Protein Sequence Analysis Using BLAST

Student Name: Yehya Trad

Internship: CodeAlpha Bioinformatics Internship

Protein Studied: Hemoglobin Subunit Beta (HBB_HUMAN)

UniProt ID: P68871

---

1. Introduction

Protein sequence analysis is a cornerstone of modern bioinformatics, enabling researchers to explore functional, structural, and evolutionary relationships between biological molecules. By comparing an unknown or selected protein sequence against well-curated biological databases, it is possible to identify homologous proteins, conserved regions, and infer biological function.

In this project, the **human Hemoglobin subunit beta (HBB)** protein sequence was analyzed using the **Basic Local Alignment Search Tool (BLAST)** provided by the National Center for Biotechnology Information (NCBI). Hemoglobin beta plays a critical role in oxygen transport in red blood cells and is highly conserved across species. The primary objective of this study was to identify homologous protein sequences, evaluate their similarity using statistical parameters such as **percent identity, query coverage, and E-value**, and interpret the biological significance of the observed alignments.

---

2. Materials and Methods

Data Source:

- UniProt database, Protein ID: **P68871**

Sequence Type:

- Protein sequence (FASTA format)

Analysis Tool:

- **NCBI BLAST** (Protein BLAST)

Database Used:

- Non-redundant protein sequences (**nr**)

Methodology:

1. Retrieve the protein sequence in FASTA format from UniProt.
2. Open **NCBI BLAST**, select **Protein BLAST**, and paste the sequence.

3. Keep default parameters for alignment and database selection.
4. Run BLAST and wait for results.
5. Analyze the **Top 5 homologous sequences** and select one **alignment** as a representative example.

