NAME: RACHAEL OLUWAKAMIYE ABOLADE

Project Title:

Comprehensive Sequence Analysis of the Human TNF Gene

Objective:

To apply bioinformatics skills learned in Module 1 to download, analyze, and interpret the sequence of the human TNF gene, which encode a proinflammatory cytokine call TNF.

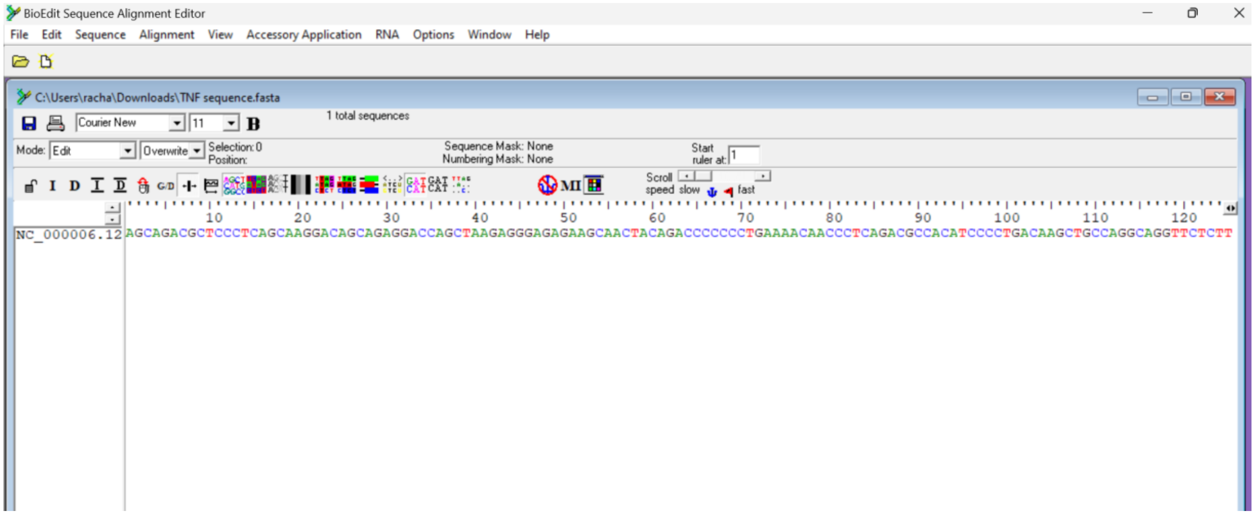
Project Overview:

In this mini project, you will perform a series of bioinformatics tasks using the human TNF gene as your sequence of interest. The project will guide you through downloading the sequence, translating it, finding ORFs, analyzing sequence composition, identifying transcription factor binding sites, searching for functional motifs, predicting coding/non-coding regions, and converting sequence file formats.

Date: AUGUST 20TH 2024

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

The human TNF gene sequence was downloaded from NCBI and viewed using BioEdit. Here is a screenshot of the sequence display:

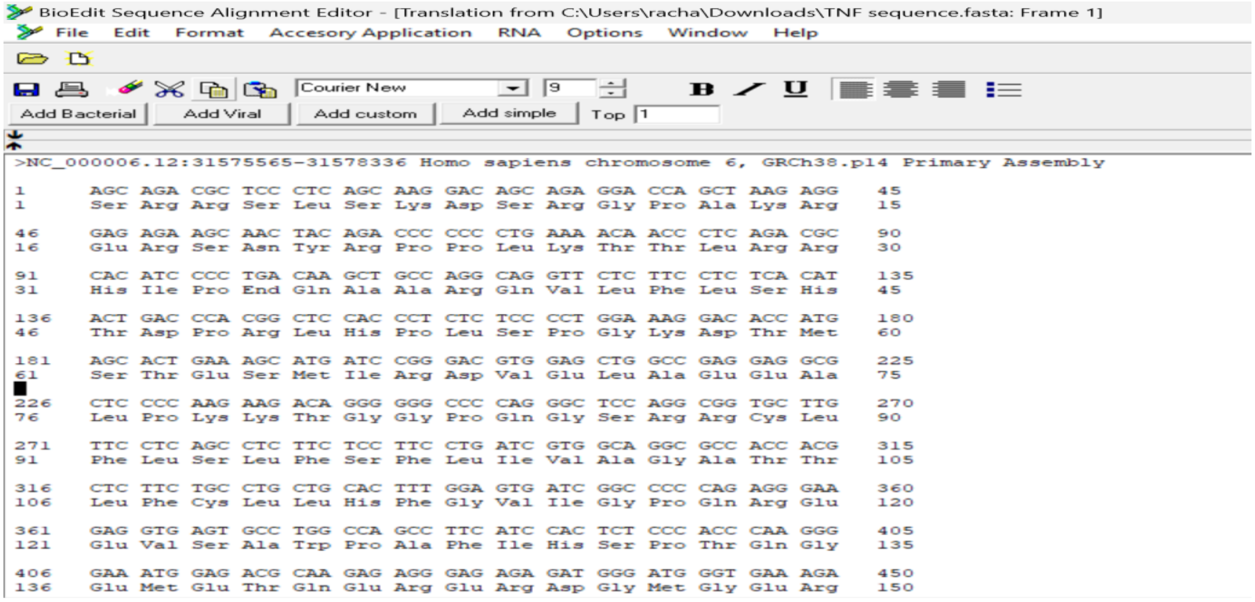


```
NC_000006.12 AGCAGACGCTCCCTCAGCAAGGACAGCAGAGGACCAGCTAAGAGGGAGAGAAGCAACTACAGACCCCCCCTGAAAACAACCCTCAGACGCCACATCCCCTGACAAGCTGCCAGGCAGGTTCTCTT
```

