Warsaw University of Technology

BIOINFORMATICS

Clustering and Phylogeny

Authors:
Abba Umar

January 3, 2024

Contents

List of Figures

1 Project Goal

The objective of this project is to conduct a comprehensive phylogenetic analysis of eight different human proteins. The analysis will involve clustering the protein sequences, constructing separate and common phylogenetic trees, and comparing the results to draw meaningful conclusions about evolutionary relationships

2 Dataset Description

The project's datasets are a carefully chosen collection of proteins that represent a range of biological functions. Every dataset contains a variety of protein sequences derived from various organisms. Eight different human proteins were selected to serve as focal points in the dataset construction process. Using internet resources like BLAST, sequences from seven different organisms were found for each of these proteins. These proteins include:

2.1 Alpha-amylase

Alpha-amylase, also known as alpha-1,4-glucan 4-glucanohydrolase, is an enzyme that hydrolyzes alpha-1,4-glycosidic bonds in starch and glycogen, yielding shorter chains thereof, dextrins, and maltose. It is the primary enzyme involved in the digestion of starch in the human digestive system. Alpha-amylase is produced by the salivary glands, pancreas, and small intestine.

2.2 Hemoglobin

Hemoglobin subunit gamma-1 is a protein that is found in the fetal and neonatal red blood cells. It is a component of fetal hemoglobin (HbF), which is the primary form of hemoglobin in the fetus. HbF has a higher affinity for oxygen than adult hemoglobin (HbA), which means that it can more easily bind to oxygen in the mother's blood and transport it to the fetus.

2.3 Insulin

Insulin is a peptide hormone produced by beta cells in the islets of Langerhans of the pancreas. It plays a crucial role in regulating blood sugar levels by facilitating the uptake of glucose from the bloodstream into cells. Insulin is essential for the proper functioning of various tissues in the body, including muscle, liver, and fat cells.

2.4 Thymosins

Thymosins are a family of over 40 peptides, each with unique amino acid sequences. They are typically around 10 amino acids long and are highly conserved across species. Thymosins are produced by thymic epithelial cells, which are the main type of cell in the thymus. They are released into the thymic environment, where they interact with developing T cells to promote their maturation and differentiation.

2.5 Kallikrein

Kallikreins are a group of serine proteases that are involved in a variety of physiological processes, including blood pressure regulation, inflammation, and wound healing. They are produced in many different tissues, including the kidney, pancreas, salivary glands, and skin.

2.6 Calsyntenins

Calsyntenins (Csts, CLSTN) are a family of type I transmembrane proteins that belong to the cadherin superfamily. Their name comes from their ability to bind calcium. They are found in various tissues throughout the animal kingdom, including humans.

2.7 Collagen

Collagen is a long, triple-helical molecule that is made up of three polypeptide chains. Each polypeptide chain is made up of repeating units of glycine, proline, and hydroxyproline. The triple helix is stabilized by hydrogen bonds and covalent crosslinks.

2.8 Adenosine deaminase

Adenosine deaminase (ADA) is an enzyme that catalyzes the deamination of adenosine to inosine. This reaction is part of the purine salvage pathway, which is responsible for recycling purine bases that are not needed for DNA and RNA synthesis.

3 Organism Description

The above proteins were meticulously chosen from a range of organisms, each contributing unique insights into diverse biological processes. The selected organisms include:

(a) Pan troglodytes

(b) Pan paniscus

(c) Gorilla gorilla

(a) Pongo abelii

(b) Theropithecus gelada

(c) Macaca thibetana thibetana

(d) Papio anubis

4 ClustalW as local blast

Upon completion of the download process for sequences corresponding to each protein across the selected eight organisms (including seven organisms and humans), I proceeded to conduct a multiple sequence alignment (MSA) for each protein. The alignment was executed using the ClustalW application, which was sourced from http://www.clustal.org/download/current/. This organized approach facilitated the comparison of sequences from different organisms for enhanced analytical insights.

