Genome Function Phylogenetics

Dennis Psaroudakis

April 18th, 2019

When building phylogenetic trees, often the sequences of certain genes serve as the basis for the tree. They are the direct substrate of evolution, which makes sense, but often are too abstract to be helpful.

Let’s say for example, that there’s a patient infected with a novel pathogen. The identity of the pathogen can be found and place it somewhere in the bacterial taxonomy, but unfortunately, none of the treatments available 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 phenotypically. Therefore, 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. Additionally, They can

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

be considered quite arbitrary1. What’s the solution? A tree that’s

exhaustive, clearly defined, and reproducible, but also is more meaningful than just a DNA sequence. And they did it [3]!

This paper is similar to the dilemma previously described except it focuses on plants as opposed to of bacteria. To summarize, the purpose of this study is to build a phylogenetic tree for plant species, based not on their genetic sequence, but instead on the functions executed by the plants.

The Gene Ontology

Historically, the function or role a gene plays in an organism is described in natural language, however the researcher characterizing that gene deemed best. While this is nice to read, it is not useful for computational analysis , as computers struggle with understanding natural language and the underlying connotation. Since different people describe the same thing with different words, the potential for misunderstanding amplifies.

Ontologies try to alleviate these problems by providing strictly organized and controlled vocabulary and clarifies the relation between terms, to control for a statement’s meaning regarless of the context or author. Additionally, computers can understand clearly defined relationships and terms with proper use of ontologies..

”The mission of the GO Consortium is to develop an up-to-date, comprehensive, compu- tational model of biological systems, from the molecular level to larger pathways, cellular and organism-level systems.”

-- GO Consortium (geneontology.org)

The Gene Ontology is an ontology describing genes by the properties of their product. For this paper, gene products are proteins and are characterized in three different aspects:

1. What biological processes is this protein part of? (e.g. photosynthesis or autophagy)
2. What molecular functions does the protein carry out? (e.g. ethylene binding or RNA ligation)
3. 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), ranging from both extremes of the spectrum: very general to very specific:

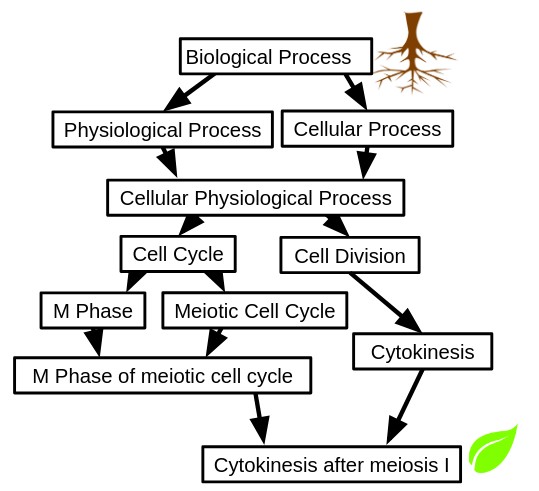


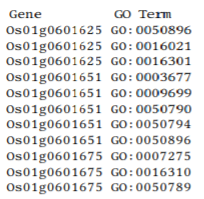
Figure 1: Subtree of the Biological Process ontology. The terms are organized in a way that general terms are always true for any of their more specific child terms. For example, any protein that is part of Cytokinesis after Meiosis I, must also be apart of Cytokinesis, the Cell Cycle etc. That way, a gene annotated with the term Cytokinesis after Meiosis I (leaf term), has implicitly been annotated with all its parent terms 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 are the functions of protein A compared to protein B? How many steps in the GO graph is required from A to B? (The fewer steps indicates more similarity between the functions)

• If protein A is involved in Biological Process X, 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



...

The general idea of using the Jaccard Distance is to measure the overlap between two subtrees in the GO hierarchy. Say, that we’re looking at two genomes (Gold Standard and Prediction) that each contain one GO term (marked by a leaf). Here is how the Jaccard distance is calculated:

First, all ancestors of that leaf term are added to each subtree. Then, the overlap between them is determined (which corresponds to Sa ∩ Sb respectively).Finally divide the number of nodes within this overlap by the number of nodes in either of the two subtrees (Sa ∪ Sb).