The output shows that the human TNF gene with accession number NC_00006, region: 31575565..31578336 is 2722 bp long.

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

The DNA sequence of the TNF gene was translated into an amino acid sequence. Here is a screenshot of the amino acid sequence generated:

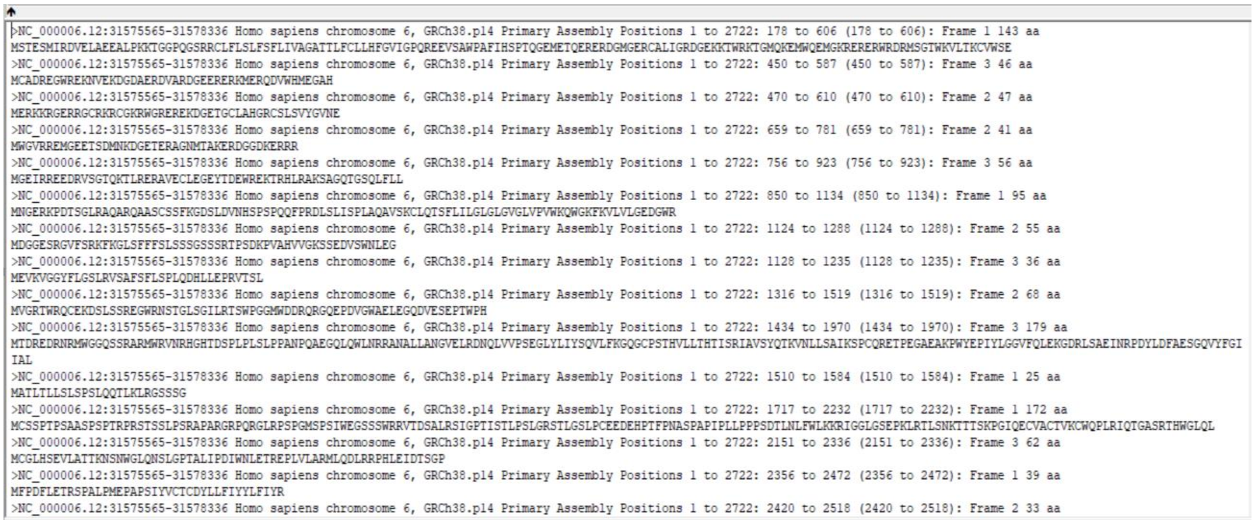

BioEdit Sequence Alignment Editor - [Translation from C:\Users\racha\Downloads\TNF sequence.fasta: Frame 1]
File Edit Format Accesory Application RNA Options Window Help
Courier New
9
Add Bacterial
Add Viral
Add custom
Add simple
Top 1

```
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly

1       AGC AGA CGC TCC CTC AGC AAG GAC AGC AGA GGA CCA GCT AAG AGG    45
1       Ser Arg Arg Ser Leu Ser Lys Asp Ser Arg Gly Pro Ala Lys Arg    15

46      GAG AGA AGC AAC TAC AGA CCC CCC CTG AAA ACA ACC CTC AGA CGC    90
16      Glu Arg Ser Asn Tyr Arg Pro Pro Leu Lys Thr Thr Leu Arg Arg    30

91      CAC ATC CCC TGA CAA GCT GCC AGG CAG GTT CTC TTC CTC TCA CAT    135
31      His Ile Pro End Gln Ala Ala Arg Gln Val Leu Phe Leu Ser His    45

136     ACT GAC CCA CGG CTC CAC CCT CTC TCC CCT GGA AAG GAC ACC ATG    180
46      Thr Asp Pro Arg Leu His Pro Leu Ser Pro Gly Lys Asp Thr Met    60

181     AGC ACT GAA AGC ATG ATC CGG GAC GTG GAG CTG GCC GAG GAG GCG    225
61      Ser Thr Glu Ser Met Ile Arg Asp Val Glu Leu Ala Glu Glu Ala    75

226     CTC CCC AAG AAG ACA GGG GGG CCC CAG GGC TCC AGG CGG TGC TTG    270
76      Leu Pro Lys Lys Thr Gly Gly Pro Gln Gly Ser Arg Arg Cys Leu    90

271     TTC CTC AGC CTC TTC TCC TTC CTG ATC GTG GCA GGC GCC ACC ACG    315
91      Phe Leu Ser Leu Phe Ser Phe Leu Ile Val Ala Gly Ala Thr Thr    105

316     CTC TTC TGC CTG CTG CAC TTT GGA GTG ATC GGC CCC CAG AGG GAA    360
106     Leu Phe Cys Leu Leu His Phe Gly Val Ile Gly Pro Gln Arg Glu    120

361     GAG GTG AGT GCC TGG CCA GCC TTC ATC CAC TCT CCC ACC CAA GGG    405
121     Glu Val Ser Ala Trp Pro Ala Phe Ile His Ser Pro Thr Gln Gly    135

406     GAA ATG GAG ACG CAA GAG AGG GAG AGA GAT GGG ATG GGT GAA AGA    450
136     Glu Met Glu Thr Gln Glu Arg Glu Arg Asp Gly Met Gly Glu Arg    150
```

