MERYLIN OGUNLOLA

BioinformHer Mini Project – Module 2 Capstone

Title: Tracking the Evolution of the Hemoglobin Beta (HBB) Gene Across Species

Supervised by: BioinformHer

1. **Task 1: Sequence Retrieval & BLAST Search**

Tools Used

A. NCBI

B. Nucleotide BLAST (blastn)

C. Organisms used : chimpanzee, chicken, zebrafish, cow, mouse, sheep, dolphin, pig, and whale

The human HBB – *homo sapiens* gene was searched on NCBI (https://www.ncbi.nlm.nih.gov/), and the "HBB – hemoglobin subunit beta (NCBI Reference Sequence: NC_000011.10, Gene ID: 3043)" was selected, as it was the first "hit" encountered **Fig1**.

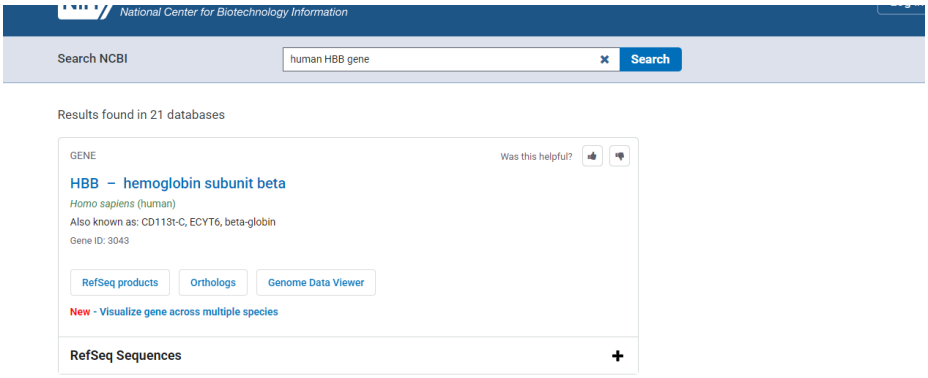


Fig1: Human HBB gene result from NCBI search

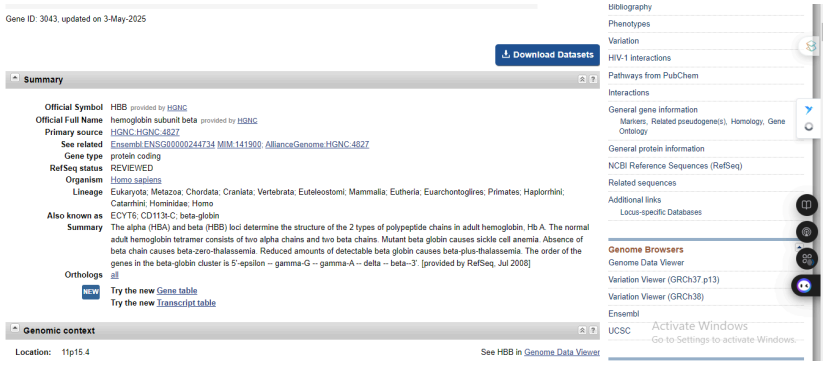


Fig2: Summary of the HBB – hemoglobin subunit beta (NCBI Reference Sequence: NC_000011.10, Gene ID: 3043)

The FASTA sequence for the human HBB – *homo sapiens* gene was copied and pasted in the "Query Sequence" box in BLAST (blastn) (https://www.ncbi.nlm.nih.gov/geo/query/blast.html) **Fig3-4**. The "nr" non – redundant database was used to reduce duplicated hits, and improve speed and the interpretation of our result **Fig4**.

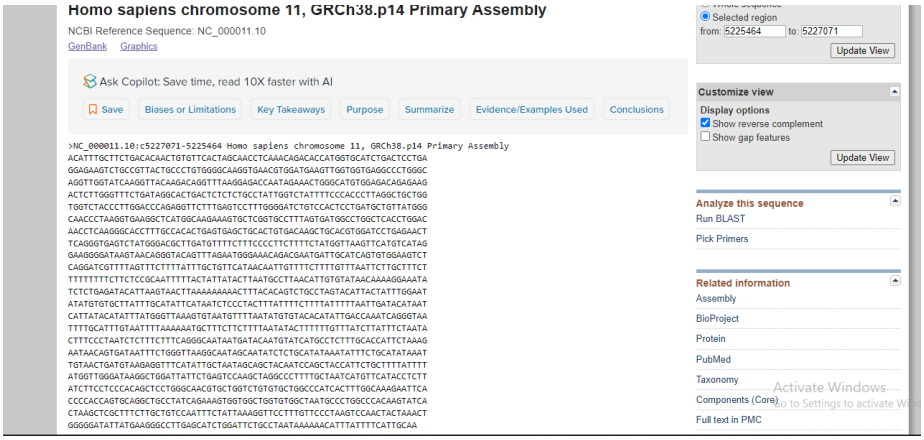


Fig3: FASTA sequence of the HBB - hemoglobin subunit beta gene

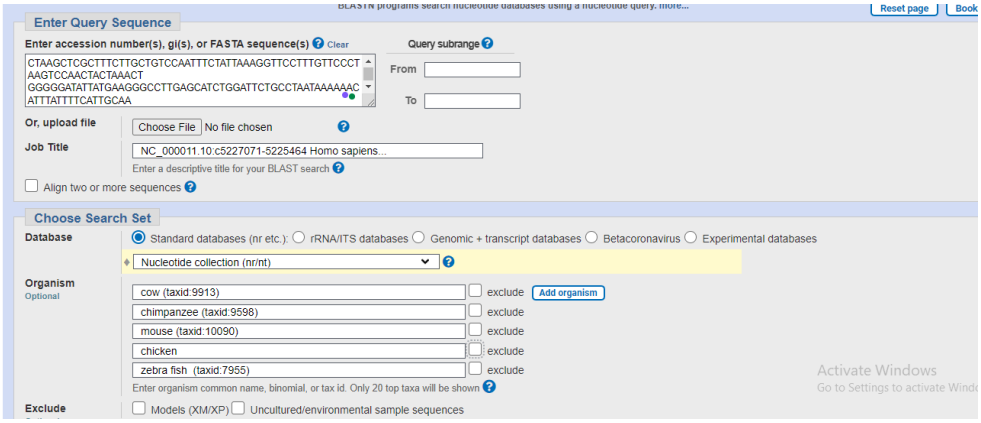


Fig4: NCBI BLAST search for chimpanzees, chicken, zebrafish, cow, and mouse.

BLAST search returned 100 significant results for only three species (chimpanzee, cow, and mouse) from the five (chimpanzee, cow, mouse, chicken, and zebrafish) that were searched for, **Fig6**.