For this example, the Jaccard Distance of Gold Standard and Prediction is 1 – 4/ 9 = 5 /9

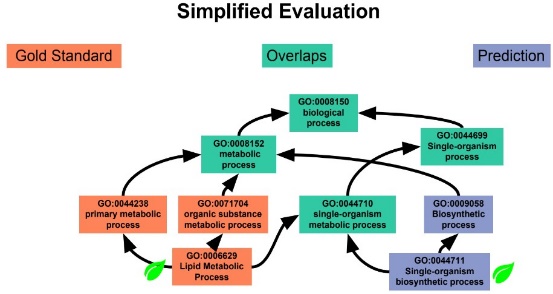
*Data*

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 developed 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 at:

[https://dill-picl.org/projects/gomap/gomap-datasets .](https://dill-picl.org/projects/gomap/gomap-datasets)

*Methods*

the First the functional annotation sets, one for each genome, for every gene in the genome are annotated with one or more GO terms. In mathematical terms, the genome annotation set is a list of tuples (G, T ) with G ∈ Genes in that genome and T ∈ Terms in the Gene Ontology.

 By utilizing the hierarchical structure of the Gene Ontology, the ancestors Ai is obtained from any term Ti. Gene Gi is not just annotated with the term Ti itself, but also with all GO terms – both general and specific terminology alike. For example, any gene apart of a metabolic process is thereby also part of a biological process. f All terms T in the dataset are annotated like that then all terms and their ancestors are combined into one big genome --wide set S, irrespectively of the gene they were originally associated with:

When this superset of annotations is created for each of the datasets, Jaccard Distance is used as a measure the similarity between any two sets., To put in biological terms, the difference between the two genomes in respect to their functional level.

Jaccard Distance (S ,S) = 1 − |Sa ∩ Sb|

*a b*

|Sa ∪ Sb|

Applying this formula to all pairwise combinations of the genomes we’re looking at yields a S × S distance matrix which then serves as the input for a neighbor joining algorithm (provided by PHYLIP). ~~I rooted~~ The resulting tree was manually rooted 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.

