Mini Project: Comprehensive Sequence Analysis of the Human TNF Gene

Project Title:
Comprehensive Sequence Analysis of the Human TNF Gene

Objective:
Applying bioinformatics skills learned to download, analyze, and interpret the sequence of the human TNF gene, which encodes a proinflammatory cytokine called TNF.

Project Overview:
In this mini-project, I performed a series of bioinformatics tasks using the human TNF gene as my sequence of interest. I began the project by downloading the sequence, translating it, finding ORFs, analyzing sequence composition, identifying transcription factor binding sites, searching for functional motifs, predicting coding/noncoding regions, and converting sequence file formats.

Task 1: Downloading a Biological Sequence from NCBI and View/Edit It

Objective:
Download the human TNF gene sequence and view it using BioEdit.

Steps:
- I accessed the NCBI homepage at NCBI.
- Searched for the human TNF gene using the term 'human TNF gene.'
- Located the correct sequence record (e.g., 'Homo sapiens TNF').
- Downloaded the sequence in FASTA format.
- Opened the sequence in BioEdit and viewed it.

Output:

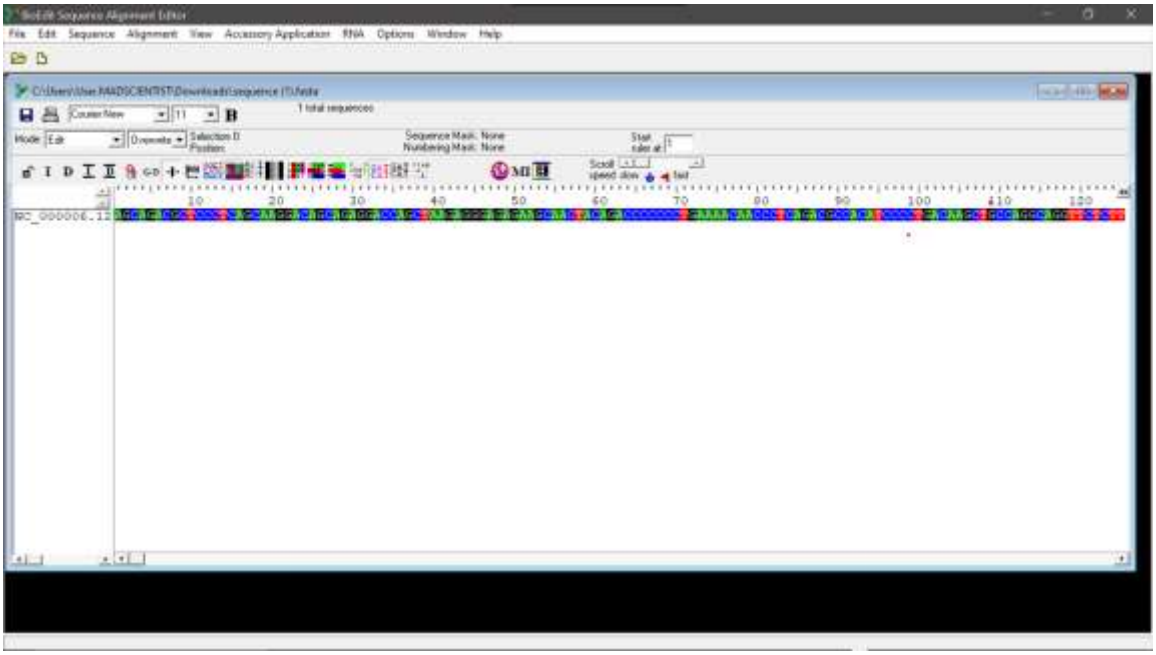

Fig 1: TNF Gene sequence display in BioEdit

Task 2: Generating a Translation of the DNA or RNA Sequence into Amino Acids

Objective:

Translate the DNA sequence of the TNF gene into an amino acid sequence.

Steps:

- Opened the downloaded TNF gene sequence in BioEdit.
- Used the 'Translate' feature in BioEdit to generate the amino acid sequence.

