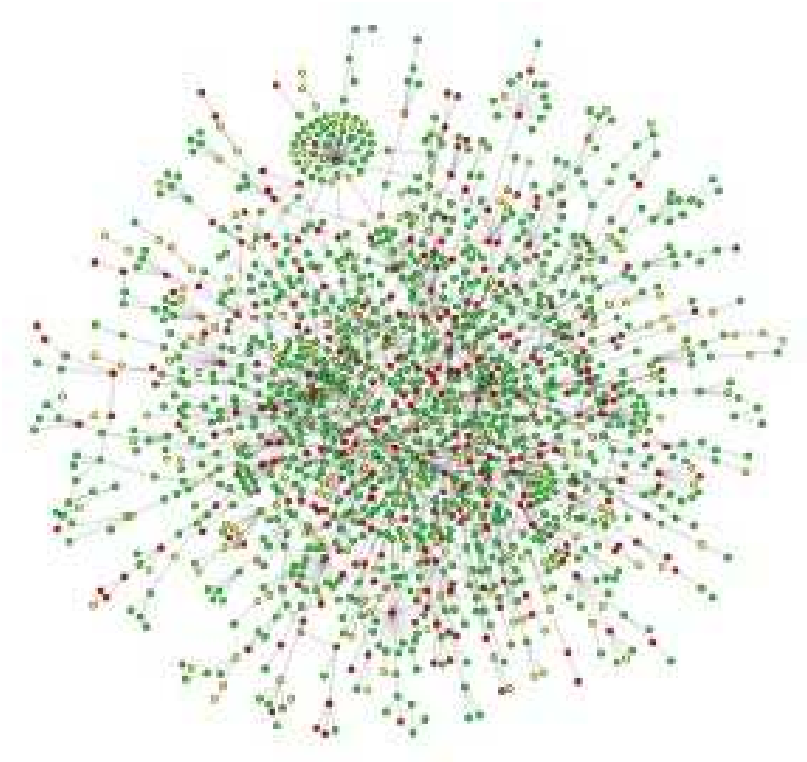
Ananth Grama[1]

[1]Center for Science of Information and
Department of Computer Science, Purdue University

Various parts of this talk involved collaborations with Jayesh Pandey, Mehmet Koyuturk, Shahin Mohammadi, Giorgos Kollias, and Shankar Subramaniam. Acknowledgements to the National Science Foundation.

Molecular Interaction Networks

- Provides a high level description of cellular organization
- Directed and undirected graph representation
- Nodes represent cellular components
  - Protein, gene, enzyme, metabolite
- Edges represent reactions or interactions
  - Binding, regulation, modification, complex membership, substrate-product relationship

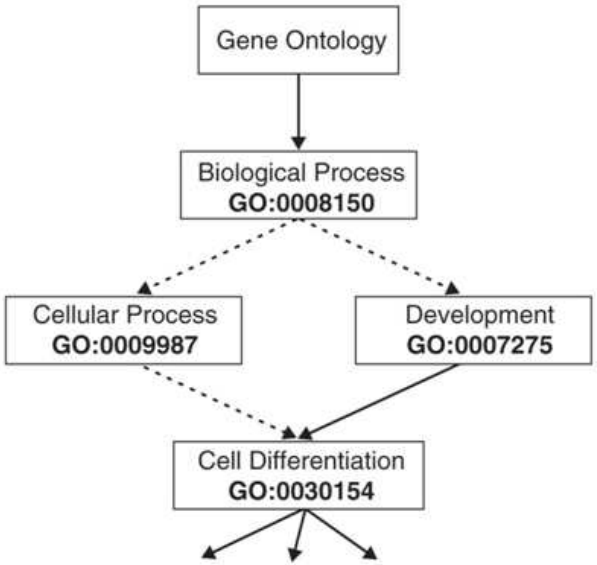


S.cerevisiae

Protein-Protein Interaction (PPI) Network

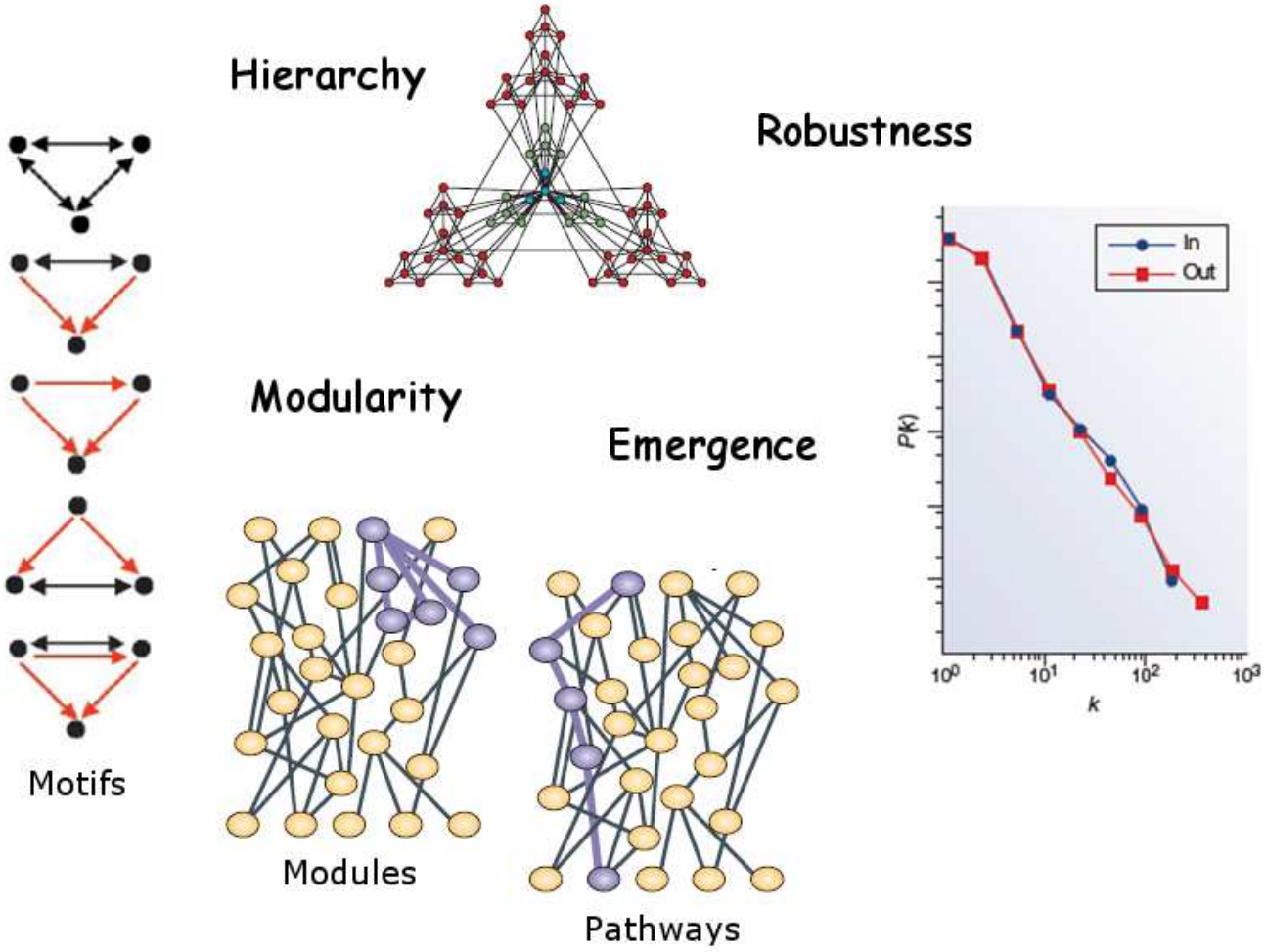
Function : Gene Ontology

- Molecular annotation provides a unified understanding of the underlying principles
- Gene Ontology: A controlled vocabulary of molecular functions, biological processes, and cellular components
- Terms (concepts) related by *is-a, part-of* relationships
- If a molecule is annotated by a term, then it is also annotated by terms on the paths towards root.



