Functional Characterization and Topological Modularity of Molecular Interaction Networks

Jayesh Pandey[1] Mehmet Koyutürk[2] Ananth Grama[1]

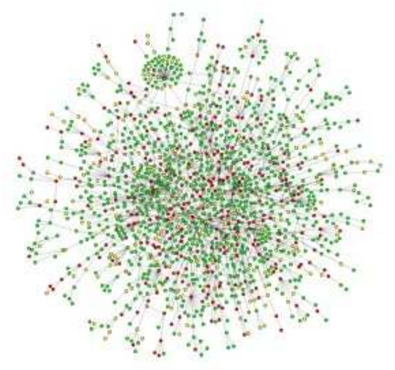
[1]Department of Computer Science
Purdue University

[2]Department of Electrical Engineering & Computer Science
Case Western Reserve University

Asia Pacific Bioinformatics Conference, 2010

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

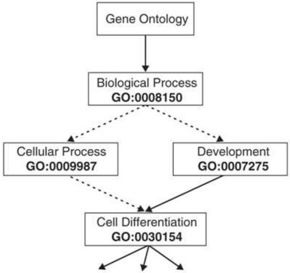


S.cerevisiae

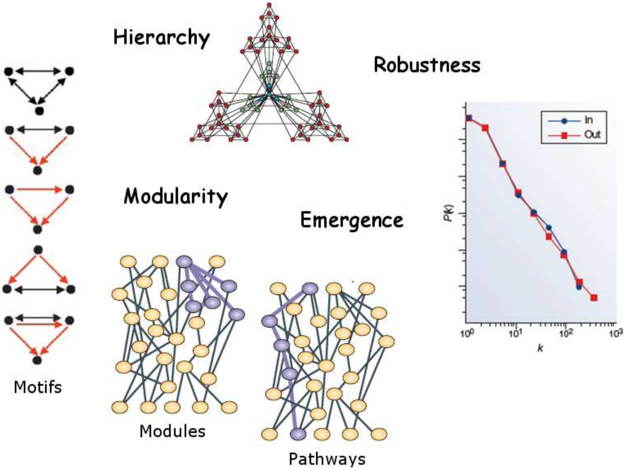
Protein-Protein Interaction (PPI) Network

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.

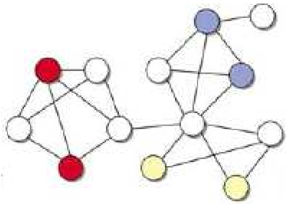


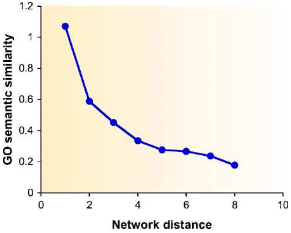
Function & Topology in Molecular Networks How does function relate to network topology?



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

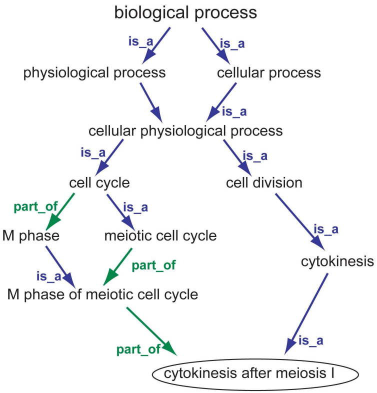




Sharan *et al.*, *MSB*, 2007

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)

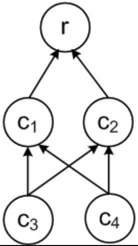


Semantic Similarity of GO Terms

- 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

- Each molecule (protein or domain) is associated with multiple GO 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)

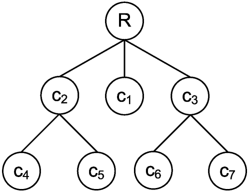
$$\Lambda(S_i, S_j) = \bigsqcup_{c_k \in S_i, c_l \in S_j} \lambda(c_k, c_l)$$

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

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

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)

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

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

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

where

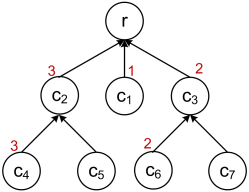
$$\Lambda(S_1, S_2, \ldots, S_n) = \bigsqcup_{c_{i_j} \in S_j, 1 \leq j \leq n} \lambda(c_{i_1}, c_{i_2}, \ldots, c_{i_n})$$