Output:

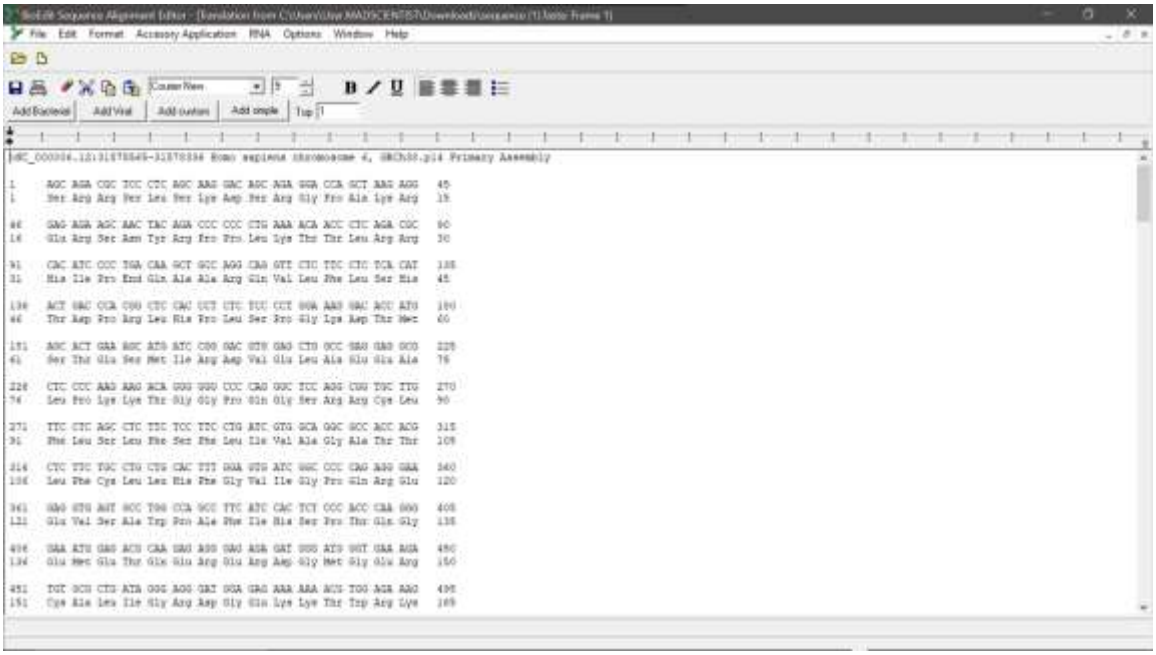

Fig 2: Translated amino acid in BioEdit.

Task 3: Finding ORFs (Open Reading Frames) in the DNA or RNA Sequence

Objective:

Identify the ORFs within the TNF gene sequence.

Steps:

- Used BioEdit's ORF Finder tool to find ORFs in the TNF gene sequence.
- Recorded the start and stop positions, lengths, and protein translations of the ORFs.

Output:

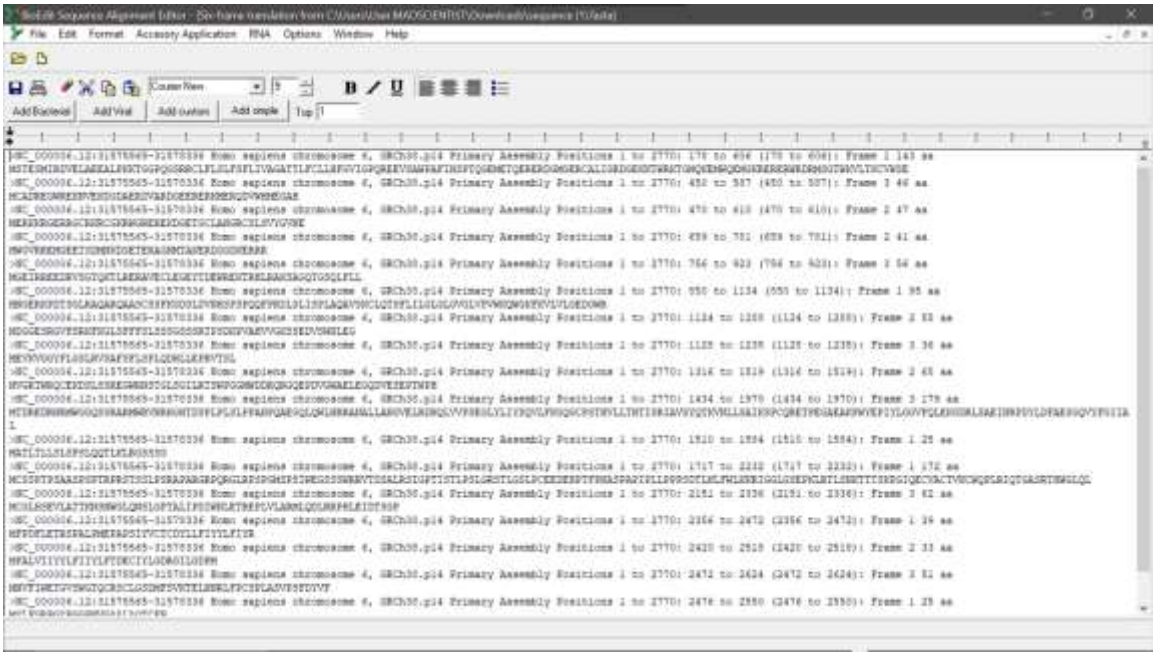

Fig 3: ORFs of the TNF gene in BioEdit

Results Interpretation

I selected 4 ORFS with the highest length which could be potential candidates for protein-coding. The ORFS selected were:

>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2770: 1434 to 1970 (1434 to 1970): Frame 3 179 aa
MTDREDRNRMWGGQSSRARMWRVNRHGHTDSPLPLSLPPANPQAEGQLQWLNRRANALL ANGVELRDNQLVVPSEGLYLIYSQVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSA IKSPCQRETPEGAEAKPWYEPIYLGGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGII AL

>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2770: 1717 to 2232 (1717 to 2232): Frame 1 172 aa
MCSSPTPSAASPSPTRPRSTSSLPSRAPARGRPQRGLRPSPGMSPSIWEGSSSWRRVTD SALRSIGPTISTLPSLGRSTLGSLPCEEDEHPTFPNASPAPIPLLPPPSDTLNLFWLKK RIGGLGSEPKLRTLSNKTTTSKPGIQECVACTVKCWQPLRIQTGASRTHWGLQL

>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2770: 850 to 1134 (850 to 1134): Frame 1 95 aa

MNGERKPDTSGLRAQARQAASCSSFKGDSLDVNHSPSPQQFPRDLSLISPLAQAVSKCL QTSFLILGLGLGVGLVPVWKQWGKFKVLVLGEDGWR

>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2770: 1316 to 1519 (1316 to 1519): Frame 2 68 aa
MVGRTWRQCEKDSLSSREGWRNSTGLSGILRTSWPGGMWDDRQRGQEPDVGWAELEGQD VESEPTWPH

I then performed SMART BLAST Analysis on the selected ORFS and the following sequences were obtained:

>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6]: 178 to 606: Frame 1 143 aa

>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6]: 850 to 1134: Frame 1 95 aa

>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6]: 1395 to 1682: Frame 3 96 aa

Conclusion: The length of codons is not a determining factor in protein-coding genes.

Task 4: Analyzing the Sequence Composition (Nucleotide or Amino Acid Frequencies)

Objective:

Analyze the nucleotide composition of the TNF gene sequence.

Steps:

- Used BioEdit to analyze the sequence composition of the TNF gene.
- Calculated the frequencies of each nucleotide and the overall GC content.
- Interpreted the results and saved the analysis.

Output:

Fig 4: Nucleotide composition of TNF gene

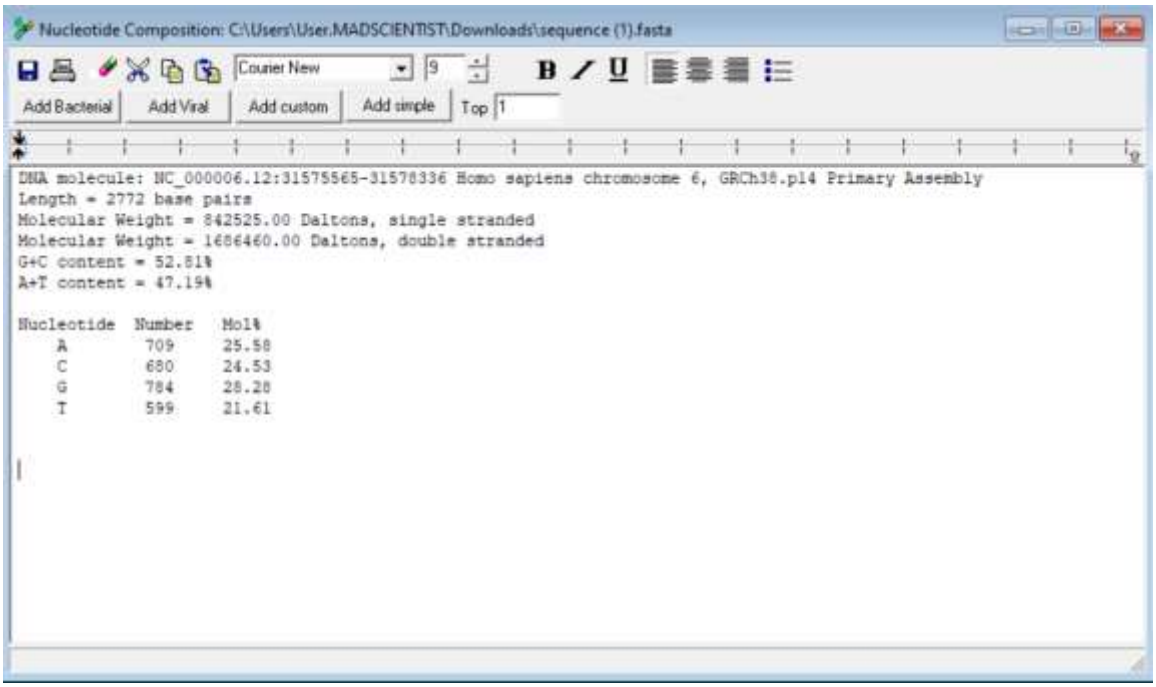

Fig 5: Detailed Nucleotide composition of TNF gene

Results Interpretation

According to their molecular weights, the TNF gene sequence has a high G+C content. The guanine-cytosine (G+C) content of a DNA sequence refers to the

percentage of nucleotides in the sequence that are either guanine (G) or cytosine (C). A higher G+C content in DNA provides increased thermal and chemical stability due to the stronger triple hydrogen bonds between G-C pairs, making the DNA more resistant to denaturation and potentially reducing mutation rates. This stability is advantageous for organisms in extreme environments and can lead to a more compact genome. However, it also requires more energy for DNA replication and transcription, potentially slowing these processes and growth rates. Additionally, high G+C content can complicate PCR amplification and sequencing and may limit codon flexibility, affecting protein expression.

Task 5: Identifying Transcription Factor Binding Sites Using the PROMO Tool

Objective:

Identify potential transcription factor binding sites in the TNF gene promoter region.

Steps:

- Accessed the PROMO tool at PROMO.
- Selected 'Homo sapiens' as the species.
- Input the entire TNF gene sequence
- Identified potential transcription factor binding sites.

Output:

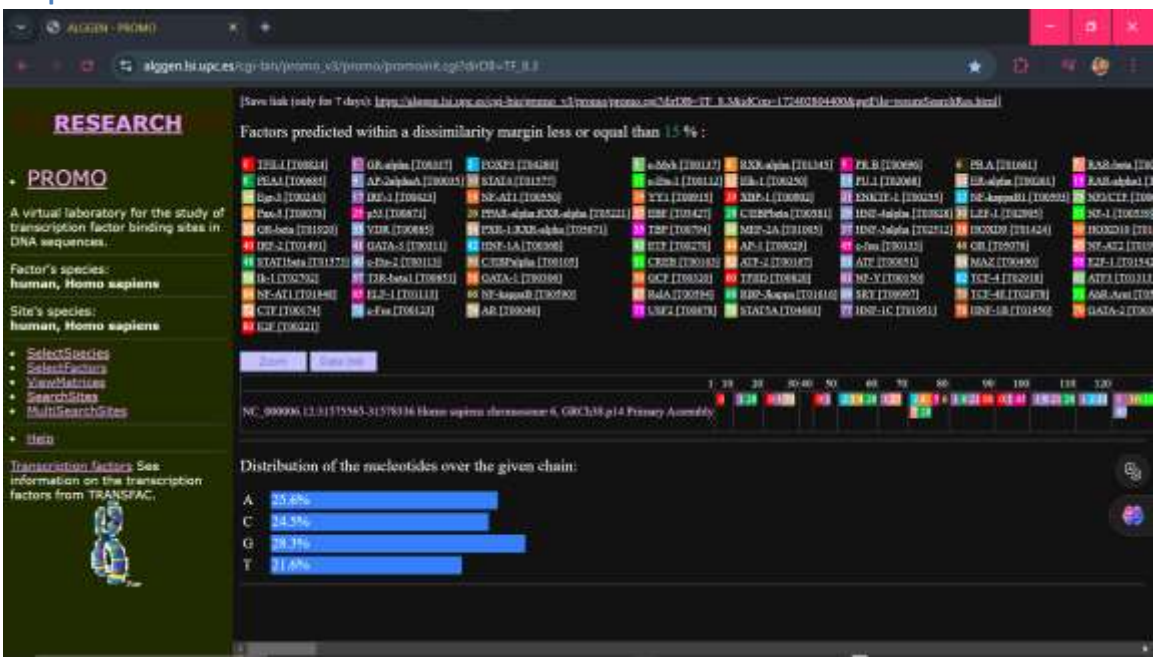

Fig 6: Transcription Factor Binding Sites (PROMO Tool)

Results interpretation

Many transcriptional binding sites were found on the TNF gene sequence. The Factors predicted within a dissimilarity margin of less than or equal to 15% as shown above. The distribution of the nucleotides over the given chain confirmed Guanine as having the highest content of 28.3% molecular weight. The figure below shows detailed results of the binding sites as a primary assembly:

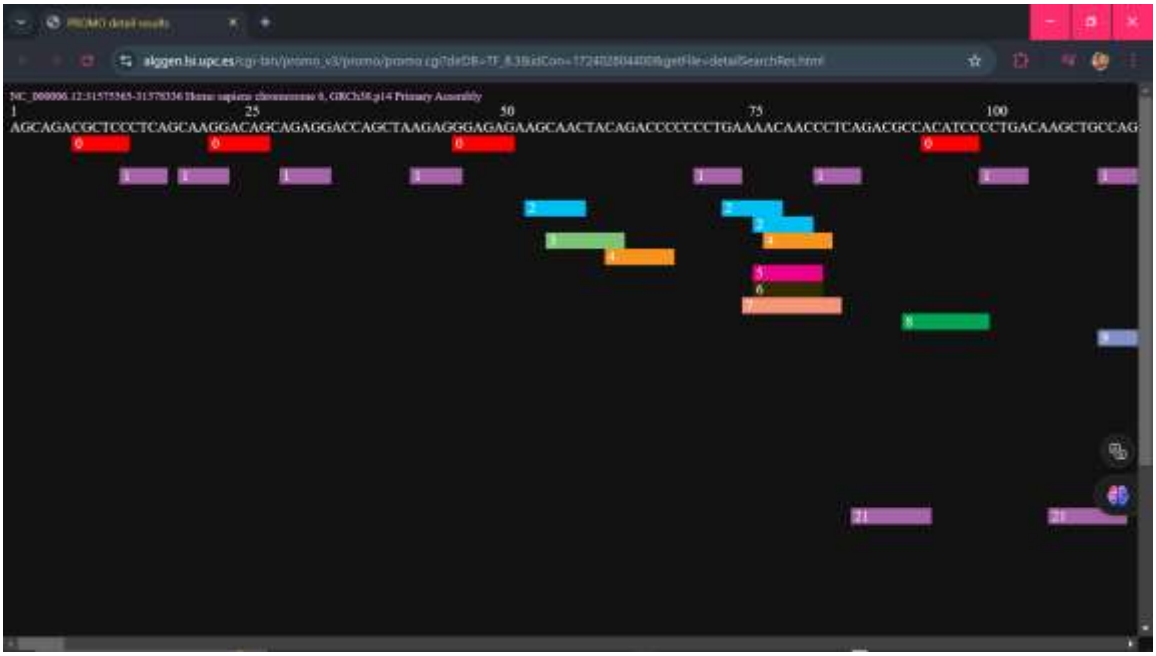

Fig 7: Transcription Binding Sites Detailed Result (PROMO)

Task 6: Searching for Functional Motifs in a Genome or Transcriptome Using MEME Suite

Objective:

Search for functional motifs in the TNF gene sequence using MEME Suite.

Steps:

- Accessed the MEME Suite at MEME Suite.
- Uploaded the TNF gene sequence in FASTA format.
- Used the default settings to search for motifs.
- Interpreted and saved the results of the motif search.

Output:

Fig 8: Functional motifs of the TNF gene (Meme Suite)

Fig 9: Detailed image of the discovered motifs (Meme Suite)

Results interpretation

A motif is an approximate sequence pattern that occurs repeatedly in a group of related sequences. MEME represents motifs as position-dependent letter-probability matrices that describe the probability of each possible letter at each position in the pattern. The output above shows three Functional motifs from the TNF gene sequence used. According to Meme Suite, these motifs are of high significance due to their high p-values. We can conclude that there is a higher biological activity starting at regions 413 to 800 due to the high clustering observed from the motif locations.

