

BioinformHer Mini Project – Module 2 Capstone

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

Objective: 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

- Retrieve the human HBB gene (nucleotide or protein) from NCBI.

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

Protein [Advanced](#) [Help](#)

FASTA

hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1
[GenPept](#) [Identical Proteins](#) [Graphics](#)

>NP_000509.1 hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPHTQRFFESFGDLSTPDVAMGNPKVKAHGKKVLG
AFSDGLAHLNLIKGTFTLSELHCDKLHVDPENFRLLGNVLCVLAHFGKEFTPPVQAAAYQKVAVGVAN
ALAHKYYH

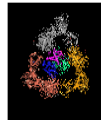
Analyze this sequence

Run BLAST

Identify Conserved Domains

Show in Genome Data Viewer

Protein 3D Structure

 Cryo-EM structure of human CD163 SRCR1-9 in complex with haptoglobin-hemoglobin
PDB: 9FNO
Source: Homo sapiens
Method: Electron Microscopy
Resolution: 5.2 Å

[See all 253 structures...](#)

Articles about the HBB gene

Hb Monza: A novel extensive HBB duplication with preserved α - β subunit interaction [Med. 2025]

HBB as a Novel Biomarker for the Diagnosis and Monitoring of L [Technol Cancer Res Treat. 2024]

Exagamlogene Autotemcel for Severe Sickle Cell Disease. [N Engl J Med. 2024]

[See all...](#)

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

blastn

blastp

blastx

tblastn

tblastx

BLAST

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) Clear

NP_000509.1

Query subrange [?](#)

From

To

Or, upload file

Choose File No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database

☒ Standard databases (nr etc.):
 ☐ ClusteredNR database **RECOMMENDED** [Learn more...](#)

Reference proteins (refseq_protein) [?](#)

Pan troglodytes (taxid:9598) ☐ exclude [Add organism](#)

Bos taurus (taxid:9913) ☐ exclude

Mus musculus (taxid:10090) ☐ exclude

Gallus gallus (taxid:9031) ☐ exclude

Danio rerio (taxid:7955) ☐ exclude

☒ exclude

Organism

Optional

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude

Optional

☐ Models (XM/XP)
 ☐ Non-redundant RefSeq proteins (WP)
 ☐ Uncultured/environmental sample sequences

Program Selection

Algorithm

☒ blastp (protein-protein BLAST)
 ☐ PSI-BLAST (Position-Specific Iterated BLAST)
 ☐ PHI-BLAST (Pattern Hit Initiated BLAST)
 ☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)

BLAST

Search database refseq_protein using Blastp (protein-protein BLAST)

☐ Show results in a new window

[◀ Edit Search](#)
[Save Search](#)
[Search Summary ▾](#)

[🔗 How to read this report?](#)
[📺 BLAST Help Videos](#)
[↶ Back to Traditional Results Page](#)

❗ Your search is limited to records that include: *Pan troglodytes* (taxid:9598), *Bos taurus* (taxid:9913), *Mus musculus* (taxid:10090), *Gallus gallus* (taxid:9031), *Danio rerio* (taxid:7955)

Job Title	ref NP_000509.1		
RID	4X6M7MSM016	Search expires on 06-16 18:03 pm	Download All ▾
Program	BLASTP 🔗 Citation ▾		
Database	refseq_protein	See details ▾	
Query ID	NP_000509.1		
Description	hemoglobin subunit beta [Homo sapiens]		
Molecule type	amino acid		
Query Length	147		
Other reports	Distance tree of results Multiple alignment MSA viewer 🔗		

