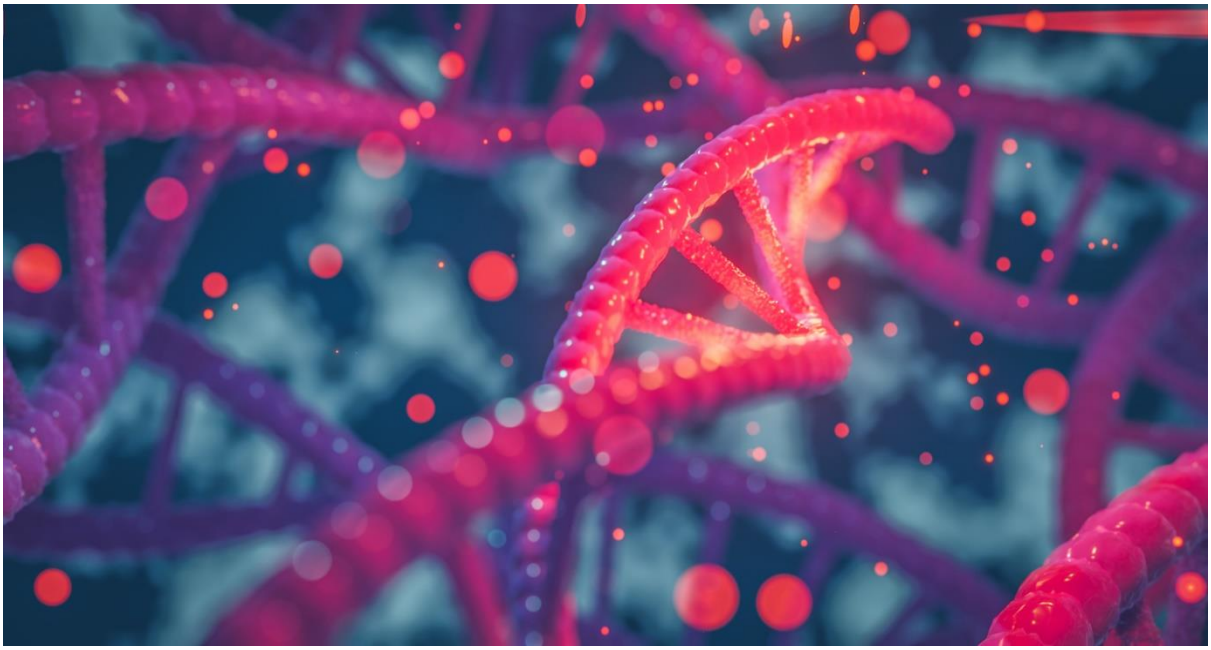


Comprehensive Sequence Analysis of the Human IL6 gene

Project Title: Comprehensive Sequence Analysis of the Human IL6 gene

Project Overview:

In this mini-project, I carried out a series of bioinformatics tasks using the human IL6 gene as my sequence of interest. I began by downloading the sequence, translating it, identifying ORFs, analyzing the sequence composition, and identifying transcription factor binding sites. I also searched for functional motifs, predicted coding or non-coding regions, and converted sequence file formats. The tools I used in this project include NCBI BLAST, BioEdit, PROMO, MEME Suite, and GENSCAN.

