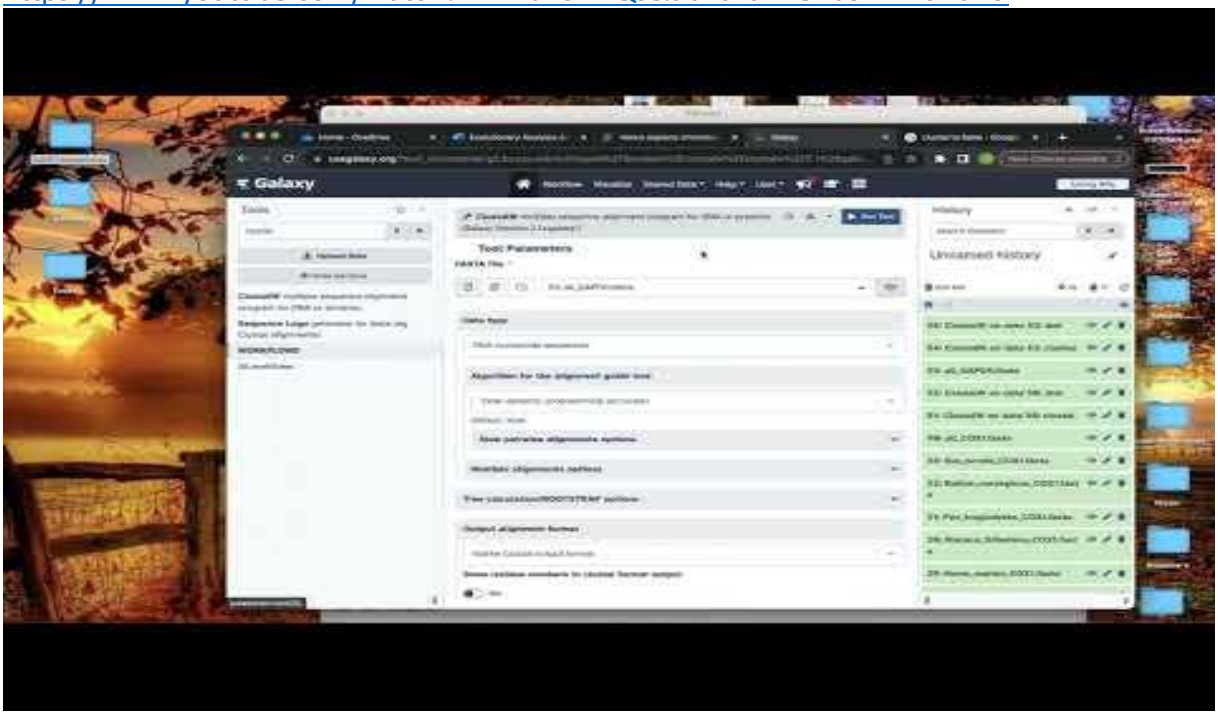
Evolutionary Analysis: Constructing Phylogenetic Trees from Reference Genome Gene Sequences

John Mariano

December 16th, 2023

Clustal to FASTA converter
http://sequenceconversion.bugaco.com/converter/biology/sequences/clustal_to_fasta.php
Galaxy site for ClustalW tool
https://usegalaxy.org/
NCBI database where data was obtained
https://www.ncbi.nlm.nih.gov/
Video Tutorial:
https://www.youtube.com/watch?v=kNiJHsYPiQs&ab_channel=JohnMariano



Purpose:

The purpose of this manual is to construct two phylogenetic trees, using the ape package in R, of eight species, *Macaca mulatta* (Macaque), *Pan troglodytes* (Chimpanzee), *Drosophila melanogaster* (Fruit fly), *Canis lupus familiaris* (Dog), *Equus caballus* (Donkey), *Sus scrofa* (Boar), *Homo sapiens* (Human), and *Rattus norvegicus* (Norway Rat), using an alignment of the genes Cytochrome c oxidase I (COX1) and Glyceraldehyde 3-phosphate dehydrogenase (GAPDH). These trees will be used to analyze the evolutionary relationship of the species and compare the rate of change in the evolution of the two genes.