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

Start and stop positions, lengths, and protein translations of the ORFs.

Output:



```
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 178 to 606 (178 to 606): Frame 1 143 aa
MSTESMIRDVELAEEALPKKTGGPQGSRRCLFLSLFSFLIVAGATTLFCLLHFGVIGPQREEVSAWPAFIHSPTQGEMETQERERDGMGERCALIGRDGEKKTWRKTGMQKEMWQEMGKRERERWRDRMSGTWKVLTKCVWSE
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 450 to 587 (450 to 587): Frame 3 46 aa
MCADREGWREKNVEKDGDAERDVARDGEERERKMERQDVWHMEGAH
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 470 to 610 (470 to 610): Frame 2 47 aa
MERKKRGERRGCRKRCGKRWGREREKDGETGCLAHGRCSLSVYGVNE
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 659 to 781 (659 to 781): Frame 2 41 aa
MWGVRREMGEETSDMNKDGETERAGNMTAKERDGGDKERRR
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 756 to 923 (756 to 923): Frame 3 56 aa
MGEIRREEDRVSGTQKTLRERAVECLEGEYTDEWREKTRHLRAKSAGQTGSQLFLL
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 850 to 1134 (850 to 1134): Frame 1 95 aa
MNGERKPDTSGLRAQARQAASCSSFKGDSLDVNHSPSPQQFPRDLSLISPLAQAVSKCLQTSFLILGLGLGVGLVPVWKQWGKFKVLVLGEDGWR
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 1124 to 1288 (1124 to 1288): Frame 2 55 aa
MDGGESRGVFSRKFKGLSFFFSLSSSGSSSRTPSDKPVAHVVGKSSEDVSWNLEG
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 1128 to 1235 (1128 to 1235): Frame 3 36 aa
MEVKVGGYFLGSLRVSAFSFLSPLQDHLLEPRVTSL
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 1316 to 1519 (1316 to 1519): Frame 2 68 aa
MVGRTWRQCEKDSLSSREGWRNSTGLSGILRTSWPGGMWDDRQRGQEPDVGWAELEGQDVESEPTWPH
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 1434 to 1970 (1434 to 1970): Frame 3 179 aa
MTDREDRNRMWGGQSSRARMWRVNRHGHTDSPLPLSLPPANPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYSQVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRETPEGAEAKPWYEPIYLGGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGI
IAL
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 1510 to 1584 (1510 to 1584): Frame 1 25 aa
MATLTLLSLSPSLQQTLKLRGSSSG
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 1717 to 2232 (1717 to 2232): Frame 1 172 aa
MCSSPTPSAASPSPTRPRSTSSLPSRAPARGRPQRGLRPSPGMSPSIWEGSSSWRRVTDSALRSIGPTISTLPSLGRSTLGSLPCEEDEHPTFPNASPAPIPLLPPPSDTLNLFWLKKRIGGLGSEPKLRTLSNKTTTSKPGIQECVACTVKCWQPLRIQTGASRTHWGLQL
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 2151 to 2336 (2151 to 2336): Frame 3 62 aa
MCGLHSEVLATTKNSNWGLQNSLGPTALIPDIWNLETREPLVLARMLQDLRRPHLEIDTSGP
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 2356 to 2472 (2356 to 2472): Frame 1 39 aa
MFPDFLETRSPALPMEPAPSIYVCTCDYLLFIYYLFIYR
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 2420 to 2518 (2420 to 2518): Frame 2 33 aa
```