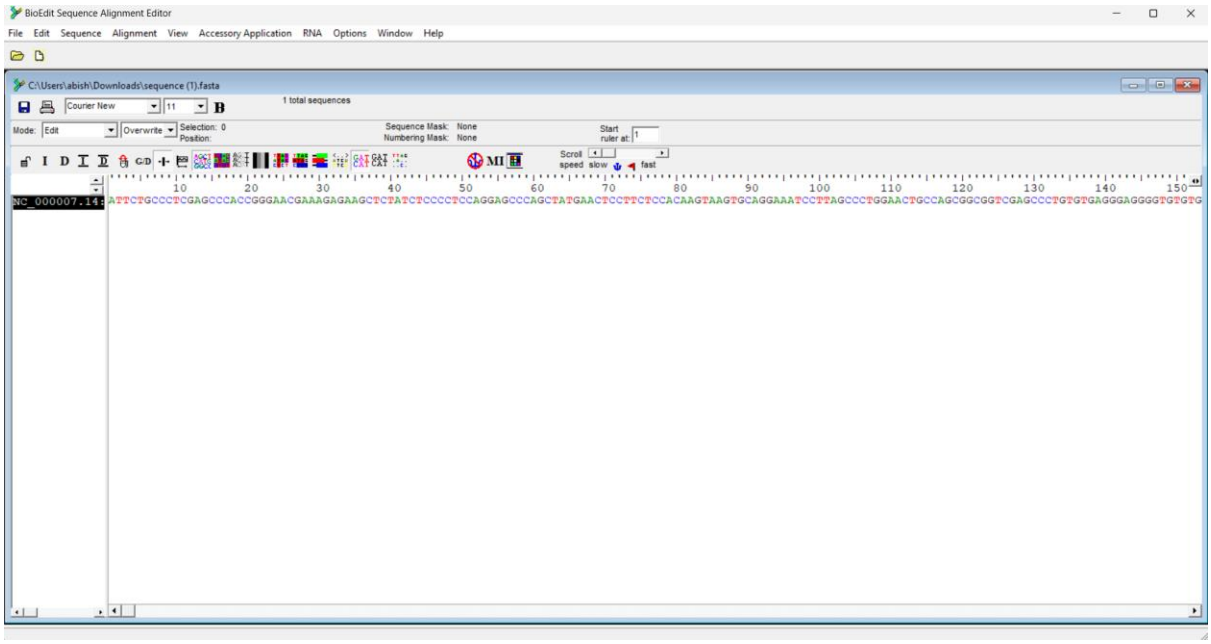
Task 1: Download a Biological Sequence from NCBI and View/Edit It

Objective: Download the human IL6 gene sequence and view it using BioEdit.

Steps:

1. Accessed the NCBI homepage.
2. Searched for the human IL6 gene using the term "human IL6 gene."
3. Identified the correct sequence record (e.g., "Homo sapiens IL6").
4. Downloaded the gene sequence in FASTA format.

5. Selected the most scientifically accepted sequence version (NCBI Reference Sequence: NC_000007.14), as it contained multiple versions.
6. Opened and viewed the sequence in BioEdit for further analysis

