BioinformHer Mini Project – Module 2
Tracking the Evolution of the Hemoglobin Beta (HBB) Gene Across Species

Prepared by : Michelle P Nkomo

Date : 15/06/2025

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

Project Objective:
To use the skills learned in Module 2 to investigate the evolutionary conservation of the HBB gene across six species. This includes sequence retrieval, alignment, logo generation, and phylogenetic tree construction.

Project Tasks

1: Sequence Retrieval & BLAST Search

Instructions:
•   Retrieve the human HBB gene (nucleotide or protein) from NCBI.

•   Use BLAST to identify HBB sequences from at least 5 other species, such as chimpanzee, cow, mouse, chicken, and zebrafish.

•   Download the FASTA format of these sequences.

•   Create a simple table that shows: Species name Accession number % identity with human HBB

Output:

Table 1 illustrating the percentage identities of different species with the human HBB

SPECIES	ACCESSION NUMBER	% IDENTITY WITH *HOMO SAPIENS* (HUMAN) (MK476504.1) **HBB**
Pan troglodytes(chimpanzee)	FJ788217.1	99.50
Bos taurus (cow)	X00376.1	81.45
Mus musculus(mouse)	EF605501.1	77.39
Gallus gallus(chicken)	V00409.1	76.96
Danio rerio(zebrafish)	U50382.1	69.74

Interpretation:
From the above table it can be observed that the chimpanzee HBB is the most closely related to the human HBB, while the zebrafish HBB is the most distantly related of the five species being compared.

2: Pairwise Sequence Alignment

Instructions:
•   Choose two species from your BLAST results: - One closely related to humans (e.g., chimpanzee) - One distantly related (e.g., zebrafish)

•   Perform pairwise alignments of: 1. Human HBB vs Closely Related Species 2. Human HBB vs Distantly Related Species

•   Use a tool like EMBOSS Needle or NCBI Pairwise Alignment. Report for each comparison: % Identity % Similarity Number of gaps A short interpretation (which one is more conserved and why)

Output :

- **1.** Pairwise alignment link for human HBB vs Chimpanzee HBB,

Emboss needle was also used for pairwise alignment. The results are below.

```
#=======================================
#
# Aligned_sequences: 2
# 1: MK476504.1
# 2: FJ788217.1
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 2234
# Identity:     599/2234 (26.8%)
# Similarity:   599/2234 (26.8%)
# Gaps:         1632/2234 (73.1%)
# Score: 2983.0
#
#
#=======================================

MK476504.1         1 --------------------------------------------------      0

FJ788217.1         1 CCAAATGAGGAGAAGATATGCTTAGAACTGAGGTAGAGTTTTCATCCATT     50

MK476504.1         1 --------------------------------------------------      0

FJ788217.1        51 CTGTCCTGTAAGTATTTTGCATATTCTGGAGACACAGGAAGAGATCCATC    100

MK476504.1         1 --------------------------------------------------      0

FJ788217.1       101 TACACATCCCAAAGCTGAATTATGGTAGACAAAACTCTTCCAYTTTTAGT    150

MK476504.1         1 --------------------------------------------------      0

FJ788217.1       151 GCATCAATTTCTTATTTGTGTAATAAGAAAATTGGGAAAACGATCTTCAA    200
```

Figure 1. Pairwise alignment for human HBB vs Chimpanzee

Identity : 26.8%

Similarity : 26.8%

Gaps : 73.1%

- **2.** Pairwise alignment link for human HBB vs Zebrafish,

```
# Length: 2617
# Identity:    1027/2617 (39.2%)
# Similarity:  1027/2617 (39.2%)
# Gaps:        1299/2617 (49.6%)
# Score: 1703.5
#
#
#=====================================

MK476504.1         1 -----------------------CCTAGGGTTGGCCAATCTACTCCCAG        26
                                            |...||||| .||.||.||||
U50382.1           1 CTGCAGAACAAGAGATCAGTTAAGCTGTGGGTT-CCCCATATACT-----       44

MK476504.1        27 GAGCAGGGAGGGCAGGAGCCAGGGCTG----GGCATAAA--------AGT        64
                             |....|.||| ||.|||||    .|.||.||        |||
U50382.1          45 --------AATTTAAGAG-CACGGCTGTTTAAGTATTAACATCATTCAGT        85

MK476504.1        65 CAGGGCAGAGCCATCTATTGCTTAC-----ATTTGCTTCTG---ACACAA       106
                      .|..||||| |.||| |.||| ||.|  .||||
U50382.1          86 TAATGCAGA-----------CGTACCCTGAAGTTG--TCGGGATCCACA-       121

MK476504.1       107 CTGTG----TTCACTA----GCAACCTCAAACAGA---CACCAT-GGTGC       144
                      ||   |||.| |   |||   |||.||.||   |..||| ||||.
U50382.1         122 ---TGCAGTTTCTC-AGAGTGCA---TCACACTGAGGGCGGCATAGGTGT       164
```
Figure 2. Pairwise alignment for human HBB vs Zebrafish

Identity : 39.2%
Similarity : 39.2%
Gaps : 49.6%

Interpretation:

The gaps - represent sites that have been deleted from the sequence in order to improve the overall alignment between sequences, in the chimpanzee there is 73.1% and in the zebrafish it is 49.6%. The dots represent mismatches a number of them can be viewed in fig 1. The vertical lines represent identity.It seems the pairwise alignment of the human HBB vs Zebrafish has more conserved regions (identity and similarity 39.2%) compared to the alignment to the Chimpanzee. The results obtained from Emboss needle illustrate that the Chimpanzee HBB gene (identity and similarity 26.8%) has less percentage identity and similarity to the human HBB compared to the Zebrafish which has a higher percentage identity. This is contrary to the results obtained in the task 1 during sequence search and BLAST. This could be owed to the shared ancestral functions in this particular gene/ region being compared and it might be crucial for a fundamental biological process, leading to stronger conservation across species, despite the overall genetic distance.

