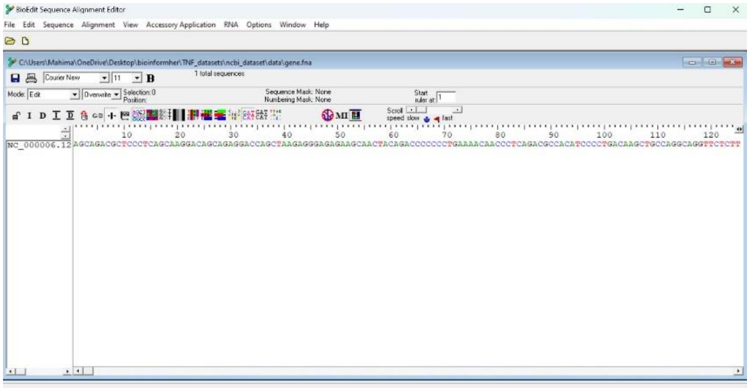
Comprehensive Analysis of the Human Tumor Necrosis Factor (TNF) Gene Sequence

Objective: This project aims to perform a comprehensive analysis of the Human Tumor Necrosis Factor (TNF) gene sequence using various bioinformatics tools such as NCBI, BioEdit, PROMO, GENSCAN, and MEME Suite. This mini-project is the first task of the BioinformHER Module 1. BioinformHER is an initiative taken by HackBio aimed at introducing young women in STEM to Bioinformatics using simple and easy-to-follow guidelines.

Tasks and interpretations

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

1. Accessed the NCBI homepage at NCBI
2. Searched for the human TNF gene using the term 'human TNF gene.'
3. Located the correct sequence record (e.g., 'Homo sapiens TNF')
4. There were multiple sequences available for the human TNF gene. Therefore, I selected the most recent sequence [NC_000006.12 (31575565-31578336)]
5. Downloaded the sequence in FASTA format
6. Opened the sequence in BioEdit and viewed it.


BioEdit Sequence Alignment Editor
File Edit Sequence Alignment View Accessory Application RNA Options Window Help
Courier New
11
B
Scroll speed slow fast

C:\Users\Mahima\OneDrive\Desktop\bioinformher\TNF_datasets\ncbi_dataset\data\gene.fna

1 total sequences

Mode: Edit, Overwrite, Selection: 0, Position:

Sequence Mask: None
Numbering Mask: None

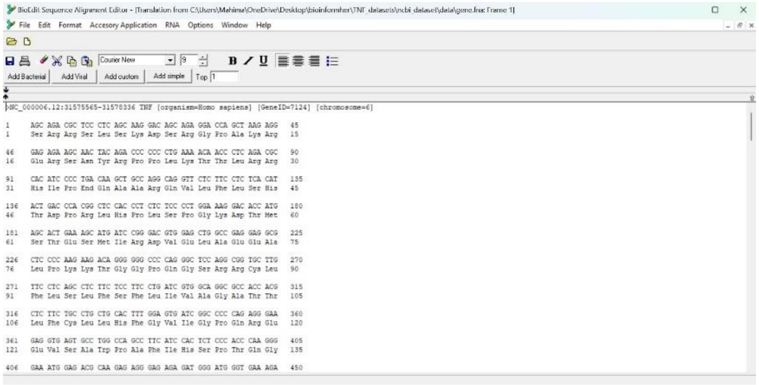
Start ruler at: 1

```
                     10        20        30        40        50        60        70        80        90       100       110       120
NC_000006.12 AGCAGACGCTCCCTCAGCAAGGACAGCAGAGGACCAGCTAAGAGGGAGAGAAGCAACTACAGACCCCCCCTGAAAACAACCCTCAGACGCCACATCCCCTGACAAGCTGCCAGGCAGGTTCTCTT
```

Task 1 Output

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

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

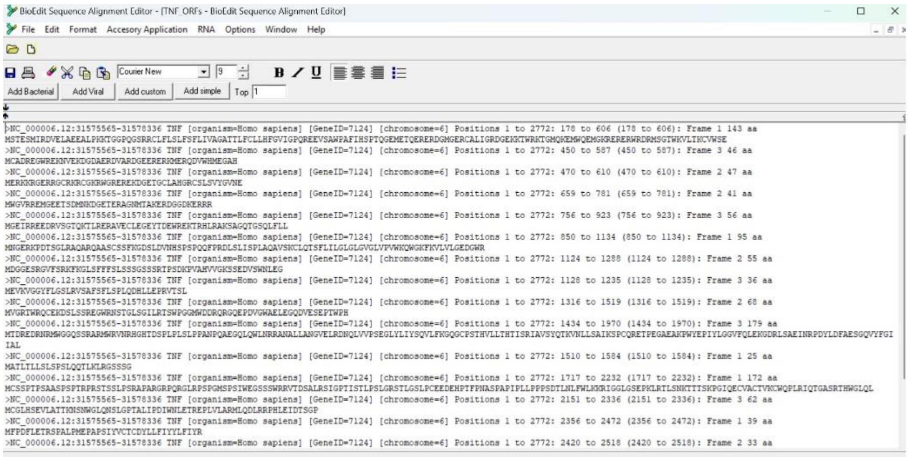

BioEdit Sequence Alignment Editor - [Translation from C:\Users\Mahima\OneDrive\Desktop\bioinformher\TNF_datasets\ncbi_dataset\data\gene.fna: Frame 1]
File Edit Format Accesory Application RNA Options Window Help
Courier New
9
B / U
Add Bacterial
Add Viral
Add custom
Add simple
Top 1

```
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6]

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
```