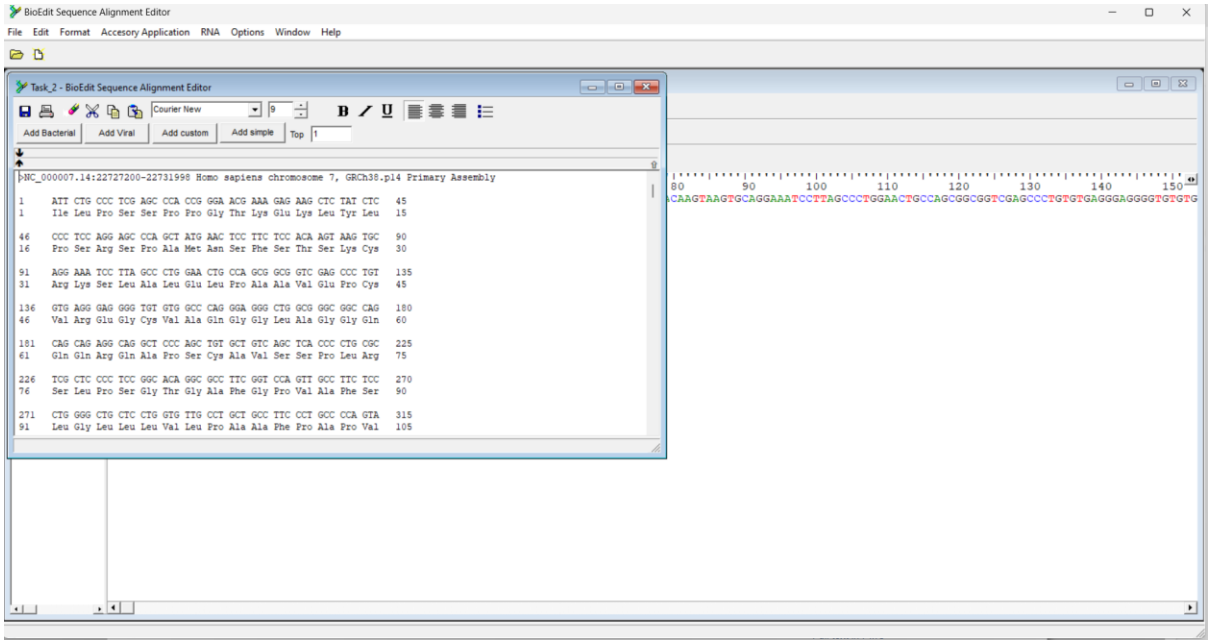


Task 2: Generate a Translation of a DNA or RNA Sequence into Amino Acids

Objective: Translate the DNA sequence of the IL6 gene into an amino acid sequence.

Steps:

1. Opened the downloaded IL6 gene sequence in BioEdit.
2. Used the 'Translate' feature in BioEdit to convert the DNA sequence into its corresponding amino acid sequence.

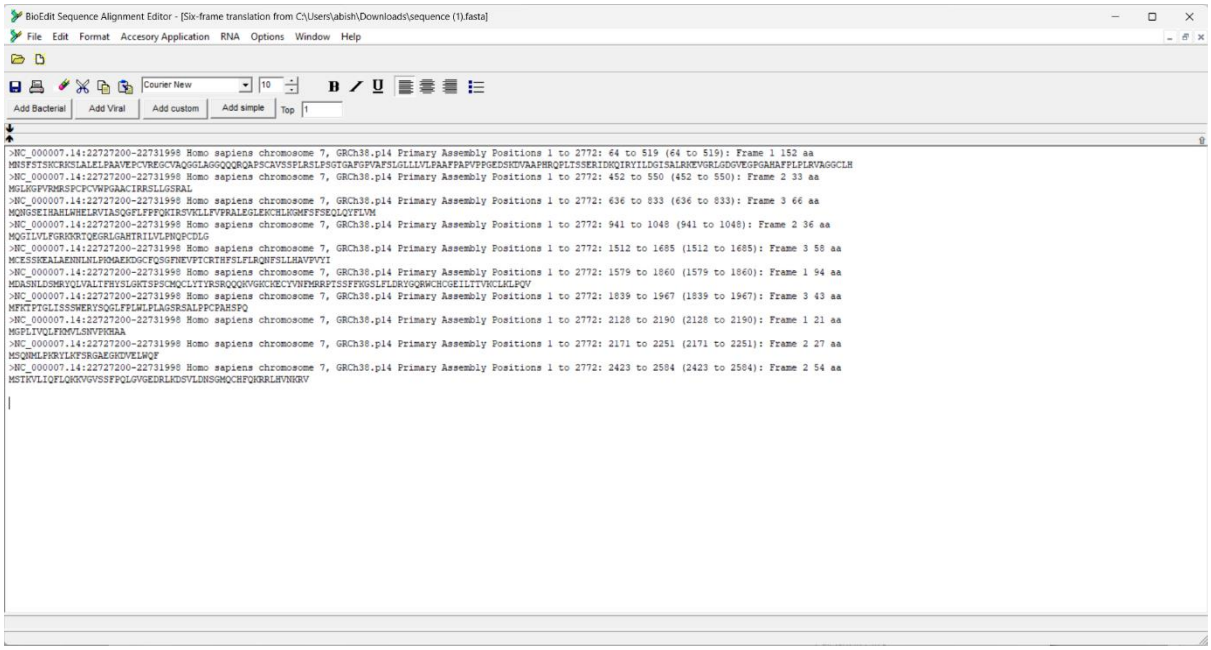


Task 3: Finding ORFs (Open Reading Frames) in a DNA or RNA Sequence

Objective: Identify the ORFs within the IL6 gene sequence.