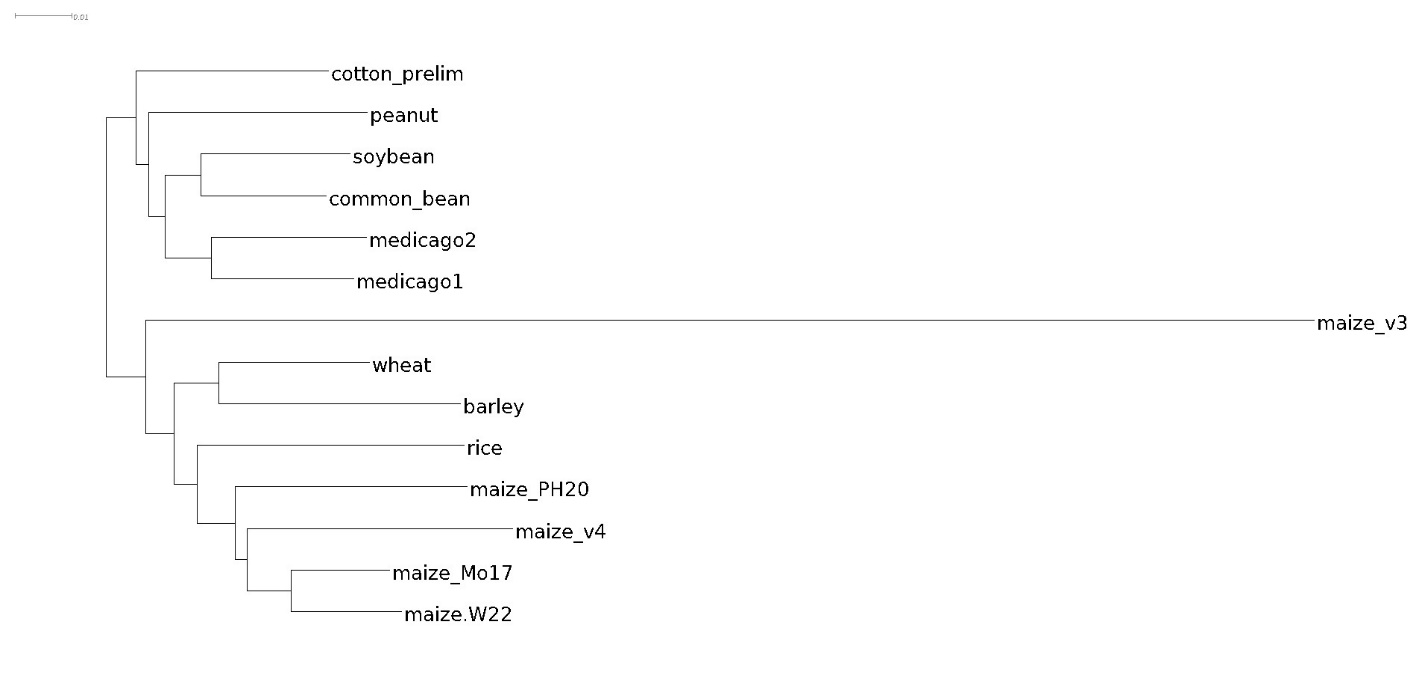
*Discussion*

Figure 2: Resulting phylogram. maize\_v3 and maize\_v4 are two assemblies of the same maize cultivar, medicago1 and medicago2 are two different genotypes.

This tree shows surprising similarity to the taxonimic tree one might expect if it had been built on sequences2. The maize cultivars are grouped together and there’s even the distinction between non-stiff stalk (W22 and Mo17), iodent (Ph207), and the stiff stalk (maize\_v4) classes. Only maize\_v3, which is

the same cultivar as maize\_v4 (B73) is way 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, an evolutionary biologist researching legumes, concludes the reason for this result from medicago having a higher rate of evolution, gene birth, and gene death than other plants [2]. ~~As you will have noticed, this is all just a work in progress now. 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!~~

This paper just begins to examine this unconventional analysis style. The following are examples of some of the concepts to explore in our future research:

• 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? If so, how would this selection process be determined?

• Is any further analysis needed to strengthen the validity of the results? (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*

Follow these steps on a Linux machine with PHYLIP to reproduce the tree in Figure 2.

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 Car- olyn 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/](http://doi.wiley.com/10.1002/pld3.52)

[10.1002/pld3.52.](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, Rox- anne Denny, Shweta Deshpande, Xinbin Dai, Jeff J. Doyle, Anne-Marie Dudez, Andrew D. Farmer, Stéphanie Fouteau, Carolien Franken, Chrys- tel 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 Nichol- son, Céline Noirot, Majesta O’Bleness, Charles R. Paule, Julie Poulain, Florent Prion, Baifang Qin, Chunmei Qu, Ernest F. Retzel, Claire Rid- dle, 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, Ke- qin Wang, Mingyi Wang, Xiaohong Wang, Jens Warfsmann, Jean Weis- senbach, Doug D. White, Jim D. White, Graham B. Wiley, Patrick Wincker, Yanbo Xing, Limei Yang, Ziyun Yao, Fu Ying, Jixian Zhai, Lip- ing 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://](http://www.nature.com/articles/nature10625)

[www.nature.com/articles/nature10625.](http://www.nature.com/articles/nature10625)

[3] Chengsheng Zhu, Tom O. Delmont, Timothy M. Vogel, and Yana Bromberg. Functional Basis of Microorganism Classification. PLOS Com- putational Biology, 11(8):e1004472, aug 2015. ISSN 1553-7358. doi:

10.1371/journal.pcbi.1004472. URL [https://dx.plos.org/10.](https://dx.plos.org/10.1371/journal.pcbi.1004472)

[1371/journal.pcbi.1004472.](https://dx.plos.org/10.1371/journal.pcbi.1004472)