For the next BLAST search, the **Max target sequence** in the **Algorithm parameter** was set to 500. The organisms selected were chimpanzee, cow, mouse, sheep, dolphin, pig, and whale **Fig7-Fig8**.

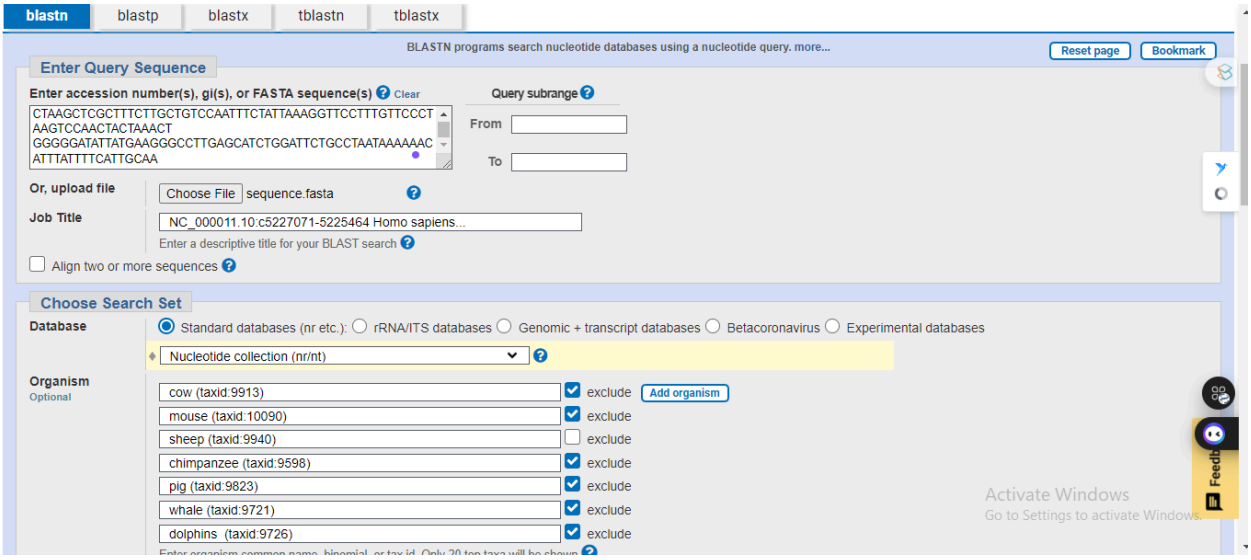


Fig5: NCBI BLAST search for chimpanzee, cow, mouse, sheep, dolphin, pig, and whale

To select the desired species, the following statistical parameters were considered:

Query Coverage: allows us to know what the percentage of our query sequence that aligns with the database hit. A high query coverage indicates that the query sequence compared spans a large portion of the database sequence, and not just conserved regions. A lower coverage means only a small portion of the query sequence matches the database sequence; it can still be significant if it has a very low e-value.

E-value (Expected value): is needed to know the number of times the scores equivalent to or better than the observed score will occur by chance. It is important to assess the likelihood of a matching occurring at random or by chance.

Percent Identity: shows the degree to which the query sequence and the matched database sequence are identical. A higher percent identity suggests a closer evolutionary relationship, and similar functions. Though only considering percent identity can be misleading, especially if the alignment is very short. For percent identities that are high, but not up to 100%, for instance a percent identity of 98%, suggests that 2% of the sequence was different because of natural variations like mutations or sequencing errors.

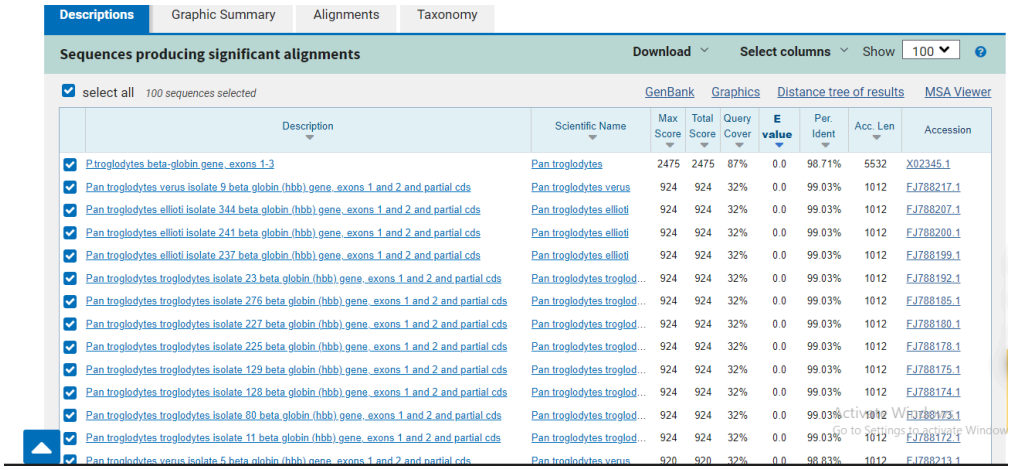


Fig6: BLAST results for chimpanzee, chicken, zebrafish, cow, and mouse.