Steps:

1. BioEdit's ORF Finder tool was used to locate the ORFs in the IL6 gene sequence.
2. Recorded the start and stop positions, lengths, and protein translations of the identified ORFs.



Interpretation of Task 3

From the identified ORFs, I selected the four with the longest lengths, as they are potential candidates for protein-coding regions.

Next, I performed SMART BLAST analysis on these ORFs and found that three protein sequences were identified from the database as significant matches, further supporting their potential role in coding for functional proteins

```
>NC_000007.14:22727200-22731998 Homo sapiens chromosome 7, GRCh38.p14 Primary
Assembly Positions 1 to 2772: 64 to 519 (64 to 519): Frame 1 152 aa
MNSFSTSKCRKSLALELPAAVEPCVREGCVAQGGLAGGQQQRQAPSCAVSSPLRSLPSGTGAFGPVAFSLGLLLVLPAAFPAP
VPPGEDSKDVAAPHRQPLTSSERIDKQIRYILDGISALRKEVGRLGDGVEGPGAHAFPLPLRVAGGCLH

>NC_000007.14:22727200-22731998 Homo sapiens chromosome 7, GRCh38.p14 Primary
Assembly Positions 1 to 2772: 636 to 833 (636 to 833): Frame 3 66 aa
MQNGSEIHAHLWHELRVIASQGFLFPFQKIRSVKLLFVPRALEGLEKCHLKGMFSFSEQLQYFLVM


>NC_000007.14:22727200-22731998 Homo sapiens chromosome 7, GRCh38.p14 Primary
Assembly Positions 1 to 2772: 1579 to 1860 (1579 to 1860): Frame 1 94 aa
MDASNLDSMRYQLVALTFHYSLGKTSPSCMQCLYTYRSRQQQKVGKCKECYVNFMRRPTSSFFKGSLFLDRYGQRWCHCGEIL
TTVKCLKLPQV

>NC_000007.14:22727200-22731998 Homo sapiens chromosome 7, GRCh38.p14 Primary
Assembly Positions 1 to 2772: 2128 to 2190 (2128 to 2190): Frame 1 21 aa
MGPLIVQLFKMVLSNVPKHAA
```

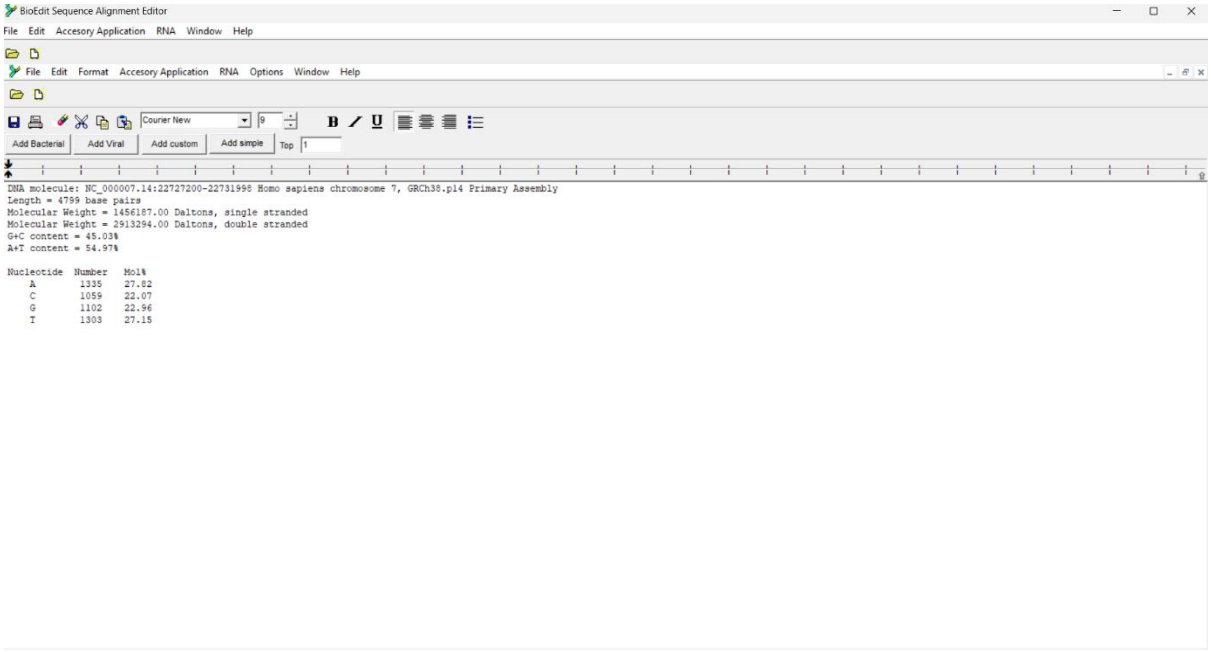
Interestingly, the longest ORF did not show any functional matches in the BLAST analysis. This observation suggests that the length of an ORF or codons alone does not always guarantee that it codes for a functional protein. Other factors, such as sequence conservation and functional domains, are crucial in determining whether a gene truly codes for a protein.