Task 7: Predicting Coding/Non-Coding Regions in a Genome Using GENSCAN

Objective:

Predict the coding and non-coding regions within the TNF gene sequence.

Steps:

- Accessed the GENSCAN tool or run it locally if installed.
- Input the TNF gene sequence in the appropriate format.
- Ran the analysis to predict coding and non-coding regions.
- Saved and interpret the results.

Output:

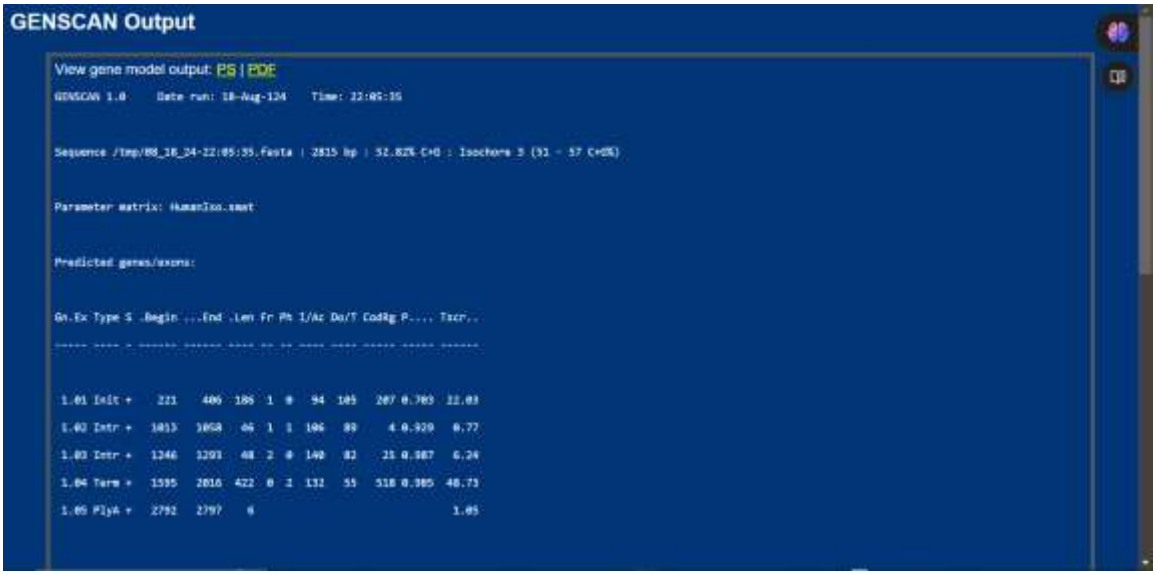

Fig 10: GENSCAN Output of the Coding & Non-coding Regions

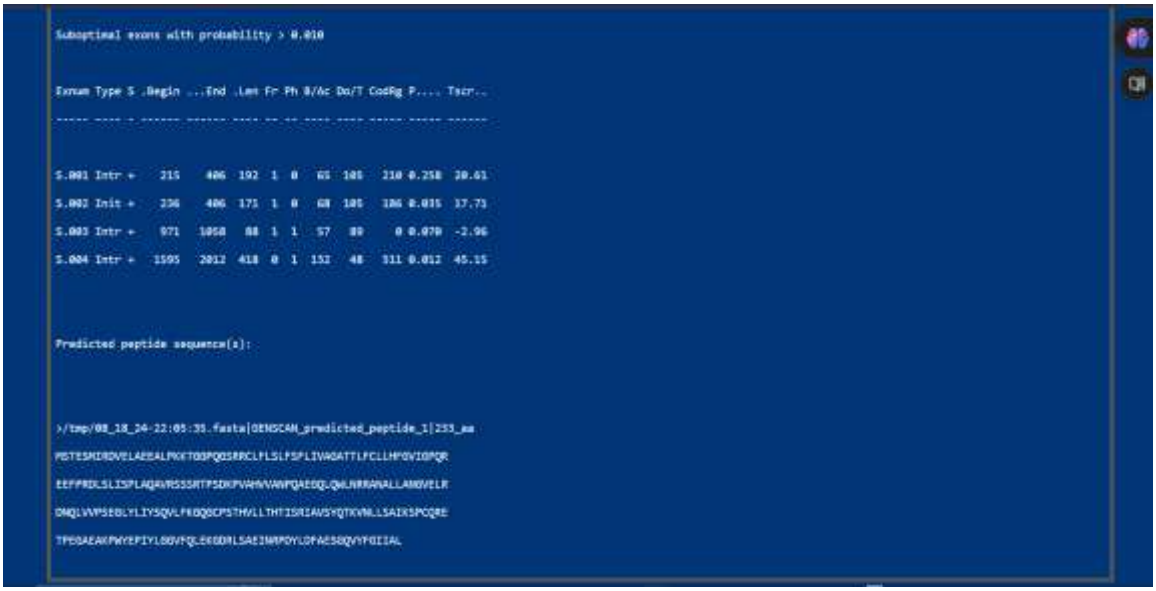

Fig 11: GENSCAN Output of the Coding & Non-coding Regions (continuation)

Results Interpretation

A gene with five exons on the positive strand can be predicted based on the GENSCAN output seen above. The first exon is an initial exon starting at position 221 and ending at position 406 with a length of 186 base pairs. There are two internal exons, one spanning position 1013 to 1058 and another from 1246 to 1293. The gene ends with a terminal exon from position 1595 to 2016, followed by a polyadenylation signal at positions 2792 to 2797. All predicted exons have high coding region and transcript scores, indicating strong confidence in the gene predictions.

Task 8: Conversion of Sequence File Formats Using BioEdit (FASTA to PHYLIP)

Objective:

Convert the TNF gene sequence from FASTA format to PHYLIP format.

Steps:

- Opened the TNF gene sequence in BioEdit.
- Used the 'Save As...' feature to convert the file to PHYLIP format.
- Verified the conversion by opening the PHYLIP file in a text editor.

Output:

Fig 12: TNF gene sequence in PHYLIP format.

Discussion

This project presents a detailed analysis of the human TNF gene, offering new insights into its structure, function, and regulatory mechanisms. The identification of multiple open reading frames (ORFs) within the gene, and their translation, points to potential protein-coding regions that may be crucial for the gene's biological functions. Although the functionality of the longest ORF has not been confirmed, the presence of other notable ORFs supports the idea that the TNF gene is involved in protein synthesis.

