Post Lab Questions: Week 5

OneCard: 195579 - October 2017 - Bioinformatics

I. Figures

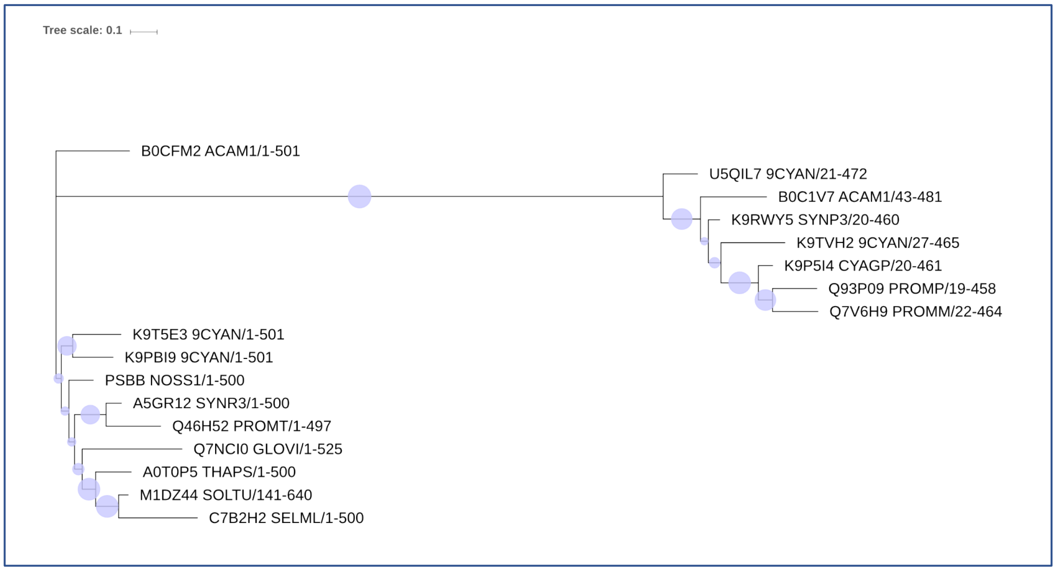


Figure 1. Maximum likelihood tree of toy dataset PSII protein, made with RAxML.

