Title: Tools and Analysis of PANTHER Genomic Data Annotations

Authors: M. Fong[1], I. Holmes[1]

Affiliations: [1]Department of Bioengineering, University of California, Berkeley

*Correspondence to: mfong92@berkeley.edu

The PANTHER (Protein ANalysis THrough Evolutionary Relationships) database contains gene families, trees, and functional annotations and is located at http://pantherdb.org. This following discussion is a description of a tool created for retrieving these annotations and an analysis of the number and quality of the annotations that are provided through the PAINT tool. The results indicate a possible deficiency in the annotations that are provided through the Gene Ontology (GO).

Two scripts were created for this analysis: the **Branch Generation Script**, which parses Newick tree files and returns a list of branches (as tuples containing two AN numbers and a branch length), and the **Annotation Mapping Script**, which translates each AN number to the corresponding GO annotations that are curated on each node. This was done using the resources from PANTHER and GO online repositories.

The summary of the analysis can be seen in Figure 1. This gives a general idea about the prevalence of GO annotations within the PANTHER data set. Of the 17761 total nodes over the 53 different families that were surveyed, only about 25% of them were annotated, including nodes that had annotations solely because they were propagated from their ancestors.

Figure 2 brings into question the actual usefulness of the PANTHER annotations for specific use cases that assume that they are a random sampling of all nodes. Given the strategy for prioritizing annotations that appears to have been using in constructing PANTHER, this data set might not be useful as a training set for evolutionary models, as well as downstream applications of those models. The left columns show the unique number of annotation sets across entire gene families, where a set of annotations is counted as the same as another set if all of the annotations of one compose of all of the annotations of the other. Only 5% of the total nodes have unique sets of annotations, which suggest that the majority of annotations come from the propagation down towards a node's descendants. The right columns demonstrate a potential problem with the above claim. Logically, it would be assumed that nodes that share an annotation would be closer together, on average, because of the effect that distance has on evolution. However, the opposite is true of the gene families shown in Figure 3, as the nodes that share an annotation are further apart than nodes that don't share an annotation. Here, distance is calculated as distance that is necessary to have been traveled through the tree. Although this is not a perfect calculation of genetic distance, this finding does point again towards the possible lack of quality of these annotations.

As it stands, the annotations that are curated to the PANTHER trees appear not to be on a random selection of nodes, but on groups of nodes where the annotations are the same. More analysis needs to be done to see whether this currently is a useful data set to conduct evolutionary comparison analyses.