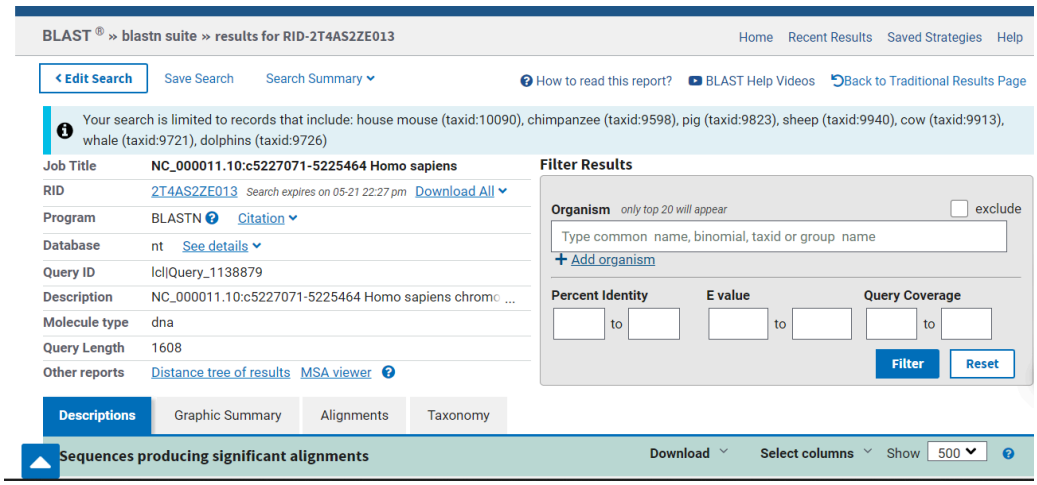


Fig7: BLAST result

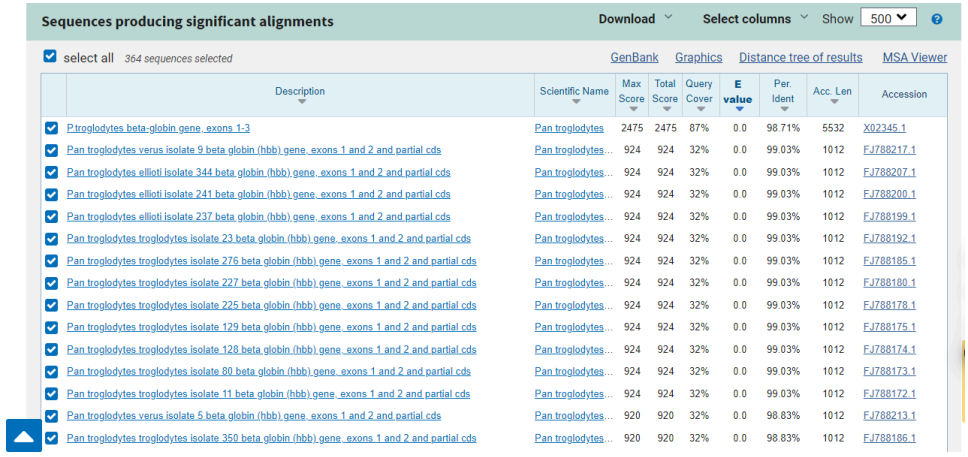


Fig8: BLAST result for chimpanzee, chicken, zebrafish, cow, mouse, sheep, dolphin, pig, and whale.

Table1: Table of the summary of the BLAST results from NCBI

S/N	Species Common Name	Species Scientific Name	Accession number	% identity with human HBB
1	Cow	*Bos taurus*	X00376.1	81.09%
2	Mouse	*Mus musculus*	V00722.1	79.03%
3	Chimpanzee	*Pan troglodytes*	X02345.1	98.71%
4	Whale	*Balaenoptera borealis*	MK622932.1	84.23%
5	Wild Sheep	*Ovis aries musimon*	DQ352468.1	80.99%
6	Pig	*Sus scrofa*	X86791.1	81.94%

The *Bos taurus*–-Cow "Bovine adult beta-globin gene" was selected because of its high percent identity of 81.09%, high e-value of 9e-126. Though its query coverage appears low, it has a relatively higher coverage than other sequences from the same species. Another factor that was considered was its Accession length, which may not be significant in interpreting the result, but because of some factors like very similar results, the accession length can help in evolutionary studies; aligning sequences of similar lengths can be important in providing more accurate phylogenetic analyses to avoid potential biases.

The *Mus musculus* –Mouse"Mouse gene for beta-1-globin" was also selected following the same reasons stated for the "Bovine adult beta-globin gene".

The *Pan troglodytes*–Chimpanzee"P.troglodytes beta-globin gene, exons 1-3" was selected majorly because of its high percent identity (98.71%), query coverage of (87%), and its e-value. Its e-value of 0.0 indicates that the alignment between the query sequence and the database sequence is not due to random chance. It often means that the actual value is extremely small, smaller than the precision limits of the database, signifying the most statistically significant hits possible.

The *Balaenoptera borealis*–Whale "Balaenoptera borealis hemoglobin B (HBB) gene, complete cds" was selected for its low e-value 4e-149, and percent identity of 84.23%.

The *Ovis aries musimon*–Wild sheep "Ovis aries musimon beta globin chain (HBB) gene" was selected for low e-value of 1e-126, and percent identity of 80.99%.

The *Sus scrofa*–Pig "Sus scrofa beta-globin gene" was selected for its e-value of 1e-113 and percent identity of 81.94%.

Task 2 : Pairwise Sequence Alignment

The species were chosen based on observation from the phylogeny in **Fig 9-10.** The species selected were Chimpanzee because it is closely related to humans **Fig9**, and cows because it is distantly related to human **Fig10.**

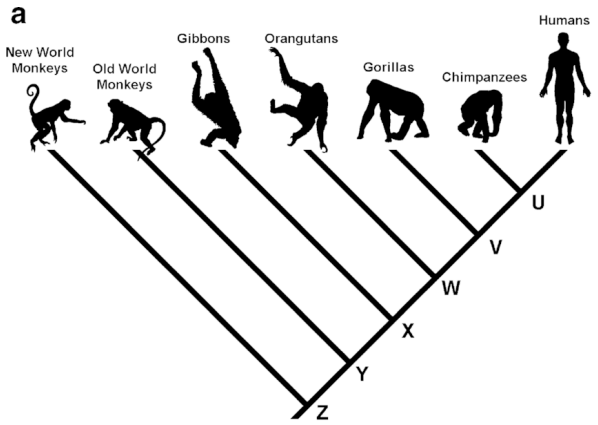


Fig9: Phylogenetic tree showing relatedness of different ape species with humans

Source: (Gregory, 2008)

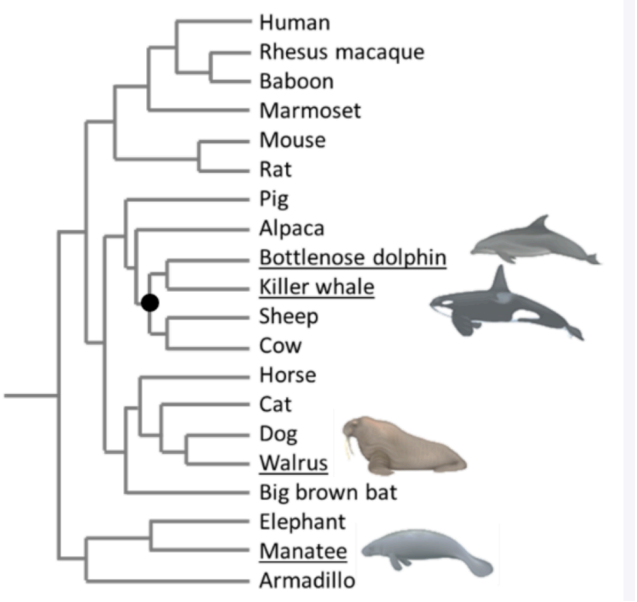


Fig10: Phylogenetic tree showing relatedness of different species with humans

Source: Chegg.com

The sequences of the selected organisms were copied as FASTA format from NCBI, and were pasted in the query box of the pairwise alignment, using EMBOSS needle **Fig11.**