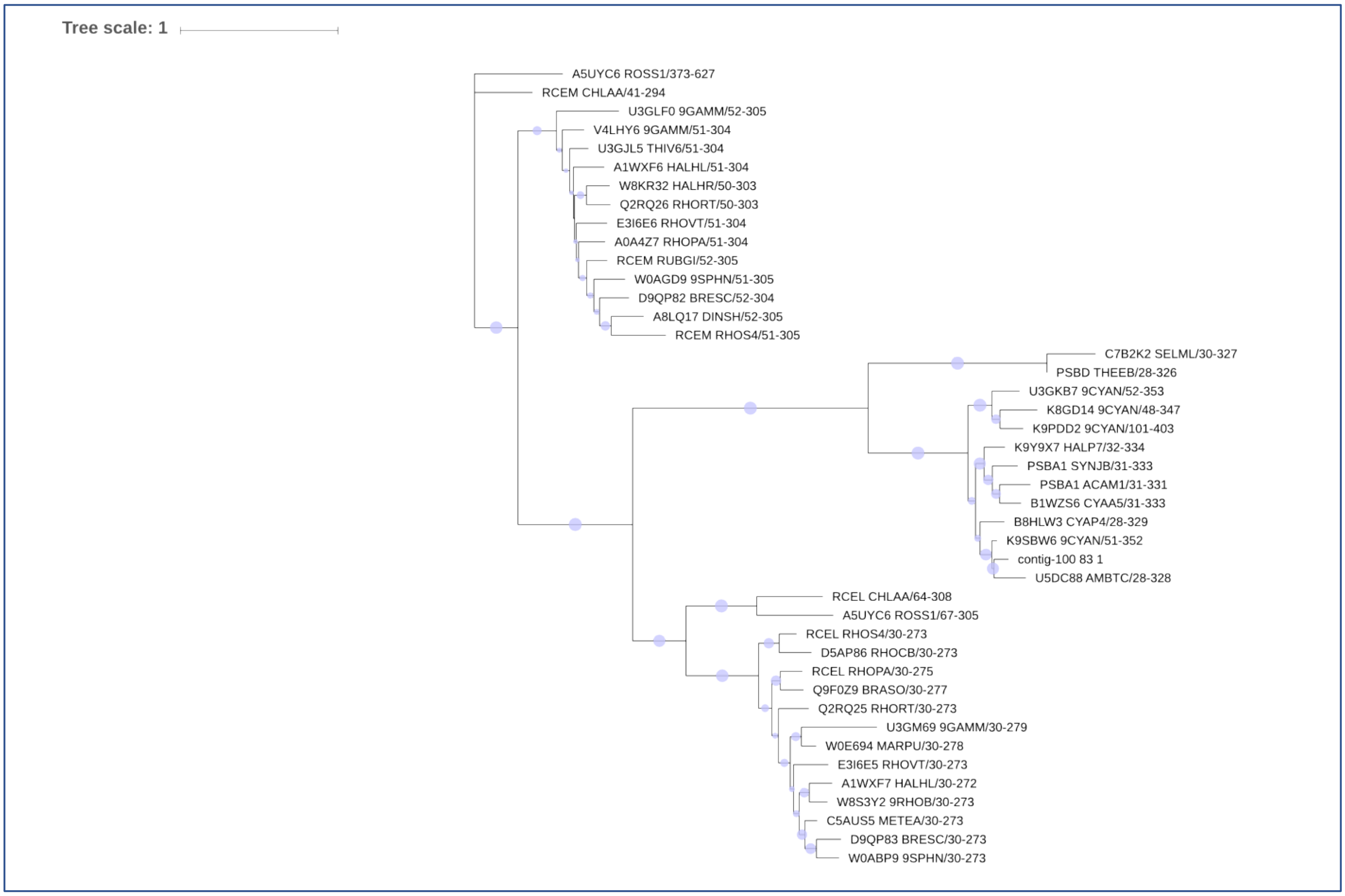


Figure 2. Maximum likelihood tree of psbA proteins, from a PFAM sample, plus "contig-100 31 1" from Tara Oceans sample ERR598983, made with RAxML.

II. Check for Understanding

1. Which organisms have the three most closely related psbA proteins to your ORF? (Hint: you can use UniProt, as we learned last week, to figure out what organisms these proteins come from.)

The closely match is to U5DC88, which comes from *Amborella trichopod*a, a “monotypic genus of understory shrubs or small trees endemic to the main island, Grand Terre, of New Caledonia” (Wikipedia). The next closet version of the protein is K9SBW6, which evidently belongs to the organism *Geitlerinema* sp. PCC 7407, a type of cyanobacteria. The third most closely related version is B8HLW3, which belongs to *Cyanothece* sp. (strain PCC 7425 / ATCC 29141), also a type of cyanobacteria. While the bacteria seem reasonable, the shrub may reflect some kind of database error.

2. Based on these results, what might you infer about the photosynthetic organisms in your sample? Are they eukaryotes, archaea, bacteria? Do you think you can make broad conclusions about the whole community of photosynthetic organisms in your sample? Why or why not?

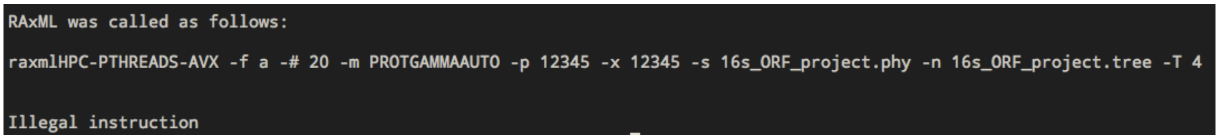
There is decent evidence that the photosynthetic organisms in the sample are bacteria, since two of most closely related proteins came from bacteria, and the third closest protein may be erroneous, as a French shrub has no business dwelling in the middle of the ocean. There was only one contig in the Tara ocean sample matching the psbA protein seed file, and a contig generally reflects a single organism, suggesting there was only a single photosynthetic organism in the Tara ocean sample (and it was a bacterium.) However, we only searched for genes related to a single protein (psbA) in the photosynthetic pathway, and it is possible that searching for other proteins would have uncovered other photosynthetic organisms.

