Functional Coherence in Molecular Interaction Networks

Mehmet Koyutürk

Case Western Reserve University
Department of Electrical Engineering & Computer Science

Computer Science Colloquium
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
"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?
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
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
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)
Assessment of pairwise interactions is simple, but not adequate.
**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](image1)

![Functional Attribute Network](image2)
A pathway of functional attributes occurs in various contexts in the gene network.

Multipath in the functional attribute network.

Frequency of Multipath

Frequency: 4

Frequency: 0
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
Emphasize modularity of pathways

- Condition on frequency of building blocks
- Evaluate the significance of the coupling of building blocks

\[
\phi( \text{green} \rightarrow \text{red} \rightarrow \text{blue} ) = \phi( \text{green} \rightarrow \text{red} \rightarrow \text{yellow} ) = 4 \\
\phi( \text{green} \rightarrow \text{red} ) = \phi( \text{red} \rightarrow \text{yellow} ) = 2 \\
\phi( \text{red} \rightarrow \text{blue} ) = 5
\]

\[
P( \text{green} \rightarrow \text{red} \rightarrow \text{yellow} ) < P( \text{green} \rightarrow \text{red} \rightarrow \text{blue} )
\]
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

\[ p_{ij} = P(\Phi_{ij} \geq \phi_{ij} | B) = \sum_{\ell=\phi_{ij}}^{\min\{\beta_i\delta_j,n\}} \frac{\binom{\beta_i\delta_j}{\ell} \binom{m-\beta_i\delta_j}{n-\ell}}{\binom{m}{n}}. \]

- \( \beta_i = \) in-degree and \( \delta_i = \) out-degree
- \( m = \) pool of potential edges, \( n = \) number of edges in network
We denote each frequency random variable by $\phi$, their observed value by $\varphi$

$$\pi_{123}: \quad \Phi_1 \xrightarrow{} \Phi_{12} \xrightarrow{} \Phi_2 \xrightarrow{} \Phi_{23} \xrightarrow{} \Phi_3 \xrightarrow{} \Phi_{123}$$

Significance of pathway $\pi_{123}$ ( $p_{123}$ ) is defined as

$$P(\phi_{123} \geq \varphi_{123}|\phi_{12} = \varphi_{12}, \phi_{23} = \varphi_{23}, \phi_1 = \varphi_1, \phi_2 = \varphi_2, \phi_3 = \varphi_3)$$
Assume that interactions are independent

There are $\varphi_{12}\varphi_{23}$ possible pairs of $\pi_{12}$ and $\pi_{23}$ edges

The probability that a pair of $\pi_{12}$ and $\pi_{23}$ edges go through the same gene (corresponds to an occurrence of $\pi_{123}$) is $1/\varphi_2$

The probability that at least $\varphi_{123}$ 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
\[\prod_{j=1}^k \left| \bigcup_{g_\ell \in T_j} F(g_\ell) \right| \]
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
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
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 \( (p < 0.01) \)

<table>
<thead>
<tr>
<th>Pathway length</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em></td>
<td>143</td>
<td>753</td>
<td>1328</td>
</tr>
<tr>
<td><em>B. subtilis</em></td>
<td>22</td>
<td>78</td>
<td>202</td>
</tr>
<tr>
<td>Common</td>
<td>10</td>
<td>54</td>
<td>157</td>
</tr>
</tbody>
</table>
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!
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!
Annotation of Regulatory Pathways

Short-Circuiting Mediator Processes

Motility & response subnetworks reappear

DNA metabolism: gateway for response

Biosynthetic processes regulate metabolism & catabolism

Signaling regulates biosynthesis, transport, motility

New hubs: phosphorylation, macromolecule formation

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
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"?

![Graph showing GO semantic similarity vs. network distance]

Sharan et al., MSB, 2007
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:

\[ I(c) = -\log_2\left(\frac{|G_c|}{|G_r|}\right) \]

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

- \( G_c \): Set of molecules that are associated with term \( c \)
- \( r \): Root term
- \( A_i \): Ancestors of term \( C_i \) in the hierarchy
- \( \lambda(c_i, c_j) = \arg\max_{c \in A_i \cap A_j} I(c) \): Minimum common ancestor of \( c_i \) and \( c_j \)
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?
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
Generalize the concept of minimum common ancestor to sets of terms (Pandey et al., ECCB, 2008)

\[
\Lambda(S_i, S_j) = \bigcup_{c_k \in S_i, c_l \in S_j} \lambda(c_k, c_l)
\]

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

\[
G_{\Lambda(S_i, S_j)} = \bigcap_{c_k \in \Lambda(S_i, S_j)} G_{c_k}
\]

is the set of molecules that are associated with all terms in the MCA set
Illustration of Information Content Based Measure

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

- \( \lambda(c_4, c_4) = c_4, \)
- \( \lambda(c_6, c_4) = \lambda(c_7, c_4) = R \)

- \( \Lambda(S_1, S_2) = \{ c_4 \} \Rightarrow \rho_I(S_1, S_2) = \) 
  \[ - \log_2(|G_{c_4}|/|G_R|) = \log_2(5/4) \]

- \( \Lambda(S_1, S_3) = \{ c_4, c_6 \} \Rightarrow \rho_I(S_1, S_3) = \log_2(5/2) \)
Symmetry: Trivially, $\rho_I(S_i, S_j) = \rho_I(S_j, S_i)$ for all $S_i, S_j$.

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 $\rho_I(S_i, S_j) \leq \rho_I(S_i, S_i)$.

Monotonicity: Since $c_k \sim c_n$ for all $c_n \in S_i \cup S_j$, we have $\Lambda(S_i \cup c_k, S_j \cup c_k) = \Lambda(S_i, S_j) \cup \Lambda(S_i \cup S_j, \{c_k\}) \cup \{c_k\} \supseteq \Lambda(S_i, S_j) \cup \{c_k\}$, 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, $\rho_I(S_i \cup c_k, S_j \cup c_k) \geq \rho_I(S_i, S_j)$.

Generality:

$\Lambda(S_i, S_j \cup S_k) = \Lambda(S_i, S_j) \cup \Lambda(S_i, S_k) \supseteq \Lambda(S_i, S_j)$. 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
Functional Coherence & Network Proximity

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
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)
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 $j$

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

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

- Resnik Max
- Resnik Avg
- Resnik Avg Max
- Proposed

Comparison of Shortest Path and Proximity metrics.
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!

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

- \( \phi(j) \): 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

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