Gene Function Phylogenetics

Dennis Psaroudakis

April 18th 2019

Introduction, Data...

Method

Starting point of our method is a functional annotation set for each genome S_a, consisting of tuples (G_i, T_i) with $G_i \in$ Genes in genome S_a and $T_i \in$ Terms in the Gene Ontology. So the annotation set maps each gene of the genome to n GO terms.

Now we use the structure of the Gene Ontology to obtain all ancestors A_i of GO term T_i in the GO hierarchy and combine all these annotations and their ancestors, irrespective of the Gene they are associated with: $\bigcup_{i=1}^{x}(T_i \cup A_i)$ The resulting set is a set of all the GO terms present in the genome annotation and their ancestors.

After this is done for all annotation sets, we introduce the Jaccard distance as a metric of (dis-)similarity between the annotation sets:

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

Then finally I can apply Neighbor-Joining to the distance matrix I generated this way.

```
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
...
```

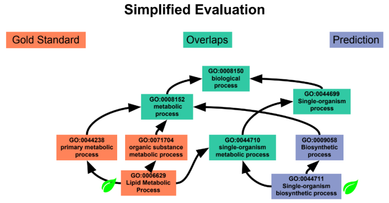
Simplified Evaluation
Gold Standard
Overlaps
Prediction
GO:0008150 biological process
GO:0008152 metabolic process
GO:0044699 Single-organism process
GO:0044238 primary metabolic process
GO:0071704 organic substance metabolic process
GO:0044710 single-organism metabolic process
GO:0009058 Biosynthetic process
GO:0006629 Lipid Metabolic Process
GO:0044711 Single-organism biosynthetic process

Figure 1: **(Zoom in)**. Graphic of how we use the Gene Ontology tree to calculate the overlap (= intersection) of two GO terms (the ones marked with a leaf)