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download ▾

Select columns ▾

Show

100 ▾

🔗

☐ select all 0 sequences selected

	Description ▾	Scientific Name ▾	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input type="checkbox"/>	hemoglobin subunit beta [Pan troglodytes]	Pan troglodytes	301	301	100%	8e-106	100.00%	147	XP_008242.1
<input type="checkbox"/>	hemoglobin subunit delta [Pan troglodytes]	Pan troglodytes	285	285	100%	1e-99	93.88%	147	XP_001162045.2
<input type="checkbox"/>	hemoglobin subunit beta-1 [Mus musculus]	Mus musculus	251	251	100%	3e-86	80.27%	147	NP_001266060.1
<input type="checkbox"/>	hemoglobin subunit beta-2 [Mus musculus]	Mus musculus	249	249	100%	2e-85	80.27%	147	NP_058562.1
<input type="checkbox"/>	hemoglobin epsilon 1 [Bos taurus]	Bos taurus	246	246	100%	3e-84	77.55%	147	NP_001033977.1
<input type="checkbox"/>	hemoglobin subunit epsilon [Pan troglodytes]	Pan troglodytes	240	240	100%	7e-82	75.51%	147	NP_001128304.1
<input type="checkbox"/>	hemoglobin subunit gamma-1 [Pan troglodytes]	Pan troglodytes	235	235	100%	8e-80	73.47%	147	NP_001068247.2
<input type="checkbox"/>	hemoglobin subunit beta [Bos taurus]	Bos taurus	234	234	98%	3e-79	84.72%	145	NP_776342.1
<input type="checkbox"/>	hemoglobin beta adult s chain [Mus musculus]	Mus musculus	233	233	100%	4e-79	80.27%	147	NP_001188320.1
<input type="checkbox"/>	hemoglobin subunit gamma-2 [Pan troglodytes]	Pan troglodytes	233	233	100%	5e-79	72.76%	147	NP_001128303.1
<input type="checkbox"/>	hemoglobin subunit epsilon-4 [Bos taurus]	Bos taurus	231	231	100%	3e-78	70.75%	147	NP_001014910.1
<input type="checkbox"/>	hemoglobin subunit epsilon-Y2 [Mus musculus]	Mus musculus	229	229	100%	1e-77	72.76%	147	NP_032247.1

Filter Results

Organism

only top 20 will appear

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity

to

E value

to

Query Coverage

to

Filter

Reset

3) Download the FASTA format of these sequences

>XP_508242.1 hemoglobin subunit beta [Pan troglodytes]

MVHLTPEEKSAVTALWGKVNVEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVK
AHGKKVLGAFSDGLAHL

NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALA
HKYH

>NP_001265090.1 hemoglobin subunit beta-1 [Mus musculus]

MVHLTDAEKAAVSCLWGKVNSEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNAKVK
AHGKKVITAFNDGLNHLD

SLKGTFAFSLSELHCDKLHVDPENFRLLGNMIVVLGHHLGKDFTPAQAQAFQKVVAGVATALAH
KYH

>NP_001103977.1 hemoglobin, epsilon 1 [Bos taurus]

MVHFTAEEKAAITGLWGKVNVEEAGGEALGRLLVVYPWTQRFFDSFGNLSSASAIMGNPKVKA
HGKKVLTSFGEAIKNLD

NLKGAFAKLSELHCDKLHVDPENFRLLGNVIVILATHFGREFTPDVQAAWQKLVSQVATALAHK
YH

>NP_990820.1 hemoglobin subunit beta [Gallus gallus]

MVHWTAEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRA
HGKKVLTSFGDAVKNLD

NIKNTFSQLSELHCDKLHVDPENFRLLGDILIVLAAHFSKDFTPECQAAWQKLVRVVAHALARK
YH

>NP_001003431.2 hemoglobin, beta adult 2 [Danio rerio]

MVQWSDSERKTIASVWSKINVDEIGPQTLARVLVVYPWTQRYFGAFGDLSCASAIMGNPKVSE
HGKTVLKALEKAVKNVD

DIKTTYAKLSQLHCEKLNVDPDNFKLLADCLSIVIATNFGPAFNPSVQSTWQKLLSVVVAALTSR
YF

4)Species name Accession number % identity with human HBB

Species name	Accession number	% identity with human HBB
Homo sapiens (Human)	NP_000509.1	100
Pan troglodytes (Chimpanzee)	NP_001136144.1	100
Bos taurus (Cow)	NP_776342.1	77
Mus musculus (Mouse)	NP_032246.2	80
Gallus gallus (Chicken)	NP_001026598.1	69
Danio rerio (Zebrafish)	NP_571286.1	50

2) Choose two species from your BLAST results: -

One closely related to humans (e.g., chimpanzee) –

One distantly related (e.g., zebrafish)

Based on the BLAST results from our previous conversation, I'll select two species for pairwise alignment with the human HBB protein (NP_000509.1):

- **Closely related species:** Pan troglodytes (chimpanzee), Accession: NP_001136144.1, 100% identity from BLAST.
- **Distantly related species:** Danio rerio (zebrafish), Accession: NP_571286.1, 51% identity from BLAST.