Function & Topology in Molecular Networks

How does function relate to network topology?



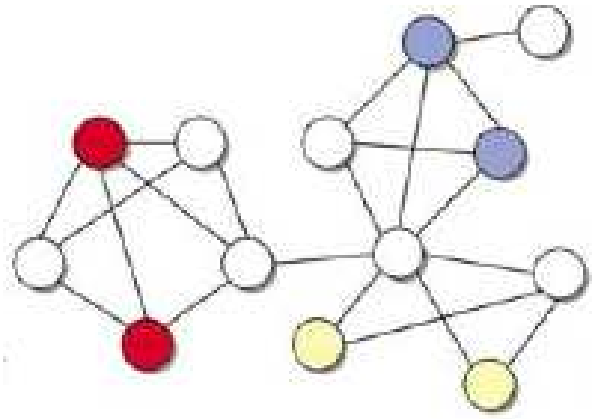
Prior Work on Topology and Function

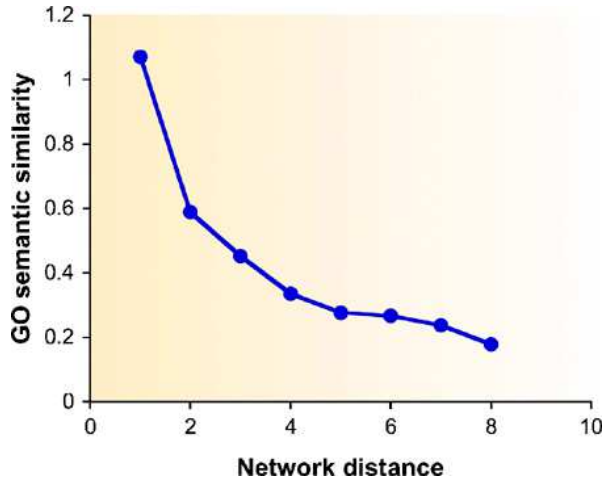
Understanding functional composition of biochemical networks

- Conservation [ISMB 04/Bioinf. 04]
- Alignment [RECOMB 05/JCB 06]
- Modularity [RECOMB 06/JCB 07]
- Inference [Bioinf. 06]
- Pathway Annotation [ISMB 07/Bioinf. 07, PSB 08]
- Network Abstractions/ Annotations [ECCB 08/ Bioinf. 08]
- Modularity and Domain Interactions [APBC 10/ BMC Bioinf. 10]
- Pathway Interaction Maps [PSB 12, Submitted]

Functional Coherence in Networks

- Modularity manifests itself in terms of high connectivity in the network
- Functional association (similarity) is correlated with network proximity
- A measure for annotation proximity of nodes (semantic similarity)
- A measure for network distance

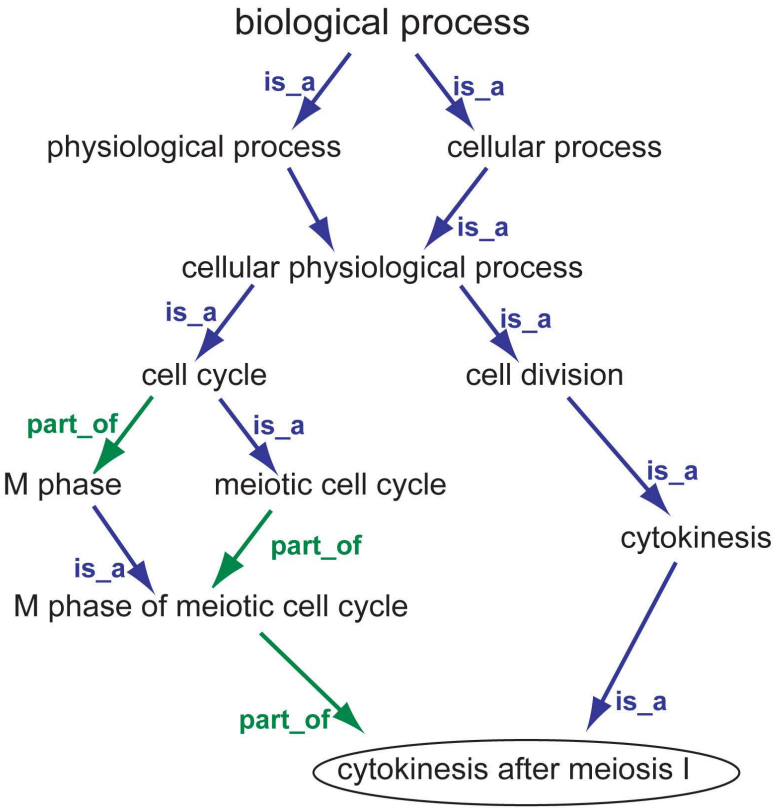




Sharan *et al.*, *MSB*, 2007

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)

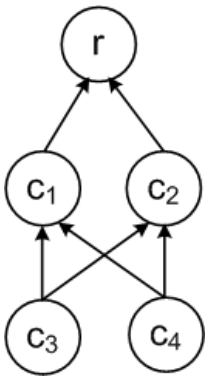


Semantic Similarity of GO Terms

- 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



Resnik(c_3, c_4) = Max(IC(c_1), IC(c_2))

Functional Similarity of Molecules with Sets of Terms

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

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

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

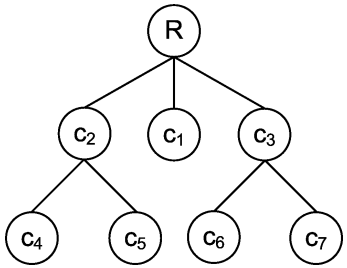
$$\Lambda(S_i, S_j) = \bigsqcup_{c_k \in S_i, c_l \in S_j} \lambda(c_k, c_l)$$

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

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

Functional Coherence of Module

- A set of molecules that participates in the same biological processes or functions
- sub-network with dense intra-connections and sparse interconnections
- 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 $\lambda(c_{i_1}, \ldots, c_{i_n}) = \operatorname{argmax}_{c \in \cap_{k=1}^{n} A_{i_k}} I(c)$

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

where

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

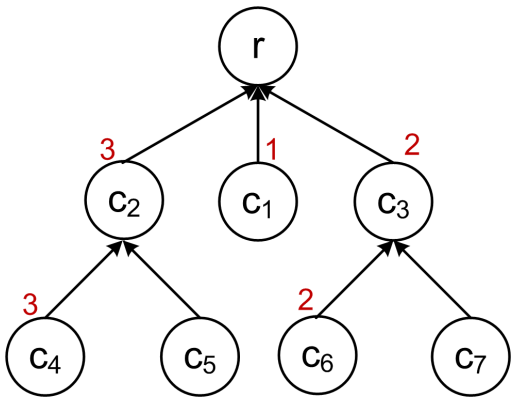
Example: $\sigma_I(S_1, S_2, S_3, S_4) = I(r) = 0$, no common ancestor!

