TRAINING REPORT

Tracking the Evolution of the Hemoglobin Beta (HBB) Gene across Species

A Comparative Study Using Sequence Alignment, Logo Visualization, and Phylogenetic Analysis

Aprajita Gupta

21-May-25

BioinformHer Mini Project – Module 2 Capstone Project

1. Introduction

The ***hemoglobin beta (HBB)*** gene plays a crucial role in oxygen transport across vertebrates. It encodes the beta subunit of hemoglobin—the iron-containing protein in red blood cells responsible for delivering oxygen to tissues. Hemoglobin functions as a heterotetramer, composed of two alpha and two beta chains, each housing a heme group that binds oxygen molecules (1).

The HBB gene is medically significant, as mutations in it lead to blood disorders such as sickle cell anemia and β-thalassemia (2, 3). Due to its essential function and evolutionary significance, the HBB gene is highly conserved across many species. This makes the HBB gene an ideal candidate for studying molecular evolution and genetic divergence. In this project, the evolutionary conservation of the HBB gene was analyzed across six representative species: *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Bos taurus* (cow), *Mus musculus* (mouse), *Gallus gallus* (chicken), and *Danio rerio* (zebrafish). These organisms span both mammalian and non-mammalian lineages, offering a wide evolutionary perspective.

Using a set of publicly available bioinformatics tools, including BLASTp, EMBOSS Needle, Clustal Omega, Skylign, and MEGA X—this study involved retrieving protein sequences from the NCBI database, conducting pairwise and multiple sequence alignments, generating sequence logos to visualize conserved residues, and constructing a phylogenetic tree to infer evolutionary relationships. The goal of this project is to apply sequence analysis techniques to investigate the degree of conservation of the HBB gene and explore how evolutionary distance affects sequence similarity. Understanding these patterns not only deepens our knowledge of molecular evolution but also demonstrates the power of computational tools in modern biology (4).

2. Objective

The objective of this mini project is to investigate the evolutionary conservation of the hemoglobin beta (HBB) gene across six vertebrate species: *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Bos taurus* (cow), *Mus musculus* (mouse), *Gallus gallus* (chicken), and *Danio rerio* (zebrafish), using fundamental bioinformatics tools and techniques.

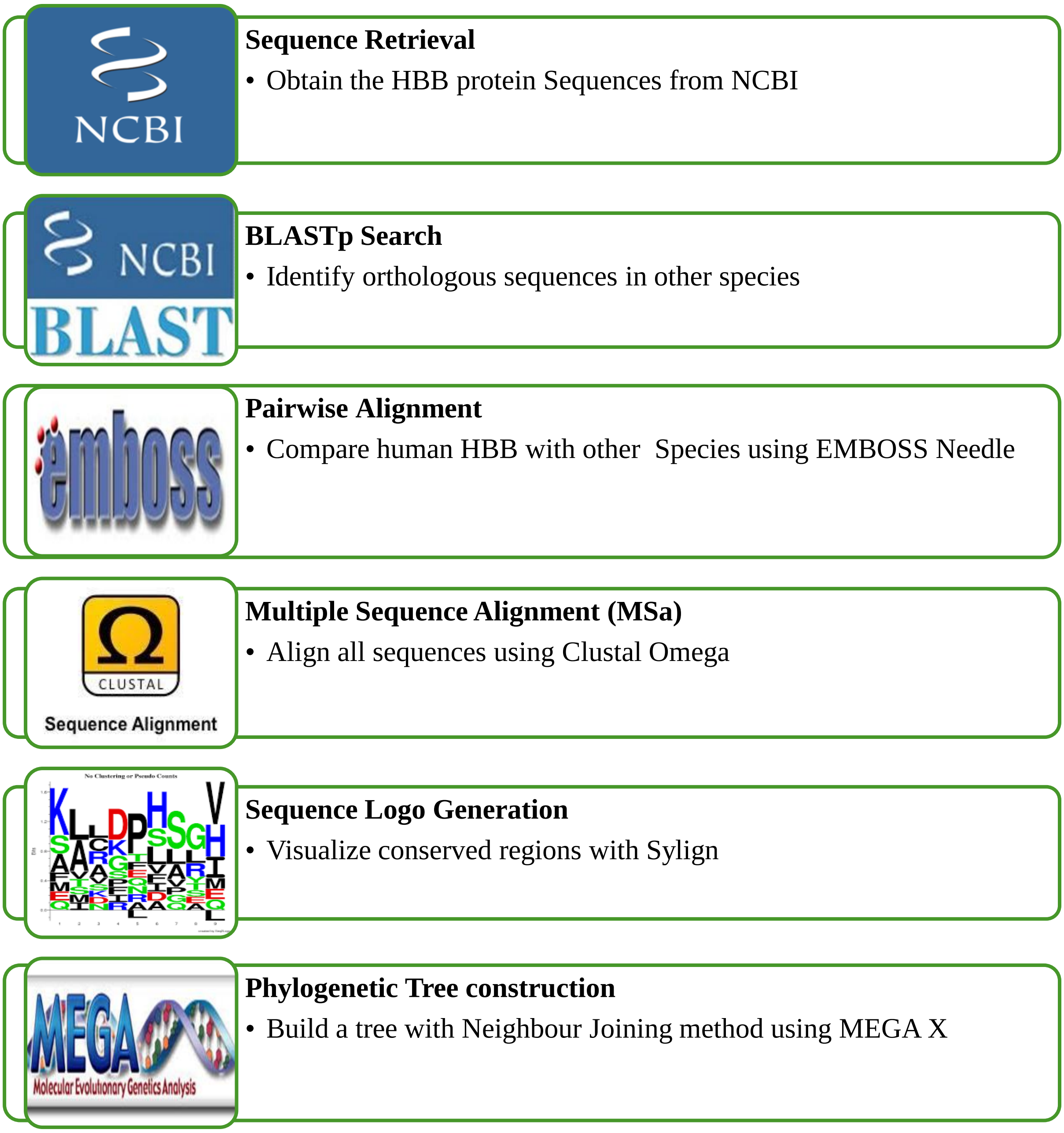
This is achieved through the following:

- Retrieval of HBB protein sequences from the NCBI database
- Pairwise sequence alignment between human and other selected species
- Multiple sequence alignment to detect conserved regions
- Sequence logo generation to visualize residue conservation
- Phylogenetic tree construction to assess evolutionary relationships

The overall aim is to evaluate how conserved the HBB gene is across species and what this reveals about their evolutionary relationships.

