**CSI 5V93, Project 1, Developing and Evaluating Similarity Measures for Hierarchical Data**

**Report Submission Due Date: 11:59pm, March 5, 2019**

|  |
| --- |
| **Background** |
| Gene Ontology (GO) is one of the most widely used ontology databases, and provides a framework to elucidate biological roles of genes by semantic analysis. GO contains terms in a structured format within three domains. GO also provides extensive annotation data across most model species. Recent studies have proposed that functional similarity between genes can be measured by semantic similarity between GO terms in a hierarchy. |

|  |
| --- |
| **Description** |
| 1. Download the most recent version of human annotation data. Label the nodes in BP and MF ontologies using annotating genes, but exclude the annotations of the "IEA" evidence code and the annotations "not" qualified. 2. Read the filtered human PPI (protein-protein interaction) data provided - [August 2018 version in BioGRID](http://web.ecs.baylor.edu/faculty/cho/5V93/biogrid_human_ppi_cln.txt). If an interacting pair was not annotated together to BP terms or MF terms, then remove the pair. 3. **Implement four semantic similarity measures**:    * normalized depth-based measure (edge-based)    * normalized common anscestors-based measure by Jaccard index (node-based)    * normalized information theoretic measure    * integrative measure of node-based and information theoretic. 4. Apply semantic similarity measures to quantify functional similarity of each PPI.    * For pairwise approaches, aggregate term-to-term semantic similarity scores by best-match averaging.    * Measure semantic similarity on BP and MF ontologies, respectively, and select a larger score between BP and MF for each PPI. 5. **Develop a new semantic similarity measure** for PPIs. 6. **Evaluate the performance of your measure** and the four existing methods of semantic similarity.    * Use the human protein complex data as ground truth - [combined version of human complexome](http://web.ecs.baylor.edu/faculty/cho/5V93/human_protein_complexes.txt). Assign a PPI to a positive data set if the two proteins occur together in at least one protein complex. Assign a PPI to a negative data set otherwise.    * Measure the area under the ROC curve (AUC) by changing the semantic similarity threshold to predict whether each PPI is positive or negative. Compare the AUC values of five semantic similarity measures. |

|  |
| --- |
| **Submission** |
| * A report in a research paper style by email attachment (due March 5, Tuesday) |

|  |
| --- |
| **Presentations** |
| * Presentation of a recent research paper (previous works) related to this project (on February 5, Tuesday) * Presentation of your proposed measure, evaluation design, and results (on March 5, Tuesday) |

|  |
| --- |
| **Related papers** |
| * Required reading   + Lord et al., "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation", *Bioinformatics* (2003)   + Pesquita et al., "Semantic similarity in biomedical ontologies", *PLoS Computational Biology* (2009)   + Wang et al. "Revealing and avoiding bias in semantic similarity scores for protein pairs", *BMC Bioinformatics* (2010)   + Cho et al. "M-Finder: Uncovering functionally associated proteins from interactome data integrated with GO annotations". *Proteome Science* (2013) * Optional reading   + Alvarez et al. "A shortest-path graph kernel for estimating gene product semantic similarity", *Journal of Biomedical Semantics* (2011)   + Benabderrahmane et al. "IntelliGO: a new vector-based semantic similarity measure including annotation origin", *BMC Bioinformatics* (2010)   + Teng et al. "Measuring gene functional similarity based on group-wise comparison of GO terms", *Bioinformatics* (2013)   + Vafaee et al. "Novel semantic similarity measure improves an integrative approach to predicting gene functional associations", *BMC Systems Biology* (2013)   + Wu et al. "Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge- and IC-based hybrid method", *PLoS One* (2013)   + Song et al. "Measure the semantic similarity of GO terms using aggregate information content", *IEEE/ACM TCBB* (2013)   + Jain et al. "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology", *BMC Bioinformatics* (2010) |

1 If one node is annoted as gene, all its parents is also annoted..