BioEdit searches the TNF sequence for open reading frames (ORFs) in 6 reading frames (3 in either direction). ORFs are sequences that contain start codons and stop codons with a minimum distance between them. BioEdit returns the range of each ORF, along with its protein translation. Each ORFs may potentially be protein coding regions. Protein-coding regions are usually of a significant length and contain multiple amino acids. Short sequences or those that do not align well with known proteins might be non-coding. Potential protein-coding regions from the output may include:

>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 1434 to 1970 (1434 to 1970): Frame 3 179 aa

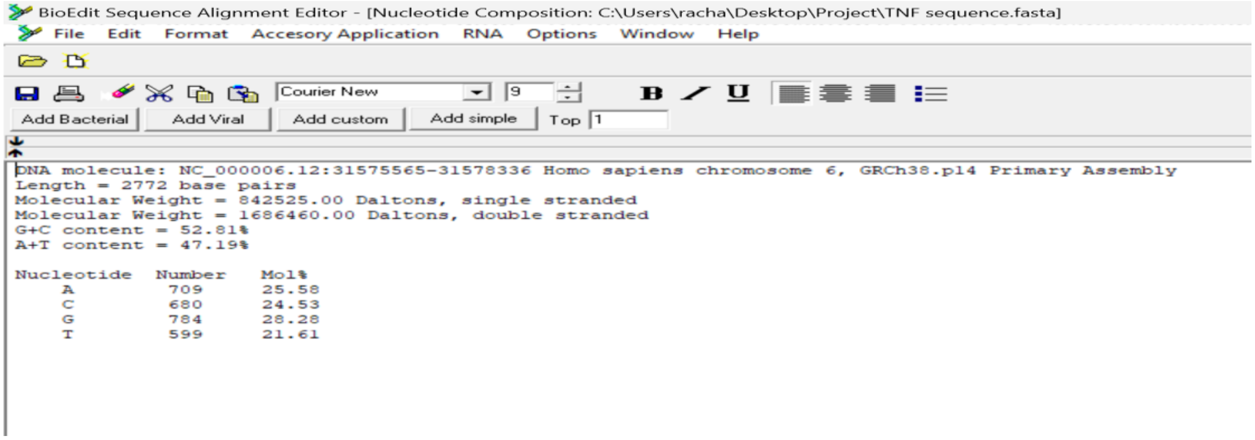
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 1717 to 2232 (1717 to 2232): Frame 1 172 aa

>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2722: 178 to 606 (178 to 606): Frame 1 143 aa

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

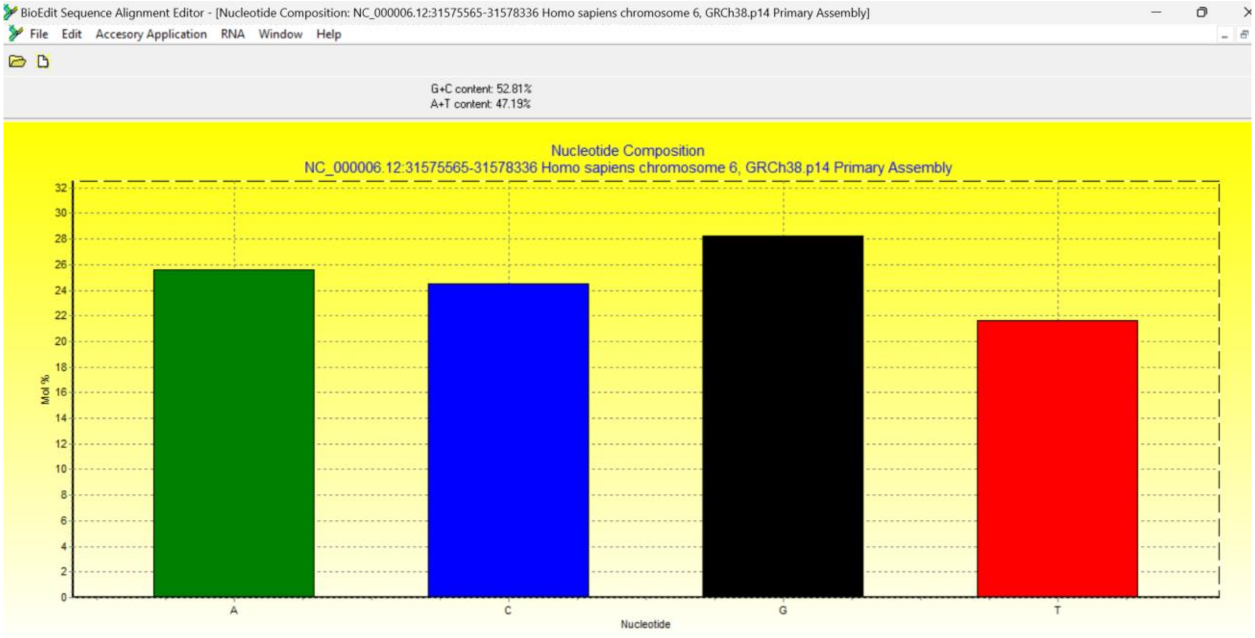
The nucleotide composition of the TNF gene sequence was analyzed:
The output shows that the Human TNF gene sequence (2722bp) contains 25.58%, 24.53%, 28.28%, 21.61% of Nucleotide A, C, G, T respectively.
The sequence has a high overall G+C content of 52.81%. A higher GC content often correlates with higher DNA stability due to the three hydrogen bonds between G and C compared to the two hydrogen bonds between adenine (A) and thymine (T).
This information is useful for various applications, such as understanding the genetic content of the region, designing primers for PCR, or studying the DNA's structural properties.

