Gene Ontology-driven similarity for supporting the prediction of integrated functional networks

Francisco Azuaje

University of Ulster, UK
Goals of this research

- To propose a method to incorporate GO-driven information into the inference of functional networks
- To study their properties and relationships with other predictive resources
- To estimate its statistical and biological relevance

**Our hypotheses:**
- GO-driven similarity networks (GOSN) represent significant features of real functional networks
- These networks, in combination with other relevant predictive resources, may improve the overall predictive ability of integrated networks
Rationale:
Post-genome biology (systems biology)

- Networks of functional relationships between genes and proteins based on different properties or resources, e.g. gene co-expression.
- A node in a network represents a gene. A connection is established if the nodes are significantly associated.
- Overlaps between different types of relationships support the idea of combining them to build more meaningful networks.
- For example, physically interacting proteins are more likely to have similar gene expression patterns, etc.
Rationale:
The role of functional annotations

- Functional annotations of gene products (e.g. annotations derived from GO-driven databases) have been recently proposed to support network inference.

- The application of GO-derived information to support the prediction of functional networks of genes has not been rigorously investigated.

- Comprehensive studies on the predictive properties of such networks have not been reported.
This remaining of this presentation

- Brief introduction to the Gene Ontology (GO) and its applications
- Estimating functional similarity with the GO
- Constructing GO-driven similarity networks (GOSN)
- Integrating GOSN and other single-source networks
- Some relevant results
- Current/future work and conclusions
The Gene Ontology

- Provides structured, controlled vocabularies that can be used to describe gene products in different organisms

- GO hierarchies: *Molecular function* (MF), *biological process* (BP), and *cellular component* (CC).

- MF: The role played by individual gene products, e.g. *G-protein coupled receptor activity*.

- BP: Objective accomplished by one or more ordered assemblies of molecular function, e.g. *signal transduction*.

- CC: Cellular localization of the gene product, e.g. *nucleus* or *anaphase-promoting complex*. 
The GO

- GO terms and their relationships within each hierarchy form a network in which each term has one or more parent terms.
- The relationship between a child and its parent can be either “is a” (is a kind of) or “part of”.

Partial view of the GO Biological Process hierarchy
The GO and its applications

- Incorporation of GO annotations into gene expression data clustering analysis (significance of over-represented terms)
- Inference of gene-phenotype associations
- Assignment of new annotations to genes using gene expression and GO annotations
- Gold-standard in network prediction studies
- Predictive source for integrated network prediction
The GO and its applications (II)

- Estimating functional similarity using the GO and model organism databases annotated to GO (SGD, MGD, WB, etc.)

- Relationships between GO-driven similarity and sequence similarity, gene co-expression, functional interactions.

- We propose to build GOSN using non-traditional similarity assessment methods.
Approaches to computing GO-driven similarity

◆ Edge counting
  - Intuitive
  - Requires density to be homogeneous in the taxonomy

◆ Information-theoretic metrics
  - Grounded in information theory
  - Compensates for heterogeneity in the taxonomy

Slide courtesy of O. Bodenreider
Information-theoretic approaches

- **Information content (IC):** nodes high in the hierarchy have a small IC
- The information shared by two nodes can also be represented by their common ancestors (*least common subsumer*)
- The more information two terms share, the more similar they are
“Taxonomy”: hierarchy (DAG) of *is a* + *part of* relations

Frequency distribution of GO terms: annotation databases

Slide courtesy of O. Bodenreider
**GO-driven similarity**

- Based on the **information content of the least common subsumer (LCS)**
- Several variants
  - Resnik (1995)
  - Lin (1998)
  - Jiang & Conrath (1997)

Slide courtesy of O. Bodenreider

\[ \text{sim}(A, B) = \max_{LCS \in S(A,B)} \left[ -\log p(LCS) \right] \]

\[ \text{sim}(A, B) = \frac{2 \times \log p(LCS)}{\log p(A) + \log p(B)} \]

\[ \text{dist}(A, B) = \log p(A) + \log p(B) - 2 \times \log p(LCS) \]
GO-driven similarity among gene products

\[
SIM(g_i, g_j) = \frac{1}{m \times n} \times \sum_{c_k \in A_i, c_p \in A_j} \text{sim}(c_k, c_p)
\]

Slide courtesy of O. Bodenreider
Constructing GOSN (I)

- GO annotations from the SGD
- Annotations encoded in the GO Biological Process hierarchy
- 57,367 pairs of genes with significant mRNA expression correlations originating from a comprehensive compendium of microarray data
Constructing GOSN (II)

- *Low similarity network (LSN)*: a connection between a pair of genes was established if their GOS was larger than 0 under the Biological Process hierarchy.