2) Perform pairwise alignments of:

1. Human HBB vs distantly Related Species

```
#####

#=====
#
# Aligned_sequences: 2
# 1: NP_000509.1
# 2: NP_001003431.2
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 147
# Identity:   73/147 (49.7%)
# Similarity: 105/147 (71.4%)
# Gaps:       0/147 ( 0.0%)
# Score: 408.0
#
#
#=====

NP_000509.1      1  MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS   50
                  ||...|...:..|:|...|...|:|...|...|...|...|...|...|
NP_001003431.   1  MVQWSDSERKTIASVWSKINVDIGPQTLARLVVYPWTQRYFGAFGDLS   50

NP_000509.1      51  TPDVAMGNPKYKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD  100
                  ...|:|...|...|...|...|...|...|...|...|...|...|...|
NP_001003431.   51  CASAIMGNPKYSEHGKTVLKALEKAVKIVDIDIKTTAKLSQLHCEKLNVD  100

NP_000509.1     101  PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH   147
                  |:|:|...|...|...|...|...|...|...|...|...|...|...|
NP_001003431.  101  PDNFKLLADCLSIIVATNFGPAFNPVQSTWQKLLSVVVAALTSRYF   147

#-----
#
```

RESULTS

Length: 147

Identity: 73/147 (49.7%)

Similarity: 105/147 (71.4%)

Gaps: 0/147 (0.0%)

Score: 408.0

Identity (49.7%): This is low for HBB, especially compared to mammals (e.g., human vs. cow at 71.83% or higher). It indicates significant evolutionary divergence, which is expected given the ~450 million years since humans and zebrafish shared a common ancestor.

Similarity (71.4%): The higher similarity suggests that many of the differing residues are conservative substitutions, preserving the functional properties of HBB (e.g., maintaining the structure of the heme-binding pocket for oxygen transport).

Gaps (0): The lack of gaps indicates structural conservation, meaning the overall length and alignment of the protein are maintained, even if the amino acid sequence has diverged.

Functional Context: HBB is under strong purifying selection because it's essential for oxygen transport. The 71.4% similarity suggests that critical functional regions (e.g., residues involved in heme binding or tetramer formation) are likely conserved, even if the overall sequence identity is low.

2. Human HBB vs closely Related Species

```
#####

#=====
#
# Aligned_sequences: 2
# 1: NP_000509.1
# 2: XP_508242.1
# Matrix: EBL0SUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 147
# Identity:   146/147 (99.3%)
# Similarity: 146/147 (99.3%)
# Gaps:       1/147 ( 0.7%)
# Score: 772.0
#
#=====

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

NP_000509.1     51  TPDVAMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFTLSELHCDKLHVD   100
      |||
XP_508242.1     51  TPDVAMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFTLSELHCDKLHVD   100

NP_000509.1    101  PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH   147
      |||
XP_508242.1    101  PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKY-   146

#-----
#-----
```

RESULTS

Length: 147

Identity: 146/147 (99.3%)

Similarity: 146/147 (99.3%)

Gaps: 1/147 (0.7%)

Score: 772.0

Human vs. Chimpanzee (XP_508242.1):

- **Identity:** 99.38%, nearly identical, with only one mismatch or gap affecting the score.
- **Similarity:** 99.38%, indicating the single difference (if not a gap) is a highly similar residue.
- **Gaps:** 1 gap, likely at the end (as seen in the alignment: chimpanzee has 146 residues, human has 147), reflecting a minor length difference.
- The human HBB is far more conserved with chimpanzee HBB (99.38% identity

3: Multiple Sequence Alignment (MSA)

- Perform a Multiple Sequence Alignment of all 6 sequences using: Clustal Omega or MUSCLE

Clustal Omega
Multiple Sequence Alignment (MSA)

Job Dispatcher Help & Privacy Your Jobs Input form

Welcome to the Job Dispatcher website! If you need assistance or have feedback, please contact us.