The sequence composition analysis reveals a higher GC content, which enhances genomic stability but could introduce difficulties in experimental procedures such as PCR. According to He et al. (2023), the discovery of transcription factor binding sites within the TNF promoter region highlights the gene’s complex regulation, underscoring its importance in immune responses and inflammatory processes. The identification of functional motifs within the TNF gene further emphasizes regions of potential biological significance. These motifs, especially those concentrated in specific regions, may play key roles in the gene's contribution to disease mechanisms. This discovery opens up new avenues for research, particularly in understanding the gene's involvement in autoimmune disorders and other inflammatory conditions.

In conclusion, this project significantly advances our knowledge of the TNF gene, providing a foundation for future studies. Such studies might focus on experimentally validating the predicted motifs and transcription factor binding sites, as well as investigating gene variants across different populations to gain a deeper understanding of the TNF gene's role in disease susceptibility and health outcomes.

References

1. Bioinformher Linkedin Resources: https://www.linkedin.com/company/bioinformher/posts/?feedView=videos

2. Burge, C. B. (1998) Modeling dependencies in pre-mRNA splicing signals. In Salzberg, S., Searls, D. and Kasif, S., eds. Computational Methods in Molecular Biology, Elsevier Science, Amsterdam, pp. 127-163.

3. Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

4. Farré, D., Roset, R., Huerta, M., Adsuara, J. E., Roselló, L., Albà, M. M., & Messeguer,
   X. (2003). Identification of patterns in biological sequences at the ALGGEN server:
   PROMO and MALGEN. Nucleic Acids Research, 31(13), 3651-3653. doi:10.1093/nar/gkg638. PROMO

5. Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and
   analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series, 41, 95-
   98. BioEdit

6. He, H., Yang, M., Li, S., Zhang, G., Ding, Z., Zhang, L., Shi, G., & Li, Y. (2023). Mechanisms and biotechnological applications of transcription factors. Synthetic and Systems Biotechnology, 8(4), 565–577. https://doi.org/10.1016/j.synbio.2023.08.006