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 practically often too abstract to be helpful. Let's say for example, that you have a patient infected with a novel pathogen. You can identify that pathogen and place it somewhere in the bacterial taxonomy, but unfortunately, none of the treatments that you usually use for closely related species show any effect against this one. How come? Just because two organisms are closely related does not *necessarily* mean that they are similar in their phenotype. So a tree that groups organisms based on their actual *in vivo* similarity would be much more helpful here. There are such trees built on phenotypic characteristics but unfortunately the choice of what characteristics to look at is not trivial and always situation-dependent. They can also be considered quite arbitrary[1]. What's needed would be something that is exhaustive, clearly defined, and reproducible, but also more meaningful than just a DNA sequence. And they did it [3]!

In this paper, I'm going to do something slightly similar. Not on bacteria, but on plants though. I know, disapointing, but I'll find a similarly fascinating reason why that's important soon. So, to summarize: I want to build a phylogenetic tree on plant species not based on their genetic sequence but on the *functions* that this plant is able to execute.

[1] Look at this (parodic) taxonomy of animals by Jorge Luis Borges: 1. those that belong to the Emperor, 2. embalmed ones, 3. those that are trained, 4. suckling pigs, 5. mermaids, 6. fabulous ones, 7. stray dogs, 8. those included in the present classification, 9. those that tremble as if they were mad, 10. innumerable ones, 11. those drawn with a very fine camelhair brush, 12. others, 13. those that have just broken a flower vase, 14. those that from a long way off look like flies.

The 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 the words. Additionally, different people will describe the same thing with different words, which has the potential for misunderstanding.

Ontologies try to alleviate these problems by providing a strictly organized and controlled vocabulary and defined relationships between the terms, so that the same statement always means the same thing, no matter the context or the author. Additionally, ontologies can be understood by computers if all relationships and terms are clearly defined.

The Gene Ontology is such an ontology that describes genes by the properties of their product. In our case these gene products are proteins, and they can be characterized in three different aspects:

"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)

- What biological processes is this protein part of? (e.g. photosynthesis or autophagy)
- What molecular functions does the protein carry out? (e.g. ethylene binding or RNA ligation)
- What cellular component is the protein active at? (e.g. outer membrane or nucleus)

Within each of these aspects, the Gene Ontology defines a huge number of terms (2,675,070 in total), that range from very general to very specific:

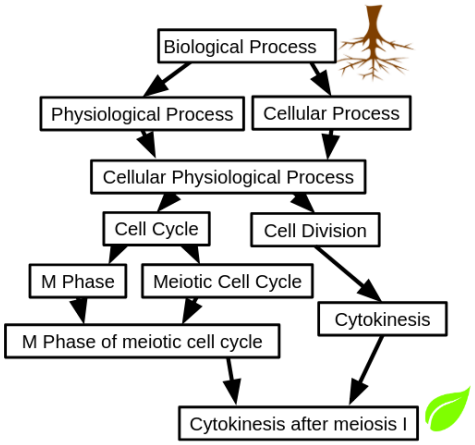


Figure 1: subtree of the *Biological Process* ontology. The terms are organized in such a way that more general terms are always true for any of their more specific child terms. For example, any protein that is part of *Cytokinesis after Meiosis I*, is also obviously part of Cytokinesis, the Cell Cycle etc. That way, a gene that has been annotated with the term *Cytokinesis after Meisosis I* (leaf term), has implicitly been annotated with all of that term's parent terms as well, all the way up to the root term.

When proteins are annotated with these terms instead of just natural language, we can now computationally answer some interesting questions, such as:

- How similar in function is protein A to protein B? (One answer would be: How many steps in the GO graph do I need from term A to B? The fewer steps, the more similar the function)
- If protein A is involved in Biological Process XYZ, what other proteins are involved in that same process?

The Gene Ontology is quite well established in the field, so you will find GO annotations for almost all relevant UniProt entries or use dedicated tools like AmiGO or QuickGO to examine a protein of interest.

Data

Annotating genes with their functions can be done experimentally (e.g. by knocking out a certain gene and seeing what processes in the cell are affected), but that is a time-consuming and expensive process, so methods have been developed that try and predict the function of a given gene. Our lab has developed such a pipeline called GOMAP which combines different prediction approaches and is able to generate high-confidence and very extensive

Functional Annotations in a reproducible manner [1]. We have been applying this pipeline to whole-genome assemblies of different plant species and generated functional annotations for every gene in each genome. These annotation sets 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 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 (provided by PHYLIP). I rooted the resulting tree manually outside of the grasses (maize, wheat, rice, barley).