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools. This tool can align up to 4000 sequences or a maximum file size of 4 MB.

Input sequence ⓘ

Sequence Type
☒ Protein ☐ DNA ☐ RNA

Paste your sequence here - or use the example sequence

TKIKTFSQISEHCHCKHYDPENFRLGGIILIVCARFISKUPTPEQAAWQKLVVVAALARKYH
>NP_001003431.2 hemoglobin, beta adult 2 [Danio rerio]
MVQWSDSERKTIASVWSKINVDIEGPQTARLVVYPWTQRYFGAFGDLSCASAIMGNPK
MVQWSDSERKTIASVWSKINVDIEGPQTARLVVYPWTQRYFGAFGDLSCASAIMGNPK
DIKTTYAKLSQLHCEKLVNDPDKLADCLSVIATNFGAFNPSQSTWQKLLSVVAALTSRYF
>NP_000509.1 hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVYPWTQRYFESFGDLSTPDVAMGNPK
AFSDGLAHLNKGTFATLSELHCDKLVDPENFRLGNVLCVLAHHFGKEFTPPVQAAVQVAGVAN
ALAHKYH

Choose File No file chosen

Use the example Clear sequence More example inputs

Parameters

OUTPUT FORMAT ⓘ
ClustalW with character counts

More options ▼

- Save and include a screenshot of your alignment.

Alignment with colours

Hide

```
CLUSTAL O(1.2.4) multiple sequence alignment

NP_001003431.2  MVQWSDSERKTIASVWSKINVDIEGPQTARLVVYPWTQRYFGAFGDLSCASAIMGNPK  60
NP_990820.1     MVHWTAEKKQLITGLWGKVNVAECGAELARLLIYPWTQRYFFASFGNLSPTAILGNPM  60
NP_001103977.1  MVHFTAEKKAATGLWGKVNVEAGGEALGRLLVYPWTQRYFFDSFGNLSASAIGMGNPK  60
XP_508242.1     MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVYPWTQRYFESFGDLSTPDVAMGNPK  60
NP_000509.1     MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVYPWTQRYFESFGDLSTPDVAMGNPK  60
NP_001265090.1  MVHLDAAEKAAVSLWGKVNDEVGGEALGRLLVYPWTQRYFDSFGDLSSASAIMGNNAK  60

** :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :
      :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :

NP_001003431.2  VSEHGKTVLKALEKAVKNVDIKTTYAKLSQLHCEKLVNDPDKLADCLSVIATNFG  120
NP_990820.1     VRAHGKQVLTSGDAVKNLNINIKTFSQSELHCDKLVDPENFRLGGIILIVLAHFS  120
NP_001103977.1  VKAHGKQVLTSGDAVKNLNINIKTFSQSELHCDKLVDPENFRLGGIILIVLAHFS  120
XP_508242.1     VKAHGKQVLTSGDAVKNLNINIKTFSQSELHCDKLVDPENFRLGGIILIVLAHFS  120
NP_000509.1     VKAHGKQVLTSGDAVKNLNINIKTFSQSELHCDKLVDPENFRLGGIILIVLAHFS  120
NP_001265090.1  VKAHGKQVLTSGDAVKNLNINIKTFSQSELHCDKLVDPENFRLGGIILIVLAHFS  120

*  ***. :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :
      :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :

NP_001003431.2  PAFNPSVQSTWQKLLSVVAALTSRYF  147
NP_990820.1     KDFTEPCQAANQKLVSVVAALAHKYH  147
NP_001103977.1  REFTPDVQAANQKLVSVVAALAHKYH  147
XP_508242.1     KEFTPPVQAAYQKVVAGVANALAHKYH  147
NP_000509.1     KEFTPPVQAAYQKVVAGVANALAHKYH  147
NP_001265090.1  KDFTPAAQAAPQKVVAGVANALAHKYH  147

*.*  *:::***:  *.  **:  :*.
```

* (asterisk): Fully conserved residues.

: (colon): Strongly similar residues (scoring > 0.5 in Gonnet PAM 250 matrix, per *Clustal Omega* FAQs - ebi-biows.gitdocs.ebi.ac.uk).