- *Medium similarity network (MSN)*: a connection between a pair of genes was established if their GOS was larger or equal to 0.5.

- *High similarity network (HSN)*: a connection between a pair of genes was established if their GOS was larger or equal to 0.8.

- *Very high similarity network (VHSN)*: a connection between a pair of genes was established if their GOS was equal to 1.
Constructing GOSN (II)

GOS networks vs. random networks (Mean similarity ± S.E)

<table>
<thead>
<tr>
<th>GOS networks</th>
<th>Random networks</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LSN</strong></td>
<td>0.374 ± 2.0E-03</td>
<td>0.150 ± 4.80E-04</td>
</tr>
<tr>
<td><strong>MSN</strong></td>
<td>0.857 ± 2.0E-02</td>
<td>0.289 ± 1.0E-03</td>
</tr>
<tr>
<td><strong>HSN</strong></td>
<td>0.98 ± 6.0E-4</td>
<td>0.48 ± 2.7E-03</td>
</tr>
<tr>
<td><strong>VHSN</strong></td>
<td>1.0 ± 0.0</td>
<td>0.594 ± 2.0E-03</td>
</tr>
</tbody>
</table>

S.E: Standard error, Sig.: Significance of the difference (Student’s t test).
Other networks integrated (SSN)

- **SGA** network (genetic interactions) (Tong et al., 2004).
- **Homol** network: protein similarity (Altschul et al., 1997) (Zhang et al., 2005).
- **Coex** network: Highly co-expressed pairs of genes (Hughes et al., 2000).
- **Physic** network: pairs of proteins belonging to the same protein complex (Mewes et al., 2002; Gavin et al., 2002; Ho et al., 2002).
- **Chip** network: transcription factor-gene interactions (Tong et al., 2004).
Different integrated networks obtained by merging all types of single-source relationships (union of networks).

Four networks were first obtained: $intLSN$, $intMSN$, $intHSN$ and $intVHSN$, which were derived from the combination of the SSN with $LSN$, $MSN$, $HSN$ and $VHSN$ respectively.
Construction of integrated networks (II)

- Reference integrated network, *intNonGOS*, which did not incorporate the GOS networks.

- *Multiple-support integrated networks*, i.e. edges supported by at least two types of functional interactions; e.g. *intMSN-MS* is a multiple-support, integrated network that incorporates the *MSN*. 
Detection of potential functional modules through network clustering

<table>
<thead>
<tr>
<th>Network</th>
<th>NC</th>
<th>AC-score</th>
<th>AID</th>
<th>ANP</th>
<th>NC-score-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSN</td>
<td>51</td>
<td>3.83</td>
<td>91.56</td>
<td>12.41</td>
<td>9</td>
</tr>
<tr>
<td>HSN</td>
<td>36</td>
<td>3.85</td>
<td>84.64</td>
<td>10.97</td>
<td>8</td>
</tr>
<tr>
<td>VHSN</td>
<td>32</td>
<td>3.99</td>
<td>90.25</td>
<td>11.25</td>
<td>7</td>
</tr>
<tr>
<td>intLSN</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>intMSN</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>intHSN</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>intVHSN</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>intLSN-MS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>intMSN-MS</td>
<td>53</td>
<td>3.96</td>
<td>99.26</td>
<td>15.54</td>
<td>11</td>
</tr>
<tr>
<td>intHSN-MS</td>
<td>51</td>
<td>3.39</td>
<td>75.18</td>
<td>13.16</td>
<td>9</td>
</tr>
<tr>
<td>intVHSN-MS</td>
<td>52</td>
<td>3.30</td>
<td>71.90</td>
<td>12.11</td>
<td>9</td>
</tr>
<tr>
<td>intNonGOS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>intNonGOS-MS</td>
<td>38</td>
<td>2.92</td>
<td>41.23</td>
<td>10.21</td>
<td>5</td>
</tr>
</tbody>
</table>

