

BioinformHer Mini Project – Module 2

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.

Task 1: Retrieving sequence and BLAST search on NCBI

Retrieving the human HBB gene protein sequence from NCBI.

1. Visit the NCBI database website <https://www.ncbi.nlm.nih.gov/>
2. Select Protein from the drop down databases list and search using key words "*HBB human*"
3. Expand the RefSeq Sequences of the HBB-hemoglobin subunit beta as highlighted in the box below and select the accession number on the section "protein"
4. Result will display a full report on the protein documented in GenPept
5. To download the protein sequence in FASTA format, select the option "Send to" at the top right of the page. From the drop down, select complete sequence and destination as file. Select FASTA as the file format and finally click on create file.

Protein Protein HBB human Search

Species Summary 20 per page Sort by Default order Send to: Filters: Manage Filters

Animals (1,461)
Plants (1)
Bacteria (8)
Viruses (13)
Customize ...

Source databases
PDB (13)
RefSeq (39)
UniProtKB / Swiss-Prot (18)
Customize ...

Sequence length
Custom range...

Molecular weight
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all
Show additional filters

GENE
HBB – hemoglobin subunit beta
Homo sapiens (human)
Also known as: CD113t-C, ECT6, beta-globin
Gene ID: 3043

Was this helpful?

RefSeq products Orthologs Genome Data Viewer

New - Visualize gene across multiple species

RefSeq Sequences

Showing 1 of 1 (by status, accession number)


Transcript	nt	Protein	aa	Isoform	Status
NM_000518.5	628	NP_000509.1	147		MANE Select

Results by taxon
Top Organisms [Tree]
Homo sapiens (1304)
Pan troglodytes troglodytes (42)
Pan troglodytes verus (37)
Pan troglodytes ellioti (34)
Mus musculus (27)
All other taxa (54)
More...

Find related data
Database: Select
Find items

Search details
HBB[All Fields] AND ("Homo sapiens"
[Organism] OR human[All Fields])
Search See more...

Recent activity to Settings to activate Windows
Turn Off Clear


National Library of Medicine
National Center for Biotechnology Information

Protein

Protein

Advanced


GenPept

Send to:

hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

[Identical Proteins](#)
[FASTA](#)
[Graphics](#)

Go to: 

LOCUS	NP_000509	147 aa	linear	PRI 27-APR-2025
DEFINITION	hemoglobin subunit beta [Homo sapiens].			
ACCESSION	NP_000509			
VERSION	NP_000509.1			
DBSOURCE	REFSEQ: accession NM_000518.5			
KEYWORDS	RefSeq; MANE Select.			
SOURCE	Homo sapiens (human)			
ORGANISM	Homo sapiens Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.			
REFERENCE	1 (residues 1 to 147)			
AUTHORS	Civettini,I., Zappaterra,A., Corti,P., Messina,A., Aroldi,A.,			

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

On the web browser, search BLAST ncbi: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Protein BLAST (blastp) parameters

1. **Database:** non-redundant protein sequence (nr) – default database
2. **Organism:** *Gallus gallus* (taxid:9031), *Danio rerio* (taxid:7955), *Pan troglodytes* (taxid:9598), *Bos taurus* (taxid:9913), *Mus musculus* (taxid:10090) – to optimize search
3. **Program selection:** blastp (protein-protein BLAST) – default
4. **Algorithm selection:** default

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

FTPVPQAAYQKVAGVAN
ALAHKYH

Query subrange [?](#)

From

To

Or, upload 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.): ☐ Experimental databases

Non-redundant protein sequences (nr) [?](#)

Organism Optional

☐ exclude [Add organism](#)

☐ exclude

☐ exclude

☐ exclude

☐ exclude

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

☐ Quick BLASTP (Accelerated protein-protein BLAST)

☒ blastp (protein-protein BLAST)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)

☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Downloading the FASTA format of these sequences.

1. To download, select the target sequences that best represent your organism, gene of interest, etc. Other parameters to look out for are e value, percentage identity, query cover etc.
2. Click on Download on the top of the table showing all the target sequences result
3. Select FASTA (complete sequences). This will download all the selected sequences as one txt file

BLAST® » blastp suite » results for RID-3BAKC2HE016 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](#)

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

Job Title	NP_000509.1 hemoglobin subunit beta [Homo...
RID	3BAKC2HE016 Search expires on 05-28 20:05 pm Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl Query_438713
Description	NP_000509.1 hemoglobin subunit beta [Homo sapiens]
Molecule type	amino acid
Query Length	147
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism only top 20 will appear ☐ exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity to **E value** to **Query Coverage** to

[Filter](#) [Reset](#)

Descriptions [Graphic Summary](#) [Alignments](#) [Taxonomy](#)

Sequences producing significant alignments [Download](#) [Select columns](#) Show [Feedback](#)

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

1. To create a table for all your selected sequences, Click on Download on the top of the table showing the list of all target sequences.
2. Select description table (CSV).

