Gene Function Phylogeny

Dennis Psaroudakis

March 26th 2019

In my research, my colleagues and I have generated (or are currently generating) genome-wide functional annotation datasets for wheat, rice, cotton, soy, barley, and several maize genome assemblies using the Gene Ontology[1]. It would be interesting to find out whether a phylogenetic tree built on the functional annotation of the genes instead of the DNA sequence would yield a tree that is similar to the currently used tree for the evolutionary relationship between these species, i.e. whether their evolution can be retraced on a functional level.

For now this is largely explorative and it is possible that I will discover interesting but different patterns than what I expected. Maybe it's also all just nonsense, we will find out.

[1] `http://geneontology.org/`

Data

The functional annotation data is being generated with the GOMAP pipeline[2] which uses several different approaches to predict the functions of all genes in a given genome (publication in the works, see [2] for the precursor developed on maize). All datasets are or will be published at `https://dill-picl.org/projects/gomap/gomap-datasets/`.

[2] see `https://dill-picl.org/projects/gomap` for more info and `https://github.com/Dill-PICL/GOMAP-singularity` for the source code

Method

I am planning to build a distance matrix between all annotation sets using the hF_1 metric defined in [1]. hF_1 is the harmonic mean of the hierarchical precision hPr and hierarchical recall hRc[3]:

$$hF_1 = 2 \cdot \frac{hPr \cdot hRc}{hPr + hRc}$$

The hierarchical precision hPr is defined as follows:

$$hPr = \frac{\text{predicted GO term} \cup \text{true GO term}}{\text{predicted GO term}}$$

and the hierarchical recall correspondingly:

$$hRc = \frac{\text{predicted GO term} \cup \text{true GO term}}{\text{true GO term}}$$

To go from single annotation hF_1 scores, those are commonly averaged to get gene or annotation set wide hF_1 metrics.

$$\text{overall } hF_1 = \sum_{i=1}^{n} \frac{hF_i}{n}$$

[3] The hF_1 is basically the tree version of the regular F_1 score.

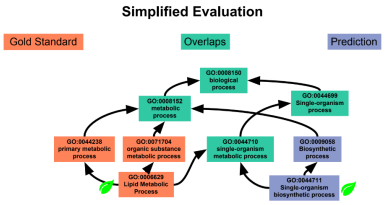

Figure 1: **(Zoom in)**. Graphic of how hPr and hRc use the Gene Ontology tree to calculate the overlap of two GO terms (the ones marked with a leaf)

The regular use case is that you have one set of GO predictions and one set of experimentally deduced GO annotations for the same genome which you can use as a gold standard to evaluate the quality of your predictions (as we did in [2]). In my case, I only want to compare two datasets without regards to any quality. Since hPr and hRc are both used in commutative contexts for the hF_1 metric, I can treat either of the datasets as the gold standard and the other one as the prediction. So my hF score between sets A and B[4] would be:

[4] $A, B \in$ {wheat, rice, cotton, soy, barley, $maize_1$, $maize_2$...}

$$\mathrm{hF}_{A,B} = 2 \cdot \frac{\frac{A \cup B}{A} \cdot \frac{A \cup B}{B}}{\frac{A \cup B}{A} + \frac{A \cup B}{B}}$$

$\mathrm{hF}_{A,B}$ will range between 0 and 1, with 1 being the score for two completely identical datasets. To turn this into a distance $d_{A,B}$ I can simply subract it from 1, making the distance between identical datasets 0 and completely disjunct datasets 1:

$$d_{A,B} = 1 - \mathrm{hF}_{A,B}$$

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

Approach

In the initial attempt, I want to ignore any genes and simply look at all functions present in any gene within the genome. That means I will combine all gene GO trees of a species into one big species-wide function tree and compare these function trees. Possibly I will limit it to the *Molecular Function* aspect of the Gene Ontology or at least exclude *Cellular Component* because I don't expect that one plant species will have, e.g. a nucleus and the other ones don't.

Then I'll see if what I have generated makes any sense and go from there.

References

[1] Michael Defoin-Platel, Matthew M. Hindle, Artem Lysenko, Stephen J. Powers, Dimah Z. Habash, Christopher J. Rawlings, and Mansoor Saqi. Aigo: Towards a unified framework for the analysis and the intercomparison of go functional annotations. *BMC Bioinformatics*, 12(1): 431, Nov 2011. ISSN 1471-2105. DOI: 10.1186/1471-2105-12-431. URL https://doi.org/10.1186/1471-2105-12-431.

[2] Kokulapalan Wimalanathan, Iddo Friedberg, Carson M. Andorf, and Carolyn J. Lawrence-Dill. Maize go annotation—methods, evaluation, and review (maize-gamer). *Plant Direct*, 2(4):e00052, 2018. DOI: 10.1002/pld3.52. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pld3.52.