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

Objective: Analyze the nucleotide composition of the IL6 gene sequence.

Steps:

1. Used BioEdit to perform a sequence composition analysis on the IL6 gene.
2. Calculated the frequencies of each nucleotide (A, T, C, G) and the overall GC content.
3. Interpreted the nucleotide distribution and GC content, and saved the analysis results for further reference.



Interpretation of Task 4

The IL6 gene sequence shows a higher G+C content. This presents both advantages and disadvantages depending on the biological context and the level of G+C content:

Advantages:

- **Genomic stability**: Higher G+C content tends to make the DNA more stable, especially under high temperatures.
- **Protection against mRNA degradation**: Increased G+C content may offer protection from exonuclease degradation.
- **Regulation of gene expression**: G+C-rich regions can form secondary structures (e.g., hairpins) that may play roles in regulating gene expression.

Disadvantages:

- **PCR/Cloning difficulties**: The increased thermal stability due to high G+C content can complicate PCR amplification and cloning processes.
- **Genomic instability**: The secondary structures formed by G+C-rich regions may contribute to genomic instability under certain conditions.
- **Transcriptional and translational hindrance**: Secondary structures could interfere with efficient transcription and translation, potentially affecting protein synthesis.

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

Objective: Identify potential transcription factor binding sites in the IL6 gene promoter region.

Steps:

1. Accessed the PROMO tool online.
2. Selected "Homo sapiens" as the species for analysis.

3. Inputted the promoter region of the IL6 gene (or used the entire gene sequence).
4. Ran the analysis to identify potential transcription factor binding sites in the IL6 gene.