---

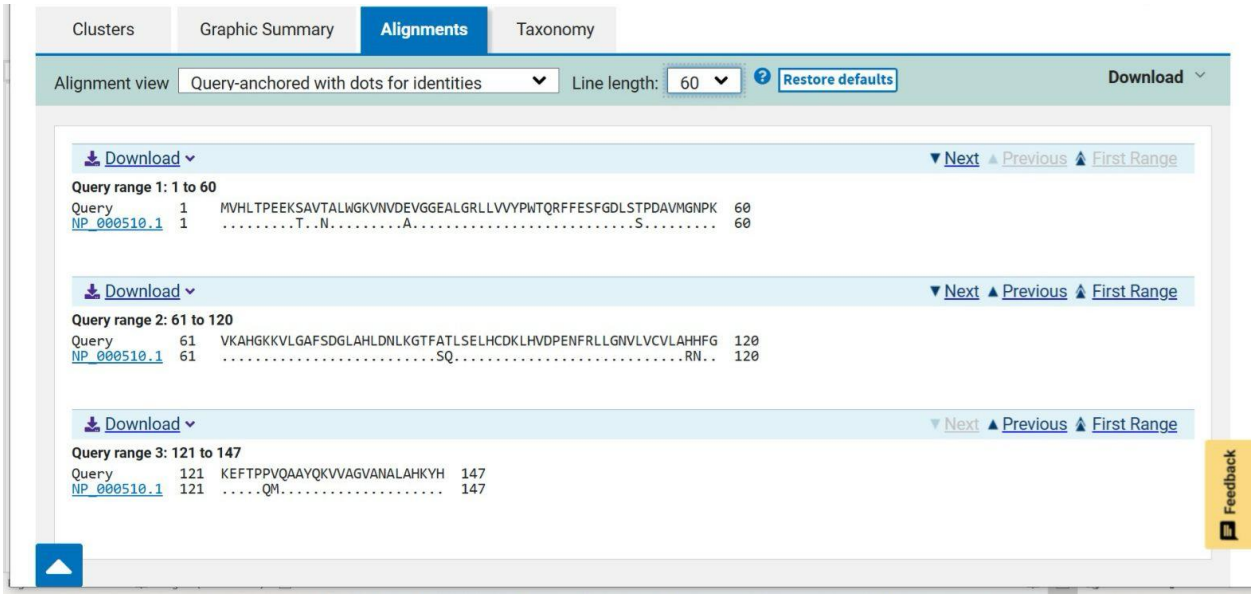
3. Results

3.1 Top 5 Homologous Proteins

select all *100 clusters selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Cluster Composition (Click the ⊞ to see the cluster contents)	Cluster Ancestor	Cluster Representative Sequence	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
⊞1 member(s), 1 organism(s)	human	Hb Monza protein [Homo sapiens]	288	288	100%	1e-97	86.47%	170	XEF57200.1
⊞180 member(s), 74 organism(s)	placentals	hemoglobin subunit delta [Homo sapiens]	284	284	100%	2e-96	93.20%	147	NP_000510.1
⊞1 member(s), 1 organism(s)	human	beta-globin [Homo sapiens]	277	277	100%	2e-93	93.20%	147	AAA35597.1
⊞19 member(s), 16 organism(s)	placentals	hemoglobin subunit beta-like [Ailuropoda melanoleuca]	273	273	100%	6e-92	89.80%	147	NP_001291812.1
⊞1 member(s), 1 organism(s)	giant panda	hypothetical protein PANDA_015769 [Ailuropoda melanol...	275	446	100%	3e-91	90.48%	258	EFB18430.1

3.2 Representative Sequence Alignment



Clusters | Graphic Summary | **Alignments** | Taxonomy

Alignment view: Query-anchored with dots for identities | Line length: 60 | Restore defaults | Download

Download ▼Next ▲Previous ⏫First Range

Query range 1: 1 to 60

```
Query       1   MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
NP_000510.1 1   .........T..N.........A............................S.........  60
```

Download ▼Next ▲Previous ⏫First Range

Query range 2: 61 to 120

```
Query       61  VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
NP_000510.1 61  ........................SQ.........................RN..  120
```

Download ▼Next ▲Previous ⏫First Range

Query range 3: 121 to 147

```
Query       121 KEFTPPVQAAYQKVVAGVANALAHKYH  147
NP_000510.1 121 .....QM.....................  147
```

Feedback

The alignment shows identical and conserved amino acids between the query protein and the homologous protein, highlighting regions of functional importance.

---

4. Discussion

The BLAST analysis revealed a strong similarity between the query protein, Hemoglobin subunit beta, and homologous proteins from multiple organisms. The **top five BLAST hits** demonstrated **high percent identity values** and **near-complete query coverage**, indicating that a large portion of the query sequence aligns with the subject sequences. Such high coverage suggests that the compared proteins share similar lengths and structural organization.

The **very low E-values** (for example, values in the range of 1e-50 to 1e-97 or equal to 0.0) indicate that the observed alignments are statistically highly significant and extremely unlikely to have occurred by chance. This confirms that the identified proteins are true homologs rather than random matches.

Additionally, the representative alignment analysis showed extensive regions of conserved amino acids between the query and subject sequences. These conserved regions are often associated with essential functional or structural roles in the protein, such as maintaining protein stability or facilitating oxygen-binding activity in hemoglobin. The presence of such conserved motifs highlights the evolutionary conservation of hemoglobin beta and underscores its critical biological function.

Overall, the BLAST results provide insights into the evolutionary relationships among hemoglobin proteins and demonstrate how sequence similarity analysis can be used to infer functional and biological relevance.

---

5. Conclusion

In this study, the protein sequence of **human Hemoglobin subunit beta (HBB)** was successfully analyzed using the NCBI BLAST tool. The analysis identified multiple homologous proteins with high sequence similarity, strong query coverage, and extremely low E-values, confirming the evolutionary conservation and functional importance of this protein.

The representative sequence alignment further emphasized the presence of conserved amino acid regions, supporting the conclusion that hemoglobin beta maintains a critical and preserved role across different species. This project demonstrates the effectiveness of BLAST as a fundamental bioinformatics tool for protein sequence comparison, functional annotation, and evolutionary analysis, and highlights its importance in biological and biomedical research.