- To initiate the alignment process, ClustalW requires a file as input. This file should contain three or more protein sequences. For instance, the input file can be "Alphaamylase_BLAST_sequence.txt" located in the project folder. This file includes protein sequences of adenosine from various organisms.
- ClustalW, the selected alignment application, aligns the input sequences and generates an output file in FASTA format. Users can choose alternative output formats within the application. Using the example of "Alphaamylase_BLAST_sequence.txt," the application produces a new file named "Alphaamylase_msa_output.fasta" for the alignment and "Alphaamylase_msa_tree" for the tree. This file contains the aligned sequences, facilitating comparative analysis.
- Upon completing multiple sequence alignments for protein sequences, the generated files, like "Alphaamylase_msa_output.fasta" undergo refinement through renaming e.g."Alphaamylase_msa_output_edited.fasta". Notably, a crucial enhancement involves adding the organism's name to the header of each sequence within protein groups. This modification enhances contextual reference, proving valuable for future analyses such as constructing phylogenetic trees and conducting comparative studies, contributing to dataset clarity and accuracy.

```
>XP_004050475.2
MGSETIKPVGTQQPSALQDRLHQKRPSSRSVPRAFAS----------
----------DHCPSAMALWMRLLPLLALLALWGPDPAAAFV
NQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSL
QKRGIVEQCCTSICSLYQLENYCN
>ABI63346.1
------------------------------------------------------------
--------------------------------MALWMRLLPLLALLALWGPDPAAAFV
NQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQ-----------GSLQPLALEGSL
QKRGIVEQCCTSICSLYQLENYCN
>XP_034787832.1
MGSETIKPAGTQQPSALQDRLHQKRPSSRSVPRAFASGGLRVPGWLDPRPQLCSREDVAG
LLKHVGVSPGAPRQGTWPSAGLSPACLPDHCPSAMALWMRLLPLLALLALWGPDPASAFV
NQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSL
QKRGIVEQCCTSICSLYQLENYCN
```

(a) Before

```
>[Gorilla gorilla gorilla]  XP_004050475.2
MGSETIKPVGTQQPSALQDRLHQKRPSSRSVPRAFAS----------
----------DHCPSAMALWMRLLPLLALLALWGPDPAAAFV
NQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSL
QKRGIVEQCCTSICSLYQLENYCN
>[Homo sapiens] ABI63346.1
------------------------------------------------------------
--------------------------------MALWMRLLPLLALLALWGPDPAAAFV
NQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQ-----------GSLQPLALEGSL
QKRGIVEQCCTSICSLYQLENYCN
>[Pan paniscus] XP_034787832.1
MGSETIKPAGTQQPSALQDRLHQKRPSSRSVPRAFASGGLRVPGWLDPRPQLCSREDVAG
LLKHVGVSPGAPRQGTWPSAGLSPACLPDHCPSAMALWMRLLPLLALLALWGPDPASAFV
NQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSL
QKRGIVEQCCTSICSLYQLENYCN
```

(b) After

5 Clustering

With the acquired files containing aligned sequences representing diverse organisms for each protein, the ensuing phase involves leveraging this dataset as input for clustering methodologies. This approach endeavors to delineate clusters or groups of sequences pertaining to individual proteins. The chosen clustering method involves the utilization of the Biopython Cluster Library.

Specifically, the application of interest within the Cluster Library is the "kcluster" function. This function orchestrates the k-means clustering algorithm on the numerical values present in the data. The output of this function encompasses cluster assignments, the within-cluster sum of distances corresponding to the optimal k-means clustering solution, and the count of occurrences wherein the optimal solution was ascertained.

The k-means clustering methodology systematically segregates sequences into distinct clusters predicated on their intrinsic similarities, thereby furnishing a systematic and quantifiable framework for subsequent analytical endeavors. This methodological rigor is paramount for the meticulous organization and interpretation of the biological sequences under investigation.

5.1 Input Parameters for kcluster()

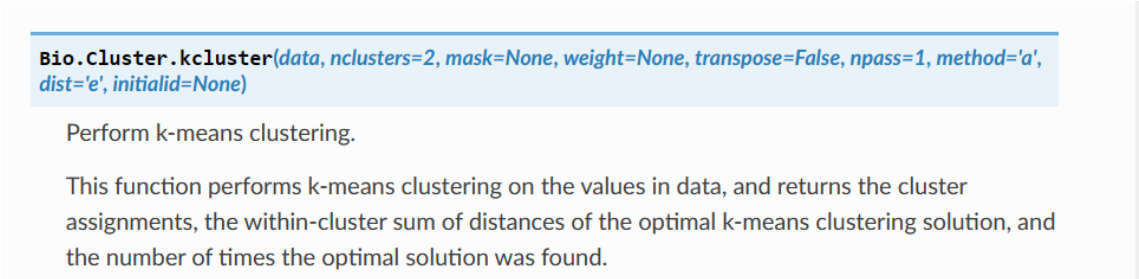

Figure 4: Kcluster
UNI (2023)

