Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation

by P. W. Lord, R. D. Stevens, A. Brass and C. A. Goble
Bioinformatics 19(10) 1275–1283
http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/10/1275

presented by Christopher Maier for INLS 279: Bioinformatics Research Review
2006-02-01
Overall Concept

• Use the addition of ontological annotations to create a new search layer on top of biological databases: semantic querying, to find entries that “mean” the same thing
What is an Ontology?
“A Conceptualization of a Specification”

- Originally a tool from philosophy to convey the existence and relationships of all that exists
- Now used as a formal method to define important concepts and relationships in a particular domain
- More powerful than controlled vocabularies due to added logical infrastructure; more powerful than taxonomies due to additional relationships
The Gene Ontology

• Contains three different “sub-ontologies”: molecular function, cellular component, and biological process

• 20,349 total terms as of December 2005

• Annotations in numerous databases

Defining and Validating Semantic Similarity
Approaches to Ontological Similarity

- Path Distance
- Depth

These approaches don’t seem to perform well in the biological domain
Figure 1

GO Fragment

- Molecular function: GO:0003674 p = 1
  - signal transducer: GO:0004871 p = 0.208
    - chaperone: GO:003754 p = 0.0102
  - receptor-associated protein: GO:0016962 p = 0.00159
  - receptor signaling protein: GO:0005057 p = 0.0281
  - receptor: GO:0004872 p = 0.124
  - ligand: GO:0005102 p = 0.0460

- Transmembrane receptor: GO:0004888 p = 0.0997
- Photoreceptor: GO:0009881 p = 0.000433
Our Definition of Similarity

- Count number of times a term appears (including implicit appearances due to subsumption relationships)
- The less frequent a term, the more informative it is
- Probability of the minimum subsumer for multiple parentage
- Similarity is a negative log function
Validation of Semantic Similarity

• Hard to use traditional validation approaches

• See if sequence similarity tracks with semantic similarity
Why Sequence Similarity?

• Properties of biological macromolecules such as DNA and proteins ultimately derive from their sequence

• Thus, proteins with very similar sequence will generally fold into a very similar 3D shape, allowing them to perform similar functions

• This serves as an empirical measure of similarity, against which our ontological measure can be proven
Adapting to SWISS-PROT

- Orphan Terms
  - “part-of” terms do not participate in “is-a” relationships!
  - Link these back to the ontology root, despite semantic impoverishment

- Link Type Bias
  - Large majority of “molecular function” is “is-a”; over half of “cellular component” is “part-of”

- Multiple Annotations
  - Take average
Figure 2  Similarity Correlations in GO
Figure 3: Similarity and Evidence Codes
Figure 4: Correlation with links removed
Outliers

• Polymorphic groups: different proteins participate in the same process

• Hyper-variable families

• Mis-annotations

• Under-annotation
Application: Semantic Search
Search

- Utilize semantic similarity to provide alternative search axes
- Each of the three sub-ontologies of GO retrieves a different kind of “similar” proteins
<table>
<thead>
<tr>
<th>Switzerland ID</th>
<th>Description</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Molecular Function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPSG.HUMAN</td>
<td>Green-sensitive opsin (Green cone photoreceptor pigment).</td>
<td>8.15</td>
</tr>
<tr>
<td>OPN4.HUMAN</td>
<td>Opsin 4 (Melanopsin).</td>
<td>7.23</td>
</tr>
<tr>
<td>OPSB.HUMAN</td>
<td>Blue-sensitive opsin (Blue cone photoreceptor pigment).</td>
<td>4.92</td>
</tr>
<tr>
<td>5H6.HUMAN</td>
<td>5-hydroxytryptamine 6 receptor (Serotonin receptor)</td>
<td>3.92</td>
</tr>
<tr>
<td>A1AA.HUMAN</td>
<td>Alpha-1A adrenergic receptor (Alpha 1A-adrenoceptor)</td>
<td>3.92</td>
</tr>
<tr>
<td>A1AB.HUMAN</td>
<td>Alpha-1B adrenergic receptor (Alpha 1B-adrenoceptor).</td>
<td>3.92</td>
</tr>
<tr>
<td>(b) Biological Process</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIPL1.HUMAN</td>
<td>Aryl-hydrocarbon interacting protein-like 1.</td>
<td>2.89</td>
</tr>
<tr>
<td>CNCG.HUMAN</td>
<td>Retinal cone rhodopsin-sensitive cGMP</td>
<td>2.89</td>
</tr>
<tr>
<td>CNRA.HUMAN</td>
<td>Rod cGMP-specific 3',5'-cyclic phosphodiesterase</td>
<td>2.89</td>
</tr>
<tr>
<td>CNRC.HUMAN</td>
<td>Cone cGMP-specific 3',5'-cyclic phosphodiesterase</td>
<td>2.89</td>
</tr>
<tr>
<td>CNRD.HUMAN</td>
<td>Retinal rod rhodopsin-sensitive cGMP</td>
<td>2.89</td>
</tr>
<tr>
<td>CRB1.HUMAN</td>
<td>Beta crystallin B1.</td>
<td>2.89</td>
</tr>
<tr>
<td>(c) Cellular Component</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A01.HUMAN</td>
<td>HLA class I histocompatibility antigen</td>
<td>1.86</td>
</tr>
<tr>
<td>5H1A.HUMAN</td>
<td>5-hydroxytryptamine 1A receptor (5-HT-1A)</td>
<td>1.86</td>
</tr>
<tr>
<td>A1A2.HUMAN</td>
<td>Sodium/potassium-transporting ATPase alpha-2 chain</td>
<td>1.86</td>
</tr>
<tr>
<td>A1AA.HUMAN</td>
<td>Alpha-1A adrenergic receptor</td>
<td>1.86</td>
</tr>
<tr>
<td>A33.HUMAN</td>
<td>Cell surface A33 antigen precursor</td>
<td>1.86</td>
</tr>
<tr>
<td>ACHA.HUMAN</td>
<td>Acetylcholine receptor protein</td>
<td>1.86</td>
</tr>
</tbody>
</table>

**Table 4**

Semantic Search Results
Conclusion
What have we learned?

• Semantic similarity is valid concept
• Ontology structure adds value above controlled vocabulary
• Possible uses: semantic search, error detection
The Future

• As GO grows both in size and in use, the value of semantic searching on GO annotations will increase

• What other similarity functions could be used?

• Are there other measures with which cellular component and biological process similarity are correlated?