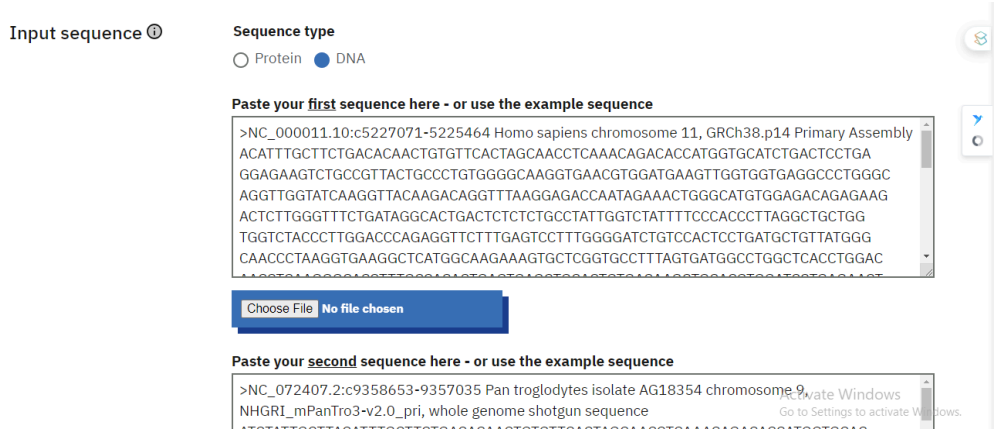
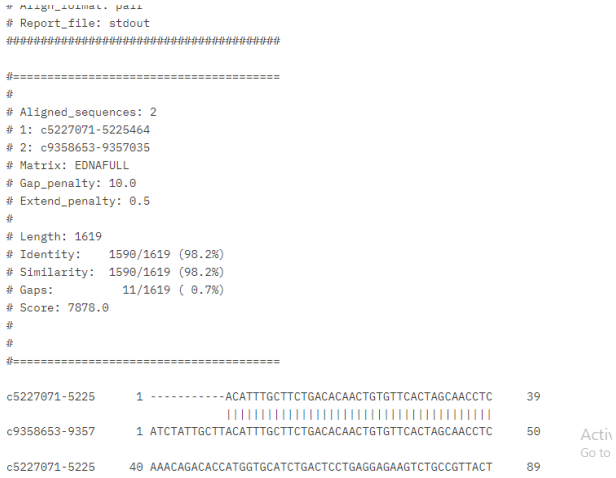


Fig11: A figure of the sequence alignment from EMBOSS needle

2a. Results after aligning the human HBB gene and the chimpanzee HBB gene, with a gap opening penalty of 10.0 and a gap extension penalty of 0.5 were applied **Fig12**.



```
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: c5227071-5225464
# 2: c9358653-9357035
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1619
# Identity:    1590/1619 (98.2%)
# Similarity:  1590/1619 (98.2%)
# Gaps:          11/1619 ( 0.7%)
# Score: 7878.0
#
#
#=======================================

c5227071-5225      1 ----------ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTC     39
                             |||||||||||||||||||||||||||||||||||||||||||
c9358653-9357      1 ATCTATTGCTTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTC     50

c5227071-5225     40 AAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACT     89
```

Fig 12: Result of sequence alignment from EMBOSS needle between *homo sapien* HBB gene and its ortholog in chimpanzees.

Length: 1619 bases

Identical Matches: 1590/1619 (98.2%)

Similarity: 1590/1619 (98.2%)

Gaps: 11/1619 (0.7%)

Alignment Score: 7878.0

From the result we can infer that from the high % identity and similarity of 98.2%, the two sequences have highly conserved regions in the HBB gene, a close evolutionary relationship. The aligned sequences have a gap of 0.7% indicating that there are minimal insertions or deletions between the sequences, and the alignment score of 7878.0 shows strong alignment this shows a better match). This indicates that the two genes evolved from a common ancestry, also indicating the close relatedness between the two organisms; *homo sapiens* and chimpanzees.

2b. Results after aligning the human HBB gene and the cow HBB gene, with a gap opening penalty of 10.0 and a gap extension penalty of 0.5 were applied **Fig13**.

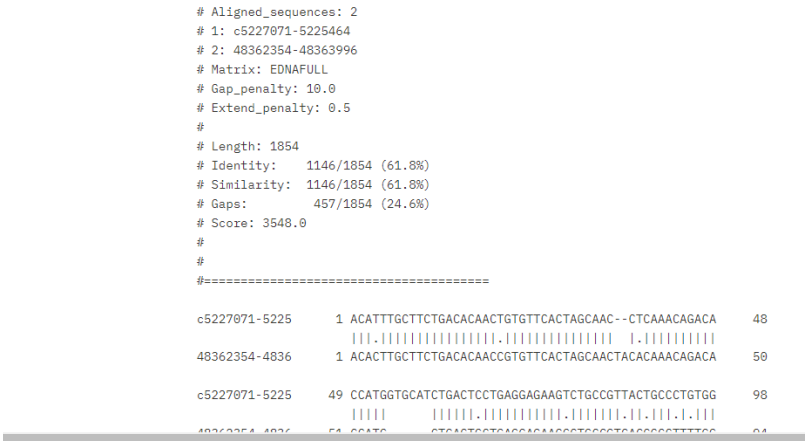


Fig 13: Result of sequence alignment from EMBOSS needle between *homo sapien* HBB gene and its ortholog in cow.

Length: 1854 bases

Identical Matches: 1146/1619 (61.8%)

Similarity: 1146/1619 (61.8%)

Gaps: 457/1854 (24.6%)

Alignment Score: 3548.0

From the result we can infer that with a 61.8% identity and similarity that approximately two-thirds of the bases are identical. This suggests only a moderate level of conservation, which suggests that comparison between genes between distantly related species. The gaps, 24.6%, suggest significant insertions or deletions between the sequences. This could be due to evolutionary divergence, or alternative splicing events. The alignment score, while not as high as the alignment between humanHBB and chimpanzees, still suggests a meaningful level of similarity.