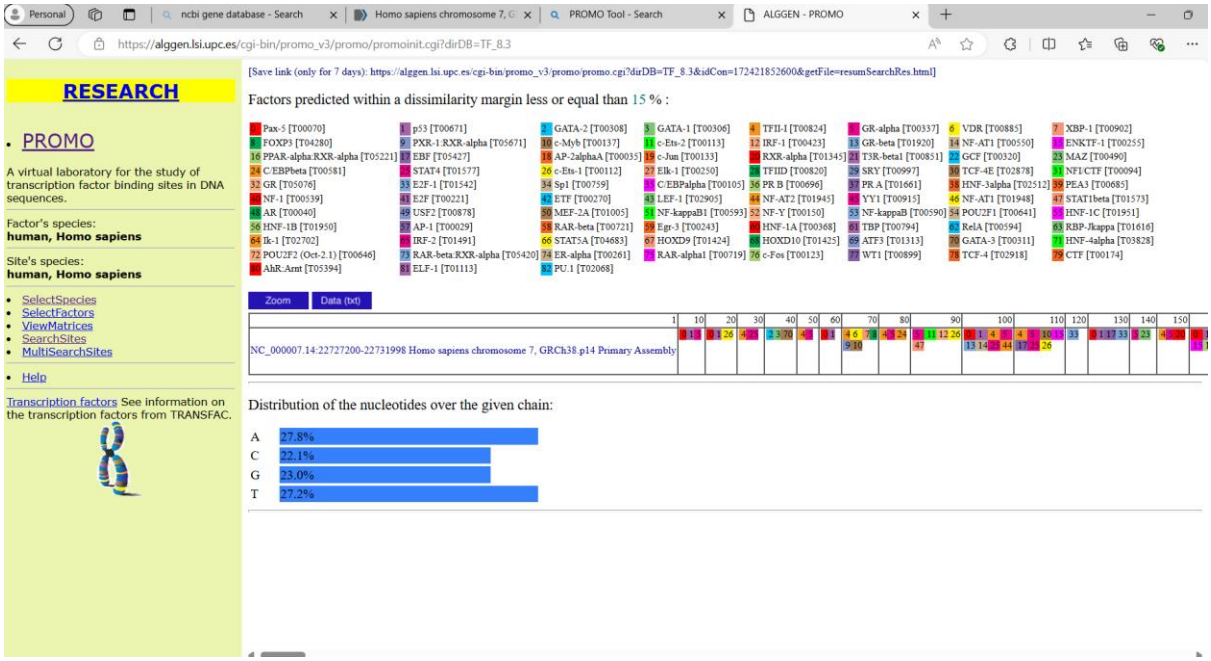


Figure 5: Transcription factor binding sites (PROMO)

Interpretation of Task 5

Found many transcriptional binding sites along the sequence

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

Objective: Search for functional motifs in the IL6 gene sequence using MEME Suite.

Steps:

1. Accessed the MEME Suite online.
2. Uploaded the IL6 gene sequence in FASTA format.
3. Used the default settings to perform the motif search.
4. Interpreted the results to identify functional motifs and saved the analysis for further review.

https://meme-suite.org/meme//opal-jobs/appMEME_5.5.61724218653090-1416992502/meme.html

MEME

Multiple Em for Motif Elicitation

For further information on how to interpret these results please access https://meme-suite.org/meme/doc/meme.html.
To get a copy of the MEME software please access https://meme-suite.org.

If you use MEME in your research, please cite the following paper:
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. [full text]

DISCOVERED MOTIFS | MOTIF LOCATIONS | INPUTS & SETTINGS | PROGRAM INFORMATION | RESULTS IN TEXT FORMAT | RESULTS IN XML FORMAT

DISCOVERED MOTIFS

	Logo	E-value	Sites	Width	More	Submit/Download
1.		2.8e-001	5	41		
2.		6.1e+001	5	19		
3.		5.4e+001	4	48		

https://meme-suite.org/meme//opal-jobs/appMEME_5.5.61724218653090-1416992502/meme.html

MEME
Multiple Em for Motif Elicitation

For further information on how to interpret these results please access https://meme-suite.org/meme/doc/meme.html.
To get a copy of the MEME software please access https://meme-suite.org.

If you use MEME in your research, please cite the following paper:
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. [full text]

DISCOVERED MOTIFS | MOTIF LOCATIONS | INPUTS & SETTINGS | PROGRAM INFORMATION | RESULTS IN TEXT FORMAT | RESULTS IN XML FORMAT

DISCOVERED MOTIFS

1.
E-value: 2.8e-001 **Site Count:** 5 **Width:** 41

Standard | Reverse Complement

Log Likelihood Ratio: 169 **Information Content:** 49.1 **Relative Entropy:** 48.8 **Bayes Threshold:** 9.893