- data: The input should be a 2-dimensional array of arrays representing sequences. However, after reading multiple sequence alignment (MSA) protein sequences files using AlignIO.read() and appending each sequence to a dataset of lists, we obtain a list of strings containing aligned sequences. To ensure compatibility with the kcluster() function, the data needs to be converted into a 2-dimensional array of integers. This conversion can be achieved using the fromstring() function from the numpy library.

- nclusters: This parameter denotes the number of clusters, with a default value of 2. While there is no strictly correct value for nclusters, it should be greater than 1 and less than the number of sequences used as input. Choosing the appropriate nclusters requires careful consideration of the dataset. In this project, given the selection of families for organisms in addition to humans, the default value of 2 is retained.
- npass: This parameter represents the number of times the clustering algorithm is executed, each time with a different initial condition. To enhance result accuracy, the value of npass is modified to a larger number, such as 1000. This adjustment contributes to obtaining a more robust and reliable clustering outcome.

For further details and code implementation regarding the aforementioned explanations, refer to the "ClusteringAndPhylogeny.ipynb" file in the Project folder.

5.2 Output of kcluster():

The output of the kcluster() function comprises a tuple representing the following components:

- Clusterid: An array containing the ID number of the cluster (e.g., 0 or 1) to which each item was assigned. For instance, if Clusterid = [1 1 1 1 1 0 0 0], it signifies that the first five sequences belong to one cluster, while the last three sequences belong to another cluster based on their similarity. It's important to note that [1 1 1 1 1 0 0 0] is equivalent to [0 0 0 0 0 1 1 1], as it represents the same two clusters with different IDs.
- Error: This value represents the within-cluster sum of distances for the clustering solution derived from the algorithm.
- nfound: The number of times this clustering solution was found during the algorithm's execution.

Post-clustering, a decision was made to write each cluster containing one or more aligned sequences from each protein into a FASTA format file. This step is crucial for subsequent tree construction for each cluster. For instance, within the project folder, two FASTA format files, namely "Alphaamylase_cluster1" and "Alphaamylase_cluster2," have been created to store the two clusters of Alphaamylase sequences. This organization facilitates further analysis and visualization of clustered sequences within the context of each specific protein.

5.3 Evaluation of Clusters

Upon performing multiple sequence alignment (MSA) followed by clustering for the Alpha-amylase protein across different organisms, the obtained cluster assignments are represented as [1 1 1 1 1 0 0 0] or [0 0 0 0 0 1 1 1]. This signifies two distinct clusters with the following observations:

- The first cluster comprises the first five sequences from the alignments, corresponding to the proteins XP_009424882.1, XP_054971359.1, AAA52279.1, XP_030858925.1, and XP_024108682.1. These sequences exhibit similarities with organisms such as Pan_troglodytes, Pan_paniscus, Homo_sapiens, Gorilla gorilla_gorilla, and Pongo_abelii.
- The second cluster includes the remaining sequences from the alignments, linked to the proteins XP_050634715.1, XP_031507222.1, and XP_025215097.1. These sequences share affinities with the organism Macaca thibetana_thibetana, Papio_anubis, and Theropithecus_gelada.

Comparison with BLAST results reveals that the clusters align well with the similarity between protein sequences from different organisms:

- Homo sapiens: 100
- Pan troglodytes: 99.02
- Pan paniscus: 98.83
- Gorilla gorilla gorilla: 98.63
- Pongo abelii: 97.65
- Theropithecus gelada: 95.11
- Macaca thibetana thibetana: 94.72
- Papio anubis: 93.93

This alignment between clusters and BLAST results is consistent across all analyzed proteins. The findings underscore the reliability of the clustering approach in capturing the underlying phylogenetic relationships among diverse organisms based on their protein sequences. Further exploration of phylogenetics is warranted to deepen our understanding of evolutionary connections among the studied organisms.