3. Methodology

This study involved a comparative sequence analysis of the hemoglobin subunit beta (HBB) protein sequence across six species: *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Bos taurus* (cow), *Mus musculus* (mouse), *Gallus gallus* (chicken), and *Danio rerio* (zebrafish). The aim was to study the evolutionary conservation of HBB using bioinformatics tools. The analysis was carried out through the steps given in the flowchart:



Sequence Retrieval
- Obtain the HBB protein Sequences from NCBI

BLASTp Search
- Identify orthologous sequences in other species

Pairwise Alignment
- Compare human HBB with other Species using EMBOSS Needle

Multiple Sequence Alignment (MSa)
- Align all sequences using Clustal Omega

Sequence Logo Generation
- Visualize conserved regions with Sylign

Phylogenetic Tree construction
- Build a tree with Neighbour Joining method using MEGA X

Figure 1: Bioinformatics Workflow for Comparative Analysis of HBB Protein Sequence

3.1 Sequence Retrieval and BLAST

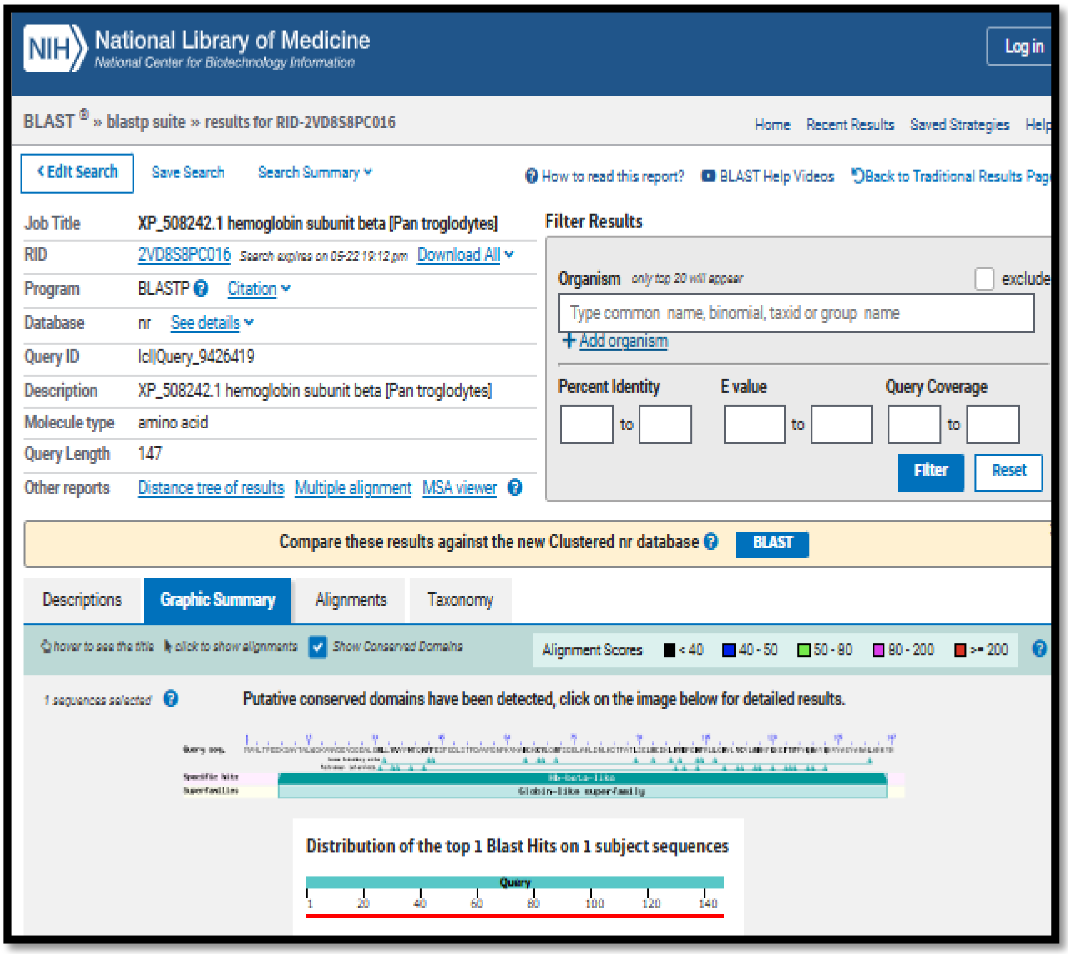
For evolutionary analysis of the Hemoglobin Beta (HBB) gene, the reference human HBB protein sequence was retrieved from the NCBI protein database in FASTA format (NCBI Accession: NP_000509.1) (5). This sequence was used as a query in a BLASTp (Basic Local Alignment Search Tool for proteins) search against the non-redundant (nr) protein database (6). The protein sequences were also retrieved for other five species: *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Bos taurus* (cow), *Mus musculus* (mouse), *Gallus gallus* (chicken), and *Danio rerio* (zebrafish), which are evolutionary diverse from each other . These species were chosen to represent different vertebrate lineages (mammals, birds, fish), providing a broad evolutionary comparison.

The BLAST search was performed using the default parameters with the goal to identify the homologous HBB protein sequences with other species. From the BLAST results, top hits corresponding to the HBB gene in each species were selected based on E-values < 1e-5 and high sequence similarity. The corresponding sequences were downloaded in FASTA format for downstream analyses. The table 1 summarizes the retrieved sequences along with their species name, accession numbers, and percentage identity with the human HBB protein:

Species	**Accession No.**	**% age Identity with Human HBB**
Human (Homo sapiens)	**NP_000509.1**	**100%**
Chimpanzee (Pan troglodytes)	**XP_508242.1**	**100%**
Mouse (Mus musculus)	**NP_001121158.1**	**57.14%**
Cow (Bos Taurus)	**NP_776342.1**	**84.72%**
Chicken (Gallus gallus)	**P02112.2**	**69.39%**
Zebrafish (Danio rerio)	**NP_571095.1**	**51.35%**

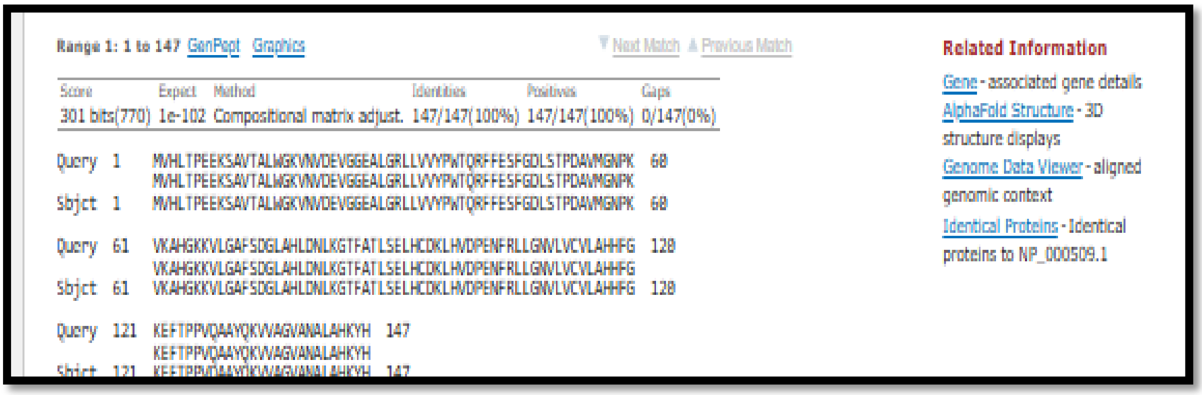
The results show that the HBB gene is highly conserved among mammals, with decreasing similarity in more distantly related species such as birds and fish. The high identity score with chimpanzee reflects the close evolutionary relationship with humans, while the lower similarity in zebrafish suggests significant divergence due to evolutionary distance.

Figure 1 and 3 shows the graphic summary of the BLASTp search. The near-complete alignment and high-scoring segment pair (HSP) indicate a very high degree of similarity between human HBB and chimpanzee HBB sequence, consistent with the 100% identity observed in the table. The red and pink bars in graphic summary results represent the degree of evolutionary closeness between the sequences. Figure 2 and 4 displays the pairwise alignment results. While a significant portion of the sequence aligns, the identity is much lower (~51.35%), and several mismatches and gaps are visible. This reflects the greater evolutionary distance between humans and zebrafish though the alignment still suggests conserved functional domains.