Result

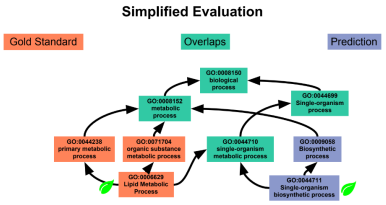
The phylogram is displayed in Figure 2. 4 of the 5 maize assemblies are grouped together in a clade, while the last one is an outlier to the remaining grasses and sits on a very long branch. In the remaining plants, cotton is an outlier to the legumes, which are all grouped in a single clade.

Discussion

This tree shows surprising similarity to the taxonimic tree one might expect if it had been built on sequences[2]: The maize cultivars are all grouped together and there's even the distinction between the non-stiff stalk (W22 and Mo17), iodent (Ph207), and stiff stalk (maize_v4) classes. Only maize_v3, which is

```
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 $1 - \frac{4}{9} = \frac{5}{9}$

[2] I should probably display that expected tree here.

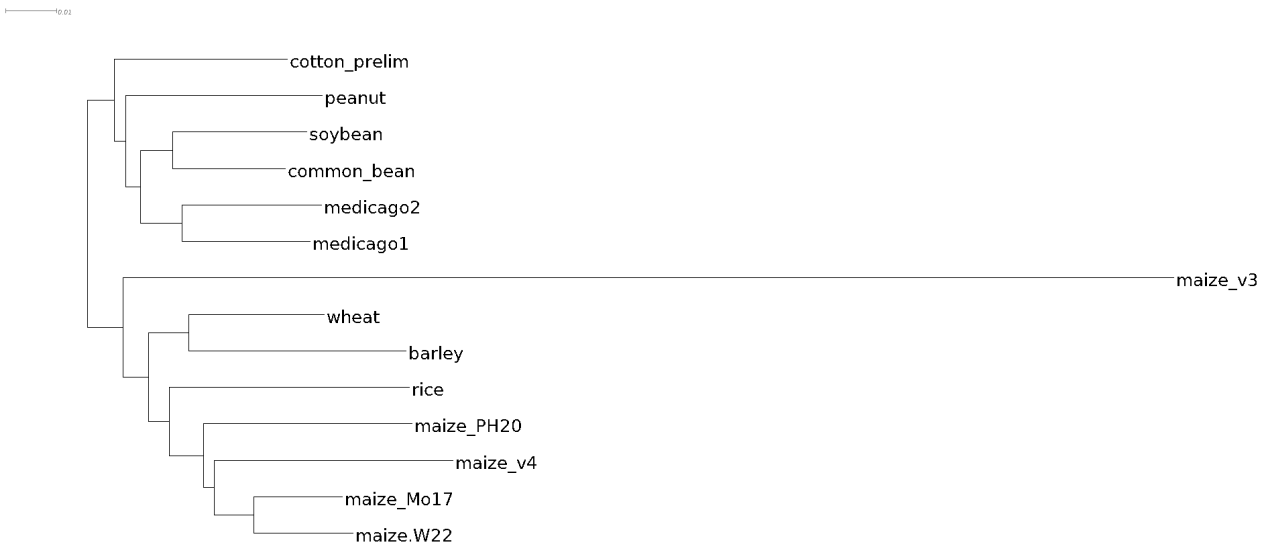


Figure 2: Resulting phylogram. maize_v3 and maize_v4 are two assemblies of the same maize cultivar, medicago1 and medicago2 are two different genotypes.

the same cultivar as maize_v4 (B73) is waaaay out there where it probably doesn't belong. We're currently investigating this but it seems that a lot went wrong with the genome assembly itself, which would naturally influence the functional annotations derived from it. The two medicago genotypes, while correctly placed together in one clade, are an outlier to the other legumes. In a taxonomic tree, their place would be switched with peanut. Steven Cannon, who is an evolutionary biologist working on legumes, assumes that the reason for that is that medicago has a much higher rate of evolution and gene birth and death than other plants [2]. As you will have noticed, this is all just a work in progress at the moment. My PI and I will meet with Steven on Monday and discuss (among other things) the following questions – feel free to comment in your review if you have any input!

- How similar is this tree to the taxonomic tree?
- What are the reasons for the differences, what are the reasons for the similarities?
- What biases in the construction process need we be aware of?
- What does the tree depict/how do we interpret it?
- In what way could it be scientifically valuable?
- Are there any other taxa that would be good to have in the tree?
- Is there any further analysis that would be good to do? (e.g. branch support, looking at each GO aspect separately, a more sophisticated measure of GAF similarity, trying a character state based method instead of distance…)
- Is rooting the tree there reasonable?

Reproducing the Tree

To reproduce this tree, follow these steps on a Linux machine that has PHYLIP available:

```
git clone https://github.com/Thyra/EEOB563.git paper_dennis # Clone the repository
cd paper_dennis/final_project # change into the directory
./build_distance_matrix > infile
module load phylip # (if you're on HPC)
neighbor # (use standard options)
```

The resulting tree is in `outtree`. Manually root it (e.g. with Dendrogram) outside of the grasses (maize, wheat, barley) and you should end up with the same tree. This will use the pre-summarized `.tree.json` files in annotation_sets, if you want to completely reproduce it from scratch, delete them and only leave the `.gaf.gz` files. But be warned, this will take a while to calculate (probably over two hours).

References

[1] 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, apr 2018. ISSN 24754455. DOI: 10.1002/pld3.52. URL `http://doi.wiley.com/10.1002/pld3.52`.

[2] Nevin D. Young, Frédéric Debellé, Giles E. D. Oldroyd, Rene Geurts, Steven B. Cannon, Michael K. Udvardi, Vagner A. Benedito, Klaus F. X. Mayer, Jérôme Gouzy, Heiko Schoof, Yves Van de Peer, Sebastian Proost, Douglas R. Cook, Blake C. Meyers, Manuel Spannagl, Foo Cheung, Stéphane De Mita, Vivek Krishnakumar, Heidrun Gundlach, Shiguo Zhou, Joann Mudge, Arvind K. Bharti, Jeremy D. Murray, Marina A. Naoumkina, Benjamin Rosen, Kevin A. T. Silverstein, Haibao Tang, Stephane Rombauts, Patrick X. Zhao, Peng Zhou, Valérie Barbe, Philippe Bardou, Michael Bechner, Arnaud Bellec, Anne Berger, Hélène Bergès, Shelby Bidwell, Ton Bisseling, Nathalie Choisne, Arnaud Couloux, Roxanne Denny, Shweta Deshpande, Xinbin Dai, Jeff J. Doyle, Anne-Marie Dudez, Andrew D. Farmer, Stéphanie Fouteau, Carolien Franken, Chrystel Gibelin, John Gish, Steven Goldstein, Alvaro J. González, Pamela J. Green, Asis Hallab, Marijke Hartog, Axin Hua, Sean J. Humphray, Dong-Hoon Jeong, Yi Jing, Anika Jöcker, Steve M. Kenton, Dong-Jin Kim, Kathrin Klee, Hongshing Lai, Chunting Lang, Shaoping Lin, Simone L. Macmil, Ghislaine Magdelenat, Lucy Matthews, Jamison McCorrison, Erin L. Monaghan, Jeong-Hwan Mun, Fares Z. Najar, Christine Nicholson, Céline Noirot, Majesta O'Bleness, Charles R. Paule, Julie Poulain,

Florent Prion, Baifang Qin, Chunmei Qu, Ernest F. Retzel, Claire Riddle, Erika Sallet, Sylvie Samain, Nicolas Samson, Iryna Sanders, Olivier Saurat, Claude Scarpelli, Thomas Schiex, Béatrice Segurens, Andrew J. Severin, D. Janine Sherrier, Ruihua Shi, Sarah Sims, Susan R. Singer, Senjuti Sinharoy, Lieven Sterck, Agnès Viollet, Bing-Bing Wang, Keqin Wang, Mingyi Wang, Xiaohong Wang, Jens Warfsmann, Jean Weissenbach, Doug D. White, Jim D. White, Graham B. Wiley, Patrick Wincker, Yanbo Xing, Limei Yang, Ziyun Yao, Fu Ying, Jixian Zhai, Liping Zhou, Antoine Zuber, Jean Dénarié, Richard A. Dixon, Gregory D. May, David C. Schwartz, Jane Rogers, Francis Quétier, Christopher D. Town, and Bruce A. Roe. The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, 480(7378):520–524, dec 2011. ISSN 0028-0836. DOI: 10.1038/nature10625. URL http://www.nature.com/articles/nature10625.

[3] Chengsheng Zhu, Tom O. Delmont, Timothy M. Vogel, and Yana Bromberg. Functional Basis of Microorganism Classification. *PLOS Computational Biology*, 11(8):e1004472, aug 2015. ISSN 1553-7358. DOI: 10.1371/journal.pcbi.1004472. URL https://dx.plos.org/10.1371/journal.pcbi.1004472.