Identification of Unknown Biological Sequences Using BLAST

Objective

- To identify unknown biological sequences using the NCBI BLAST tool and interpret the results to determine sequence type, organism of origin, and biological function.

Sequence Classification

- Based on the nucleotide and amino acid composition of the given sequences:
  - Sequence 1 is DNA
  - Sequence 2 is DNA
  - Sequence 3 is Protein

Sequence 1 Analysis

BLAST Details

- BLAST Tool Used: BLASTn
- Database: core_nt (GenBank)
- Query Length: 220 bp

Top Hit Identification

- Gene/Sequence Name: Ebola virus genome fragment
- Organism: *Zaire ebolavirus*
- Query Coverage: 100%
- Percent Identity: 100%
- E-value: 4e-109
- Accession ID: MH464888.1

Alignment Verification

- The alignment shows a continuous matching region with no significant gaps or mismatches. The extremely low E-value and complete query coverage strongly
  support that the unknown sequence corresponds to the identified viral genome fragment.

Biological Function

- This sequence is part of the Ebola virus genome, which encodes proteins essential for viral replication, host cell infection, and viral assembly. Such sequences play a critical role in viral pathogenicity and disease progression.

Classification & Interpretation

- Sequence Type: Gene (DNA)
- Organism Type: Viral
- Conservation: Highly conserved among Ebola virus strains

Conclusion

- The given unknown sequence is identified as a genomic fragment of Zaire ebolavirus. BLAST analysis shows a 100% identity with an E-value of 4e-109, indicating a strong and reliable match.

Sequence 2 Analysis

BLAST Details

- BLAST Tool Used: BLASTn
- Database: core_nt (GenBank)
- Query Length: 241 bp

Top Hit Identification

- Gene Name: Insulin (INS) gene, complete coding sequence
- Organism: *Homo sapiens*
- Query Coverage: 100%
- Percent Identity: 99.59%
- E-value: 2e-118
- Accession ID: AH002844.2

Alignment Verification

- The alignment demonstrates a high level of sequence similarity with minimal mismatches and no major gaps. The very low E-value confirms that the sequence corresponds to the human insulin gene.

Biological Function

- The insulin gene encodes the insulin hormone, which plays a vital role in regulating blood glucose levels by facilitating cellular glucose uptake. It is essential for maintaining metabolic homeostasis in humans.

Classification & Interpretation

- Sequence Type: Gene (DNA)
- Organism Type: Eukaryotic
- Conservation: Highly conserved across mammalian species

Conclusion

- The given unknown sequence is identified as the insulin (INS) gene from Homo sapiens. BLAST analysis shows a 99.59% identity and an E-value of 2e-118, indicating a highly significant match.

Sequence 3 Analysis

BLAST Details

- BLAST Tool Used: BLASTp
- Database: ClusteredNR
- Query Length: 100 amino acids

Top Hit Identification

- Protein Name: Envelope glycoprotein gp160
- Organism: Viral
- Query Coverage: 92%
- Percent Identity: 80.43%
- E-value: 2e-44
- Accession ID: 7AEJ_A

Alignment Verification

- The protein alignment shows substantial similarity across most of the sequence length, with acceptable variation typical of viral proteins. The low E-value supports a confident protein-level identification.

Biological Function

- Envelope glycoprotein gp160 is a viral surface protein involved in host cell attachment and membrane fusion. It plays a crucial role in viral entry into host cells and is a key determinant of viral infectivity.

Classification & Interpretation

- Sequence Type: Protein
- Organism Type: Viral
- Conservation: Conserved among related viral species

Conclusion

- The given unknown sequence is identified as the viral envelope glycoprotein gp160. BLAST analysis shows an 80.43% identity with an E-value of 2e-44, indicating a strong and biologically meaningful match.

Final Interpretation

- BLAST analysis successfully identified all three unknown sequences by comparing them against curated biological databases. The results demonstrate the effectiveness of BLAST in distinguishing between DNA and protein sequences, determining organismal origin, and inferring biological function. This activity highlights the importance of sequence similarity analysis in genomic and proteomic research.

Name: Aaditya Kashyap

aadikun2007@gmail.com