Job Title	XP_508242.1 hemoglobin subunit beta [Pan troglodytes]
RID	2VD8S8PC016 Search expires on 05-22 19:12 pm Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_9426419
Description	XP_508242.1 hemoglobin subunit beta [Pan troglodytes]
Molecule type	amino acid
Query Length	147
Other reports	Distance tree of results Multiple alignment MSA viewer

Figure 2: BLASTp graphic summary of human HBB aligned with chimpanzee HBB sequence



Range 1: 1 to 147 GenPept Graphics ▼ Next Match ▲ Previous Match

```
Score          Expect  Method                        Identities      Positives       Gaps
301 bits(770)  1e-102  Compositional matrix adjust.  147/147(100%)   147/147(100%)   0/147(0%)

Query  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
            MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
Sbjct  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60

Query  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
            VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
Sbjct  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120

Query  121  KEFTPPVQAAYQKVVAGVANALAHKYH  147
            KEFTPPVQAAYQKVVAGVANALAHKYH
Sbjct  121  KEFTPPVQAAYQKVVAGVANALAHKYH  147
```

Related Information

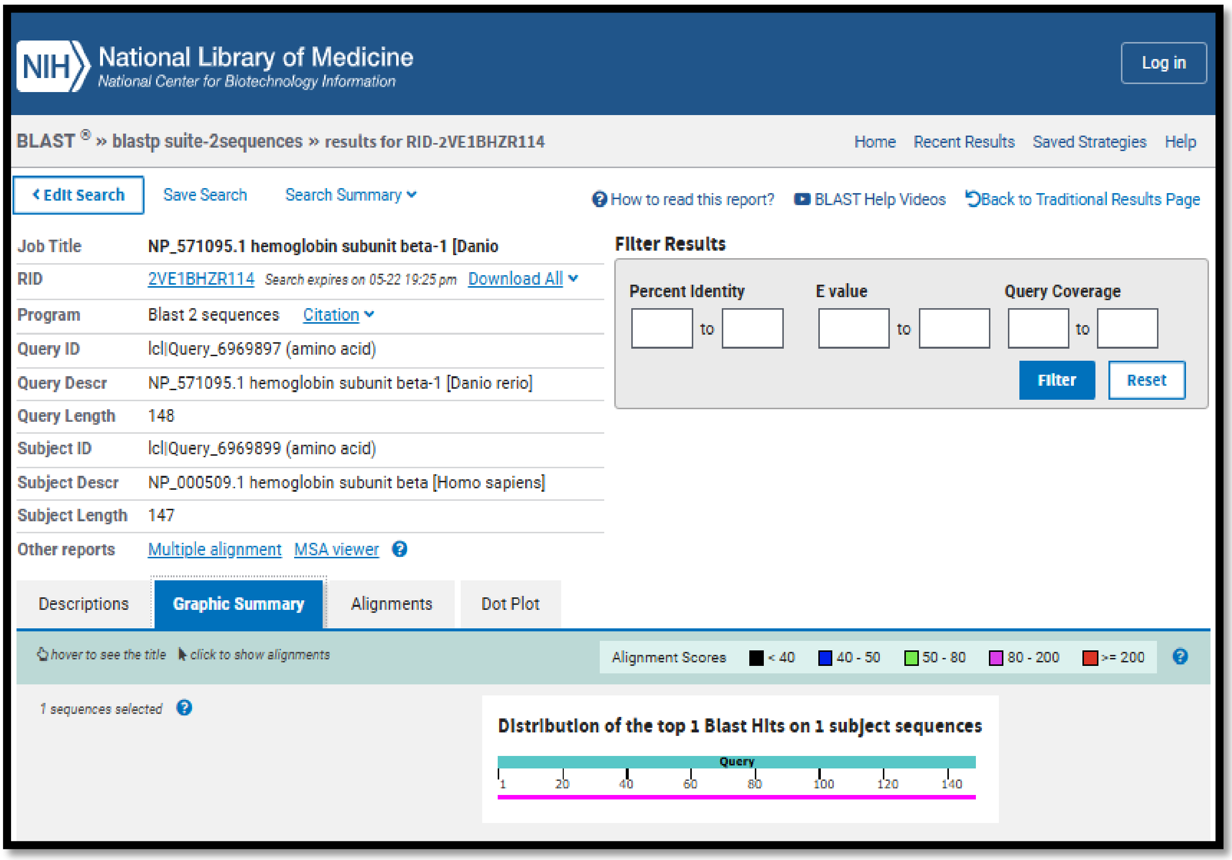
Gene - associated gene details

AlphaFold Structure - 3D structure displays

Genome Data Viewer - aligned genomic context

Identical Proteins - Identical proteins to NP_000509.1

Figure 3: Pairwise BLASTp alignment between human HBB and chimpanzee HBB sequence



NIH National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST ® » blastp suite-2sequences » results for RID-2VE1BHZR114

Home Recent Results Saved Strategies Help

‹ Edit Search Save Search Search Summary ˅ ❓How to read this report? BLAST Help Videos Back to Traditional Results Page

Job Title	NP_571095.1 hemoglobin subunit beta-1 [Danio
RID	2VE1BHZR114 Search expires on 05-22 19:25 pm Download All ˅
Program	Blast 2 sequences Citation ˅
Query ID	lcl|Query_6969897 (amino acid)
Query Descr	NP_571095.1 hemoglobin subunit beta-1 [Danio rerio]
Query Length	148
Subject ID	lcl|Query_6969899 (amino acid)
Subject Descr	NP_000509.1 hemoglobin subunit beta [Homo sapiens]
Subject Length	147
Other reports	Multiple alignment MSA viewer ❓

Filter Results

Percent Identity □ to □ E value □ to □ Query Coverage □ to □

Filter Reset

Descriptions | **Graphic Summary** | Alignments | Dot Plot

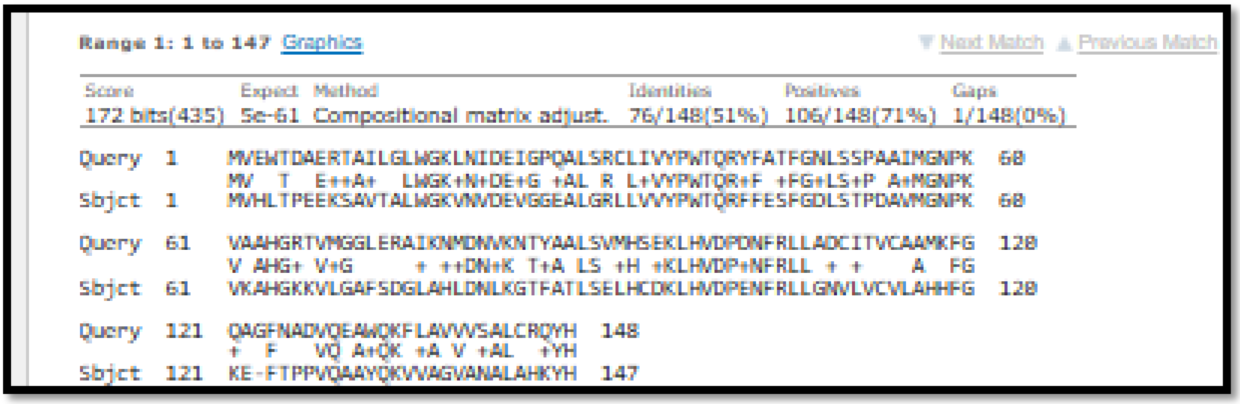
hover to see the title click to show alignments