NC: Number of clusters; AC-score: Average MCODE cluster score; AID: Average interaction density per cluster; ANP: Average number of proteins per cluster; NC-score-5: Number of clusters with MCODE cluster scores greater than 5.
## Linking networks to significant functional categories and pathways (I)

*intNonGOS-MS*: Linking clusters to MIPS functional categories and KEGG pathways

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Sample of significantly-represented MIPS functional categories (number of proteins)</th>
<th>Associations with KEGG pathways (number of proteins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Stress response (9); Extracellular/secretion protein (1); Cell membrane (1); Unclassified (12)</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Regulation of splicing (1); Ribosome biogenesis (25); Ribosomal proteins (25); Nucleic acid binding (6); RNA binding (6).</td>
<td>Ribosome (25)</td>
</tr>
<tr>
<td>3</td>
<td>Transposable elements, viral and plasmid proteins (19)</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Transcription (9); rRNA processing (8); Ribosome biogenesis (13); Ribosomal proteins (5); Nucleic acid binding (6); RNA binding (6).</td>
<td>Ribosome (1)</td>
</tr>
<tr>
<td>5</td>
<td>DNA processing (6); DNA synthesis and replication (6); DNA topology (6); DNA recombination (6); Stress response (6); Biogenesis of nucleus (6); Organization of chromosome structure (6)</td>
<td>-</td>
</tr>
</tbody>
</table>
Linking networks to significant functional categories and pathways (II) (5 out of 11 clusters)

*intMSN-MS*: Linking clusters to MIPS functional categories and KEGG pathways

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Examples of significantly represented MIPS functional categories (number of proteins)</th>
<th>Associations with KEGG pathways (number of proteins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Extracellular/secretion proteins (1); Cell membrane or cell wall attached (1); Unclassified proteins (32)</td>
<td>Galactose metabolism (2); Starch and sucrose metabolism (2);</td>
</tr>
<tr>
<td>2</td>
<td>rRNA processing (35); ribosome biogenesis (17); ribosomal proteins (8); nucleic acid binding (18); RNA binding (18); Nucleotide binding (7); ATP binding (7);</td>
<td>Ribosome (2)</td>
</tr>
<tr>
<td>3</td>
<td>Ribosome biogenesis (34); ribosomal proteins (34); nucleic acid binding (5); RNA binding (5).</td>
<td>Ribosome (36)</td>
</tr>
<tr>
<td>4</td>
<td>Amino acid metabolism (29); Assimilation of ammonia (6); Metabolism of glutamine (1); Degradation of glutamine (1); Metabolism of arginine (5); Biosynthesis of arginine (5); Metabolism of urea cycle (2); Metabolism of the aspartate family (9); Metabolism of threonine (3); Metabolism of methionine (4); Metabolism of serine (3); Metabolism of the pyruvate family (5); C-compound, carbohydrate anabolism (7); Secondary metabolism (7); Complex cofactor/cosubstrate binding (6)</td>
<td>Valine, leucine and isoleucine biosynthesis (4); Lysine biosynthesis (5); Phenylalaine, tyrosine and triptophan biosynthesis (8)</td>
</tr>
<tr>
<td>5</td>
<td>Unclassified proteins (20)</td>
<td>Galactose metabolism (1); Pentose and glucorate interconversion (1),</td>
</tr>
</tbody>
</table>
Future work

- Improve cluster interpretation and validation
- Other similarity assessment methods
- Cluster-based assignment of function to uncharacterized genes
- Other integration (machine learning) methods
- Different model organisms
- Other applications of GO-driven similarity: Co-expression validity assessment, relationship with other functional properties.
Summary

- A method to reconstruct networks using similarity information extracted from the GO and the *Saccharomyces* Genome Database (SGD).
- GOSN represent significant features of real functional networks.
- These networks, in combination with other relevant predictive resources, have the potential to improve the overall predictive ability of integrated networks.
- Integrated networks comprising GOS relationships contain more meaningful clusters than those ignoring GOS-based evidence.
Acknowledgments

◆ LHNCBC, NLM, NIH:
Dr. Donald King (Acting Director)
May Cheh and Rob Logan (Program Coordinators)

Research collaborators:
◆ Olivier Bodenreider, NLM, NIH
◆ Haiying Wang and Huiru Zheng, UU
◆ Alban Chesneau, EMBL-Grenoble

Contact, additional information:
fj.azuaje@ieee.org
http://ijsr32.infj.ulst.ac.uk/~e10110731