6 Phylogenetics

In the upcoming phase, Phylo.TreeConstruction from Biopython will be employed to generate phylogenetic trees. The procedure involves creating trees for each group of proteins following alignment (MSA), with a dedicated tree for each cluster. Additionally, a final tree encompassing all downloaded sequences together will be constructed.

6.1 Trees for Each Group of Protein

In the process of constructing phylogenetic trees for each group of proteins, the following steps are undertaken:

- Utilize the AlignIO class module with its read() function to read the aligned sequences for each group of proteins.
- Apply the DistanceCalculator() function from Phylo to compute the distances between the aligned sequences.
- Use the DistanceTreeConstructor() function to construct trees utilizing the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algorithm. UPGMA is a clustering method that sequentially merges clusters of similar items to build a phylogenetic tree.
- An example of a tree for the Alpha-amylase protein will be generated, visually representing the evolutionary relationships among the aligned sequences.

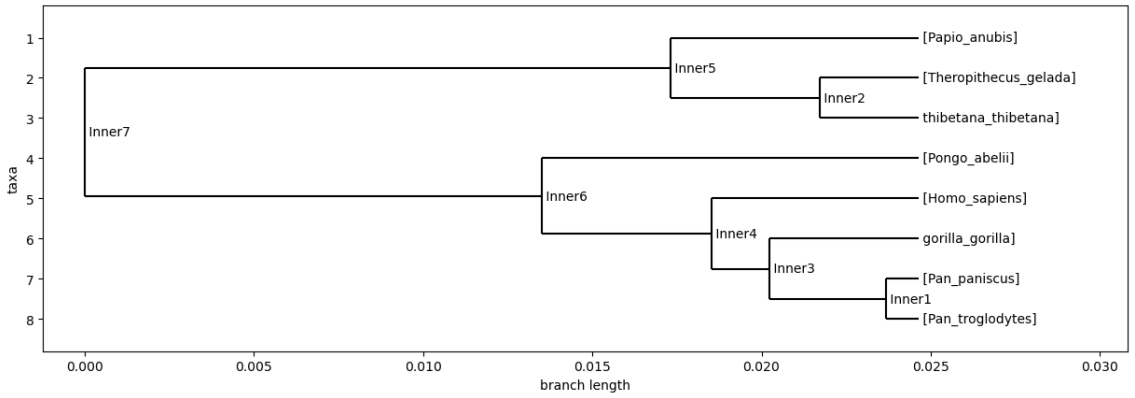

Figure 5: Tree for Alpha-amylase protein

- After constructing and drawing each tree, append it to a Newick format file named "trees.dnd". This file consolidates all the trees for subsequent use in

building a consensus tree.

By systematically applying these steps for each protein group and cluster, the resulting trees will offer insights into the evolutionary connections among organisms based on their protein sequences. The "trees.dnd" file will serve as a comprehensive repository of all generated trees, facilitating the construction of a consensus tree in subsequent analyses.

6.2 Trees for Each Cluster

To construct phylogenetic trees for each cluster, the above procedure was implemented, focusing on the Alpha-amylase protein as an example:

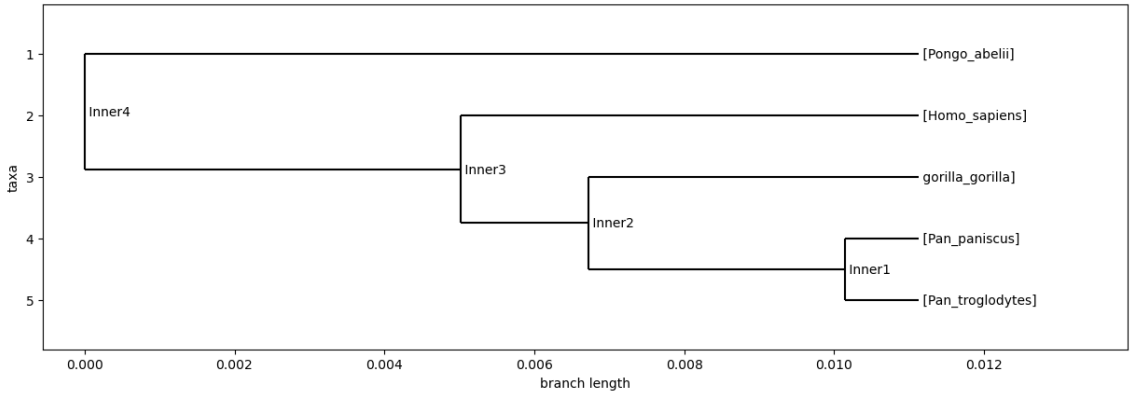

Figure 6: Tree for Alpha-amylase protein cluster1

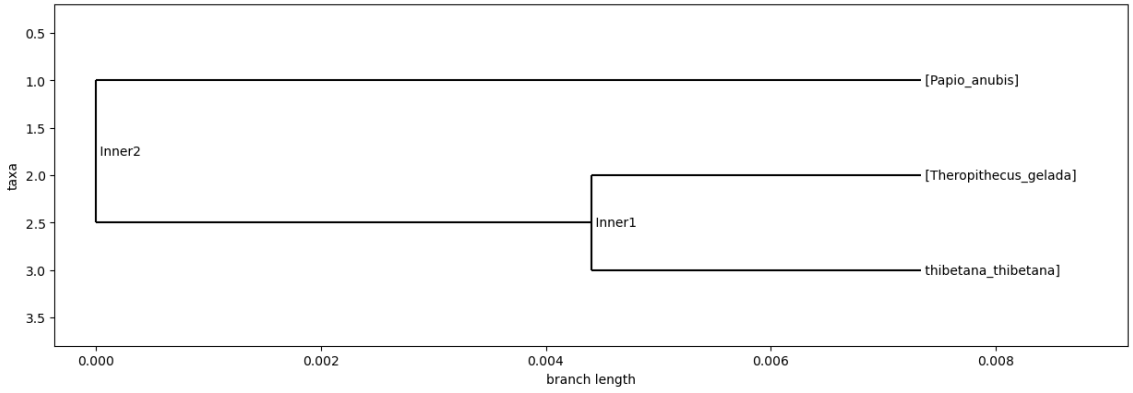

Figure 7: Tree for Alpha-amylase protein cluster2

These trees provide detailed insights into the evolutionary patterns within distinct

groups of sequences, enhancing our understanding of the relationships among organisms based on the Alpha-amylase protein.

6.3 Common Tree for All Sequences

To construct a unified phylogenetic tree for all downloaded sequences, the following approach was employed:

- All sequences from diverse organisms were compiled into a single fasta format file. This consolidated file encompassed sequences from various proteins and organisms.
- ClustalW was utilized to perform MSA on the compiled sequences. This step is crucial as it ensures that the sequences are aligned, facilitating the subsequent construction of a meaningful phylogenetic tree.
- The Phylo.TreeConstruction module, along with the UPGMA algorithm, was employed to build a comprehensive phylogenetic tree. This algorithm systematically merges clusters of similar sequences, generating a tree that represents the overall evolutionary relationships among all downloaded sequences.

The resulting tree, displayed below, provides a holistic view of the evolutionary connections across different organisms based on the aligned sequences. This unified tree is instrumental in understanding the broader phylogenetic patterns present within the entire dataset.