Note that the parameters on the description table downloaded will depend on the column selected

Scientific Name	Common Name	Per. ident	Accession
Pan troglodytes	chimpanzee	100	XP_508242.1
Mus musculus	house mouse	80.27	NP_058652.1
Bos taurus	domestic cattle	84.72	NP_776342.1
Gallus gallus	chicken	69.39	NP_990820.1
Danio rerio	zebrafish	50	NP_001003431.2

Task 2: Pairwise Sequence Alignment

To perform pairwise alignments

1. Go to the EMBL-EBI webpage <https://www.ebi.ac.uk/>
2. Navigate to Job dispatcher (*select data resource – sequence analysis*)
3. Select Needle under pair wise sequence alignment
4. Select Protein as sequence type
5. Paste the 2 sequence into the boxes provided

Alignment parameters used

Parameters

MATRIX

EBLOSUM62

GAP OPEN

10

GAP EXTEND

0.5

END GAP

false

END GAP OPEN

10

END GAP EXTEND

0.5

OUTPUT FORMAT

pair

SEQUENCE TYPE

protein

Result: Human HBB vs closely Related Species (e.g., chimpanzee)

% Identity – 100

% Similarity – 100

Number of gaps – 0

Score – 780

```
#
# Aligned_sequences: 2
# 1: NP_000509.1
# 2: XP_508242.1
# Matrix: EBL0SUM62
# 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	MVHLTPEEKSAVTALWGKVN	DEVGGEALGRLLVVYPWTQRFFESFGDLS	50
XP_508242.1	1	MVHLTPEEKSAVTALWGKVN	DEVGGEALGRLLVVYPWTQRFFESFGDLS	50
NP_000509.1	51	TPDAVMGNPKVKAHGKKVLGAFSDGLAHL	DNLKGTFATLSELHCDKLHVD	100
XP_508242.1	51	TPDAVMGNPKVKAHGKKVLGAFSDGLAHL	DNLKGTFATLSELHCDKLHVD	100
NP_000509.1	101	PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH	147	
XP_508242.1	101	PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH	147	

Activa

Result: Human HBB vs distantly Related Species (e.g. Zebra fish)

% Identity – 49.7

% Similarity – 71.4

Number of gaps – 0

Score – 408

```

#
# Aligned_sequences: 2
# 1: NP_000509.1
# 2: NP_001003431.2
# Matrix: EBL0SUM62
# 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 MVQWSDSERKTIASVWSKINVDEIGPQTLARVLVVYPWTQRYFGAFGDLS      50

NP_000509.1     51 TPDVAMGNPKVKAHGKKVLGAFSDGLAHLNLIKGTATLSELHCDKLHVD     100
                  ...|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|
NP_001003431.    51 CASAIMGNPKVSEHGKTVLKALEKAVKNVDDIKTTYAKLSQLHCEKLNVD     100