Output:


BioEdit Sequence Alignment Editor - [Nucleotide Composition: C:\Users\racha\Desktop\Project\TNF sequence.fasta]
File Edit Format Accesory Application RNA Options Window Help
Courier New
9
Add Bacterial
Add Viral
Add custom
Add simple
Top 1

```
DNA molecule: NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly
Length = 2772 base pairs
Molecular Weight = 842525.00 Daltons, single stranded
Molecular Weight = 1686460.00 Daltons, double stranded
G+C content = 52.81%
A+T content = 47.19%

Nucleotide   Number    Mol%
    A         709      25.58
    C         680      24.53
    G         784      28.28
    T         599      21.61
```

Here is also a graphical illustration of the TNF Nucleotide composition

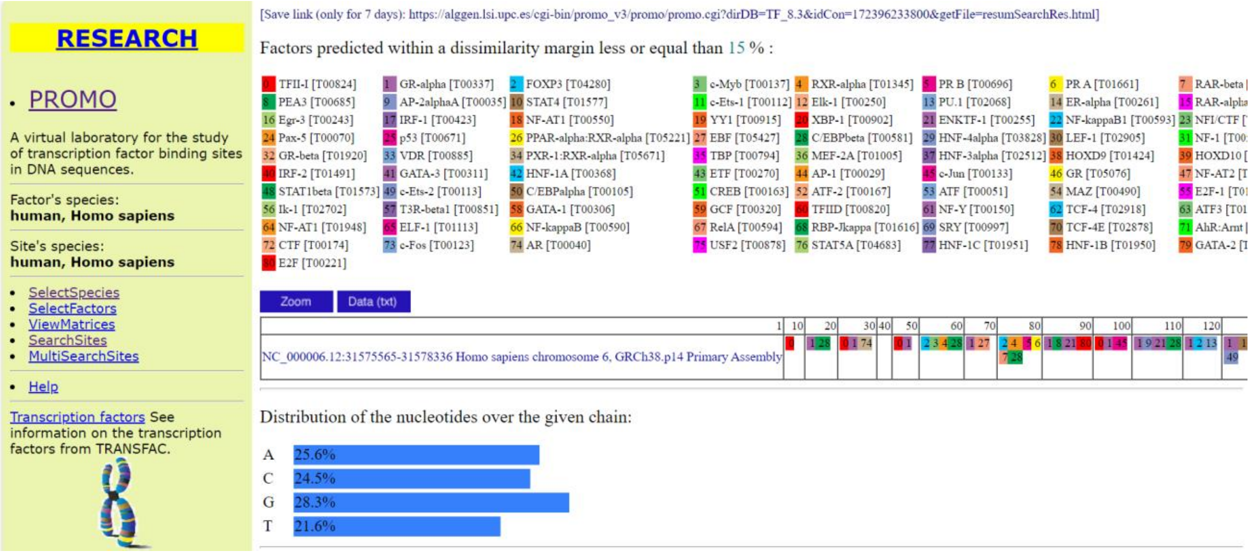


G+C content: 52.81%
A+T content: 47.19%

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

The potential transcription factor binding sites in the TNF gene promoter region were Identified.

Output:



RESEARCH

- PROMO

A virtual laboratory for the study of transcription factor binding sites in DNA sequences.

Factor's species:
human, Homo sapiens

Site's species:
human, Homo sapiens

- SelectSpecies
- SelectFactors
- ViewMatrices
- SearchSites
- MultiSearchSites

- Help

Transcription factors See information on the transcription factors from TRANSFAC.