Weighted Information Content

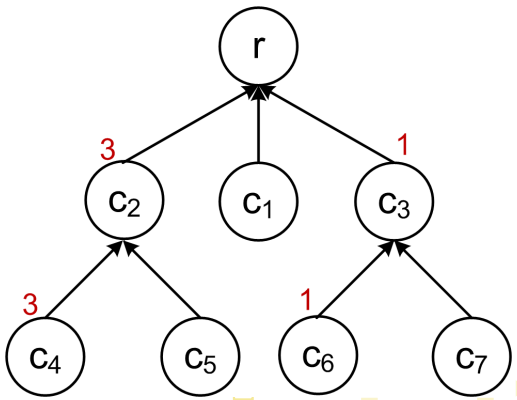
Weigh the information content of shared functionality by the number of molecules that contribute to the shared functionality

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

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

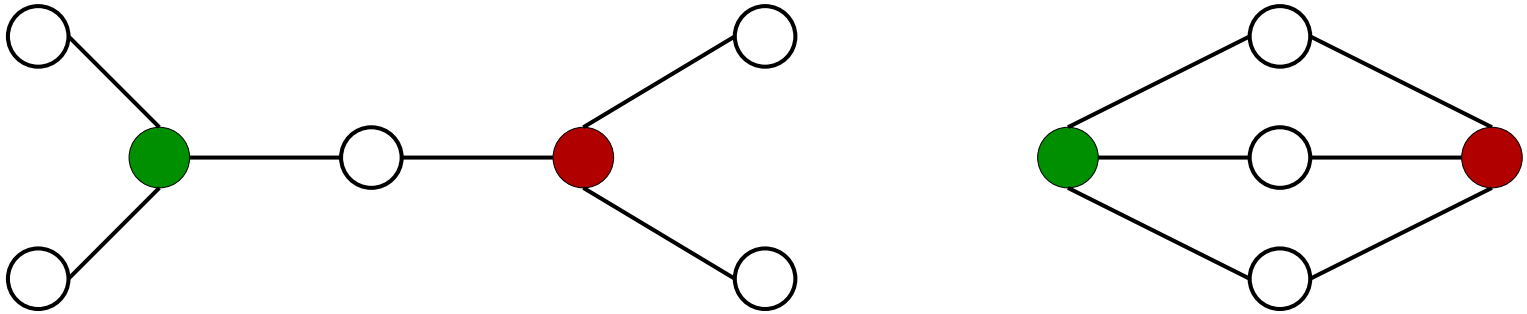


$\sigma_W(S_1, S_2, S_3) = 0.75$



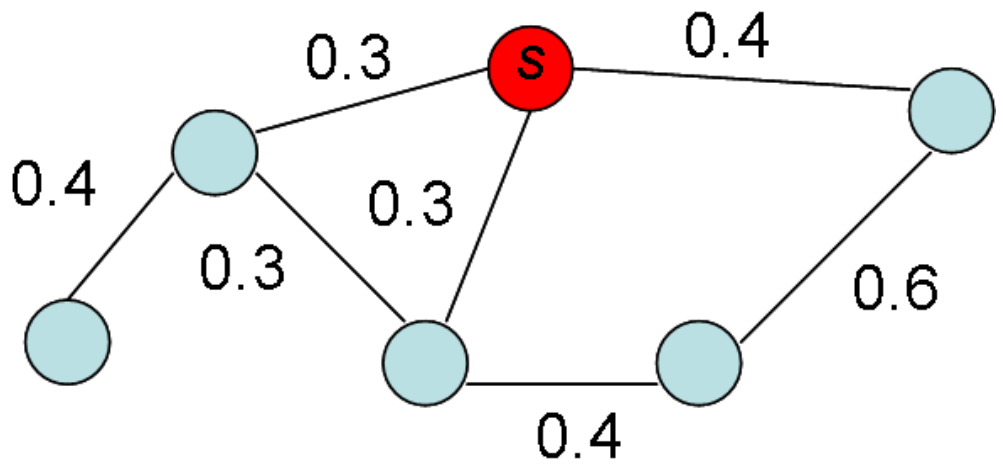
Accounting for Multiple Paths

- 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)



Random walks with restarts

- Consider a random walker that starts on a source node *s*. At every tick, the walker chooses randomly among available edges or goes back to node *s* with probability *c*.