Background:

Phylogenetics, the study of evolutionary relationships among species, lies at the heart of modern biology. It provides a framework for understanding the intricate web of evolution and illustrates the historical paths species have taken. One fundamental aspect of phylogenetics involves constructing phylogenetic trees, which are graphical representations of these evolutionary relationships. The sequence of a gene can be used to construct a phylogenetic tree to determine the evolutionary divergence of that gene among species. The choice of genes used for tree construction is critical, as it directly influences our evolutionary reconstructions. Different genes may have different rates of evolution and modes of inheritance. Choosing genes with various evolutionary properties can lead to variations in the inferred phylogenetic relationships among species, influencing the accuracy and reliability of the evolutionary reconstructions. (Kumar, S., & Hedges, S. B. 2011). In this manual, two genes, Cytochrome c oxidase I (COX1) and Glyceraldehyde 3-phosphate dehydrogenase (GAPDH), are the genes selected for phylogenetic tree construction.

COX1, a mitochondrial gene encoding a critical component of the respiratory chain, offers invaluable insights into the genetic divergence and evolutionary history of diverse organisms. This mitochondrial gene's utility stems from its relatively conserved nature within species and variability between species, making it an excellent molecular marker for phylogenetic analysis (National Center for Biotechnology Information. (n.d.). Cytochrome c oxidase subunit 1). Similarly, GAPDH, an essential enzyme involved in glycolysis, provides a different perspective on evolutionary relationships. It has a well-documented presence across life forms, allowing for the study of ancient divergences. GAPDH is a gene located on the nuclear genome as opposed to the mitochondrial genome (National Center for Biotechnology Information. (n.d.). Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) [GeneID 2597]).

Understanding the evolutionary relationships among species is not just a matter of academic curiosity but holds immense importance in various aspects of biology. It aids in species classification, ecological studies, and conservation efforts. Moreover, it informs us about the genetic basis of species-specific traits and adaptations. For instance, when studying a group of species that share a common ancestor, a phylogenetic tree can illuminate when and how specific

traits or characteristics evolved. This knowledge has profound implications for medicine, agriculture, and environmental conservation. It can guide the development of new drugs, reveal the origins of diseases, and enhance our ability to preserve biodiversity and manage ecosystems effectively. In this manual, I will be using the genetic sequences of seven mammals and one insect outgroup. I will be using the trees to both compare the differences in the phylogenetic relationships between the two genes, as well as analyze the differences in the evolutionary rate of the two genes (Degnan, J. H., & Rosenberg).

However, constructing phylogenetic trees from molecular data is a complex task that requires sophisticated computational and bioinformatics approaches. The enormous amount of genetic data generated by modern sequencing technologies necessitates powerful algorithms and computational tools to process and analyze it. Bioinformatics techniques are indispensable for managing, aligning, and interpreting genetic sequences. These approaches allow us to extract meaningful information from DNA or protein sequences and convert it into evolutionary insights. In the context of the COX1 and GAPDH genes, bioinformatics techniques are essential for phylogenetic tree reconstruction (Molecular Evolution and Phylogenetics).

In the realm of phylogenetic analysis, a variety of methods exist for reconstructing the evolutionary relationships among biological sequences. While more sophisticated approaches often incorporate complex evolutionary models, I will utilize the Neighbor-Joining (NJ) method for its simplicity and efficiency in handling large datasets. It is important to note that alternative methods, such as Maximum Likelihood (ML) or Bayesian inference, are available and may be preferred in certain contexts due to their ability to account for more intricate evolutionary processes. In this manual, I opted for the NJ method as a pragmatic choice, recognizing that the choice of phylogenetic method can impact the interpretation of results. The availability of various methods allows researchers to tailor their analyses to the specific requirements of their projects, and we acknowledge that alternative methods, including but not limited to ML and Bayesian methods, provide valuable avenues for further exploration and refinement in future studies (Molecular Evolution and Phylogenetics).

Tools:

ClustalW sequence alignment tool on Galaxy.

https://usegalaxy.org/

Clustal format to FASTA format converter

http://sequenceconversion.bugaco.com/converter/biology/sequences/clustal_to_fasta.php

Ape R package for constructing phylogenetic trees in R. Additional information about the Ape package can be found at the link below

https://cran.r-project.org/web/packages/ape/ape.pdf

Manual:

Step 1 – Data acquisition
All data for this manual was obtained from NCBI using the nucleotide database (https://www.ncbi.nlm.nih.gov/). Under the "All Databases" tab, select "Nucleotide" then search for the gene and species you are looking for. All data was obtained in the FASTA format.