Name	Strand	Start	p-value	Sites
1. NC_000007.14:22727200-22731998	+	3616	6.35e-24	GAGTTCAAGA CCAGCCTGGATAACATAGCAAGACCCCATCTCTACAAAAAA CCAAAACCAA
1. NC_000007.14:22727200-22731998	+	3839	5.68e-19	GAGTTTAAGA CCAGCCTGGTCAACATGGTAAAACCCCATCTCTACTTAAAA ATACAAAAAA
1. NC_000007.14:22727200-22731998	+	3996	1.40e-15	CCACTGCACT CCAGCCTGGGTGACAGAACGAGACCTTGACTCAAAAAAAAA AAAAAGAAGT
1. NC_000007.14:22727200-22731998	+	46	1.58e-14	GCTCTATCTC CCCTCCAGGAGCCCAGCTATGAACTCCTTCTCCACAAGTAA GTGCAGGAAA
1. NC_000007.14:22727200-22731998	-	3085	2.51e-14	CAGGGAATTC CCATGCATGTTAACATATGAGAAACCCTGGCTTAAGTAGAA ATGGCAGAGG

MOTIF LOCATIONS

Only Motif Sites | Motif Sites+Scanned Sites | All Sequences | Download PDF | Download SVG

Name	p-value	Motif Locations
1. NC_000007.14:22727200-22731998	1.81e-39	

INPUTS & SETTINGS

Sequences

Role	Source	Alphabet	Sequence Count	Total Size
Primary Sequences	sequence_1.fasta	DNA	1	4799

Background Model

Source: built from the (primary) sequences

Order: 0

Name	Freq.	Bg.			Bg.	Freq.	Name
Adenine	0.275	0.275	A	T	0.275	0.275	Thymine
Cytosine	0.225	0.225	C	G	0.225	0.225	Guanine

Other Settings

Motif Site Distribution	ANR: Any number of sites per sequence
Objective Function	E-value of product of p-values
Starting Point Function	E-value of product of p-values
Site Strand Handling	Sites may be on either strand
Maximum Number of Motifs	3
Motif E-value Threshold	no limit
Minimum Motif Width	6
Maximum Motif Width	50
Minimum Sites per Motif	2
Maximum Sites per Motif	5

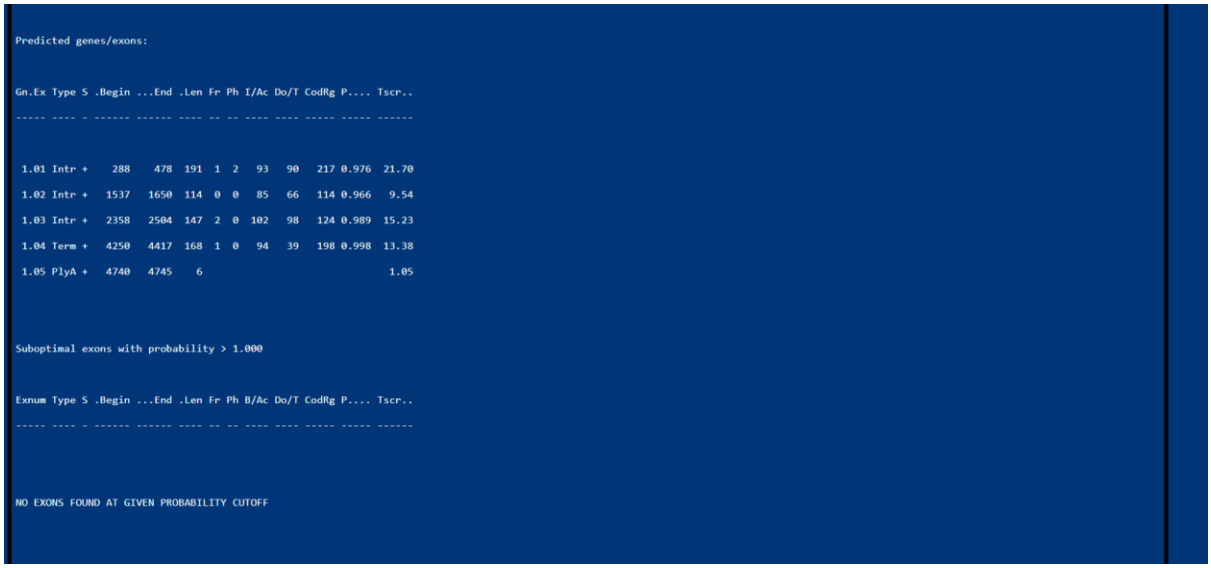
Show Advanced Settings

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

Objective: Predict the coding and non-coding regions within the IL6 gene sequence.