Alignment Scores ■ < 40 🟦 40 - 50 🟩 50 - 80 🟪 80 - 200 🟥 >= 200 ❓

1 sequences selected ❓

Distribution of the top 1 Blast Hits on 1 subject sequences

Figure 4: BLASTp graphic summary of human HBB aligned with zebrafish HBB sequence



Range 1: 1 to 147 Graphics ▼ Next Match ▲ Previous Match

```
Score          Expect  Method                        Identities   Positives      Gaps
172 bits(435)  5e-61   Compositional matrix adjust.  76/148(51%)  106/148(71%)   1/148(0%)

Query  1    MVEWTDAERTAILGLWGKLNIDEIGPQALSRCLIVYPWTQRYFATFGNLSSPAAIMGNPK  60
            MV  T  E++A+  LWGK+N+DE+G +AL R L+VYPWTQR+F +FG+LS+P A+MGNPK
Sbjct  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60

Query  61   VAAHGRTVMGGLERAIKNMDNVKNTYAALSVMHSEKLHVDPDNFRLLADCITVCAAMKFG  120
            V AHG+ V+G     + ++DN+K T+A LS +H +KLHVDP+NFRLL + +    A  FG
Sbjct  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120

Query  121  QAGFNADVQEAWQKFLAVVVSALCRQYH  148
            +  F   VQ A+QK +A V +AL  +YH
Sbjct  121  KE-FTPPVQAAYQKVVAGVANALAHKYH  147
```

Figure 5: Pairwise BLASTp alignment between human HBB and zebrafish HBB sequence

3.2 Pairwise Sequence Alignment

To assess the degree of evolutionary conservation in the Hemoglobin Beta (HBB) gene, pairwise sequence alignments were conducted between the human HBB protein sequence and two orthologous sequences from other vertebrate species. The selection of species was based on phylogenetic proximity to humans, as determined from the BLASTp search results. One species, the chimpanzee (*Pan troglodytes*), represents a closely related primate, while the other, zebrafish (*Danio rerio*), represents a distantly related fish species. The alignments were carried out using the EMBOSS Needle tool, which performs global alignments based on the Needleman-Wunsch algorithm (7). This method ensures that the entire length of both sequences is aligned, providing comprehensive measures of identity, similarity, and gap distribution across the full sequence. The purpose of this step was to quantify the extent of sequence conservation and to infer evolutionary divergence between species in relation to humans.

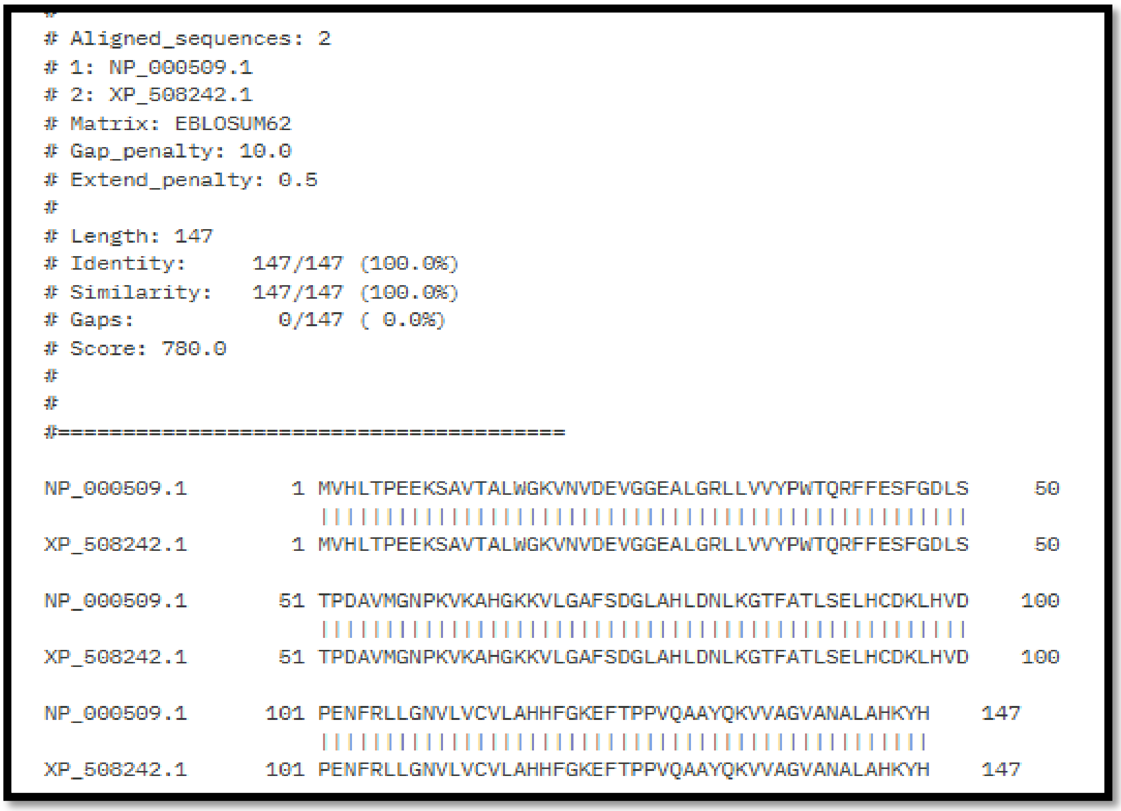
3.2.1 Alignment between Human and Chimpanzee HBB Sequences

The pairwise alignment between human and chimpanzee HBB protein sequences revealed an extremely high level of conservation. The alignment results indicated a percentage identity of 100.00% and a percentage similarity of 100.00%, with no gaps observed throughout the alignment. This high degree of sequence conservation is consistent with the known evolutionary relationship between humans and chimpanzees, who share a common ancestor approximately 6–7 million years ago (8). The lack of mutations or insertions/deletions suggests strong selective pressure to maintain this protein’s function in both species. Given that HBB encodes a critical component of the hemoglobin complex responsible for oxygen transport in blood, functional constraints are expected to limit variation at the protein level (9). These results support the hypothesis that the HBB gene is evolutionarily conserved among primates due to its essential physiological role.