III. Mini Research Question

I’m not sure I things correctly here, but I wanted to investigate the relationships between organisms in my sample by searching for 16s ribosomal RNAs in my ORF file, and then creating a phylogenetic tree containing all the organisms I could identify via their 16s barcodes.

I searched for “16s rRNA” on PFAM and selected the top hit, which was described as *Methyltr_RsmB-F* (PF01189). I downloaded the seed, created a database from my sample’s ORF file, and did a blastp using the seed file as the query. This yielded 194 hits, many of them with very small e-values. Next, I wrote some python code that took the hits from this blast and searched for them in my original ORF file, to construct a new fasta file. I would then create a phylogentic tree from this fasta file, which hopefully would contain a number of different versions of the 16s rRNA in my sample.

I transferred the fasta file to the server, and ran a muscle alignment, then converted “.afa” to “.phy” with Rika’s script. Unfortunately, I couldn’t get RAxML to run! I kept getting an error:



```
RAxML was called as follows:

raxmlHPC-PTHREADS-AVX -f a -# 20 -m PROTGAMMAAUTO -p 12345 -x 12345 -s 16s_ORF_project.phy -n 16s_ORF_project.tree -T 4


Illegal instruction
```

I tried downloading RAxML on my computer and running, and I think this was finally successful. Then I was able to make a tree. Perhaps the tree would be more meaningful if I had saved the names from the query file rather than the database file into my new fasta file I made? Then I might be able to look up the organisms, instead of simply showing the names of ORFs. Nonetheless, I believe all the ORFs in this tree represent 16s proteins, and thus the branches represent evolutionary closeness between different organisms. However, it is not terribly easy to see what organisms they are (unless you cross-reference this tree with my blastp results). Additionally, I'm unsure if the 16s rRNA is a good indicator of differences between organisms, or merely used to identify organisms, which are then compared on the basis of other genes? Well, here's the tree and the commands I entered. You can also see the code I wrote in a Jupyter notebook online:

https://nbviewer.jupyter.org/github/dustinmichels/biol338-genomics/blob/master/lab-5/data/blastp_to_fasta.py.ipynb

