Comparative and Regulatory Genomics-class assignment 1

Kristen Michelle Nader

November 3, 2020

Search Mpn515 in UniProt

Entry	Entry name		Protein names	Gene names	Organism	Length
P75271	RPOC_MYCPN		**DNA-directed RNA polymerase subunit...**	**rpoC** MPN_515, MP327	Mycoplasma pneumoniae (strain ATCC 29342 / M129)	1,290

Figure 1: Uniprot result

From what species is this protein?

This species corresponds to *Mycoplasma pneumoniae*.

What is the function of this protein?

This protein is an DNA-dependent RNA polymerase that catalyses the transcription of DNA to RNA.(for the gene rpoC)

Find homologs in at least 25 distinct species (model organisms)

We can BLAST the protein sequence either directly on BLAST or through the BLAST extension on UNIPROT. From this we find many matches in the *Mycoplasma* genus.

Make a fasta file in a text editor and give human-readable names to your proteins

We can extract different species that the protein exists in, and save them in a fasta file. Then we can use bash commands to do some pre-processing. This will allow us to change the header lines so a name that is more readable.

Make a multiple alignment , use at least 2 alignment programs.

I chose T-COFFEE and CLUSTALW to do multiple sequence alignment and then to visualize phylogenetic trees.

Create a phylogenetic tree for each of your alignments, compare them. Visualize your trees in iTol

These trees have small differences in terms of branch length but this difference is to a small degree. In addition, there are some differences in terms of the ordering of the sequences in the tree. However, in phylogenetic trees, branches can be rotated around the notes and the order does not hold significance and therefore there is little difference bewteen figures 2 and 3.