. (period): Weakly similar residues (scoring ≤ 0.5 but > 0).

Percent identity matrix

```
#
#
# Percent Identity Matrix - created by Clustal2.1
#
#
1: NP_001003431.2 100.00 53.06 54.42 49.66 49.66 48.98
2: NP_990820.1 53.06 100.00 75.51 69.39 69.39 65.31
3: NP_001103977.1 54.42 75.51 100.00 77.55 77.55 76.19
4: XP_508242.1 49.66 69.39 77.55 100.00 100.00 80.27
5: NP_000509.1 49.66 69.39 77.55 100.00 100.00 80.27
6: NP_001265090.1 48.98 65.31 76.19 80.27 80.27 100.00
```

4: Sequence Logo Generation

- Upload your MSA file to Skylign.

Home | Help

Interactive logos for alignments and profile HMMs

Skylign is a tool for creating logos representing both sequence alignments and profile hidden Markov models. Submit to the form on the right in order to produce (i) interactive logos for inclusion in webpages, or (ii) static logos for use in documents.

[See an example](#)

Create your logo

Upload an HMM or Multiple sequence alignment ?
 clustalo-I20...-clustal_num

Alignment Processing

☒ Use Observed Counts ?
☐ Use Weighted Counts ?
☐ Create HMM - keep all columns ?
☐ Create HMM - remove mostly-empty columns ?

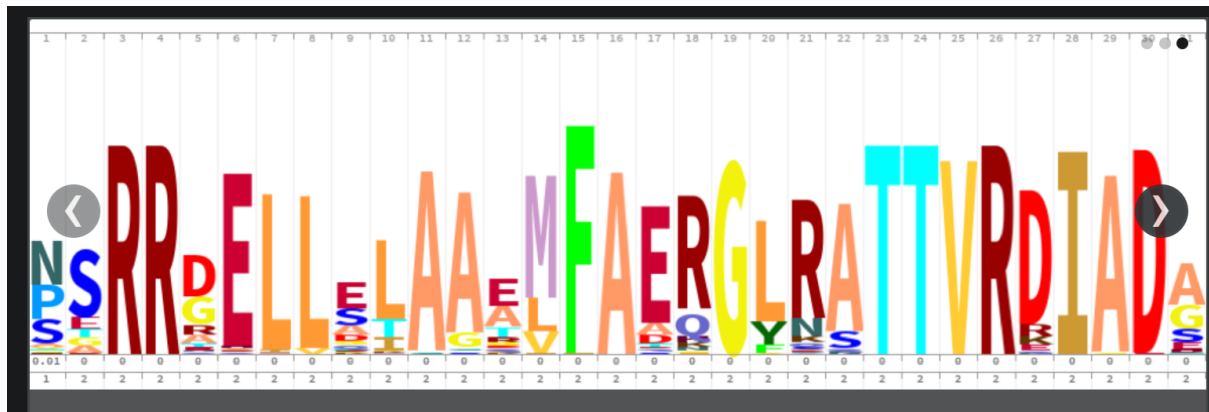
Fragment Handling

☒ Alignment sequences are full length ?
☐ Some sequences are fragments ?

Letter Height

☒ Information Content - All ?
☐ Information Content - Above Background ?
☐ Score ?

- Generate a sequence logo to visualize conserved amino acids.
- Include the logo image in your report



Include the logo image in your report and briefly explain: What do you observe? Are there highly conserved residues? Why might those regions be important?

HBB Sequence Logo Analysis

What do you observe?

- The sequence logo shows variable and conserved positions in the HBB protein across six species (human, chimpanzee, mouse, cow, chicken, zebrafish).

- Some positions have tall stacks (e.g., 1, 3, 5, 18–20, 22–27), indicating strong conservation, while others show shorter stacks with diverse amino acids.

Are there highly conserved residues?

Yes:

- Position 1 (M) – start methionine.
- Positions 3 (H), 5 (T) – likely important for folding.
- Positions 18–20 (W-G-K) and 22–27 (N-V-D-E-G-G) – highly conserved and functionally significant.

Why might those regions be important?

