Genome Function Phylogenetics

Dennis Psaroudakis

April 18th 2019

Introduction, Gene Ontology, Data...

Method

Starting point of our method are the functional annotation sets, one for each genome, which annotate every gene in the genome with one or more GO terms. In more mathematical terms the genome annotation set is a list of tuples (G, T) with $G \in$ Genes in that genome and $T \in$ Terms in the Gene Ontology.

We can use the hierarchical structure of the Gene Ontology to obtain the ancestors A_i of any term T_i; in other words the gene G_i is not just annotated with the term T_i itself but also with all GO terms that are a more general statement of that term (e.g. any gene that is part of a metabolic process is thereby also part of a biological process). We do that for all terms T in the dataset and combine all of the terms and their ancestors into one big genome-wide set S, irrespectively of the gene they were originally associated with: $S = \bigcup_{i=1}^{x}(T_i \cup A_i)$.

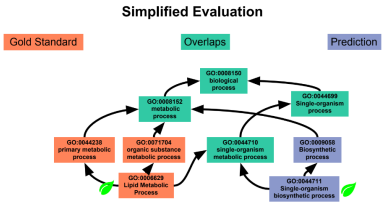
When this superset of annotations is created for each of the datasets, we can use the Jaccard Distance as a measure of how (dis-)similar any two sets are from each other, or in biological terms how different the two genomes are on a functional level:

$$\text{Jaccard Distance}(S_a, S_b) = 1 - \frac{|S_a \cap S_b|}{|S_a \cup S_b|}$$

Applying this formula to all pairwise combinations of the genomes we're looking at yields a $S \times S$ distance matrix that can then serve as the input for a neighbor joining algorithm.

Gene	GO Term
Os01g0601625	GO:0050896
Os01g0601625	GO:0016021
Os01g0601625	GO:0016301
Os01g0601651	GO:0003677
Os01g0601651	GO:0009699
Os01g0601651	GO:0050790
Os01g0601651	GO:0050794
Os01g0601651	GO:0050896
Os01g0601675	GO:0007275
Os01g0601675	GO:0016310
Os01g0601675	GO:0050789
...	

The general idea of using the Jaccard Distance in this context is to measure the overlap of two subtrees in the GO hierarchy. Say, for simplicity, that we're looking at two genomes (here called Gold Standard and Prediction) that each only contain one single GO term (marked by a leaf). First, we add all ancestors of that leaf term to each subtree. Then, we determine the overlap (which corresponds to $S_a \cap S_b$), and divide the number of nodes in this overlap by the number of nodes in either of the two subtrees ($S_a \cup S_b$).



In the case of this example, the Jaccard Distance of Gold Standard and Prediction would be $1 - \frac{4}{9} = \frac{5}{9}$