	COX1 Accession #	**GAPDH Accession #**
Macaca mulatta	NC_011519.1:5839-7377	NC_041764.1:6685663-6689624
Pan troglodytes	NC_001643.1:5321-6862	NC_072410.1:12366840-12371591
Drosophila melanogaster	NC_024511.2:1474-3009	NT_033778.4:c7793380-7791901
Canis lupus familiaris	NC_002008.4:5349-6893	NC_051831.1:c38827216-38823453
Equus caballus	NC_001788.1:5357-6901	NC_009149.3:35091754-35096110
Sus scrofa	NC_000845.1:6511-8055	NC_010447.5:c64135194-64129678
Homo sapiens	NC_012920.1:5904-7445	NC_000012.12:6534517-653837
Rattus norvegicus	NC_001665.2:5323-6867	NC_051339.1:c157967158-157962312

Step 2 – Data Wrangling

The next step is to combine all the separate FASTA files for each gene into one file. This step is necessary because the ClustalW tool used for the alignment requires a single file rather than multiple files. This can be done quickly by putting all the files for each gene in their own folder, then using the command line to combine the files. To do this, first navigate to the specific gene directory with the fasta files. From here, a command shown below can be used to combine the files

```
cat *.fasta > all_GAPDH.fasta
```

This command takes the GAPDH gene fasta files and combines them in a new file called all_GAPDH.fasta. Repeat this step for all genes that a tree will be constructed from. If you are not comfortable using command line, the files can be combined by simply copying and pasting the files into one file. Now that we have our fasta files combined we can align the sequences within the file.

Step 3 – Aligning the Genes

To align the sequences, I used the ClustalW tool on the galaxy platform (https://usegalaxy.org/). From the galaxy home page, under the tool search bar, we can select upload data, and select our combined fasta file from local files. This will upload our fasta file to galaxy so that we can use it. Next, search for the ClustalW tool on the left side of the page. From the tool page, select the fasta file that was uploaded by clicking the dropdown menu at the top of the ClustalW tool page.