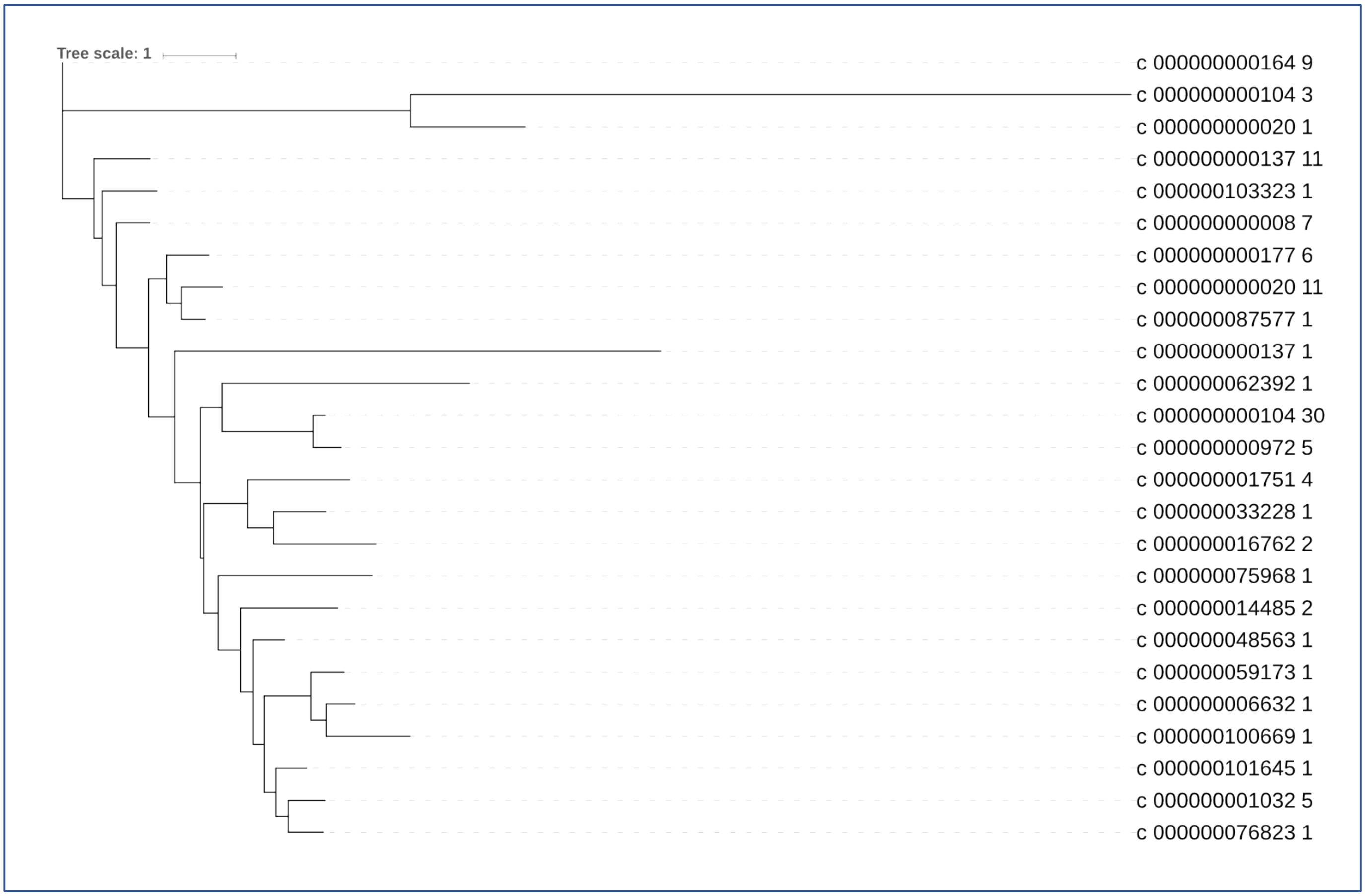


Figure 3. Phylogenetic tree containing ORFs from my Tara sample that matched a seed of 16s rRNAs from Pfam.

Commands:

```
# Rearrange files...
# (on server)
cd project_directory
mkdir alignments_and_trees
cp mapping/ERR599031_ORFs.noasterisks.faa alignments_and_trees/
mv ERR599031_ORFs.noasterisks.faa ERR599031_ORFs.faa

# Make db
# (on server)
makeblastdb -in ERR599031_ORFs.faa -dbtype prot

# blast
# (on server)
blastp -query 16s_protein_fasta.txt -db ERR599031_ORFs.faa -outfmt 6 -
evalue 1e-08 -out 16s_protein_vs_ERR599031_ORFs.blastp

# extract protein sequences for matching contigs, using python script
# (on my comptuer)
python3 blastp_to_fasta.py

# Make a multiple sequence alignment with muscle
# (on server)
muscle -in 16s_ORF_project.fasta -out 16s_ORF_project.afa

# Covert afa to phy
# (on server)
convert_afa_to_phy.py 16s_ORF_project.afa

# Make tree, take 1
# (on server --> failed)
raxmlHPC-PTHREADS-AVX -f a -# 20 -m PROTGAMMAAUTO -p 12345 -x 12345 -s
16s_ORF_project.phy -n 16s_ORF_project.tree -T 4

# Make tree, take 2
# (on my computer)
./raxml  -f a -# 20 -m PROTGAMMAAUTO -p 12345 -x 12345 -s
16s_ORF_project.phy -n 16s_ORF_project.tree -T 4
```