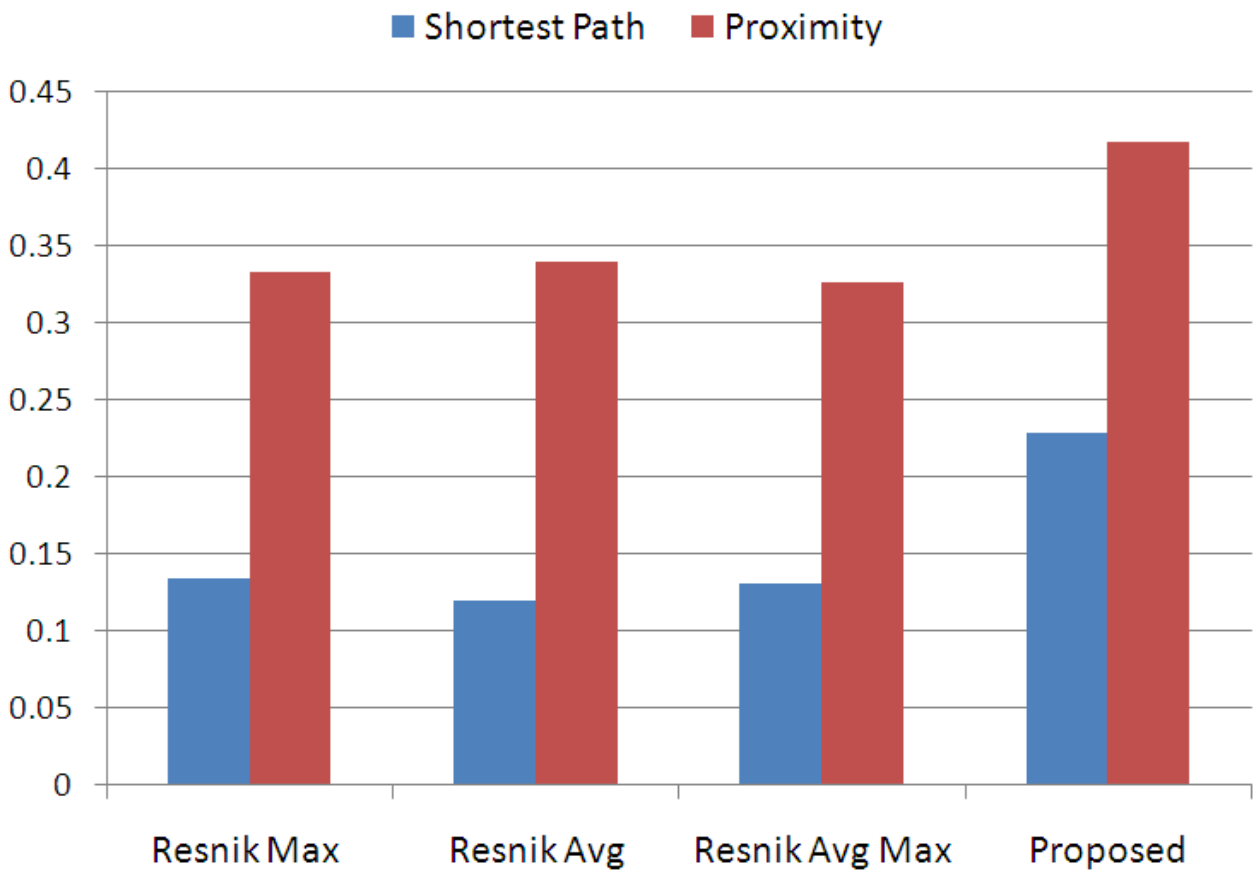
Proximity Based On Random Walks

- 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

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

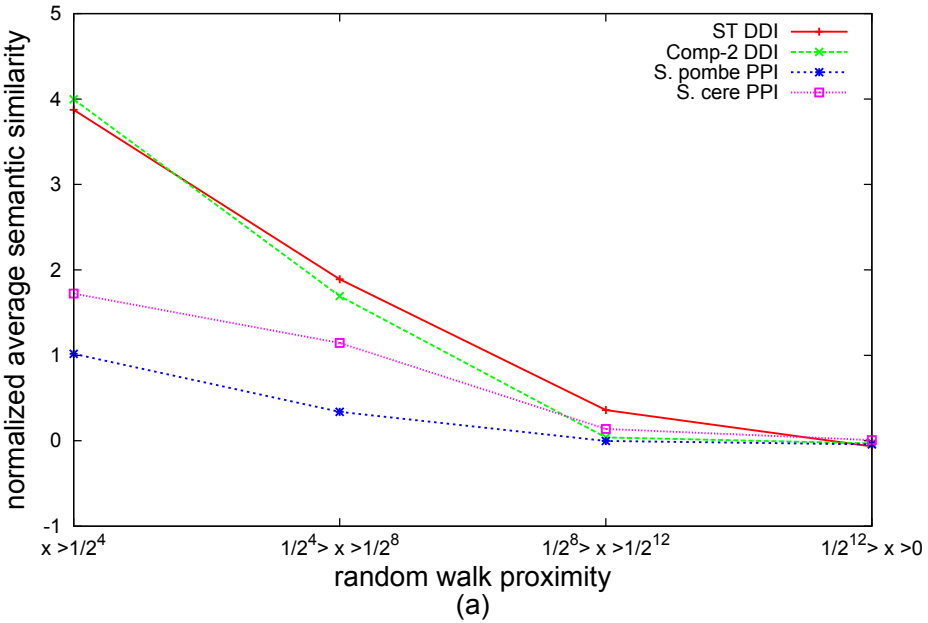
  - $\Phi(i,j)$: Network proximity between protein i and protein j
  - A: Stochastic matrix derived from the adjacency matrix of the network
  - I: Identity matrix
  - c: Restart probability
- Define proximity between proteins i and j as $\{\Phi(i,j) + \Phi(j,i)\}/2$

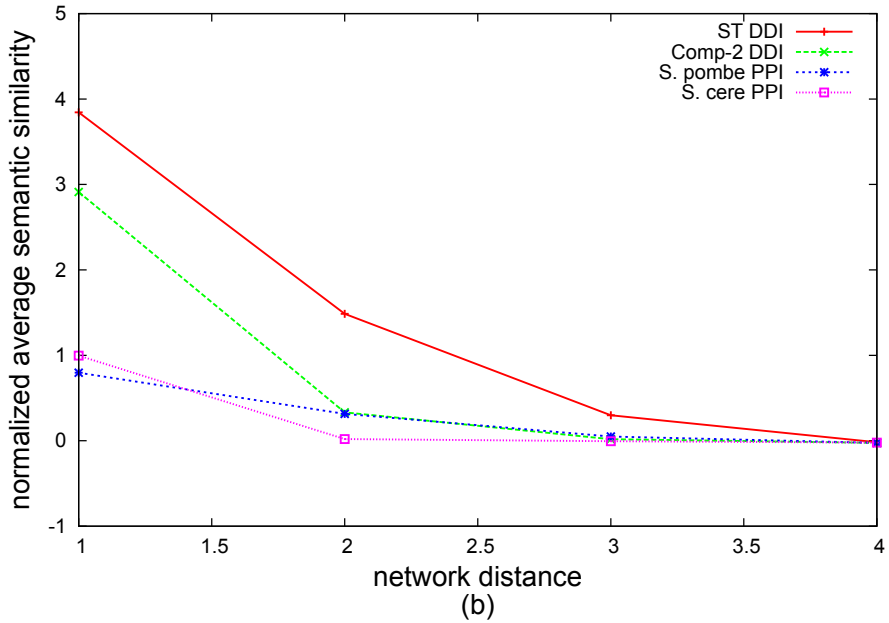
Network Proximity & Functional Similarity



Correlation between functional similarity and network proximity

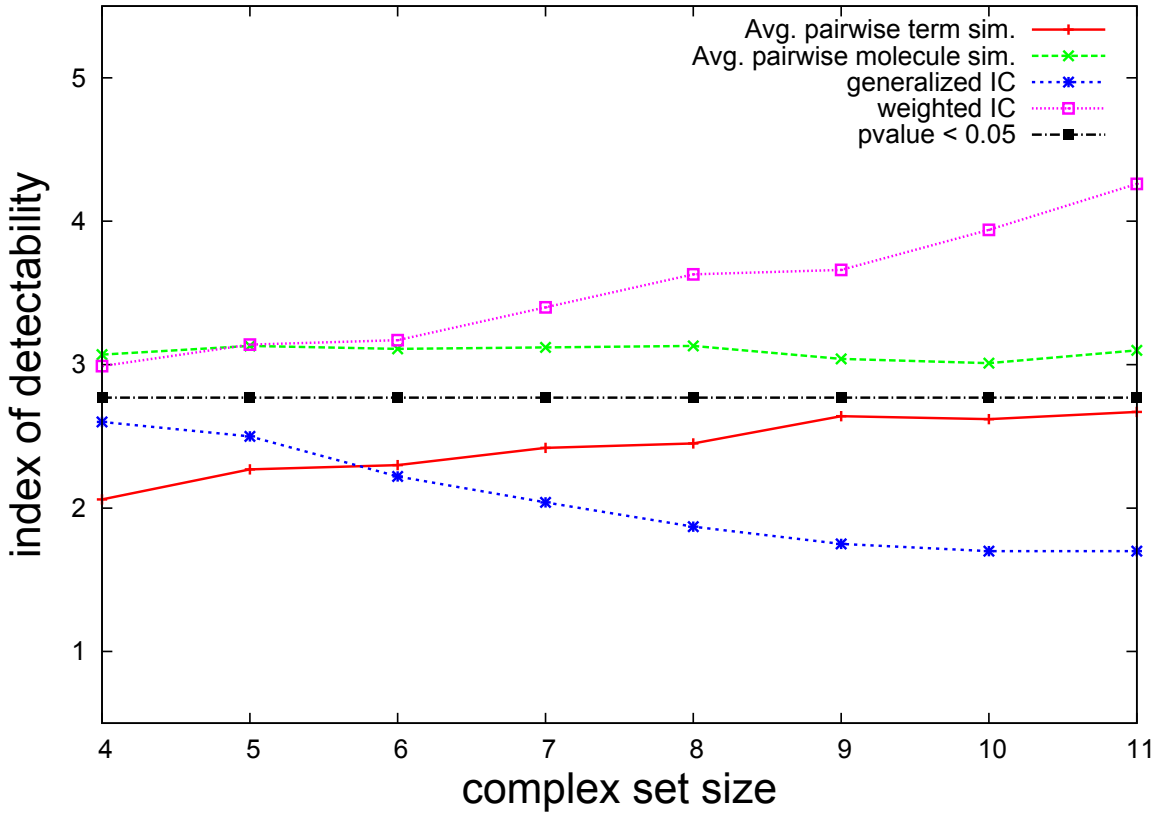
Topological Proximity and Functional Similarity