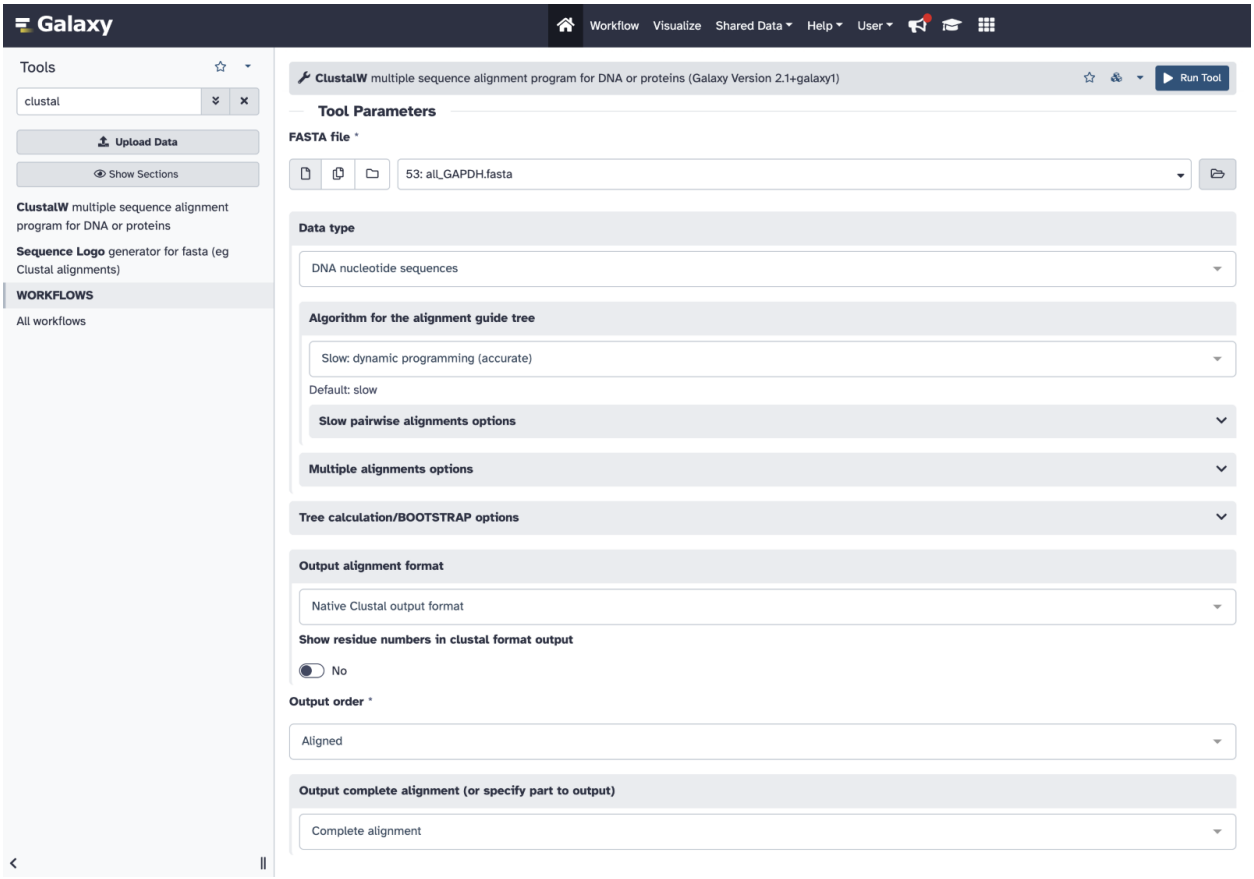
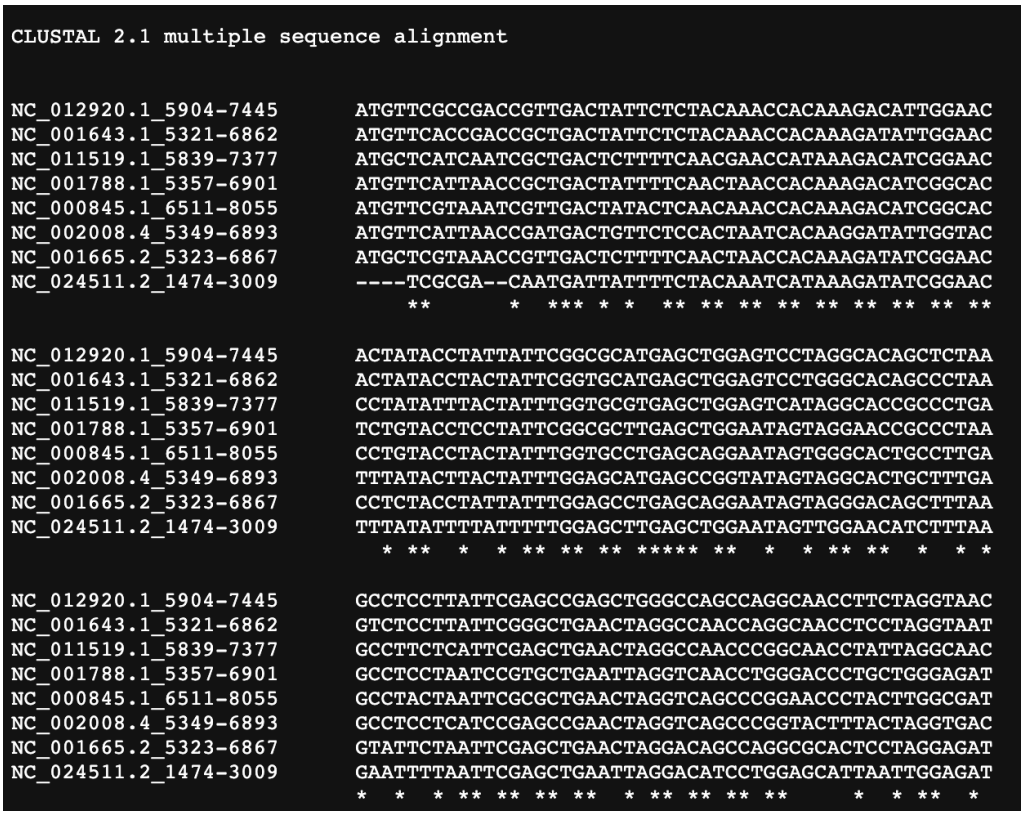