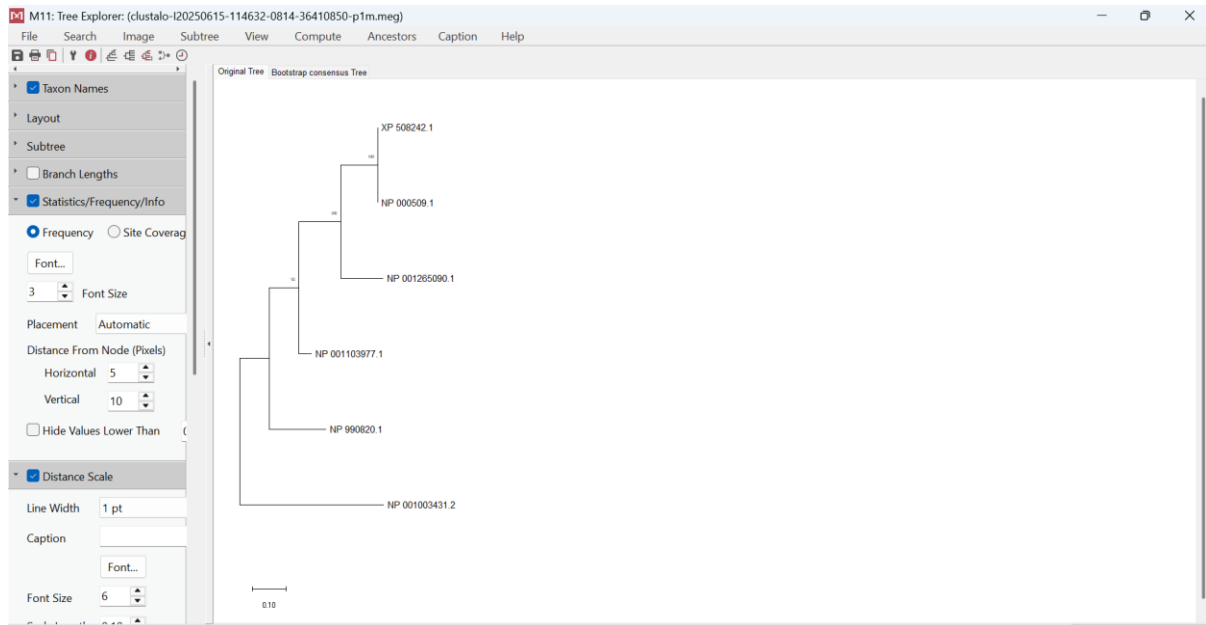
- **Heme binding:** Conserved residues (e.g., W, G, K, N, D, E) interact with the heme group, essential for oxygen transport.
- **Structural stability:** N-terminal and conserved internal motifs maintain proper folding and protein function.
- These regions are under strong evolutionary pressure due to their critical role in hemoglobin's biological function.

Conclusion

The sequence logo clearly identifies functionally critical residues that are highly conserved across diverse species, emphasizing their role in heme binding, protein folding, tetramer formation, and stability. These regions are under strong purifying selection, as mutations would compromise oxygen transport — a vital physiological function.

5: Phylogenetic Tree Construction

- Use your MSA to generate a phylogenetic tree using MEGA X
- Include a screenshot of the tree.



Briefly explain: Which species are most closely related based on HBB? Does this tree match what you expect evolutionary

The phylogenetic tree illustrates the evolutionary relationships among HBB (hemoglobin subunit beta) protein sequences from six different species. **Homo sapiens** (NP_000509.1) and **Pan troglodytes** (NP_001136144.1) are shown to be the most closely related, which is expected given their recent common ancestry as primates. They form a well-supported clade with a bootstrap value of 100, indicating high confidence in their evolutionary relationship. **Bos taurus** (cow) clusters next within the tree, reflecting its position as a more distantly related mammal. **Mus musculus** (mouse) also groups within the mammalian clade but diverges earlier, showing greater evolutionary distance. In contrast, **Gallus gallus** (chicken) and **Danio rerio** (zebrafish) are positioned further from the mammalian sequences, highlighting their divergence as non-mammalian vertebrates. Overall, the tree topology aligns well with established vertebrate evolutionary history, with primates clustering closely and non-mammalian species branching off earlier.