3. Task 3: Multiple Sequence Alignment (MSA)

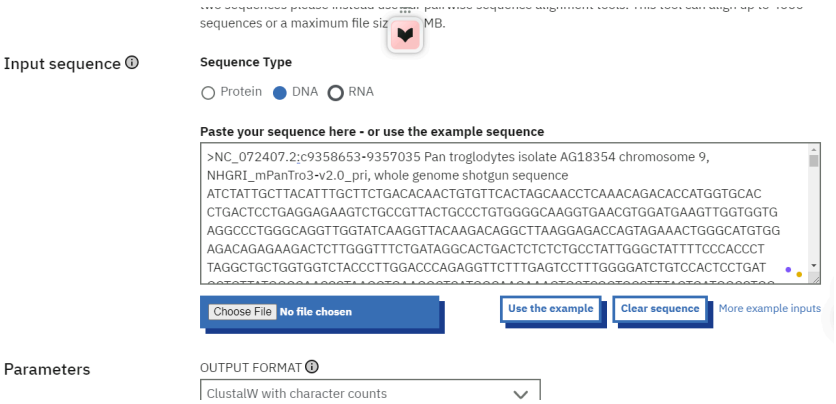


Fig14: MSA on Clustal Omega

The FASTA sequences from six organisms were pasted in the query box **Fig14**.

A B

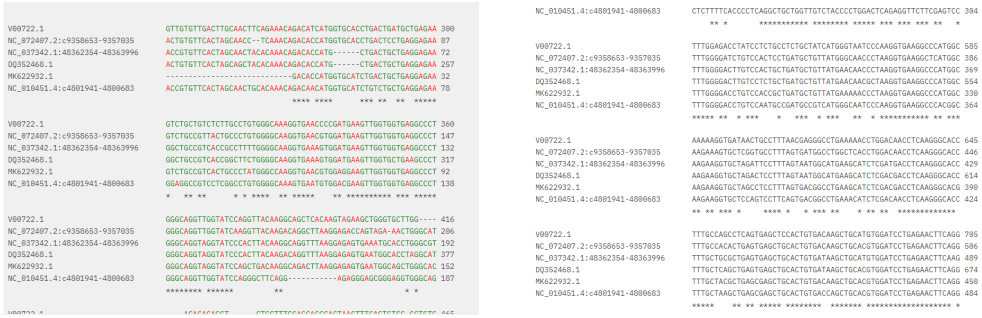


Fig 15. (A) shows the aligned sequences with colour. (B) shows the aligned sequences without colour.

Along the sequences from the six organisms, regions highlighted in yellow **Fig16**, are highly conserved along the sequences. Asides going through the aligned sequences, what was used to identify the conserved regions was the asterisks (*) indicating the conserved nucleotides across all the sequences. A high density of the asterisks means that region is highly conserved, suggesting it may be functionally important, hence no divergence in those regions.

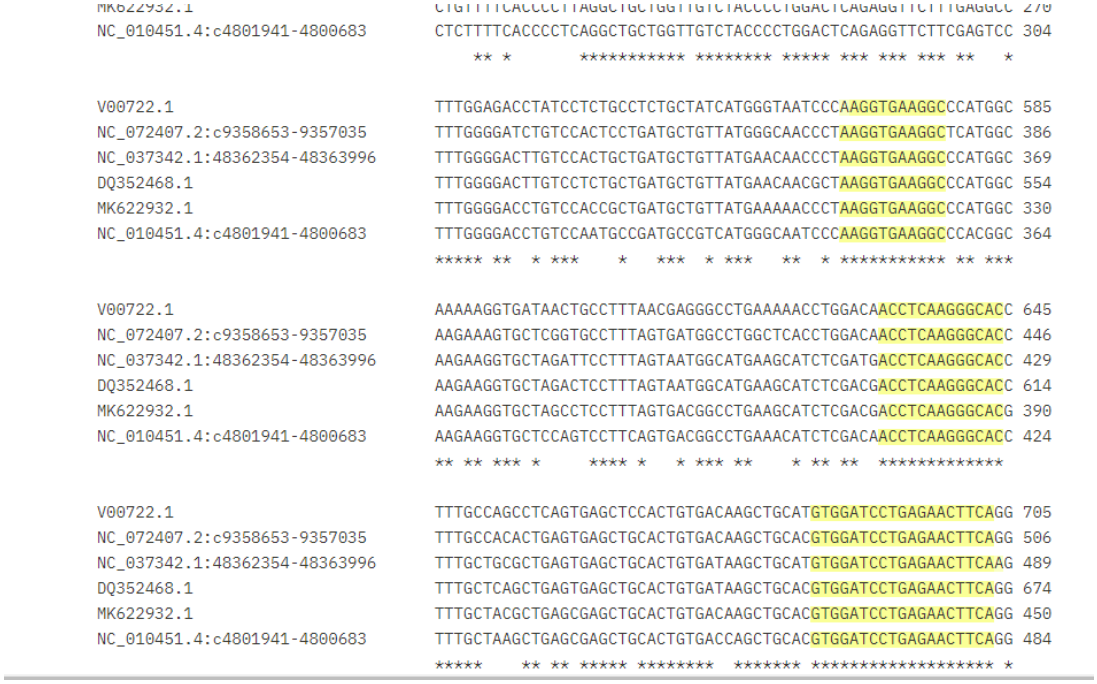


Fig16: A figure of the conserved sequences highlighted in yellow.

4. Task 4: Sequence Logo Generation

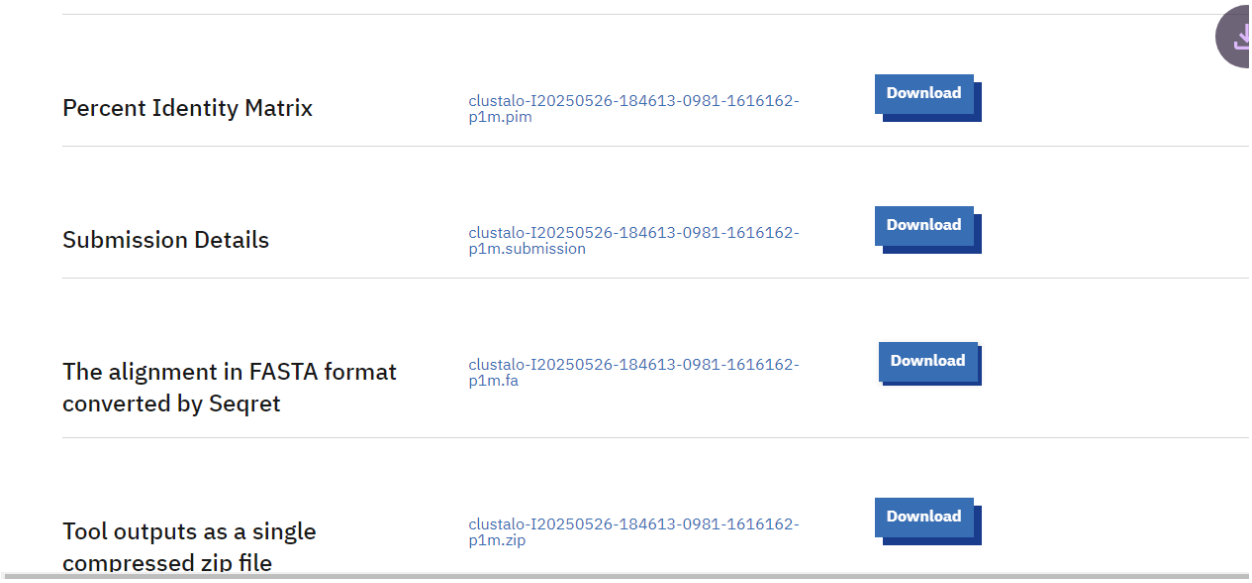


Fig17: Figure showing the selection for the download of the MSA file from Clustal Omega. The "alignment in FASTA format converted by Seqret" was downloaded.