Figure 1 – The ClustalW tool page on the Galaxy site

Once the data has been selected, the tool can be run. This will output two files, one dnd file, and one clustal file. The clustal file is what we will be using. The clustal output file will look like this.



```
CLUSTAL 2.1 multiple sequence alignment


NC_012920.1_5904-7445      ATGTTCGCCGACCGTTGACTATTCTCTACAAACCACAAAGACATTGGAAC
NC_001643.1_5321-6862      ATGTTCACCGACCGCTGACTATTCTCTACAAACCACAAAGATATTGGAAC
NC_011519.1_5839-7377      ATGCTCATCAATCGCTGACTCTTTTCAACGAACCATAAAGACATCGGAAC
NC_001788.1_5357-6901      ATGTTCATTAACCGCTGACTATTTTCAACTAACCACAAAGACATCGGCAC
NC_000845.1_6511-8055      ATGTTCGTAAATCGTTGACTATACTCAACAAACCACAAAGACATCGGCAC
NC_002008.4_5349-6893      ATGTTCATTAACCGATGACTGTTCTCCACTAATCACAAGGATATTGGTAC
NC_001665.2_5323-6867      ATGCTCGTAAACCGTTGACTCTTTTCAACTAACCACAAAGATATCGGAAC
NC_024511.2_1474-3009      ----TCGCGA--CAATGATTATTTTCTACAAATCATAAAGATATCGGAAC
                               **      *  *** * *  ** ** ** ** ** ** ** **

NC_012920.1_5904-7445      ACTATACCTATTATTCGGCGCATGAGCTGGAGTCCTAGGCACAGCTCTAA
NC_001643.1_5321-6862      ACTATACCTACTATTCGGTGCATGAGCTGGAGTCCTGGGCACAGCCCTAA
NC_011519.1_5839-7377      CCTATATTTACTATTTGGTGCGTGAGCTGGAGTCATAGGCACCGCCCTGA
NC_001788.1_5357-6901      TCTGTACCTCCTATTCGGCGCTTGAGCTGGAATAGTAGGAACCGCCCTAA
NC_000845.1_6511-8055      CCTGTACCTACTATTTGGTGCCTGAGCAGGAATAGTGGGCACTGCCTTGA
NC_002008.4_5349-6893      TTTATACTTACTATTTGGAGCATGAGCCGGTATAGTAGGCACTGCTTTGA
NC_001665.2_5323-6867      CCTCTACCTATTATTTGGAGCCTGAGCAGGAATAGTAGGGACAGCTTTAA
NC_024511.2_1474-3009      TTTATATTTTATTTTTGGAGCTTGAGCTGGAATAGTTGGAACATCTTTAA
                             * **  *  * ** ** ** ***** **  *  * ** **  *  * *

NC_012920.1_5904-7445      GCCTCCTTATTCGAGCCGAGCTGGGCCAGCCAGGCAACCTTCTAGGTAAC
NC_001643.1_5321-6862      GTCTCCTTATTCGGGCTGAACTAGGCCAACCAGGCAACCTCCTAGGTAAT
NC_011519.1_5839-7377      GCCTTCTCATTCGAGCTGAACTAGGCCAACCCGGCAACCTATTAGGCAAC
NC_001788.1_5357-6901      GCCTCCTAATCCGTGCTGAATTAGGTCAACCTGGGACCCTGCTGGGAGAT
NC_000845.1_6511-8055      GCCTACTAATTCGCGCTGAACTAGGTCAGCCCGGAACCCTACTTGGCGAT
NC_002008.4_5349-6893      GCCTCCTCATCCGAGCCGAACTAGGTCAGCCCGGTACTTTACTAGGTGAC
NC_001665.2_5323-6867      GTATTCTAATTCGAGCTGAACTAGGACAGCCAGGCGCACTCCTAGGAGAT
NC_024511.2_1474-3009      GAATTTTAATTCGAGCTGAATTAGGACATCCTGGAGCATTAATTGGAGAT
                           *  *  * ** ** ** **  * ** ** ** **     *  * **  *
```

Figure 2 – Clustal format file output from the ClustalW tool

In the clustal format, lines of each of the eight species are shown. A star at the bottom of a position indicates that the base pair is the same across all sequences.

Step 4 - Convert from clustal to fasta

The ape package in R uses an aligned fasta file to create the phylogenetic tree, so we need to convert the clustal file to fasta format. The conversion from clustal to fasta format can be done by uploading our clustal file from galaxy at this link (http://sequenceconversion.bugaco.com/converter/biology/sequences/clustal_to_fasta.php). Select the clustal format to fasta format, then upload the clustal file from galaxy. Once the file is converted, the file can be downloaded. This aligned fasta file is now ready to be used to construct a phylogenetic tree. The file will appear as shown below.