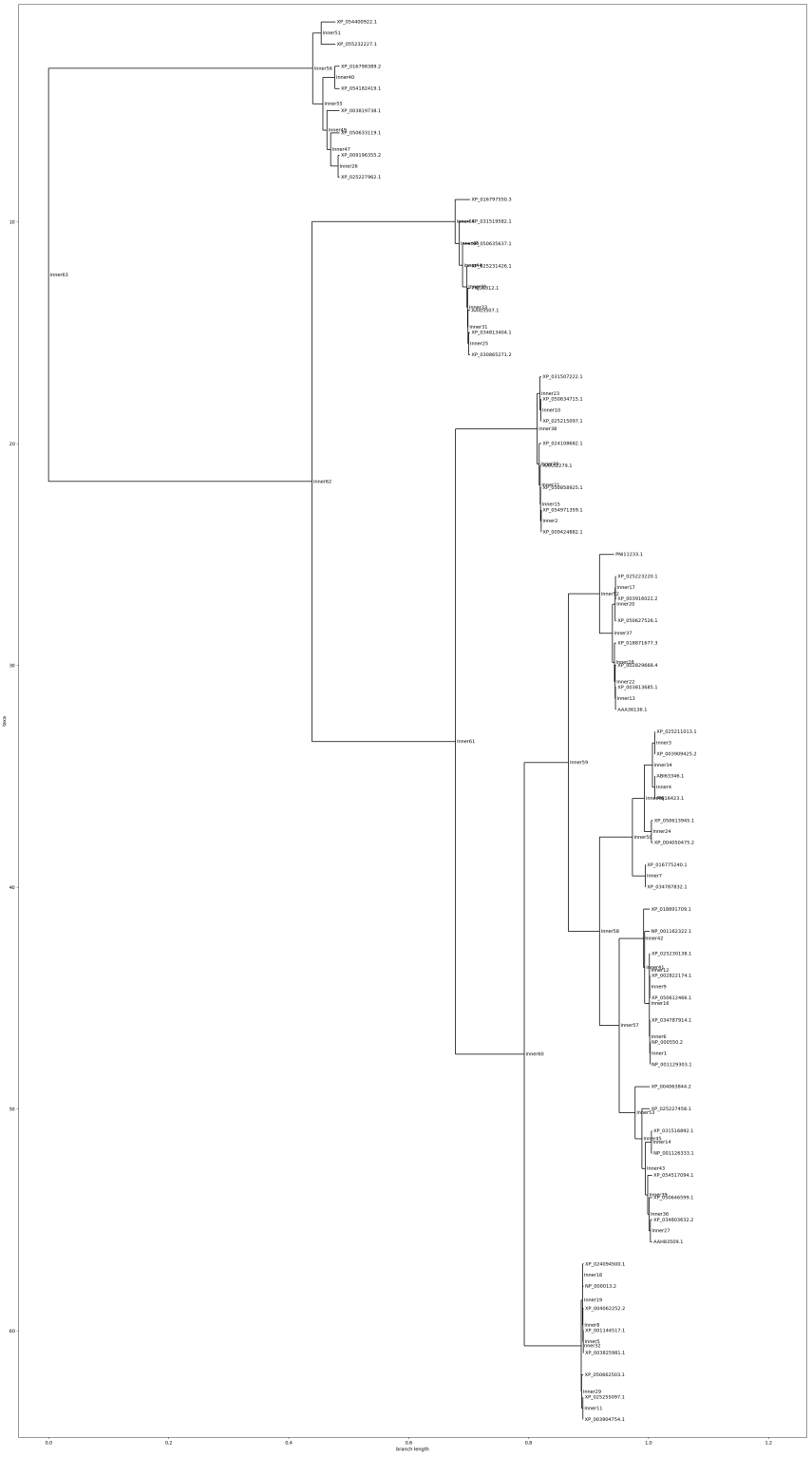

Figure 8: Tree for All sequence

This consolidated tree serves as a valuable resource for comprehending the overarching evolutionary relationships among diverse organisms, offering a global perspective on the phylogenetic structure of the downloaded sequences.

6.4 Consensus Tree Generation

Creating consensus trees is a crucial step in synthesizing multiple phylogenetic trees into a unified representation that captures common branching patterns. The process involves generating a consensus tree for each cluster by leveraging the majority_consensus function with varying thresholds. In this project, two attempts were made to build consensus trees:

- The initial attempt to construct a consensus tree from the trees of various clusters encountered challenges. Consensus trees demand a consistent taxonomy across the input trees. However, due to the presence of varying numbers of sequences within some clusters, achieving uniform taxonomy proved impractical. As a result, the task of building a consensus tree from these clusters proved to be unfeasible. Notably, the clusters corresponding to Alpha-amylase, Collagen, Insulin, Adenosine, and Thymosin were exceptions to this limitation. These specific clusters, characterized by a more uniform sequence count, successfully underwent consensus tree construction.