After the MSA file was downloaded from Skylign (https://skylign.org/) was used to generate sequence logos to visualize sequence motifs **Fig18**. The logos are used to graphically represent the conserved regions of the aligned nucleotides, where the conserved positions are represented with taller letters, and the less conserved regions are shorter **Fig19**.

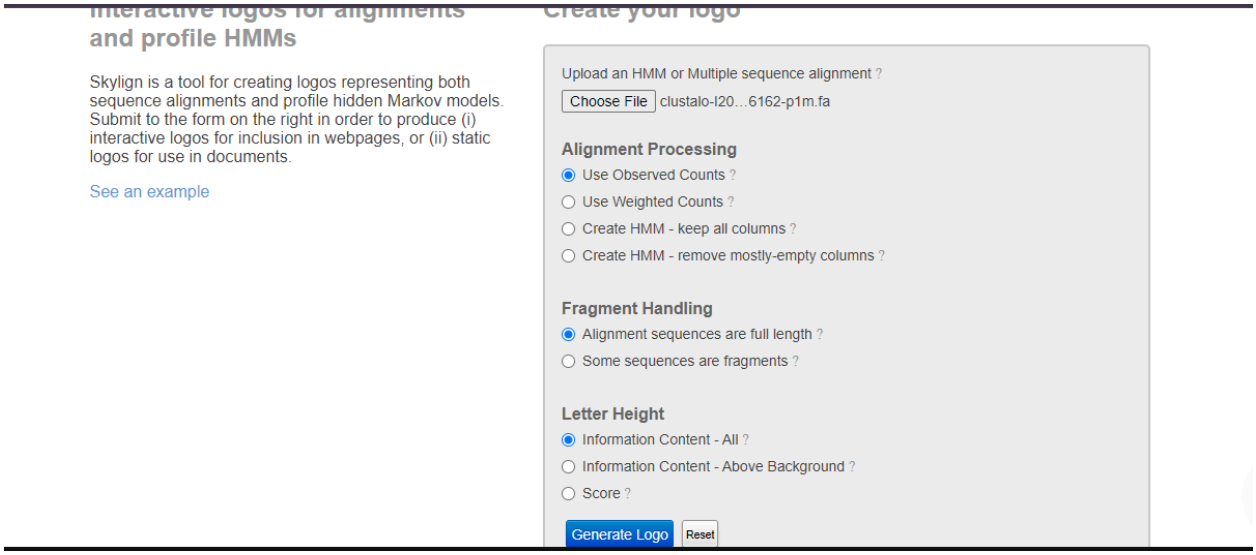


Fig18: The file was uploaded and the other sections were left as default. Then the "Generate Logo" button was selected.

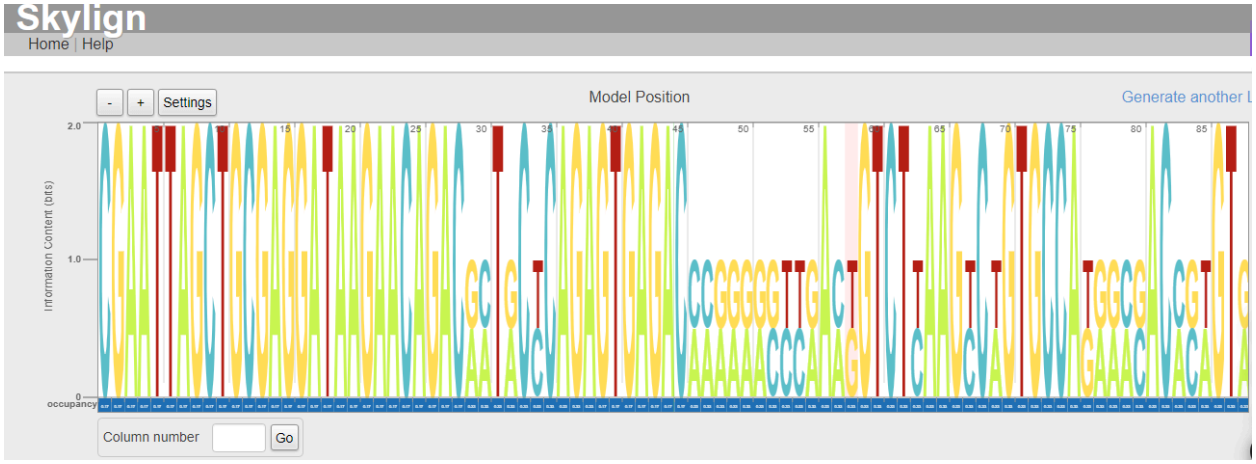


Fig19: The result of the Sequence Logo generated from Skylign

Observation

The sequence logo shows the various conserved regions across the nucleotide sequences. Some positions displaying tall, single letter stacks show strong conservation. This can be observed around positions 5 to 25, 35 to 45, 60, 65, 70). Other positions show a mixture of shorter letters, indicating more variability among the aligned sequences.

From the results, position 5, position 10-25 and 35-45 (with more conserved C and G sequences). This indicates that most of the conserved sequences share the same nucleotide, C and G being more dominant compared to A and T.

The conserved regions often represent functionally or structurally important regions. They may encode important protein-coding regions, or regulatory motifs like promoters, enhancers, or splice sites, where variation may alter the expression of the genes. The conserved regions are likely important for the maintenance of the structure or function of the β-globin protein gotten from the genes.

5. Task 5 Phylogenetic Tree Construction

The MEGA X (molecular Evolutionary Genetics Analysis) software was downloaded from https://www.megasoftware.net/downloads/dload_win_gui. And it was used to generate a phylogenetic tree of the six selected organisms to view their evolutionary relationship.