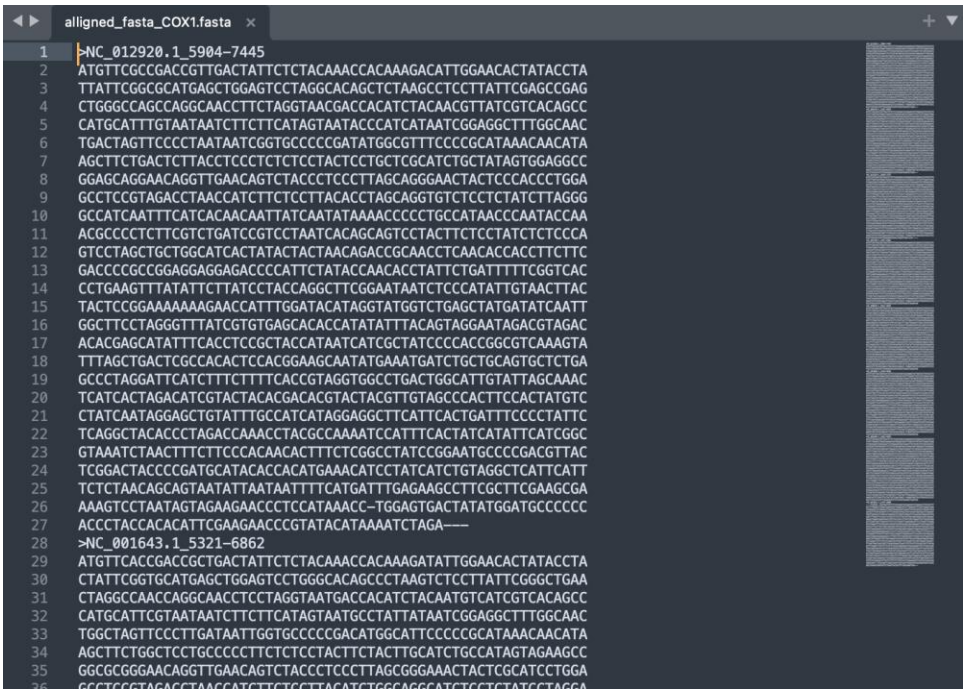


Figure 3 – Aligned fasta file

Step 5 – Constructing phylogenetic trees

To customize the labels that will appear on our tree, we can edit the file, and replace the accession numbers with what we want the labels to be. This is shown below.

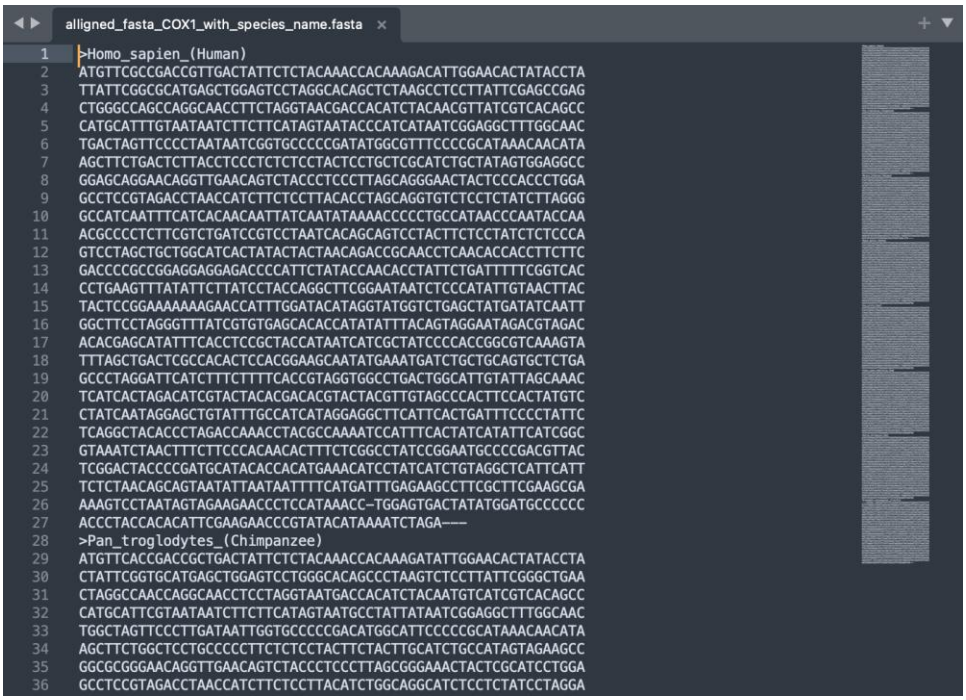


Figure 4 – Aligned fasta file with names replacing accession number

With the new file containing the species names, we can create a clearer phylogenetic tree.

