Project Title: Bioinformatics

Task 2: Multiple Sequence Alignment of Insulin Protein

Domain: Bioinformatics

Intern Name: Muhammad Daniyal Hameed **Date:** December 9, 2025

1. Introduction

Multiple Sequence Alignment (MSA) is a central technique in bioinformatics used to align three or more biological sequences (protein, DNA, or RNA) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied. MSA is essential for identifying conserved sequence regions, which often correspond to structural or functional domains.

For this task, the **Insulin** protein was selected. Insulin is a peptide hormone produced by beta cells of the pancreatic islets; it regulates the metabolism of carbohydrates, fats, and protein by promoting the absorption of glucose from the blood. Due to its critical physiological role, the core structure of insulin is highly conserved across vertebrates.

2. Methodology

- **Tool Used:** Clustal Omega (EMBL-EBI).
- **Sequences:** Protein sequences for the insulin gene were retrieved from the UniProt/NCBI database for the following five species:
    1. *Homo sapiens* (Human)
    2. *Mus musculus* (Mouse)
    3. *Bos taurus* (Cow)
    4. *Sus scrofa* (Pig)
    5. *Danio rerio* (Zebrafish)
- **Procedure:** The sequences were input in FASTA format into the Clustal Omega tool. The default parameters were used to generate the alignment.

3. Results and Interpretation

The resulting alignment reveals significant conservation across the species, particularly in the regions responsible for the protein's active structure.

3.1. Alignment Output

```
CLUSTAL O(1.2.4) multiple sequence alignment


NP_571131.1         MAVWLQAGALLVLLVVS-SVSTNPGTPQHLCGSHLVDALYLVCGPTGFFYNPKRDVEPL-        58
NP_032412.3         MALLVHFLPLLALLALWEPKPTQAFVKQHLCGPHLVEALYLVCGERGFFYTPKSRREVED        60
NP_000198.1         MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED        60
NP_001103242.1      MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAEN        60
NP_776351.2         MALWTRLAPLLALLALWAPAPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG        60
                    **:  :   **.**.:     :   . ***** ***:*******  ****.**   *

NP_571131.1         --LGFLPPKSAQETEVADFAFKDHAELIRKRGIVEQCCHKPCSIFELQNYCN       108
NP_032412.3         PQVEQLELGG--SPG--DLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN       108
NP_000198.1         LQVGQVELGG--GPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN       110
NP_001103242.1      PQAGAVELGG--GLG--GLQALALEGPPQKRGIVEQCCTSICSLYQLENYCN       108
NP_776351.2         PQVGALELAG--GPG-----AGGLEGPPQKRGIVEQCCASVCSLYQLENYCN       105
                         :   .                 :*****:*** . **:::*:****
```

Figure 1: Clustal Omega Multiple Sequence Alignment of Insulin proteins.

3.2. Analysis of Conserved Regions The alignment output uses specific symbols to indicate conservation:

- **Asterisk (*):** Indicates positions which have a single, fully conserved residue.
- **Colon (:):** Indicates conservation between groups of strongly similar properties.
- **Period (.):** Indicates conservation between groups of weakly similar properties.

Key Observations:

1. **Cysteine Conservation:** The Cysteine (C) residues (e.g., around positions 45-50 and 90-100 in the alignment) are perfectly conserved across all five species (marked with *). These residues are crucial because they form the **disulfide bridges** (links) that hold the A-chain and B-chain of the insulin molecule together. Without these specific cysteines, the protein would not fold correctly.

2. **Functional Core:** The regions corresponding to the mature insulin peptide show high similarity between the mammal species (Human, Pig, Cow, Mouse). The Zebrafish sequence (*NP_571131.1*) shows more variation (gaps and different amino acids) compared to the mammals, which is expected given the greater evolutionary distance between fish and mammals.

3. **Signal Peptide:** The beginning of the sequence (N-terminus) shows more variation. This region often contains the signal peptide, which directs the protein's transport and is cleaved off, meaning it is under less evolutionary pressure to stay exactly the same.

4. Conclusion

The Multiple Sequence Alignment successfully highlights the evolutionary conservation of the insulin protein. The strict conservation of Cysteine residues confirms their importance in maintaining the protein's 3D structure. The high degree of similarity between the pig (*Sus scrofa*) and human (*Homo sapiens*) sequences also explains why pig insulin was successfully used to treat human diabetes for many years before synthetic human insulin became available.

5. References

1. Sievers, F., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7, 539.

2. UniProt Consortium. (2023). UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), D523–D531.