[Save link (only for 7 days): https://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promo.cgi?dirDB=TF_8.3&idCon=172396233800&getFile=resumSearchRes.html]

Factors predicted within a dissimilarity margin less or equal than 15 % :

0 TFII-I [T00824]	1 GR-alpha [T00337]	2 FOXP3 [T04280]	3 c-Myb [T00137]	4 RXR-alpha [T01345]	5 PR B [T00696]	6 PR A [T01661]	7 RAR-beta [
8 PEA3 [T00685]	9 AP-2alphaA [T00035]	10 STAT4 [T01577]	11 c-Ets-1 [T00112]	12 Elk-1 [T00250]	13 PU.1 [T02068]	14 ER-alpha [T00261]	15 RAR-alpha
16 Egr-3 [T00243]	17 IRF-1 [T00423]	18 NF-AT1 [T00550]	19 YY1 [T00915]	20 XBP-1 [T00902]	21 ENKTF-1 [T00255]	22 NF-kappaB1 [T00593]	23 NFI/CTF [
24 Pax-5 [T00070]	25 p53 [T00671]	26 PPAR-alpha:RXR-alpha [T05221]	27 EBF [T05427]	28 C/EBPbeta [T00581]	29 HNF-4alpha [T03828]	30 LEF-1 [T02905]	31 NF-1 [T00
32 GR-beta [T01920]	33 VDR [T00885]	34 PXR-1:RXR-alpha [T05671]	35 TBP [T00794]	36 MEF-2A [T01005]	37 HNF-3alpha [T02512]	38 HOXD9 [T01424]	39 HOXD10 [
40 IRF-2 [T01491]	41 GATA-3 [T00311]	42 HNF-1A [T00368]	43 ETF [T00270]	44 AP-1 [T00029]	45 c-Jun [T00133]	46 GR [T05076]	47 NF-AT2 [T
48 STAT1beta [T01573]	49 c-Ets-2 [T00113]	50 C/EBPalpha [T00105]	51 CREB [T00163]	52 ATF-2 [T00167]	53 ATF [T00051]	54 MAZ [T00490]	55 E2F-1 [T0
56 Ik-1 [T02702]	57 T3R-beta1 [T00851]	58 GATA-1 [T00306]	59 GCF [T00320]	60 TFIID [T00820]	61 NF-Y [T00150]	62 TCF-4 [T02918]	63 ATF3 [T01
64 NF-AT1 [T01948]	65 ELF-1 [T01113]	66 NF-kappaB [T00590]	67 RelA [T00594]	68 RBP-Jkappa [T01616]	69 SRY [T00997]	70 TCF-4E [T02878]	71 AhR:Arnt
72 CTF [T00174]	73 c-Fos [T00123]	74 AR [T00040]	75 USF2 [T00878]	76 STAT5A [T04683]	77 HNF-1C [T01951]	78 HNF-1B [T01950]	79 GATA-2 [T
80 E2F [T00221]							

	1	10	20	30	40	50	60	70	80	90	100	110	120	
NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly		0	1 28	0 1 74		0 1	2 3 4 28	1 27	2 4 5 6 7 28	1 8 21 80	0 1 45	1 9 21 28	1 2 13	1 1 49

Distribution of the nucleotides over the given chain:

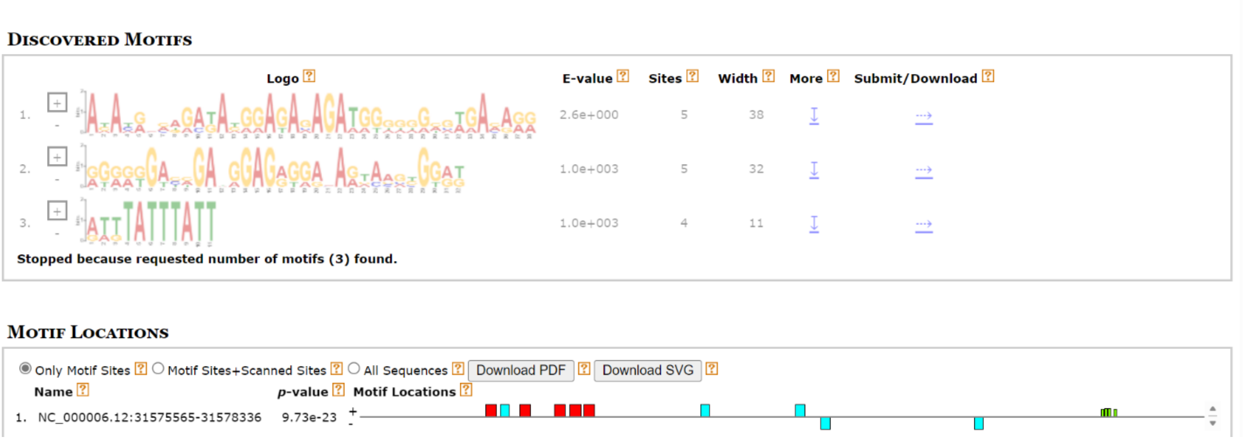
A	25.6%
C	24.5%
G	28.3%
T	21.6%

The output shows eighty (80) potential Transcription factors within a dissimilarity margin less or equal than 15%. These Transcription factors play a critical role in regulating the TNF gene expression by binding to specific sequence in the Nucleotide and either activating or repressing transcription.

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

Functional motifs in the TNF gene sequence were identified using MEME Suite.