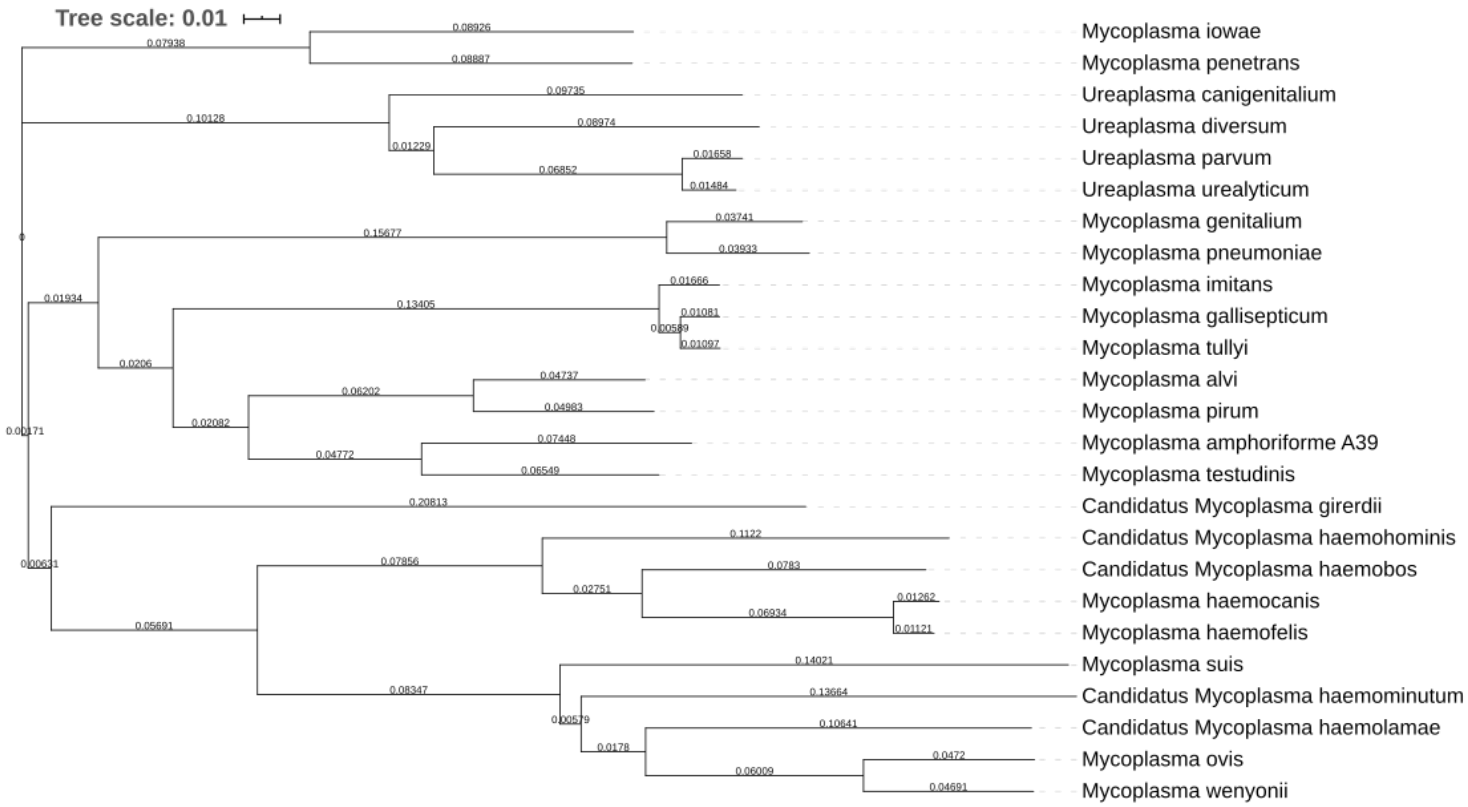

Figure 2: Phylogenetic tree visualized using iTol with T-COFFEE

Make an accurate alignment and create a tree with following options for both16S and RpoA: remove gaps from the alignment, use multiple substitution correction, use neighbor joining and number of boostraps 1000

The following commands were performed:

Listing 1: ClusatalW command for 16S

```
clustalw2 -INFILE=16S_Myco.fsa.txt -ALIGN -TREE -BOOTSTRAP -BOOTLABELS=branch
-CLUSTERING=NJ -OUTPUTTREE=nj -TOSSGAPS -NEWTREE=16S_OUTPUT.tree
```

Listing 2: ClustalW command for RpoA

```
clustalw2 -INFILE=RpoA_Myco.fsa.txt -ALIGN -TREE -BOOTSTRAP -BOOTLABELS=branch
-CLUSTERING=NJ -OUTPUTTREE=nj -TOSSGAPS -NEWTREE=RpoA_OUTPUT.tree
```

Something interesting to notes is that bootstrap values were not given even after specifying the appropriate arguments. Visualization was done using iTOL but can also be visualized using FigTree.
where the aguments: -INFILE : specify the input file
-ALIGN : A full multiple sequence alignment
-TREE : calculates a default neighbor joining tree
-BOOTSTRAP : boostrap for the default neighbor joining tree and a default n=1000
-BOOTLABELS : node or branch