3: Multiple Sequence Alignment (MSA)

Instructions
•   Perform a Multiple Sequence Alignment of all 6 sequences using: Clustal Omega or MUSCLE
•   Save and include a screenshot of your alignment.
•   Highlight any highly conserved regions.

Output:
MSA link:
https://www.ebi.ac.uk/jdispatcher/msa/clustalo/summary?jobId=clustalo-I20250613-064316-0889-71480255-p1m

Figure 3. Illustration of multiple sequence alignment results

Interpretation:

The * in the alignments represent conserved identical sequences and the – represents gaps. The gaps represent sites that have been deleted from the sequence in order to improve the overall alignment between sequences. From the results one can observe that there are a number of conserved regions. It signifies that specific stretches of DNA sequences have remained largely unchanged across different species. These conserved regions are often crucial for the proper functioning or structure of the molecule and are therefore maintained by natural selection.

4: Sequence Logo Generation

Instructions
- Upload your MSA file to Skylign.
- Generate a sequence logo to visualize conserved amino acids.
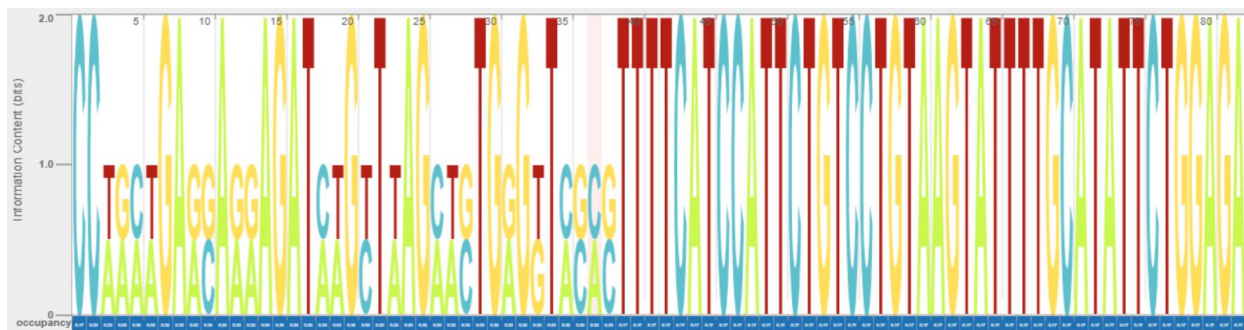
Output:

Figure 4. Screenshot of sequence logo.

Interpretation:

In fig 4 we can view tall letters which indicate highly conserved positions while, shorter stacks represent variable regions in the alignment. In the above image there are a lot of highly conserved regions. These code for essential proteins and regulatory elements. Natural selection preserves these regions due to their importance for survival and function.

5: Phylogenetic Tree Construction

Instructions:

Use your MSA to generate a phylogenetic tree using MEGA X.
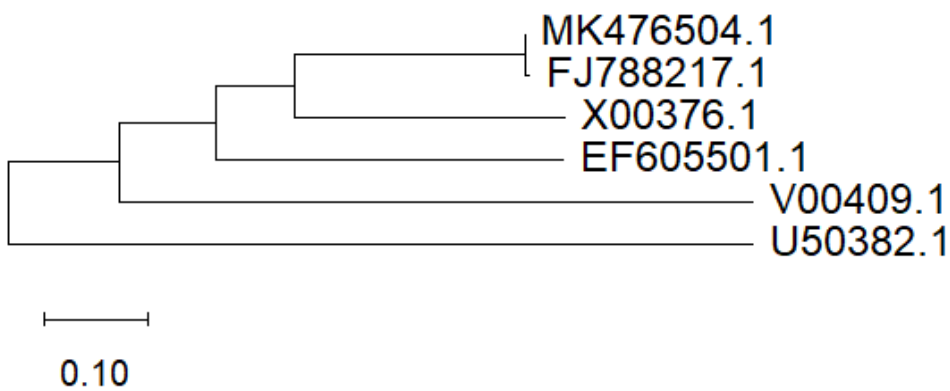
Output:

Figure 5. Screenshot of phylogenetic tree.

Interpretation:

The phylogenetic tree reveals a clear evolutionary relationship between humans and the compared species, with chimpanzee (FJ788217.1) being the closest relative due to their recent common ancestor. The mammalian clade, comprising cow (X00376.1) and mouse (EF605501.1), diverges slightly further from humans, while the avian (chicken, (V00409.1)) and fish (zebrafish, (U50382.1)) lineages diverge even earlier. The high conservation between humans and chimpanzee reflects their close evolutionary relationship, and the decreasing percentage identity values correspond to increasing evolutionary distance. This tree is consistent with expected evolutionary relationships, confirming the phylogenetic positions of these species relative to humans.