Comparison of the DDI and PPI networks with respect to the relation between semantic similarity vs proximity and network distance

Comparison of Coherence Meaures



Index of Dectectability vs. complex sizes

$$d(\sigma) = \frac{\text{mean}_{t \in T}(\sigma(t)) - \text{mean}_{t \in C}(\sigma(t))}{\sqrt{((\text{std}_{t \in T}(\sigma(t)))^2 + (\text{std}_{t \in C}(\sigma(t)))^2)/2}}$$

Lessons Learned

- Random walk based measures of topological proximity are better suited to existing interaction data
- Measures that quantify coherence among entire sets are superior to aggregares of known pair-wise measures

Part II

Building Pathway Maps Using Synthetic Lethality Networks

Genetic Interactome

Double mutants exhibit unexpected phenotypes, as compared to joint single mutations.

Definition

- **Negative Interactions**: more severe phenotype than expected
  - Also known as *aggravating* or *synergistic*
- **Positive Interactions**: Less severe phenotype than expected
  - Also known as *alleviating* or *epistatic*

Most commonly used:

Phenotype : Growth rate

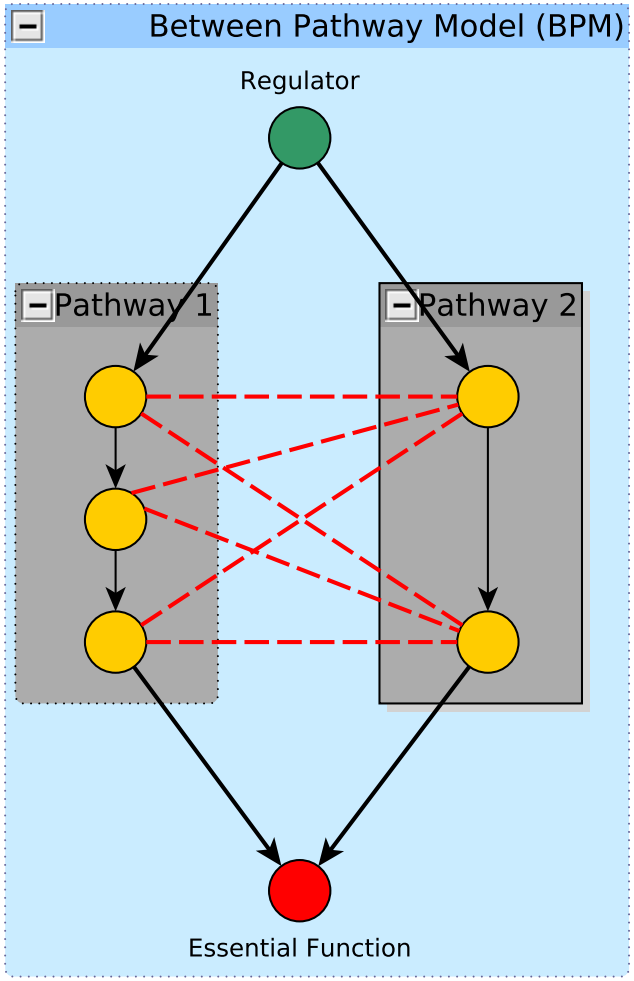
Model : Multiplicative null model

Organization of Genetic Interactions

Definition

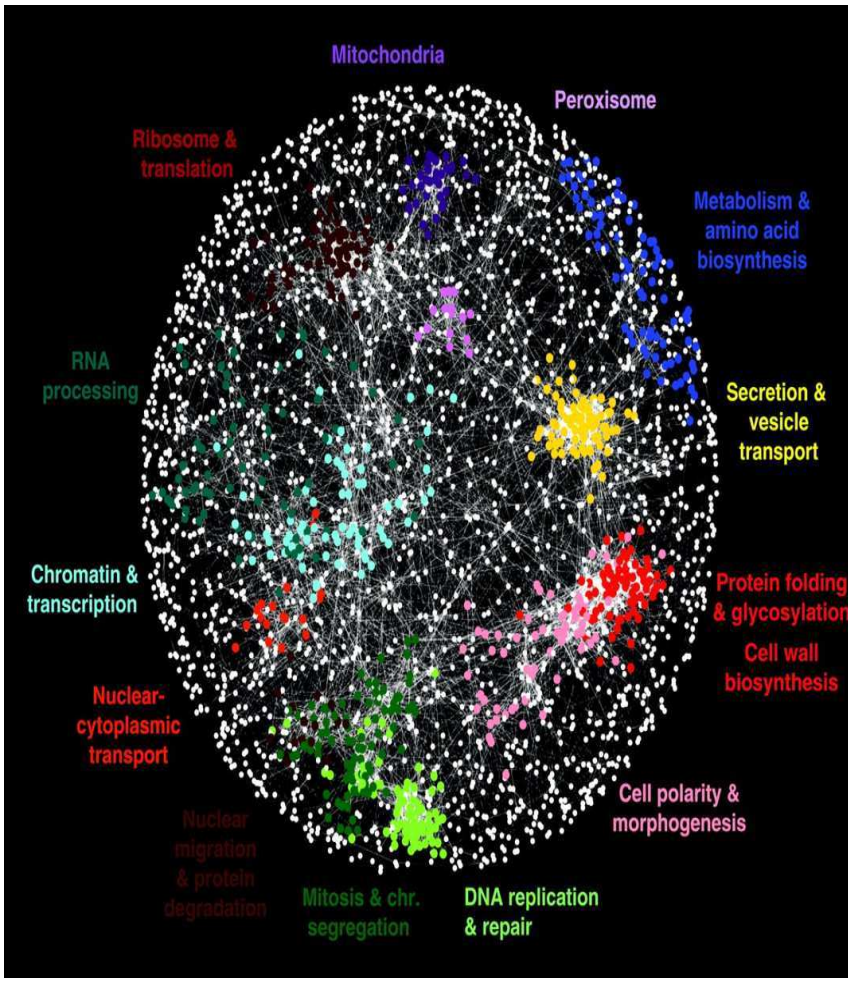
- Between-Pathway Model
  - Among genes participating in redundant functions
- Within-Pathway Model
  - Among genes with additive effect
- Indirect Effect
  - Among genes with distant functions that are not directly related