3.2.2 Alignment between Human and Zebrafish HBB Sequences

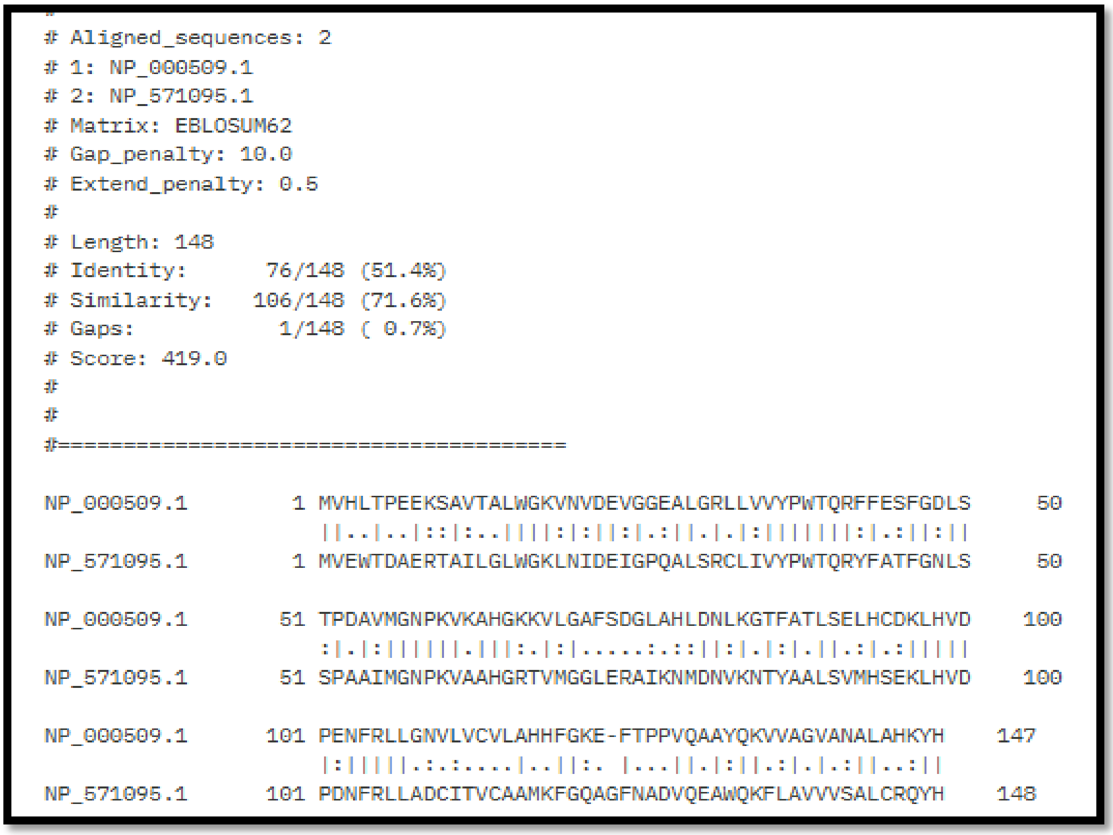
The pairwise alignment between the human and zebrafish HBB sequences exhibited moderate conservation. The results reported a percentage identity of 51.4%, percentage similarity of 71.6%, and 1 gaps distributed across the alignment. (10). Despite the lower identity, several conserved residues were observed, particularly in functionally important regions of the protein, such as those involved in heme binding and structural folding. The presence of gaps may reflect lineage-specific insertions or deletions (indels), as well as variation in sequence length or secondary structure requirements. These results highlight both the divergent evolutionary paths taken by the HBB gene across distant vertebrates and the functional importance of conserved residues maintained over time.

The differences observed between the human–chimpanzee and human–zebrafish alignments clearly reflect how sequence similarity relates to evolutionary distance. The very high identity and similarity in the human–chimpanzee comparison indicate that the HBB gene has remained highly conserved among closely related species. On the other hand, the lower identity and the presence of multiple gaps in the human–zebrafish alignment suggest that more genetic changes have accumulated over a much longer evolutionary timespan. These results align well with the principles of molecular evolution, which propose that genes involved in vital biological functions, like hemoglobin in oxygen transport, tend to remain conserved, especially among species with a recent common ancestor (11).



```
#
# Aligned_sequences: 2
# 1: NP_000509.1
# 2: XP_508242.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 147
# Identity:     147/147 (100.0%)
# Similarity:   147/147 (100.0%)
# Gaps:           0/147 ( 0.0%)
# Score: 780.0
#
#
#=======================================

NP_000509.1        1 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS     50
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
XP_508242.1        1 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS     50

NP_000509.1       51 TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
XP_508242.1       51 TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD    100

NP_000509.1      101 PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
                     |||||||||||||||||||||||||||||||||||||||||||||||
XP_508242.1      101 PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
```

Figure 6: Pairwise alignment between human and chimpanzee HBB protein sequences



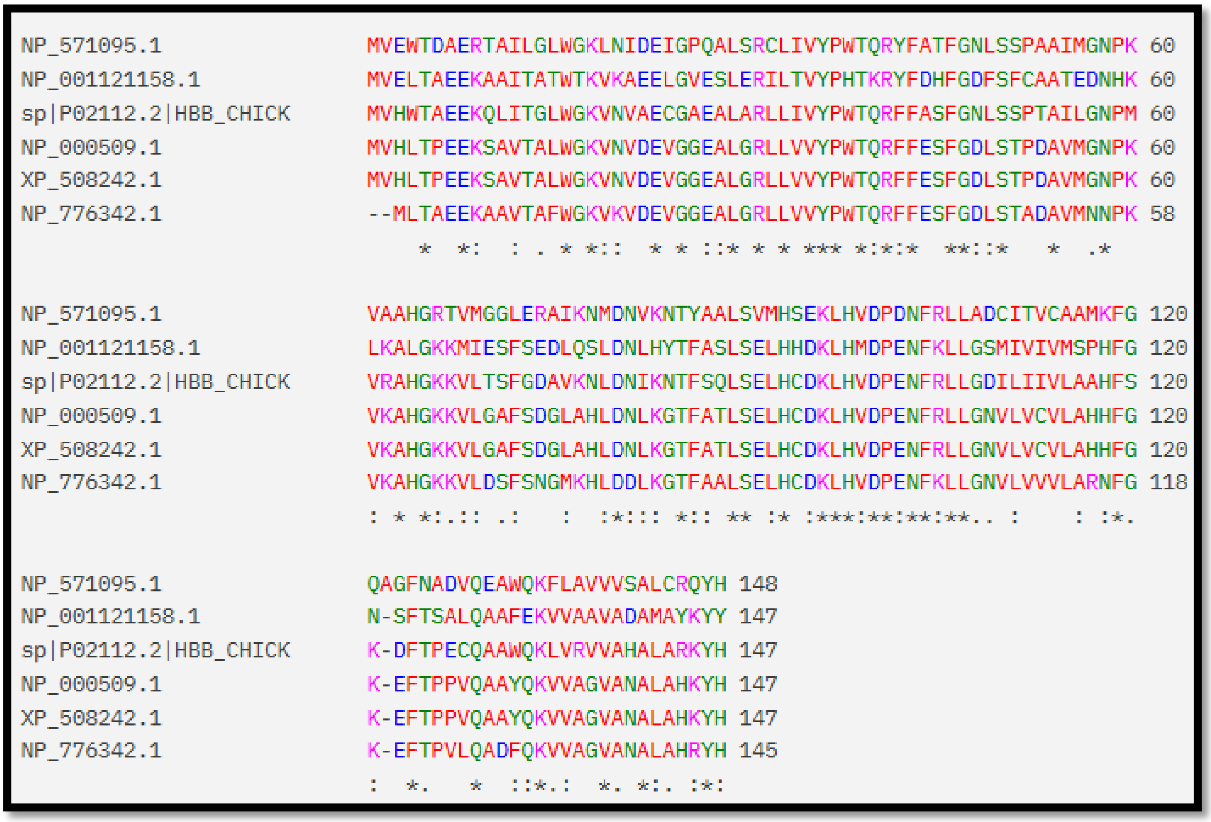
```
# Aligned_sequences: 2
# 1: NP_000509.1
# 2: NP_571095.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 148
# Identity:      76/148 (51.4%)
# Similarity:   106/148 (71.6%)
# Gaps:           1/148 ( 0.7%)
# Score: 419.0
#
#
#=======================================

NP_000509.1        1 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS     50
                     ||..|..|::|:..||||:|:||:|.:||.|.|:||||||||:|.:||:||
NP_571095.1        1 MVEWTDAERTAILGLWGKLNIDEIGPQALSRCLIVYPWTQRYFATFGNLS     50

NP_000509.1       51 TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD    100
                     :|.|:||||||.|||:.|:|.....:.::||:|.|:|.||.:|.:|||||
NP_571095.1       51 SPAAIMGNPKVAAHGRTVMGGLERAIKNMDNVKNTYAALSVMHSEKLHVD    100

NP_000509.1      101 PENFRLLGNVLVCVLAHHFGKE-FTPPVQAAYQKVVAGVANALAHKYH    147
                     |:|||||.:.:....|..||:. |...||.|:||.:|.|.:||..:||
NP_571095.1      101 PDNFRLLADCITVCAAMKFGQAGFNADVQEAWQKFLAVVVSALCRQYH    148
```