A. The first step to constructing the phylogenetic trees in R is to install and load the APE package (https://cran.r-project.org/web/packages/ape/ape.pdf) as shown below.

```
install.packages("ape")
library(ape)
```

Once the ape library has been loaded in, we can begin constructing the trees. Below is the code used to construct a phylogenetic tree using the GAPDH sequences.

B. Now we will use the read.dna() function to read in our fasta file. We must set the format equal to fasta as well.

```
GAPDH_sequences_with_names <-
read.dna("/Users/johnmariano/Desktop/GAPDH/alligned_fasta_GAPDH_with_names.fasta", format = "fasta")
```

C. Now we can use this object to create a distance matrix, which quantifies the similarity between each of the sequences. This is done using the dist.dna() command.

```
GAPDH_dist_matrix_names <- dist.dna(GAPDH_sequences_with_names)
```

D. Next, the nj() command is used to create a neighbor joining tree from the distance matrix.

```
GAPDH_nj_tree_names <- nj(GAPDH_dist_matrix_names)
```

E. Next, the root() command is used to root the tree based on an outgroup

```
GAPDH_rooted_tree <- root(GAPDH_nj_tree_names, outgroup =
"Drosophila_melanogaster_(Fruitfly)",resolve.root = TRUE)
```

F. Next, the plot() function is used to visualize the neighbor joining tree. Using this code, the GAPDH sequences used will produce a tree.

```
plot.phylo(GAPDH_rooted_tree)
```

G. Now a title can be added to our tree with the title() command.

```
title('GAPDH')
```

H. Lastly, a scale bar can be added to the tree using the add.scale.bar() command

```
add.scale.bar(ask = TRUE)
```

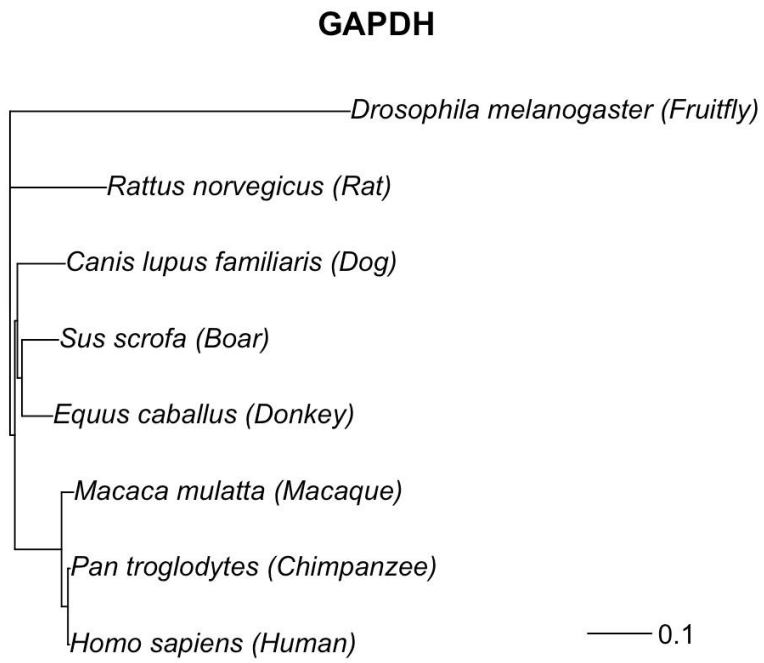


Figure 5 – GAPDH NJ tree

When the steps above are repeated for the COX1 sequences, the tree created will appear as shown below.