Between-Pathway Model (BPM)



- Bi-cliquish structure
- Have been used to:
  1. Predict co-pathway membership of gene pairs
  2. Extract redundant pathways

The Genetic Landscape of a Cell



Adopted from Costanzo et al., 2010

- Baker's yeast, *Saccharomyces cerevisiae*
- Synthetic Genetic Array (SGA)
- 1712 query genes
  1. 1378 null alleles of non-essential genes
  2. 334 hypomorphic or conditional alleles of essential genes
- 3885 array strains

Functional Annotations

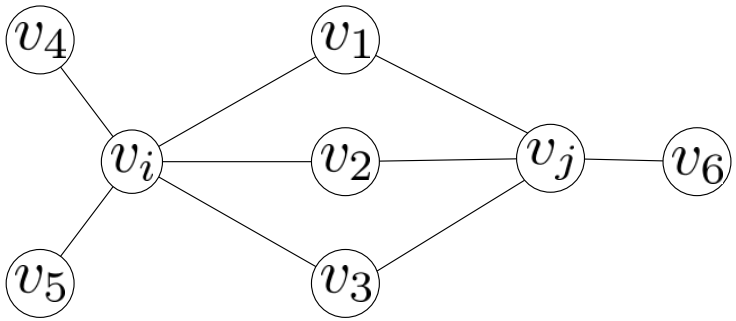
- KEGG Pathway Database
- Annotations for 1026 genes in the experiment
- 96 Pathways
  - 80 pathways after filtering pathways with less than 10 genes.

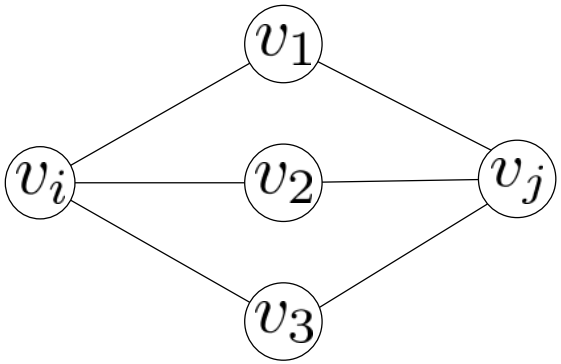
Local Neighborhood Similarity

A Predictor of Co-Pathway Membership

Similarity prediction methods

1. Number of Shared Neighbors
2. Congruence Score
3. Pearson Correlation of Interaction Profiles





Both v_i and v_j have three shared neighbors. However, in the first case their congruence score is almost 0.6, while in the second case it is approximately 2 (assuming a graph of size 10).

Evaluating Ranking Methods

Given a pathway P_A and a cut size (target set) l.

Definition

$$\begin{aligned} P-value(X=k) &= Prob(k \leq X) \\ &= HGT(k|N,N_A,l) \\ &= \sum_{x=k}^{min(N_A,l)} \frac{C(l,x)C(N-l,N_A-x)}{C(N,N_A)} \end{aligned}$$

X: Random variable denoting the number of true positives in a random sample, N: Total number of gene pairs, N_A: Number of gene pairs in pathway A, l: Size of target set

Minimum HyperGeometric (mHG) Score

Target size unknown:

Definition

The Minimum HyperGeometric (mHG) score

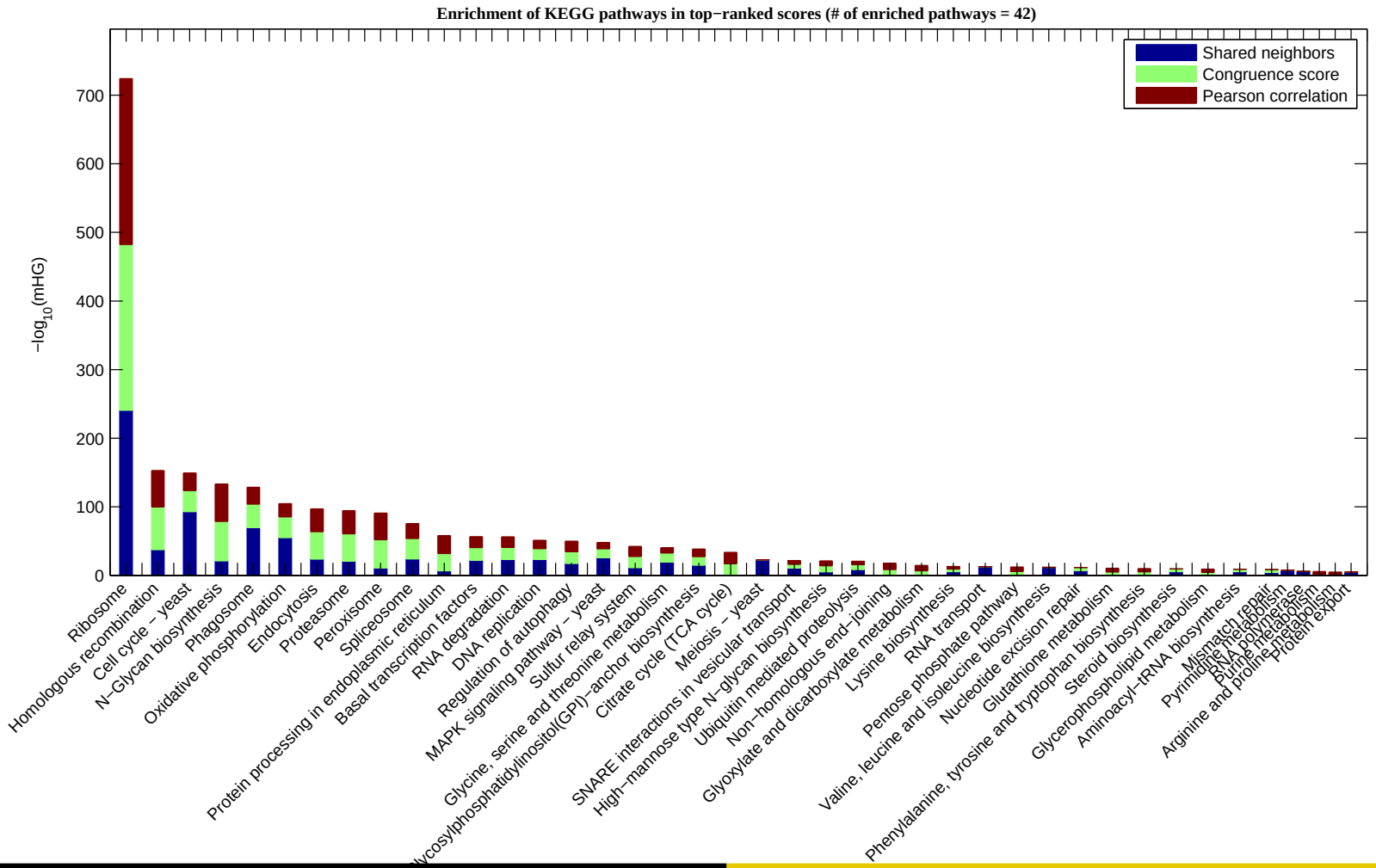
$$mHG(\lambda) = min_{1 \leq l \leq N} HGT(b_l(\lambda); N, N_A, l),$$

where $b_l(\lambda) = \sum_{i=1}^{l} \lambda_i$

λ_i is 1 if both of the genes in the i^{th} ranked gene pair are members of P_A, and 0 otherwise.