Figure 7: Pairwise alignment between human and zebrafish HBB protein sequences

3.3 Multiple Sequence Alignment (MSA)

To further investigate the evolutionary conservation of the Hemoglobin Beta (HBB) gene, a Multiple Sequence Alignment (MSA) was performed using six HBB protein sequences retrieved from BLAST results. The selected sequences included those from Homo sapiens (human), Pan troglodytes (chimpanzee), Mus musculus (mouse), *Gallus gallus* (chicken), Bos taurus (cow), and Danio rerio (zebrafish). This diverse taxonomic representation allowed for the comparison of both closely and distantly related species, offering insight into conserved functional regions and evolutionary divergence.

The alignment was conducted using Clustal Omega, an efficient and scalable multiple sequence alignment tool developed by the European Bioinformatics Institute (EMBL-EBI) [Ref. 1]. Clustal Omega uses a progressive alignment approach based on guide trees and Hidden Markov Models (HMMs), which enables it to align both closely and distantly related sequences with high accuracy (12). It is particularly well-suited for protein alignments and is widely adopted in comparative genomics and molecular evolution studies. The sequences were submitted in FASTA format, and default alignment parameters were used.



```
NP_571095.1                MVEWTDAERTAILGLWGKLNIDEIGPQALSRCLIVYPWTQRYFATFGNLSSPAAIMGNPK 60
NP_001121158.1             MVELTAEEKAAITATWTKVKAEELGVESLERILTVYPHTKRYFDHFGDFSFCAATEDNHK 60
sp|P02112.2|HBB_CHICK      MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM 60
NP_000509.1                MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60
XP_508242.1                MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60
NP_776342.1                --MLTAEEKAAVTAFWGKVKVDEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMNNPK 58
                               *  *:  : . * *::  * * ::* * * *** *:*:*  **::*   *  .*

NP_571095.1                VAAHGRTVMGGLERAIKNMDNVKNTYAALSVMHSEKLHVDPDNFRLLADCITVCAAMKFG 120
NP_001121158.1             LKALGKKMIESFSEDLQSLDNLHYTFASLSELHHDKLHMDPENFKLLGSMIVIVMSPHFG 120
sp|P02112.2|HBB_CHICK      VRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFS 120
NP_000509.1                VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 120
XP_508242.1                VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 120
NP_776342.1                VKAHGKKVLDSFSNGMKHLDDLKGTFAALSELHCDKLHVDPENFKLLGNVLVVVLARNFG 118
                           : * *:.:: .:   :  :*::: *:: ** :* :***:**:**:**.. :    : :*.

NP_571095.1                QAGFNADVQEAWQKFLAVVVSALCRQYH 148
NP_001121158.1             N-SFTSALQAAFEKVVAAVADAMAYKYY 147
sp|P02112.2|HBB_CHICK      K-DFTPECQAAWQKLVRVVAHALARKYH 147
NP_000509.1                K-EFTPPVQAAYQKVVAGVANALAHKYH 147
XP_508242.1                K-EFTPPVQAAYQKVVAGVANALAHKYH 147
NP_776342.1                K-EFTPVLQADFQKVVAGVANALAHRYH 145
                           :  *.   *  ::*.:  *. *:. :*:
```

Figure 8: Multiple sequence alignment of HBB protein sequences showing conserved and semi-conserved residues among all 6 sequences using Clustal Omega

The alignment output was saved in CLUSTAL format with residue numbering. The Figure 8 shows the multiple sequence alignment of all the 6 sequences and also gives the highly conserved regions between them. Conserved amino acid residues were observed particularly in functionally critical domains, such as those involved in heme binding and structural integrity. This suggests that evolutionary pressures have maintained these residues across vertebrates to preserve hemoglobin's oxygen-carrying function. In Figure 8 symbols indicate conservation levels across the sequences such as (*) indicates fully conserved residues, (:) denotes strongly conserved substitutions and (.) indicates weakly conserved substitutions. Several stretches in the alignment (e.g. 1-20, 40-90 and 120-140) are fully conserved or semi-conserved across all species. These regions are essential for oxygen binding or globin structural stability and are maintained due to purifying selection. The chicken HBB sequence shows multiple substitutions, especially in the C-terminal region and some stretches of the N-terminal, where conserved residues in mammals are replaced with different amino acids.

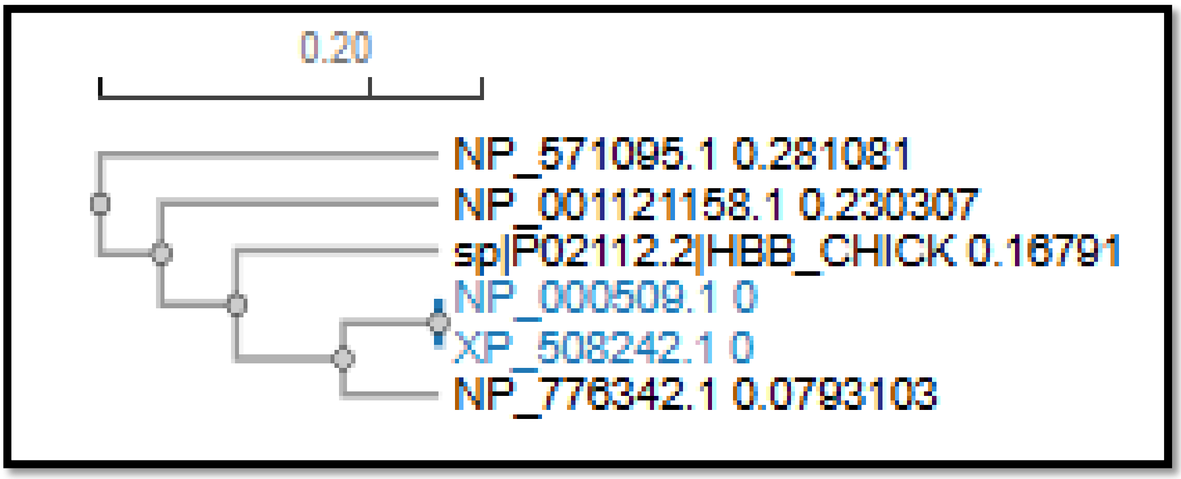


Figure 9: Guide tree (Phylogram) generated by Clustal Omega showing phylogenetic relationships among the selected species

The guided tree output gives a visual representation of the evolutionary distance between the aligned hemoglobin beta (HBB) protein sequences. Human and Chimpanzee are closely clustered with distance values of 0.281 and 0.230, respectively, consistent with their recent divergence and high sequence similarity. Zebrafish appears the most divergent with a distance of 0.079, indicative of its placement as a more distantly related vertebrate.