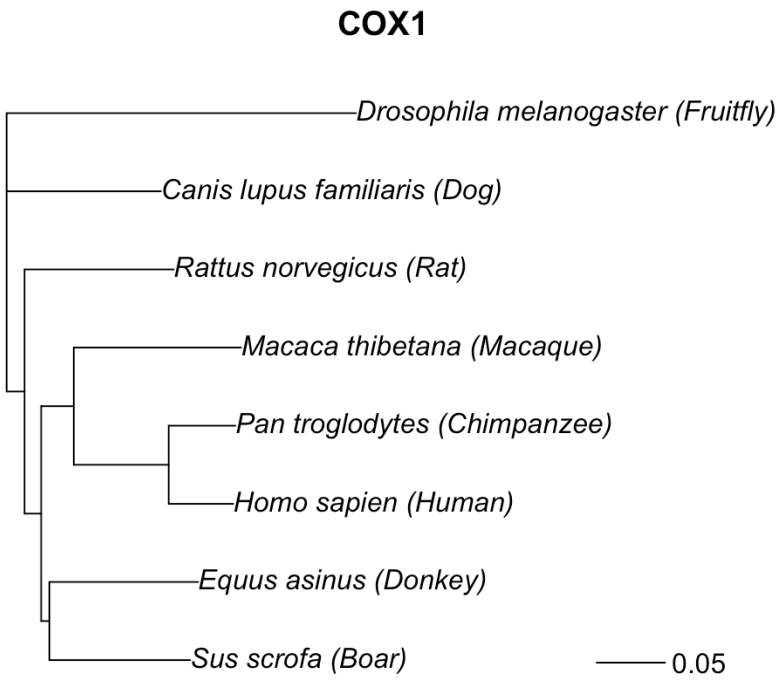


Figure 6 – COX1 NJ tree

Results:

The GAPDH NJ tree exhibits a clear clustering of species. Notably, it separates the species into distinct groups, highlighting the relationships between them. Let's explore some of the key groupings and relationships. The species Homo sapiens (Human), Pan troglodytes (Chimpanzee), and Macaca mulatta (Macaque) are found in a tight cluster, indicating a close genetic relationship. This is consistent with the known evolutionary proximity between these primates. Based on the branch lengths, the fruit fly appears as the most distant species, which is expected, given the vast genetic differences between insects and mammals. The branch lengths in the NJ tree reflect the genetic distances between the gene of each species. Shorter branches suggest close genetic relationships, while longer branches represent more distant relationships (Kumar, S., & Hedges, S. B. 2011). For example, the branch connecting the Human, Chimpanzee, and Macaque is quite short, reflecting their recent common ancestry. The GAPDH scale bar has a value of 0.1, and the fruitfly has about four scale bar lengths. This gives the fruit fly GAPDH gene a distance of about 0.4. This can be used to compare with the value for the COX1 gene.

The COX1 NJ tree has many similarities to the GAPDH tree. In the COX1 tree, the Human, Chimpanzee, and Macaque were also shown grouped together indicating a close genetic relationship. The Fruit Fly was again the most distantly related based on its branch length being the longest, but in the COX1 tree, the relative branch length of the Fruit Fly was not as pronounced as in the GAPDH tree. Also, based on the COX1 tree, the Dog is the closest relative

to the Fruit Fly, but based on the GAPDH tree, the Rat is the most closely related to the Fruit Fly. By looking at the scale bar for the COX1 gene, the fruit fly genetic distance of the COX1 can be estimated to be about 0.2. We can compare these distance values of the two genes to make inferences about the evolution of these genes. The COX1 gene has a lower distance value suggesting that the COX1 gene has been more conserved than the GAPDH gene in the evolution between fruit flies and the other mammals.

In conclusion, the phylogenetic trees constructed using the GAPDH gene and the COX1 gene exhibit both notable similarities and differences. While both trees provide valuable insights into the evolutionary relationships among the analyzed species, they reflect distinct aspects of their genetic history. The similarities between the trees validate common ancestry and known evolutionary patterns, while the differences highlight the influence of gene-specific evolutionary rates and histories. Combining the information from multiple gene trees can offer a more comprehensive understanding of the complex evolutionary relationships among the species under investigation. (Degnan, J. H., & Rosenberg).

References:

Kumar, S., & Hedges, S. B. (2011). Molecular Evolution and Phylogenetics. Oxford University Press. ISBN-13: 978-0199562597.

National Center for Biotechnology Information. (n.d.). Cytochrome c oxidase subunit 1 [GeneID 4512]. NCBI Gene Database. Retrieved November 2, 2023, from https://www.ncbi.nlm.nih.gov/gene/4512

National Center for Biotechnology Information. (n.d.). Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) [GeneID 2597]. NCBI Gene Database. Retrieved November 2, 2023, from https://www.ncbi.nlm.nih.gov/gene/2597

National Center for Biotechnology Information. (n.d.). Molecular Evolution and Phylogenetics. BMC Evolutionary Biology, 5(1), 8. https://bmcecolevol.biomedcentral.com/articles/10.1186/1471-2148-5-8

Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. \textit{Trends in Ecology & Evolution}, 24(6), 332–340. https://doi.org/10.1016/j.tree.2009.01.009