- mHG Adjusted for Multiple Comparison

Predictions Are Not Equally Accurate in Different KEGG Pathways



Highlights

Basic Idea

Heterogeneous performance of co-pathway membership predictions $\Longleftrightarrow$ Existence of specific structure around enriched pathways

- Decomposing neighborhood of each pathway
- Inferring lethal crosstalk among pathways

Modified Congruence Score (MCS)

Evaluating Neighborhood Overlap of Gene Pairs With Respect to a Given Pathway

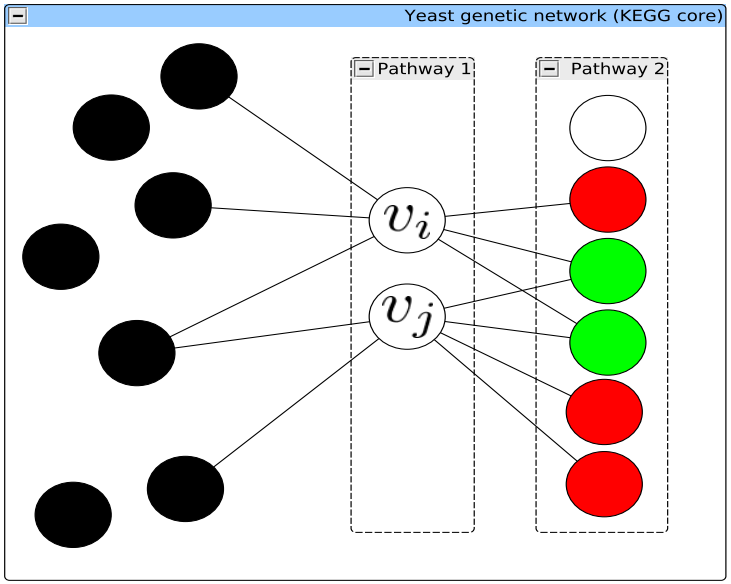
Definition

$$
\begin{aligned}
P - value(X = k_{ij}^B) &= Prob(k_{ij}^B \leq X) \\
&= HGT(k_{ij}^B | n_B, d_i^B, d_j^B) \\
&= \sum_{x=k_{ij}^B}^{\min(d_i^B, d_j^B)} \frac{C(d_j^B, x) C(n_B - d_j^B, d_i^B - x)}{C(n_B, d_i^B)}
\end{aligned}
$$

MCS is defined as $-log_{10}$ of the P-value.

Modified Congruence Score (MCS)
continued

Example



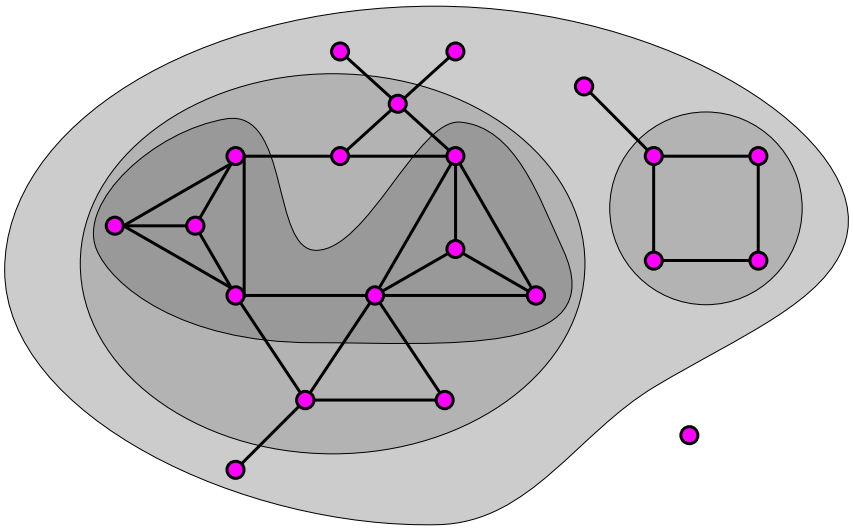
A sample neighborhood configuration for v_i and v_j. Here $n = 15, D_i = 6, D_j = 5, n_B = 6, d_i = 3, d_j = 4$ and $k = 2$.

Constructing Neighborhood Overlap Graph For a Given Pathway Pair

Definition

The neighborhood overlap graph (NOG) of a given pathway P_A with respect to pathway P_B, denoted by $H_{A \rightarrow B} = (V_H, E_H)$, is an unweighted, undirected graph defined over same vertices as P_A. In this graph, there is a link between vertices v_i and v_j if the network structure around them with respect to P_B is statistically significant .

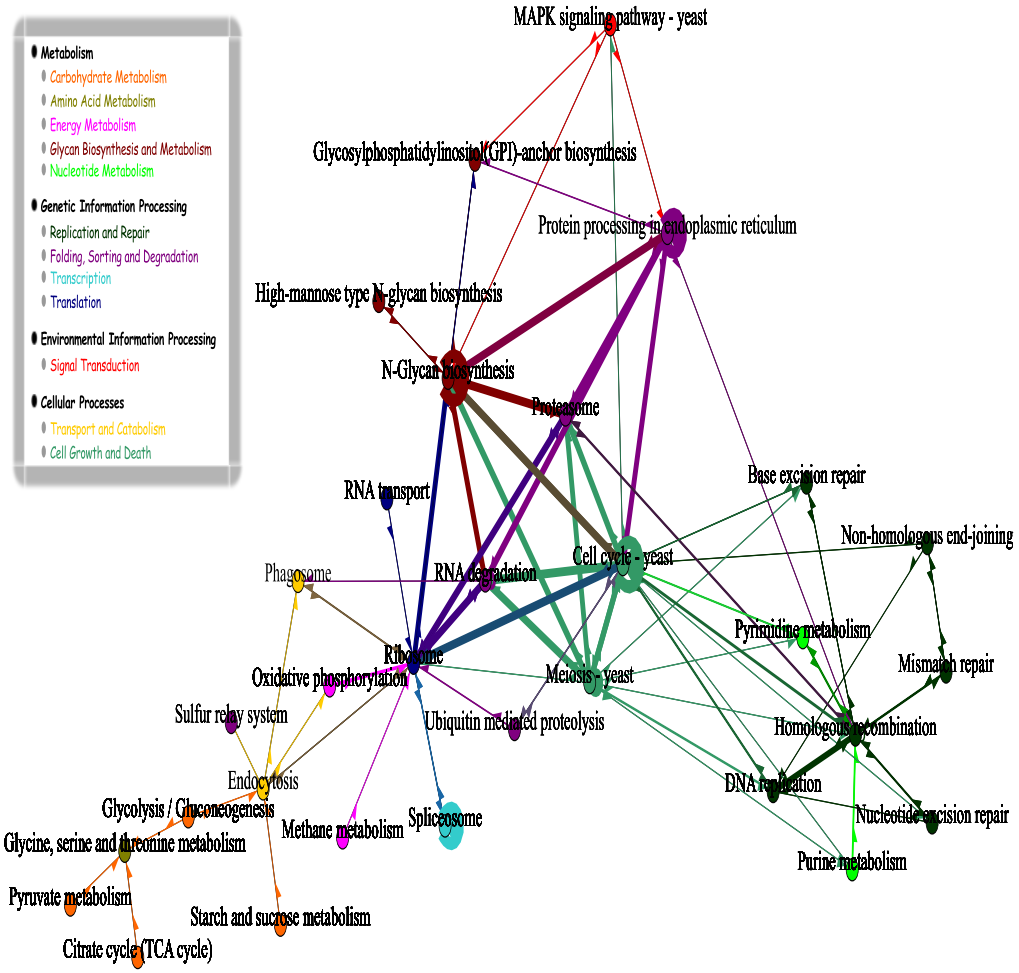
Pruning neighborhood overlap graph, finding cohesive subgraphs, and identifying interaction ports