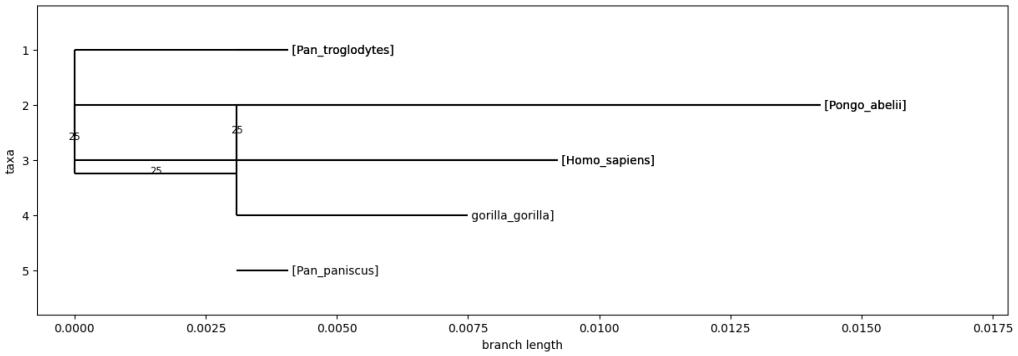

Figure 9: Consensus tree cluster

- The second consensus tree was successfully constructed from the trees of each protein group. The second consensus tree derived from protein groups offers a comprehensive overview of shared evolutionary signals within each protein category.

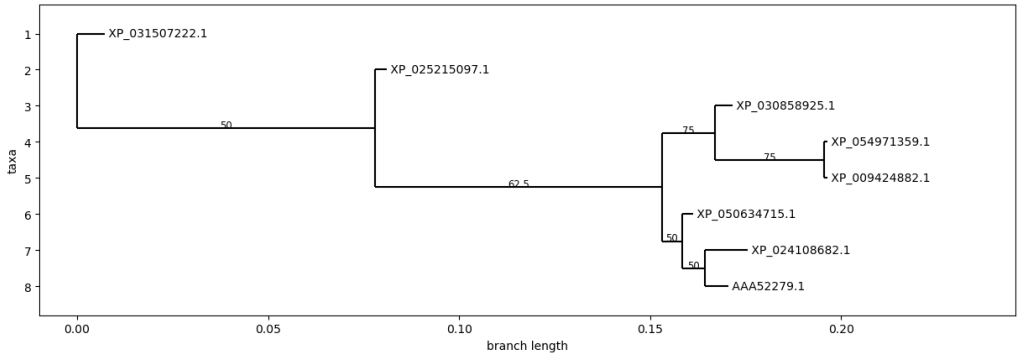

Figure 10: Consensus tree cluster

6.5 Visualization by Color

I have successfully constructed the phylogenetic tree for all the sequences using ClustalW and enhanced its visual representation by employing Geneious Prime to apply color coding. The utilization of ClustalW allowed for the alignment of sequences, facilitating a robust foundation for the tree's structure. Geneious Prime's advanced features then enabled me to intuitively colorize the branches, providing a visually informative depiction of relationships within the dataset. This integrated approach not only ensures the accuracy of the phylogenetic tree but also enhances its accessibility and interpretability through the strategic use of color in Geneious Prime. I have intricately colored the components of the constructed phylogenetic tree using Geneious Prime, employing a distinctive palette to enhance visual clarity.

Proteins:

- Adenosine: Pink (# FFC0CB)
- Alpha-amylase: Teal (#008080)
- Kallikrein: Lavender (#E6E6FA)
- Hemoglobin: Gold (#FFD700)
- Collagen: Indigo (#4B0082)
- Insulin: Slate Gray (#708090)
- Calsyntenin: Olive (#808000)
- Thymosin: Coral (#FF7F50)

Organism:

- Homo sapiens: Red (#FF0000)

- Pan troglodytes: Yellow (#FFFF00)
- Pongo abelii: Green (#00FF00)
- Gorilla gorilla gorilla: Purple (#800080)
- Pan paniscus: Blue (#0000FF)
- Macaca thibetana thibetana: Brown (#A52A2A)
- Papio anubis: Orange (#FFA500)
- Theropithecus gelada: Cyan (#00FFFF)