Steps:

1. Accessed the GENSCAN tool online.
2. Input the IL6 gene sequence in the required format.
3. Ran the analysis to predict the coding and non-coding regions.
4. Saved and interpreted the results to identify the coding and non-coding regions within the IL6 gene sequence



```
Predicted genes/exons:

Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.01 Intr +    288    478  191  1  2   93   90   217 0.976  21.70
 1.02 Intr +   1537   1650  114  0  0   85   66   114 0.966   9.54
 1.03 Intr +   2358   2504  147  2  0  102   98   124 0.989  15.23
 1.04 Term +   4250   4417  168  1  0   94   39   198 0.998  13.38
 1.05 PlyA +   4740   4745    6                             1.05


Suboptimal exons with probability > 1.000

Exnum Type S .Begin ...End .Len Fr Ph B/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------


NO EXONS FOUND AT GIVEN PROBABILITY CUTOFF
```

Interpretation of Task 7

According to the GENSCAN output, the IL6 gene has been predicted with the following features:

- **Exons**: The gene comprises five exons on the positive strand.
    - **Initial Exon**: Spans from position 288 to 478, with a length of 191 base pairs.
    - **Internal Exons**:
        - One spanning from position 1537 to 1650.
        - Another from position 2358 to 2504.
    - **Terminal Exon**: Extends from position 4250 to 4417.
- **Polyadenylation Signal**: Located from positions 4740 to 4745.

All predicted exons exhibit high coding region and transcript scores, suggesting a strong confidence in the accuracy of these gene predictions. The presence of the polyadenylation signal indicates a typical mRNA processing site, further supporting the validity of the predicted coding regions.

Task 8: Convert Between Sequence File Formats Using BioEdit (FASTA to PHYLIP)

Objective: Convert the IL6 gene sequence from FASTA format to PHYLIP format.

Steps:

1. Opened the IL6 gene sequence in BioEdit.
2. Used the "Save As..." feature to convert the file to PHYLIP format.
3. Verified the conversion by opening the PHYLIP file in a text editor to ensure the format was correctly applied.

Discussion

This project provided a thorough analysis of the human IL6 gene, yielding significant insights into its structure, function, and regulation. The identification of multiple ORFs within the gene, coupled with their translation, indicates potential protein-coding regions that may be crucial for the IL6 gene's biological roles. Although the longest ORF was not confirmed as functional, the presence of other significant ORFs supports the gene's involvement in protein synthesis and highlights its potential functional importance.

The analysis of sequence composition revealed a higher GC content in the IL6 gene. While this increased GC content contributes to genomic stability, it may also present challenges for experimental procedures such as PCR, due to its heightened thermal stability.

The identification of transcription factor binding sites within the IL6 promoter region highlights the gene's complex regulatory mechanisms, which are essential for its functions in immune responses and inflammation.

Additionally, the discovery of functional motifs within the IL6 gene points to regions of potential biological significance. These motifs, particularly those clustered in specific areas, may be critical for understanding the gene's involvement in disease processes. They could serve as valuable targets for future research, especially in investigating the gene's role in autoimmune disorders and other inflammatory conditions.

In conclusion, this project not only deepens our understanding of the IL6 gene but also opens new avenues for exploring its role in health and disease. Future studies should focus on experimentally validating the predicted motifs and transcription factor binding sites, as well as examining gene variants across different populations to gain a better understanding of its role in disease susceptibility.

References:

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. Journal of Molecular Biology, 215(3), 403-410. doi:10.1016/S0022-2836(05)80360-2. NCBI BLAST

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

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

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., ... & Noble, W. S. (2009). MEME Suite: tools for motif discovery and searching. Nucleic Acids Research, 37(suppl_2), W202-W208. doi:10.1093/nar/gkp335. MEME Suite

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology, 268(1), 78-94. doi:10.1006/jmbi.1997.0951.
GENSCAN