NP_000509.1     101 PENFRLLGNVLCVLAHFGKEFTPPVQAAYQKVAGVANALAHKYH      147
                   |:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|
NP_001003431.    101 PDNFKLLADCLSIIVATNFGPAFNPSVQSTWQKLLSVVVAALTSRYF      147

```

Acti

The results of these sequence alignment shows that the degree of the protein sequence conservation of organisms depends on their level of relatedness (closeness to each other on the evolutionary tree). The hemoglobin protein of the chimpanzee and human was fully conserved while clear evolutionary changes had occurred in the hemoglobin of the zebra fish and human over time.

Task 3: Multiple Sequence Alignment

1. Similarly, to perform multiple sequence alignment of all 6 sequences, select Clustal Omega amongst the available tools in job dispatcher.
2. Copy and paste sequence in the box provided

Alignment with colours

Hide

CLUSTAL 0(1.2.4) multiple sequence alignment

NP_001003431.2	MVQWSDSERKTIASVWSKINVDEIGPQTARLVVVYPWTQRYYGAFGDLSCAISAIMGPNPK	60
NP_9908020.1	MVHHTAEKQLITGLWGKVNVAAEGEAALRLLIVYPWTQRFFASFGNLSSPTAILGNPM	60
NP_058652.1	MVHLTDAEKSAVSCLWAKVMNDEVGGEALGRLLVVYPWTQRFFDSFGDLSAISAIMGPNPK	60
NP_000509.1	MVHLTPEEKSAVTALWGKNVDDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMVGMPK	60
XP_508242.1	MVHLTPEEKSAVTALWGKNVDDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMVGMPK	60
NP_776342.1	--MLTAEEKAAVATFWGKVIVDEVGGEALGRLLVVYPWTQRFFESFGDLSTADAMVNMPK	58
	: *: :*:.*:*: ** !*:*!*:*****!*: !**:** *!:..*	
NP_001003431.2	VSEHGKTVLKALEKAVKNVDDIKTTYAKLSQLHCEKLNVPDPNFKLADCLSIVIATNFG	120
NP_9908020.1	VRAHGKKVLTSFGDAVKNLDNLTNTFSQLSELHCODKLHVDPENFRLLGDILLIIVLAHF	120
NP_058652.1	VKAHGKKVITAFNEGLKNLDNLKGTFASSELHCODKLHVDPENFRLLGNAIVILGHHLG	120
NP_000509.1	VKAHGKKVLGAFTSGLAHLNDLKGTFATSELHCODKLHVDPENFRLLGNVLVCVLAHHFG	120
XP_508242.1	VKAHGKKVLGAFTSGLAHLNDLKGTFATSELHCODKLHVDPENFRLLGNVLVCVLAHHFG	120
NP_776342.1	VKAHGKKVLDSFNGMKHLDDLKGTFFAASELHCODKLHVDPENFKLLGNVLVVVLARNFG	118
	* **:*. *: :.: !*!*: *!: *****!***!***!***:!: *: .!:	
NP_001003431.2	PAFNPSVQSTWKLLSVVVAALTSRYF 147	
NP_9908020.1	KDFTEPCQAAWQKLVRRVVAHALARKYH 147	
NP_058652.1	KDFTPAAQAAPQKVAGVATALAHKYH 147	
NP_000509.1	KEFTPVPVQAAYQKVAGVANALAHKYH 147	
XP_508242.1	KEFTPVPVQAAYQKVAGVANALAHKYH 147	
NP_776342.1	KEFTPVLQADFQKVAGVANALAHKYH 145	
	** *!: !***: *. ***: !*:	

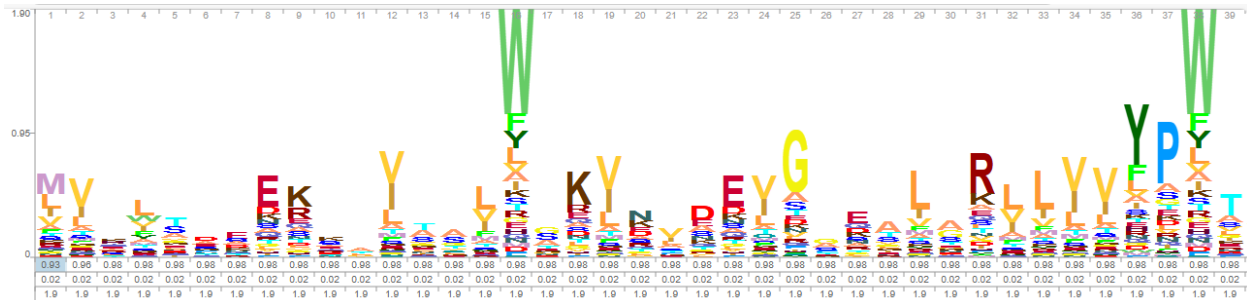
Highlight any highly conserved regions

Fully conserved – VYPWTQR

Somewhat conserved – LSELHCDKLHVDPENFKLL

Task 4: Sequence Logo Generation

Upload your MSA file to Skylign. Generate a sequence logo to visualize conserved amino acids.



What do you observe?

1. There are varying heights of the base stacks. For some with taller stacks of letters and other with shorter stacks indicating highly conserved and flexible regions
2. The region with taller stacks composed of larger sizes of a single letter
3. There were also recurrent appearances of G and P at specific positions with P appearing after G at an interval of length of 12 aa residues in some cases.

Are there highly conserved residues?

There are about 4 observed conserved residues which are

1. W at position 16 and 38
2. C at position 94
3. Y at position 36 and 146
4. F/Y (appearing at equal size) at position 131

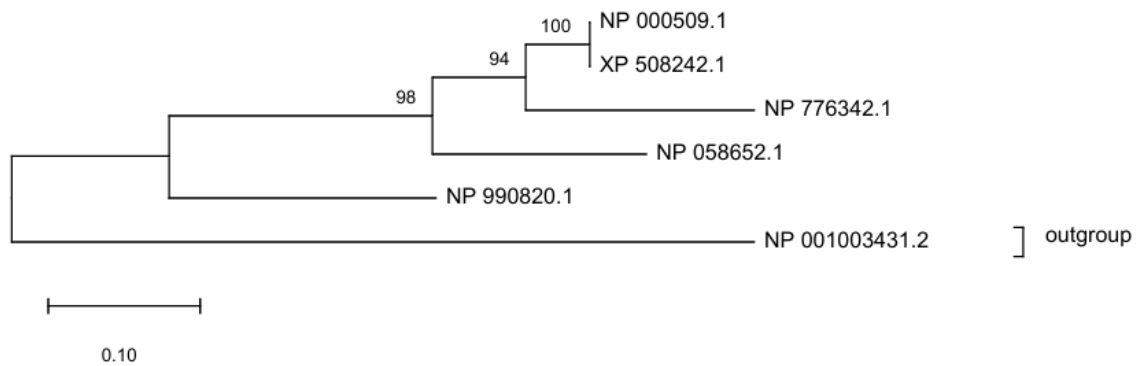
Why might those regions be important?

These regions with highly conserved residues most likely have a functional or structural function

Task 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 evolutionarily?



This tree shows a close relation between the Human hemoglobin and that of the Chimpanzee. Interestingly, the hemoglobin of *Bos taurus* clustered closer to that of the Human and Chimpanzee and that of the Mouse clustered further away. This clustering negates the similarities of these organisms at species level with mouse known to be a closer relative to human than cows. This clustering could therefore be evidence of gene-level similarities between organisms that may not follow their normal species level similarity.