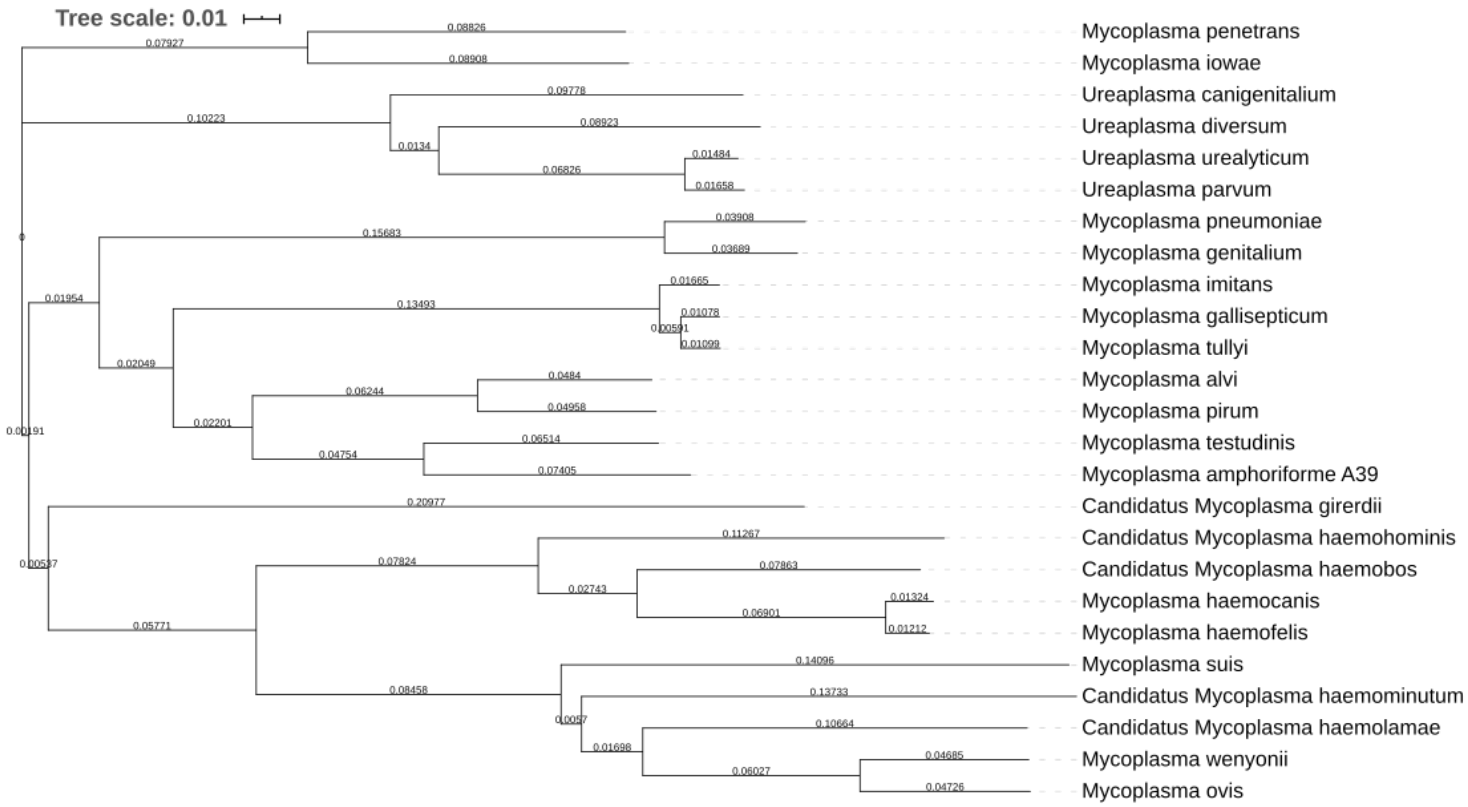

Figure 3: Phylogenetic tree visualized using iTol with clustalw

-CLUSTERING : Neighbor Joining
-OUTPUTTREE : default NJ
-TOSSGAPS : ignore positions with gaps
-NEWTREE : create a new guide tree

In addition, there are 18 sequences in both fasta files. I found this by using Bash commands:

Listing 3: Command to determine number of sequences for RpoA

```
grep ">" RpoA_Myco.fsa.txt | wc -l
```

Listing 4: Command to determine number of sequences for 16S

```
grep ">" 16S_Myco.fsa.txt | wc -l
```