Output:



DISCOVERED MOTIFS

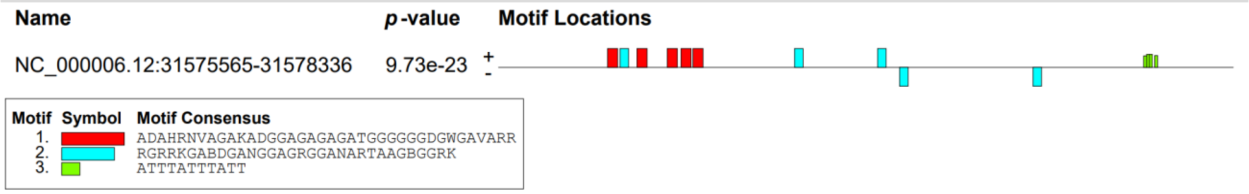
	Logo	E-value	Sites	Width	More	Submit/Download
1.		2.6e+000	5	38		
2.		1.0e+003	5	32		
3.		1.0e+003	4	11		

Stopped because requested number of motifs (3) found.

MOTIF LOCATIONS

● Only Motif Sites ○ Motif Sites+Scanned Sites ○ All Sequences

	Name	p-value	Motif Locations
1.	NC_000006.12:31575565-31578336	9.73e-23	



Name	p-value	Motif Locations
NC_000006.12:31575565-31578336	9.73e-23	

Motif	Symbol	Motif Consensus
1.	🟥	ADAHRNVAGAKADGGAGAGAGATGGGGGGDGWGAVARR
2.	🟦	RGRRKGABDGANGGAGRGGANARTAAGBGGRK
3.	🟩	ATTTATTTATT

The output shows three consensus motifs spanning the human TNF gene sequence.

Sequence motifs are short recurring patterns in DNA that are presumed to have a biological function. The above screenshot indicates which nucleotides or amino acids are most frequently observed at each position in the alignment. This is shown in a logo format, where each position shows the most common base or amino acid.

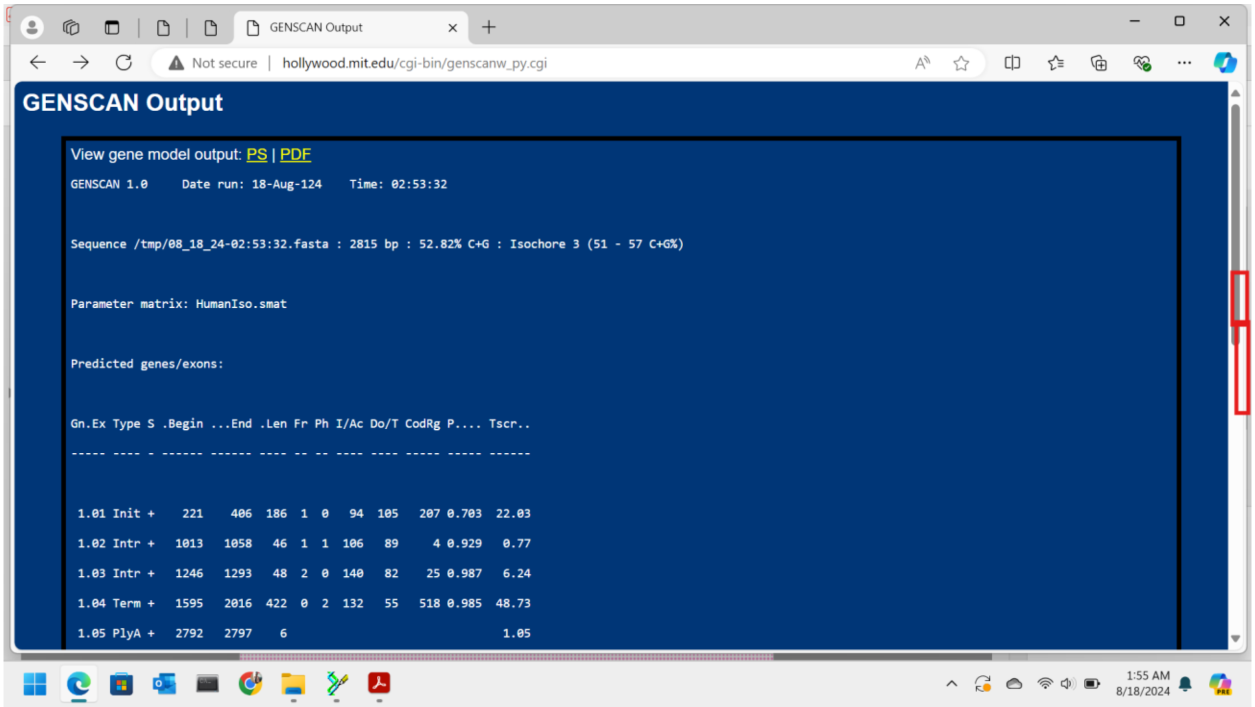
The overall p-value of 9.73e-23 of the motifs indicates that the motifs are highly statistically significant, suggesting a very low probability that the observed pattern is due to chance. In practical terms, this p-value provides strong evidence that the motif is likely to be biologically relevant.