The steps taken are as follows:

1. On the menu bar of MEGA "File" was selected, and the "Open a File/Session" was selected **Fig20**.
2. The .fas file containing the multiple sequence alignment, downloaded from the MSA from Clustal Omega, was selected
3. The pop option asking preference for opening file, "Analyze" button was selected **Fig21**.
4. The preferred sequence was selected (Nucleotide) **Fig22.**
5. Below the menu bar, "Phylogeny" was selected, and the first option "Construct/Test Maximum Likelihood Tree" was selected **Fig23.**
6. Next for the analysis preference, "Test of Phylogeny" was selected as Standard Bootstrap (slow) and "Model/Method" was the Tamura-Nei model. The other methods were left as default **Fig24.**

   Note: The "Standard Bootstrap (slow)" method was selected to ensure the reliability of the tree, as it performs the full number of replicates (500), giving more accurate and consistent support values for each branch. Although it is slower, it is more trustworthy, especially for meaningful evolutionary comparisons. The Tamura-Nei model was chosen because it is a realistic and widely accepted model for DNA evolution. It accounts for differences in mutation rates (transitions vs. transversions) and unequal base frequencies, making it more accurate for analyzing genes, which may have base composition bias and evolutionary constraints. These chosen parameters help produce a tree that is both statistically sound and biologically accurate.

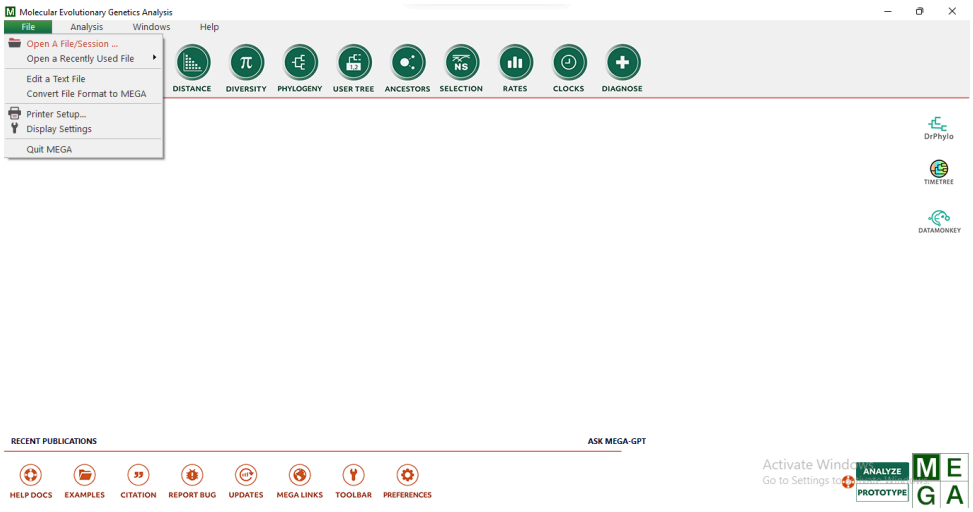


Fig20: Select File —> Open a file/ Session —> Select .fa file of the Multiple Sequence alignment.

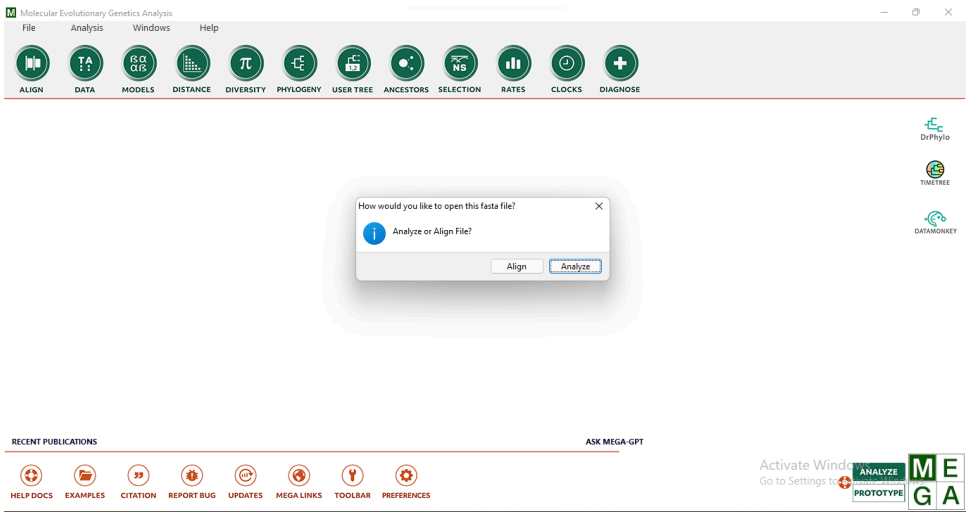


Fig21: Select “Analyze” to analyze MSA file

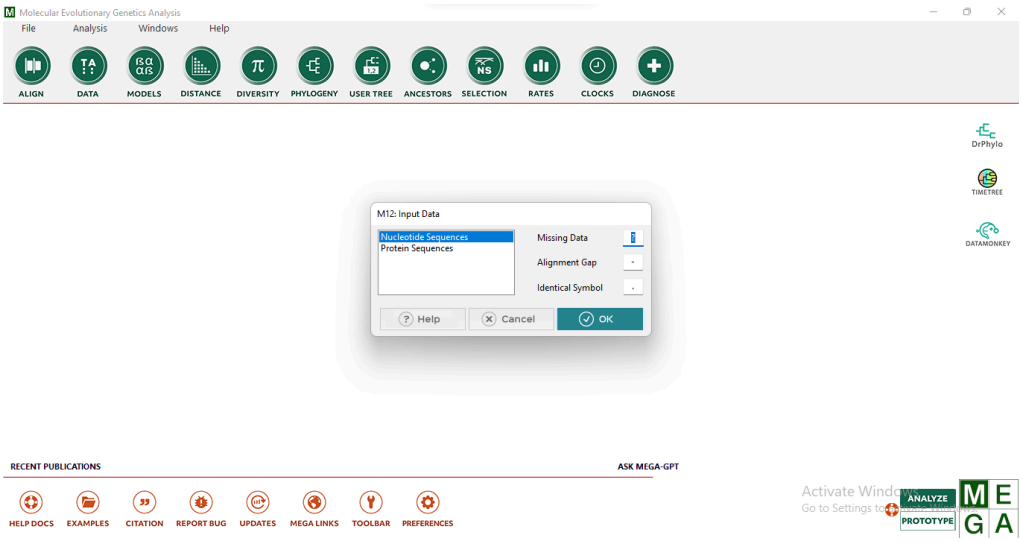


Fig22: Select preferred sequence, Nucleotide sequences.

Fig23: Select phylogeny to generate a phylogenetic tree.