3.4: Sequence Logo Generation

To visually represent conserved regions in the aligned hemoglobin beta (HBB) sequences, a sequence logo was generated using the Skylign web server (13). This tool creates graphical representations of multiple sequence alignments by displaying the degree of conservation at each position in the protein sequence. In a sequence logo, each column corresponds to an aligned position in the sequence, and the height of each letter (representing an amino acid) is proportional to its conservation at that site. The total height of the stack indicates the level of sequence conservation, while the relative height of individual letters within the stack shows the frequency of specific amino acids at that position. The FASTA format file was given as input with default parameters.

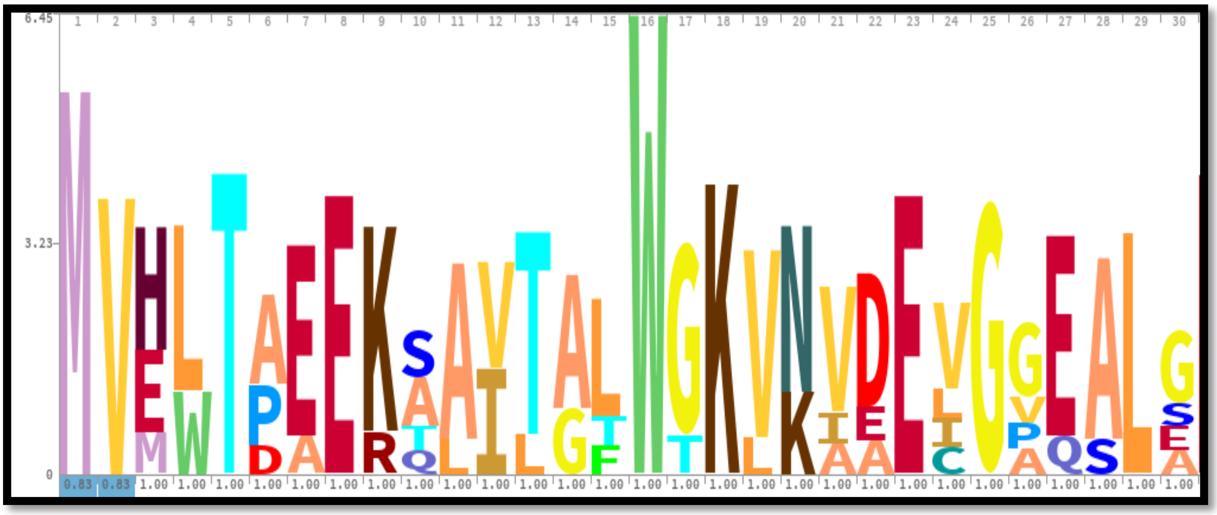
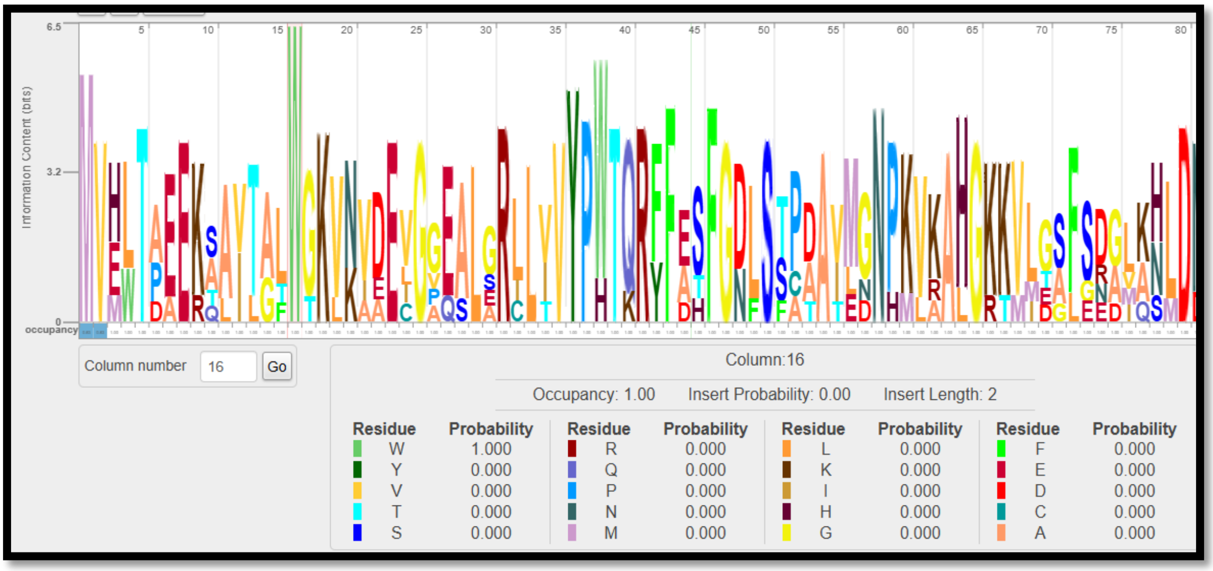


Figure 10: Sequence Logo of Conserved Amino Acid Motifs in HBB Protein (positions 1-30) across 6 species

The sequence logo highlights the evolutionary conservation and variability across the HBB gene in different vertebrates. The observed conservation patterns support the hypothesis that functionally crucial regions of this protein have been maintained over evolutionary time, aligning with principles of molecular evolution and protein structure-function relationships. In the Figure 10, the highly conserved residues are represented by the tall stacks with a single dominant amino acid for e.g. positions 40-90. These conserved residues likely correspond to functionally critical domains, such as those involved in heme binding, structural stability or subunit interactions within the hemoglobin complex (14). Moderate variability is observed in the N-terminal and C-terminal regions, though certain residues remain conserved even in these segments. Highly conserved positions are typically critical for molecular interactions and functionality, and even minor changes at these sites may impair protein activity, which explains their preservation across species. The sequence logo thus provides a powerful and intuitive visual summary of how evolutionary pressures have shaped the HBB protein across diverse vertebrates. The clear conservation of essential residues, even between distant species like humans and zebrafish, reinforces the idea that the HBB protein is subject to strong purifying selection, maintaining its vital role in oxygen transport throughout vertebrate evolution.