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

The coding and non-coding regions within the TNF gene sequence were predicted using GENSCAN.

Output:

Here is a screenshot of the output:



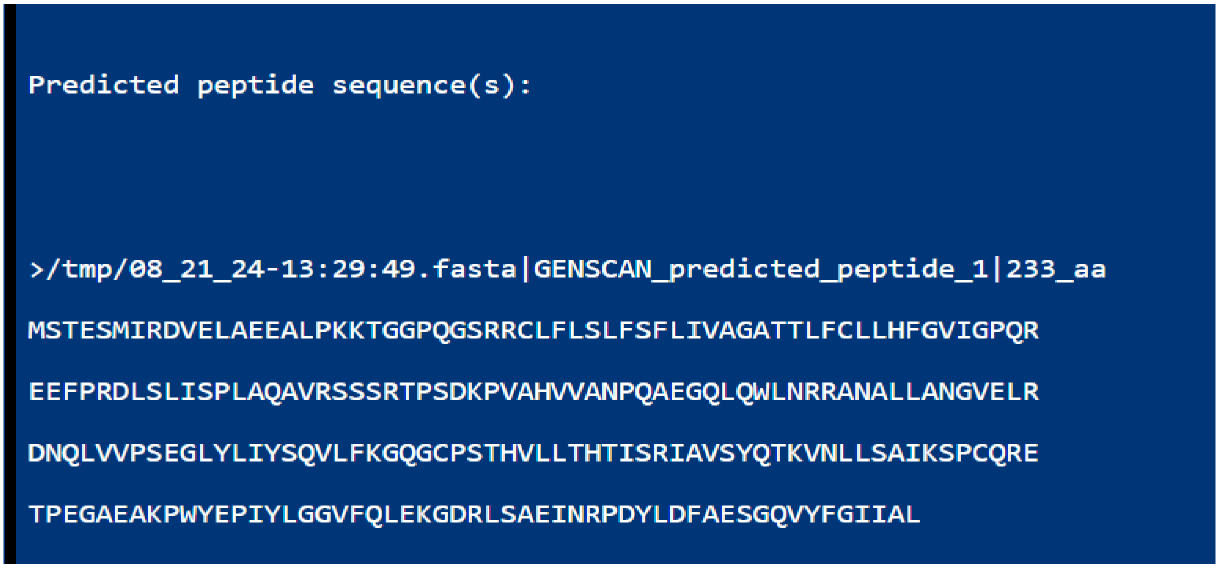
GENSCAN Output

View gene model output: PS | PDF

```
GENSCAN 1.0    Date run: 18-Aug-124    Time: 02:53:32


Sequence /tmp/08_18_24-02:53:32.fasta : 2815 bp : 52.82% C+G : Isochore 3 (51 - 57 C+G%)


Parameter matrix: HumanIso.smat


Predicted genes/exons:


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


 1.01 Init +    221    406  186  1  0   94  105   207 0.703  22.03
 1.02 Intr +   1013   1058   46  1  1  106   89     4 0.929   0.77
 1.03 Intr +   1246   1293   48  2  0  140   82    25 0.987   6.24
 1.04 Term +   1595   2016  422  0  2  132   55   518 0.985  48.73
 1.05 PlyA +   2792   2797    6                                1.05
```

Coding regions are parts of the DNA sequence that contain the instructions for building proteins, while non-coding regions do not contain these instructions.
GENSCAN output helps in identifying and mapping coding regions (exons) and non-coding regions (introns and intergenic regions) in a DNA sequence. It provides important information for gene prediction and annotation, which is crucial for understanding gene structure and function.

GENSCAN has predicted five exons on the positive strand on the gene. One initial exon, two internal exons, a terminal exon and a polyadenylation signal within the TNF DNA sequence. The confidence levels for these predictions vary, with some predictions showing high probability and others more moderate indicating strong confidence in the gene predictions. The output also includes a predicted peptide sequence 233 aa in length derived from the gene model which includes various motifs and domains that may be functionally significant, which is essential for functional annotation and further analysis.



```
Predicted peptide sequence(s):


>/tmp/08_21_24-13:29:49.fasta|GENSCAN_predicted_peptide_1|233_aa

MSTESMIRDVELAEEALPKKTGGPQGSRRCLFLSLFSFLIVAGATTLFCLLHFGVIGPQR

EEFPRDLSLISPLAQAVRSSSRTPSDKPVAHVVANPQAEGQLQWLNRRANALLANGVELR

DNQLVVPSEGLYLIYSQVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRE

TPEGAEAKPWYEPIYLGGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL
```

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

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

Output

The Fasta and Phylip format of the sequence is attached to this report.