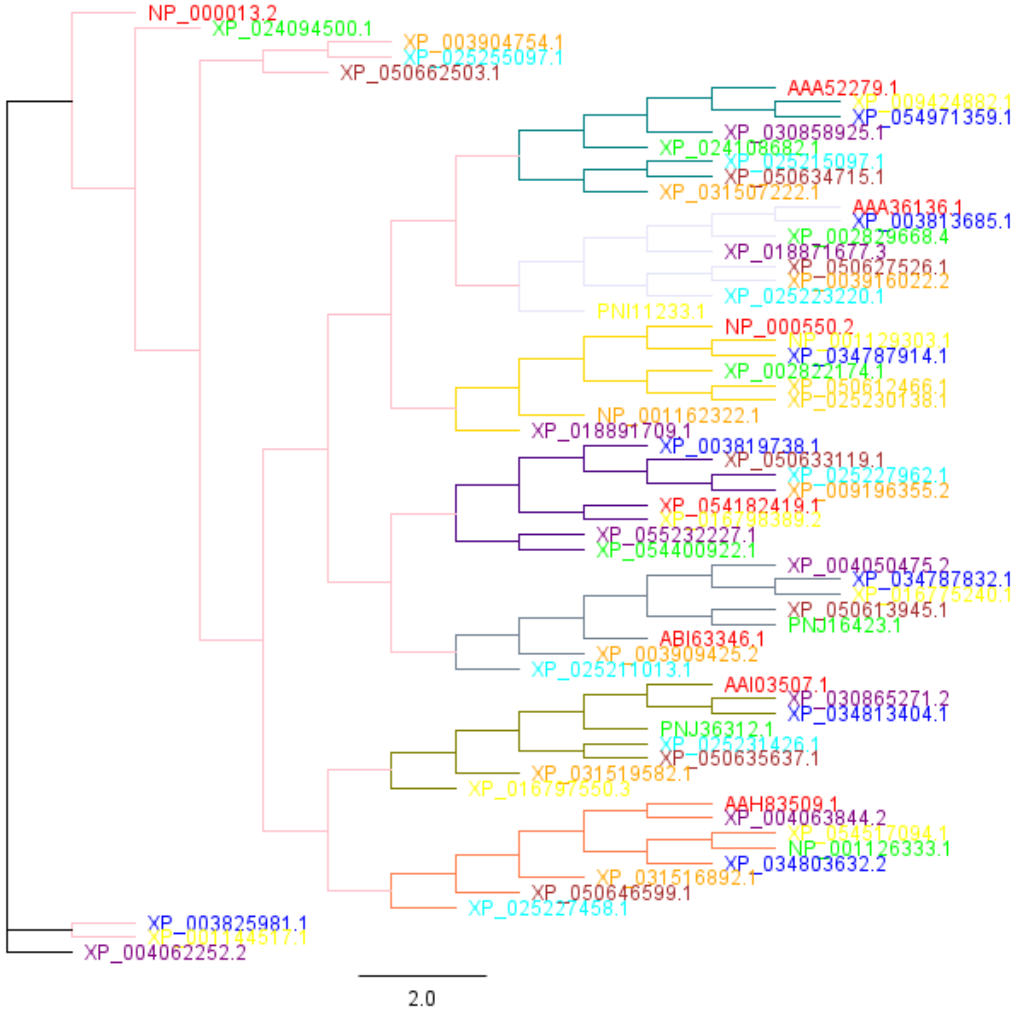

Figure 11: All sequence Tree colored

7 Conclusion

When evaluating the evolutionary relationships among the listed primate species using both the consensus tree and the common tree, distinct patterns emerge. The consensus tree strongly supports the closer relationship between Homo sapiens, Pan troglodytes, and Paniscus by emphasizing precision and clarity. This tree gives a visually cohesive picture of the evolutionary relationship of two closely related species by highlighting their shared ancestry and recent divergence.

Conversely, the common tree, although informative, may present challenges in precisely delineating the intricate relationships among these primate species. While it captures certain correlations, the consensus tree remains superior in its ability to navigate the complexities of evolutionary connections within this set of organisms.

In conclusion, the consensus tree proves invaluable in elucidating the evolutionary relationships among Homo sapiens, Pan troglodytes, Pan paniscus, Gorilla gorilla gorilla, Pongo abelii, Macaca thibetana thibetana, Papio anubis, and Theropithecus gelada. Its accuracy and straightforward representation make it the preferred tool for discerning the nuanced branches of primate evolution, shedding light on the shared ancestry and distinctive paths taken by each species.

References

UNI (2023), "Bio cluster package." URL `https://biopython.org/docs/1.75/api/Bio.Cluster.html#Bio.Cluster.kcluster`.