Genome Function Phylogenetics

Dennis Psaroudakis

April 18th 2019

When we build phylogenetic trees, we often use the sequences of certain genes as the basis for our tree. They are the direct substrate of evolution so that makes sense, but when we think of evolution on a macro level, we don't really think of the sequences, we think of phenotype. Using phenotypes for evolutionary trees is a hard thing to do though because the choice of characteristics can be kind of arbitrary (http://wiki.c2.com/?BorgesClassificationOfAnimals Chinesische Tiere (Jorge Luis Borges 1966)), because it is very far away from the material evolution actually happens on, i.e. DNA. In this paper, I tried to go a middle ground: Instead of looking at the genetic sequences or the macromolecular phenotype, I'm looking at the presence of certain biological processes, molecular functions, or cellular components, in other words: along the path of evolution, any meaningful change in an organism goes along with the development, change, or loss of such a function; what good is a change in the DNA sequence if it does not lead to a modified function of that gene?

Gene Ontology

Historically, the function/role that a gene plays in an organism has always been described in natural language, however the researcher characterizing that gene deemed best. While this is nice to read, it is not very useful if you want to do computation on it, as computers are (still) horrible at understanding natural language and determining the structure in meaning behind these words. Additionally, different people will describe the same thing with different words, which has the potential for misunderstanding. lalala Gene Ontology The function/role that a gene plays in an organism can be quite a vague expression and if you look at papers doing research on these things, you will see that every researcher describes it differently.

> "The mission of the GO Consortium is to develop an up-to-date, comprehensive, computational model of biological systems, from the molecular level to larger pathways, cellular and organism-level systems."
> — GO Consortium (geneontology.org)

Data

The data that serve as the input for this project are functional annotation datasets in the GAF file format[1] which we generated with our GOMAP pipeline[2]. They are (or will be shortly) available from https://dill-picl.org/projects/gomap/gomap-datasets.

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

[1] http://geneontology.org/docs/go-annotation-file-gaf-format-2.0/

[2] https://dill-picl.org/projects/gomap

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
...	

tuples (G, T) with $G \in$ Genes in that genome and $T \in$ Terms in the Gene Ontology.

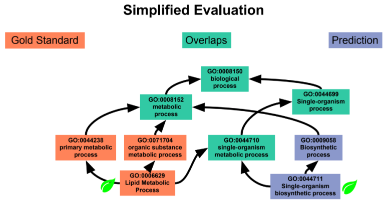
We can use the hierarchical structure of the Gene Ontology to obtain the ancestors A_i of any term T_i; in other words the gene G_i is not just annotated with the term T_i itself but also with all GO terms that are a more general statement of that term (e.g. any gene that is part of a metabolic process 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.

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 ($S_a \cup S_b$).


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

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