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

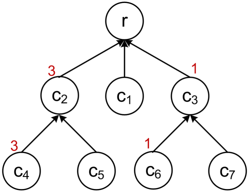
Weighted Information Content Weigh the information content of shared functionality by the number of molecules that contribute to the shared functionality

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

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

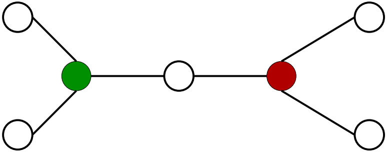


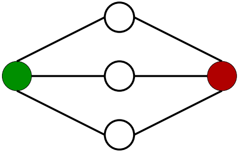
$\sigma_W(S_1, S_2, S_3) = 0.75$



Accounting for Multiple Paths

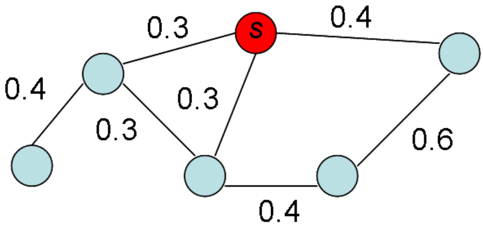
- Is "shortest path" a good measure of network proximity?
  - Multiple alternate paths might indicate stronger functional association
  - In well-studied pathways, redundancy is shown to play an important role in robustness & adaptation (*e.g.*, genetic buffering)





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*.



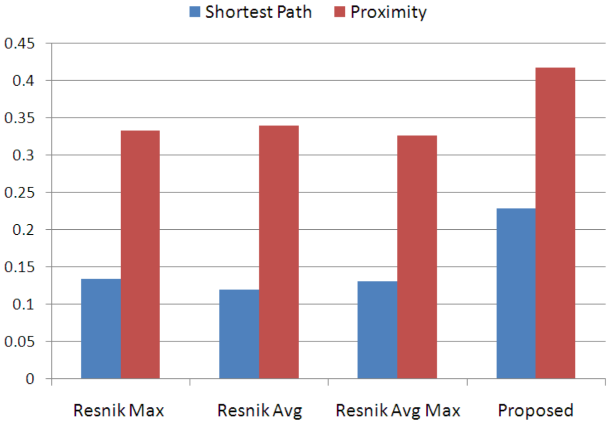
Proximity Based On Random Walks

- 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*

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

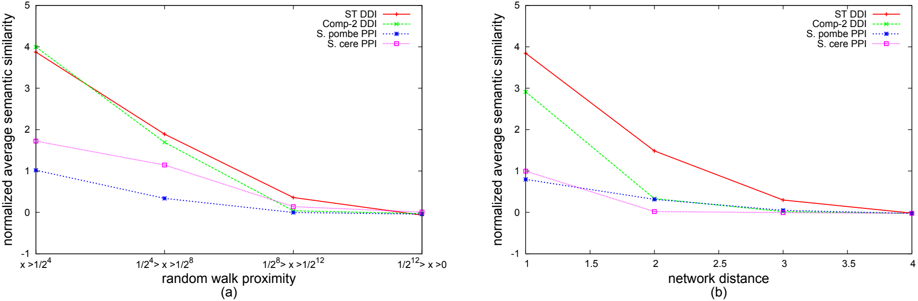
  - $\Phi(i,j)$: Network proximity between protein *i* and protein *j*
  - *A*: Stochastic matrix derived from the adjacency matrix of the network
  - *I*: Identity matrix
  - *c*: Restart probability
- Define proximity between proteins *i* and *j* as $\{\Phi(i,j) + \Phi(j,i)\}/2$

Network Proximity & Functional Similarity



Correlation between functional similarity and network proximity

Topological Proximity and Functional Similarity

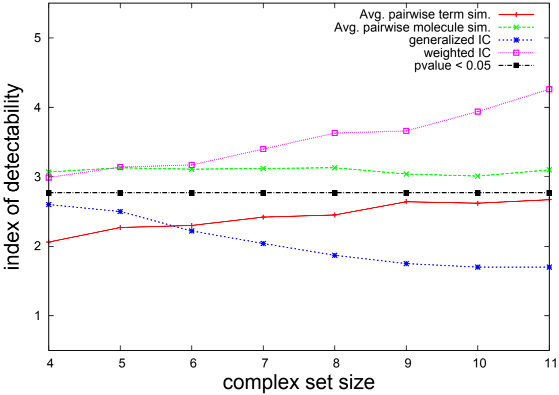


(a)

(b)

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

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

Conclusion & Ongoing work

- 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
- Future work : Using proximity measure to identify disease implicated genes in networks