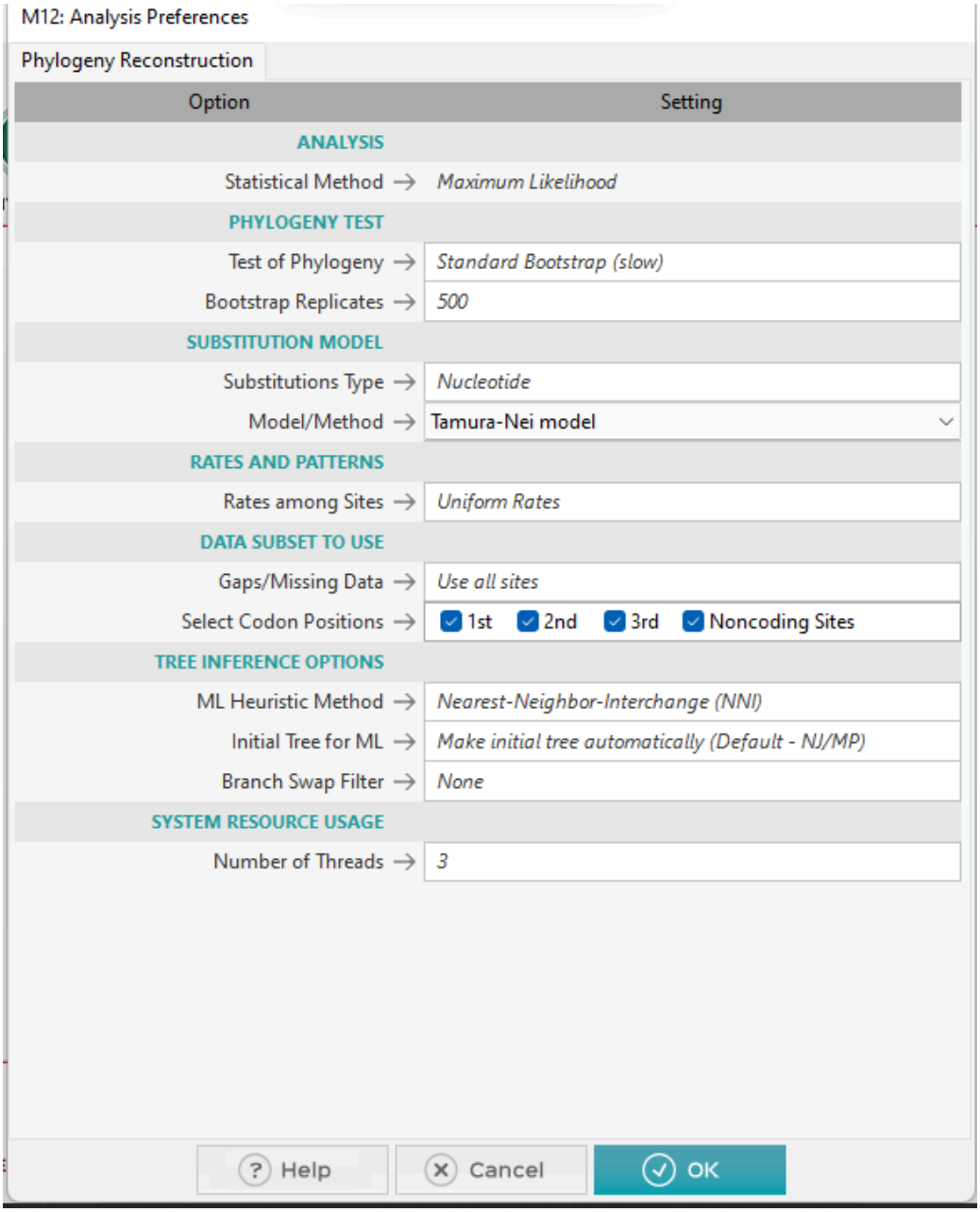


Fig24: Select the analysis preference.

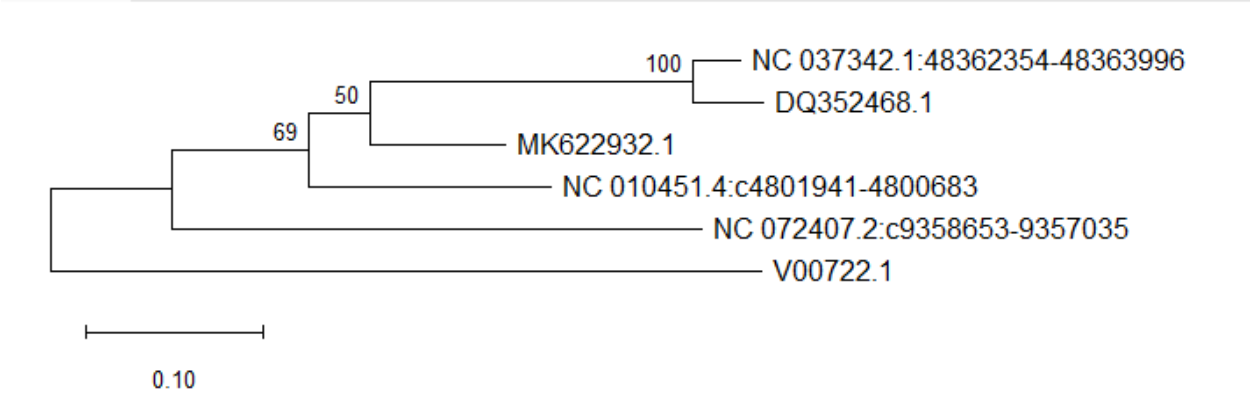


Fig25: Phylogenetic tree. NC 037342.1 - *Bos taurus* (COW); DQ352468.1 - *Movies aries* (SHEEP); MK622932.1 - *Balaenoptera borealis* (WHALE); V00722.1 - *Mus musculus* (MOUSE); NC 072407.2 - *Pan troglodytes* (CHIMPANZEE); NC 010451.4 - *Sus scrofa* (PIG)

NC_037342.1:48362354-48363996 and DQ352468.1 sequences are very closely related, with strong bootstrap support (100), indicating high confidence in their common ancestry. The clade grouping of MK622932.1 with the strongly supported pair, NC_037342.1:48362354-48363996 and DQ352468.1, has low-to-moderate bootstrap support (50), implying the uncertainty about the grouping relationship. More data or a different method may clarify this.
NC_010451.4:c4801941-4800683 and NC_072407.2:c9358653-9357035 form a moderately supported clade, indicating a probable evolutionary relationship with acceptable but not strong confidence. V00722.1 branches off earliest (i.e., basal position), suggesting it is the most divergent sequence in the tree. It likely shares the most ancestral characteristics or differs the most genetically from the others.
Lastly, the 0.10 scale bar represents 10 nucleotide substitutions per 100 nucleotides, or 10% divergence. It provides a reference to estimate how genetically similar or divergent the sequences are based on the branch lengths.