Column:16

Occupancy: 1.00 Insert Probability: 0.00 Insert Length: 2

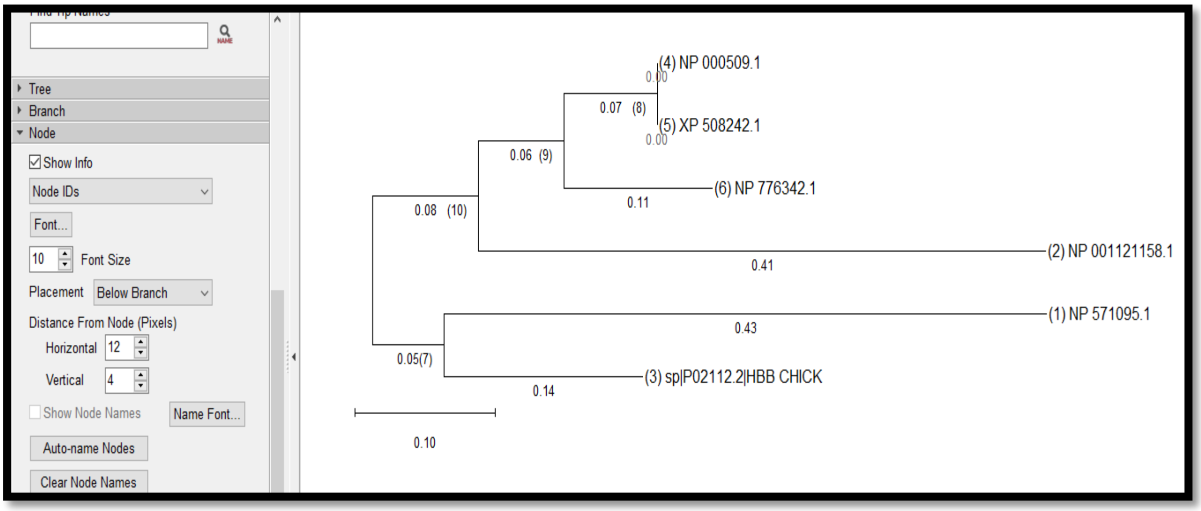
Residue	Probability	Residue	Probability	Residue	Probability	Residue	Probability
🟩 W	1.000	🟥 R	0.000	🟧 L	0.000	🟩 F	0.000
🟩 Y	0.000	🟪 Q	0.000	🟫 K	0.000	🟥 E	0.000
🟨 V	0.000	🟦 P	0.000	🟫 I	0.000	🟥 D	0.000
🟦 T	0.000	■ N	0.000	🟪 H	0.000	■ C	0.000
🟦 S	0.000	🟪 M	0.000	🟨 G	0.000	🟧 A	0.000

Figure 11: Zoomed-In View of Column 16 in the Sequence Logo of HBB Protein Showing Full Conservation of Tryptophan (W)

3.5 Phylogenetic Tree Construction

To infer the evolutionary relationships among the six selected vertebrate species based on their Hemoglobin Beta (HBB) protein sequences, a phylogenetic tree was constructed using the MEGA X software platform. MEGA X (Molecular Evolutionary Genetics Analysis) is a widely used bioinformatics tool for building and visualizing phylogenetic trees (15). It enables researchers to perform evolutionary analysis by providing multiple methods for tree construction, distance estimation, and sequence comparison. In this study, the Neighbor-Joining (NJ) method was employed, a distance-based algorithm that constructs trees by finding pairs of operational taxonomic units (OTUs) that minimize the total branch length at each stage of clustering. The evolutionary distances were calculated using the Poisson correction model, which assumes equal rates of amino acid substitution and accounts for multiple substitutions at the same site.

The aligned protein sequences of *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Bos taurus* (cow), *Mus musculus* (mouse), *Gallus gallus* (chicken), and *Danio rerio* (zebrafish) were imported into MEGA X as a .mas file. Once the NJ method was applied, MEGA X generated a rooted tree with branch lengths representing the number of amino acid substitutions per site, effectively capturing the evolutionary distances between sequences.



☑ Show Info
Node IDs
10 Font Size
Placement Below Branch
Distance From Node (Pixels)
Horizontal 12
Vertical 4
□ Show Node Names

Figure 12: Phylogenetic Tree of HBB Protein Sequences from Six Vertebrate Species Constructed Using the Neighbor-Joining Method in MEGA X

The Figure 12 revealed that *Homo sapiens* (accession NP_000509.1) and *Pan troglodytes* (XP_508242.1) clustered together with a branch length of 0.00, indicating complete identity between the human and chimpanzee HBB sequences. This strong similarity is expected due to their close evolutionary lineage and recent divergence, estimated at approximately 6 to 7 million years ago (8). The cow (*Bos taurus*, NP_776342.1) formed a sister group to this primate clade, reflecting a slightly greater evolutionary distance, but still within the mammalian lineage. The mouse (*Mus musculus*, NP_001121158.1) branched earlier than the cow, suggesting a more distant relationship with humans and chimpanzees. The chicken (*Gallus gallus*, P02112.2) and zebrafish (*Danio rerio*, NP_571095.1) were positioned on more distant branches, zebrafish appeared at the most distant point on the tree, indicating the greatest divergence from human HBB (9).

The topology of the tree accurately reflected the expected evolutionary relationships among the species, validating the use of HBB as a molecular marker for phylogenetic analysis. These also align with earlier findings from the multiple sequence alignment and sequence logo analyses, which revealed highly conserved residues among mammals and progressively lower conservation in non-mammalian species.

4. Conclusion

This project provided an integrated application of core bioinformatics tools to explore the evolutionary conservation of the Hemoglobin Beta (HBB) protein across six vertebrate species. Through systematic analysis involving sequence retrieval, pairwise and multiple sequence alignment, sequence logo generation, and phylogenetic tree construction, we were able to evaluate both sequence similarity and divergence across evolutionary lineages. The results revealed that HBB is a highly conserved protein, particularly among mammals. The 100% sequence identity between the human and chimpanzee HBB proteins underscores their close evolutionary relationship. In contrast, more distantly related species such as chicken and zebrafish exhibited lower similarity, although several functionally important residues remained conserved. These findings are supported by the sequence logo and phylogenetic tree, both of which highlight conserved domains under strong purifying selection due to their essential role in oxygen transport. This project not only deepened my understanding of evolutionary biology but also strengthened my practical skills in computational biology. By working hands-on with tools like BLAST, Clustal Omega, Skylign, and MEGA X, I gained valuable experience in handling real biological data, performing sequence-based analyses, and interpreting evolutionary patterns. It demonstrated how accessible, open-source tools can be effectively used to draw meaningful biological conclusions and emphasized the relevance of bioinformatics in modern molecular research.

References:

1. Cortez-Romero CR, Lyu J, Pillai AS, Laganowsky A, Thornton JW. Symmetry facilitated the evolution of heterospecificity and high-order stoichiometry in vertebrate hemoglobin. Proc Natl Acad Sci U S A. 2025;122(4):e2414756122.
2. Butt H, Mandava M, Jacobsohn D. Advances in Gene Therapy for Sickle Cell Disease: From Preclinical Innovations to Clinical Implementation and Access Challenges. CRISPR J. 2025.
3. Dordevic A, Mrakovcic-Sutic I, Pavlovic S, Ugrin M, Roganovic J. Beta thalassemia syndromes: New insights. World J Clin Cases. 2025;13(10):100223.
4. Kong L, Ye M, Tan J, Lai W, Liao J, Zhang Y, et al. Genome-wide identification and expression profiling of the hemoglobin gene family under hypoxia stress in Sillago sihama. Comp Biochem Physiol Part D Genomics Proteomics. 2025;55:101500.
5. Coordinators NR. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2018;46(D1):D8-D13.
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403-10.
7. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics. 2000;16(6):276-7.
8. Chen FC, Li WH. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet. 2001;68(2):444-56.
9. Hardison RC. Evolution of hemoglobin and its genes. Cold Spring Harb Perspect Med. 2012;2(12):a011627.
10. Kobayashi I, Katakura F, Moritomo T. Isolation and characterization of hematopoietic stem cells in teleost fish. Dev Comp Immunol. 2016;58:86-94.
11. Mao Y, Peng T, Shao F, Zhao Q, Peng Z. Molecular evolution of the hemoglobin gene family across vertebrates. Genetica. 2023;151(3):201-13.
12. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011;7:539.
13. Wheeler TJ, Clements J, Finn RD. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics. 2014;15(7).
14. Yan XT, Chang KL, Huang ZB, Xu YT, Li ZP, Liu WB, et al. A protein structure-dependent fluorescent probe for hemoglobin monitoring and controllable imaging in living cells. Int J Biol Macromol. 2024;283(Pt 3):137868.
15. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution. 2018;35(6):1547-9.