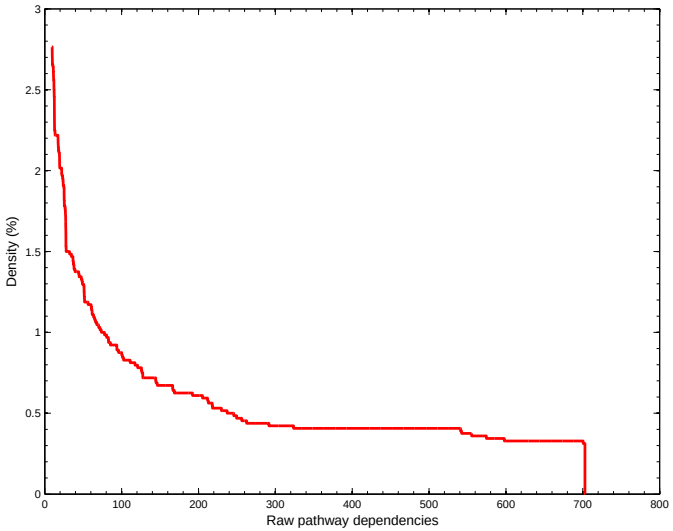


Adopted from Batagelj and Zaversnik, 2002

1. Iterative peeling of K-shells Pruning hairy components
2. Connected components in each core
3. Evaluating the significance of components
   - Evaluating significance using ER random graph model

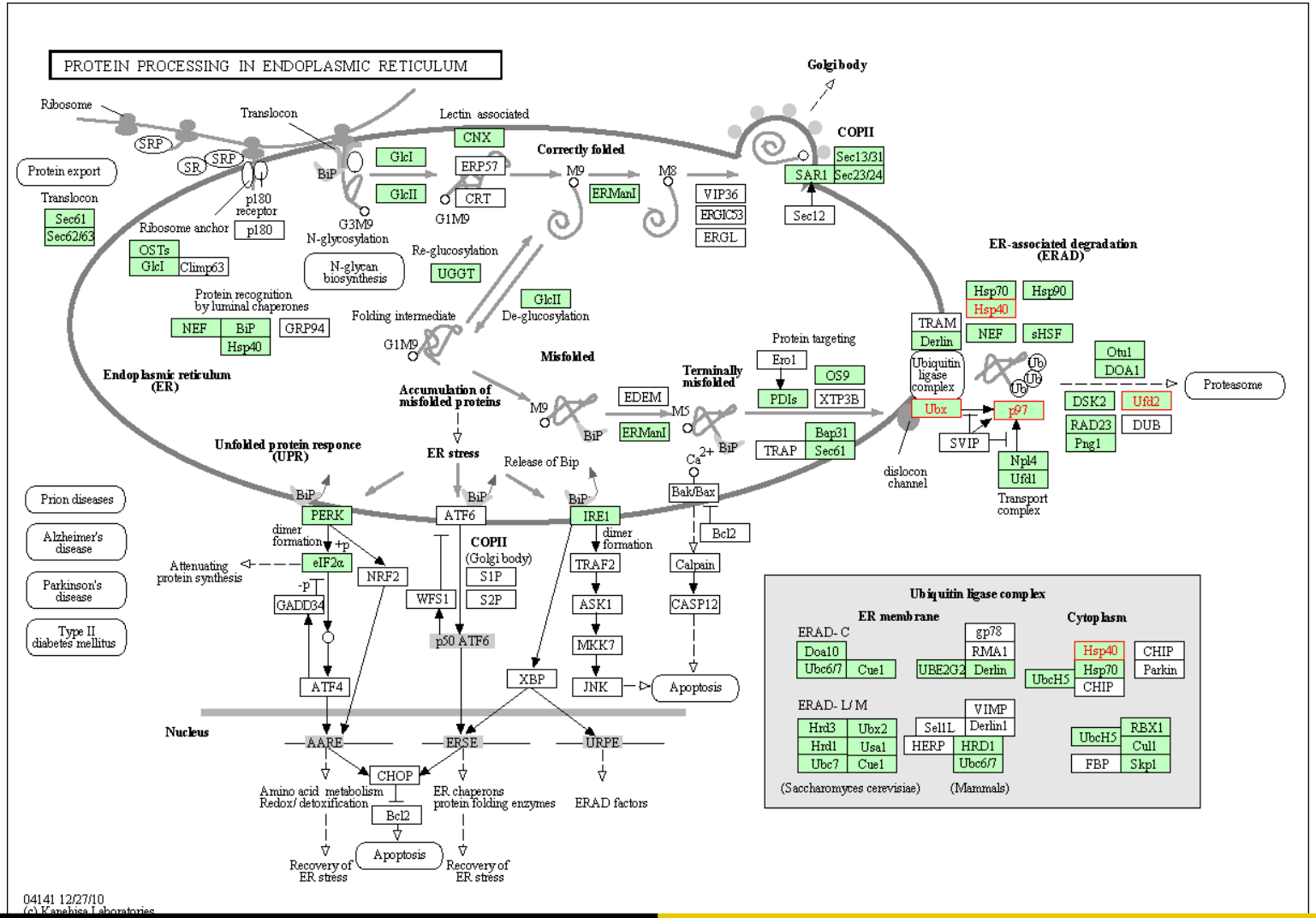
KEGG Crosstalk Map





Interaction Port Case Study

Crosstalk Between Protein Processing in ER and Proteasome



Summary

- The local neghborhood similarity gives heterogeneous performance in predicting co-pathways membership of gene pairs.
- This phenomena is due to the specific structure around enriched pathways
- Decomposing the neighborhood around each pathway sheds light on the cellular machinery.

- Future works:
  - Analysing the hierarchy of ports instead of the most significant interaction port.
  - Using our methodology to uncover dependencies among functional pathways.

For Further Reading I

M. Costanzo et al.
The Genetic Landscape of a Cell
Science, 425 2010.

SJ. Dixon et al.
Systematic Mapping of Genetic Interaction Networks
Annual review of genetics, 43, 601 2009.

D. Segre et al.
Modular Epistasis in Yeast Metabolism
Nature Genetics, 37, 77 2005.

C.L. Tucker and S. Fields
Lethal Combinations
Nature Genetics, 35, 204 2003.