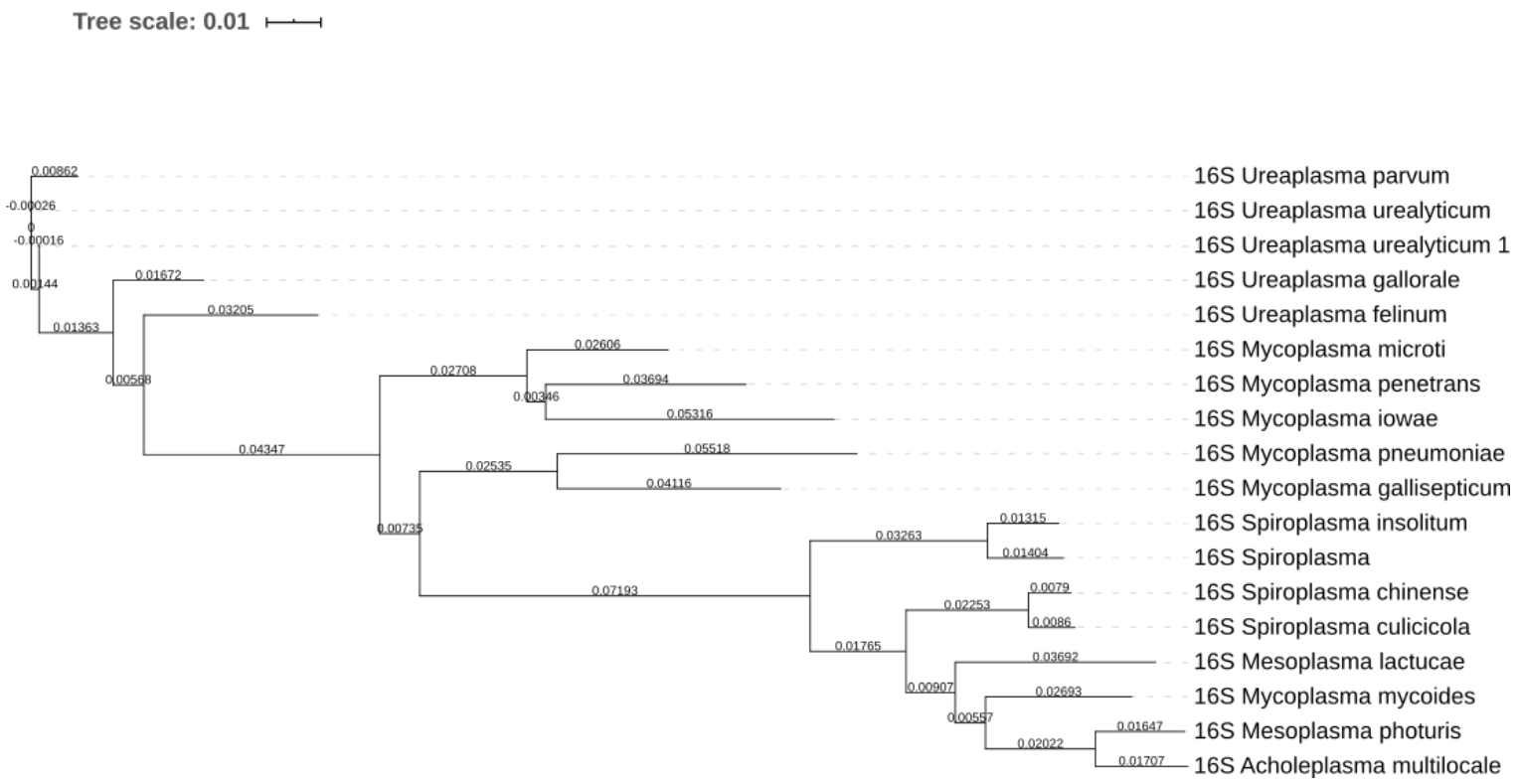

Figure 4: Phylogenetic tree visualized using iTol with ClUSTALw for 16S

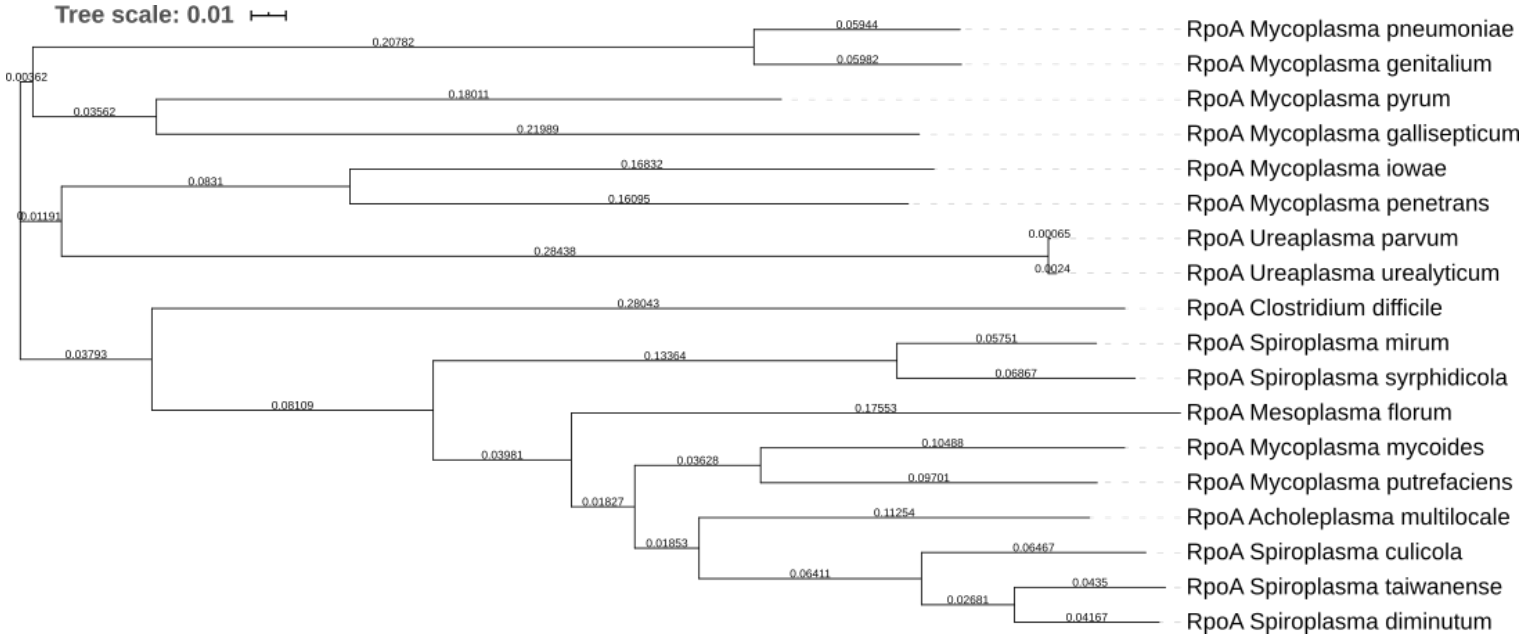

Figure 5: Phylogenetic tree visualized using iTol with CLUSTALW for RpoA