Task 2 output

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

1. Used BioEdit's ORF Finder tool to find ORFs in the TNF gene sequence
2. Recorded the start and stop positions, lengths, and protein translations of the ORFs
3. Selected the longest ORF and analyzed it using SMART BLAST to determine whether it encodes the Human Tumor Necrosis Factor (TNF)



```
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 178 to 606 (178 to 606): Frame 1 143 aa
MSTESMIRDVELAEEALPKKTGGPQGSRRCLFLSLFSFLIVAGATTLFCLLHFGVIGPQREEVSAWPAFIHSPTQGEMETQERERDGMGERCALIGRDGEKKTWRKTGMQKEMWQEMGKRERERWRDRMSGTWKVLTKCVWSE
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 450 to 587 (450 to 587): Frame 3 46 aa
MCADREGWREKNVEKDGDAERDVARDGEERERKMERQDVWHMEGAH
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 470 to 610 (470 to 610): Frame 2 47 aa
MERKKRGERRGCRKRCGKRWGREREKDGETGCLAHGRCSLSVYGVNE
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 659 to 781 (659 to 781): Frame 2 41 aa
MWGVRREMGEETSDMNKDGETERAGNMTAKERDGGDKERRR
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 756 to 923 (756 to 923): Frame 3 56 aa
MGEIRREEDRVSGTQKTLRERAVECLEGEYTDEWREKTRHLRAKSAGQTGSQLFLL
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 850 to 1134 (850 to 1134): Frame 1 95 aa
MNGERKPDTSGLRAQARQAASCSSFKGDSLDVNHSPSPQQFPRDLSLISPLAQAVSKCLQTSFLILGLGLGVGLVPVWKQWGKFKVLVLGEDGWR
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 1124 to 1288 (1124 to 1288): Frame 2 55 aa
MDGGESRGVFSRKFKGLSFFFSLSSSGSSSRTPSDKPVAHVVGKSSEDVSWNLEG
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 1128 to 1235 (1128 to 1235): Frame 3 36 aa
MEVKVGGYFLGSLRVSAFSFLSPLQDHLLEPRVTSL
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 1316 to 1519 (1316 to 1519): Frame 2 68 aa
MVGRTWRQCEKDSLSSREGWRNSTGLSGILRTSWPGGMWDDRQRGQEPDVGWAELEGQDVESEPTWPH
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 1434 to 1970 (1434 to 1970): Frame 3 179 aa
MTDREDRNRMWGGQSSRARMWRVNRHGHTDSPLPLSLPPANPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYSQVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRETPEGAEAKPWYEPIYLGGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGI
IAL
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 1510 to 1584 (1510 to 1584): Frame 1 25 aa
MATLTLLSLSPSLQQTLKLRGSSSG
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 1717 to 2232 (1717 to 2232): Frame 1 172 aa
MCSSPTPSAASPSPTRPRSTSSLPSRAPARGRPQRGLRPSPGMSPSIWEGSSSWRRVTDSALRSIGPTISTLPSLGRSTLGSLPCEEDEHPTFPNASPAPIPLLPPPSDTLNLFWLKKRIGGLGSEPKLRTLSNKTTTSKPGIQECVACTVKCWQPLRIQTGASRTHWGLQL
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 2151 to 2336 (2151 to 2336): Frame 3 62 aa
MCGLHSEVLATTKNSNWGLQNSLGPTALIPDIWNLETREPLVLARMLQDLRRPHLEIDTSGP
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 2356 to 2472 (2356 to 2472): Frame 1 39 aa
MFPDFLETRSPALPMEPAPSIYVCTCDYLLFIYYLFIYR
>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 2420 to 2518 (2420 to 2518): Frame 2 33 aa
```

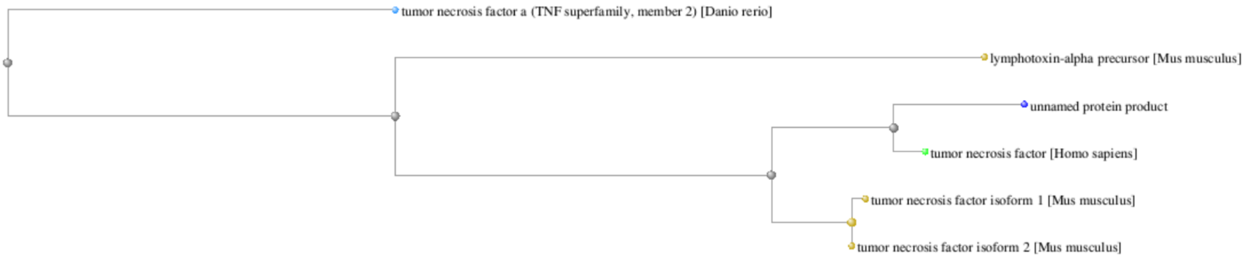
Task 3 output

Interpretation: I analyzed the longest ORF using NCBI's SMART BLAST to determine the protein-encoding region of the TNF gene. The ORF with a length of 179 residues showed similarity to TNF (*Homo sapiens*) indicating that it is the protein-encoding region of the TNF gene.

>NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6] Positions 1 to 2772: 1434 to 1970 (1434 to 1970): Frame 3 179 aa

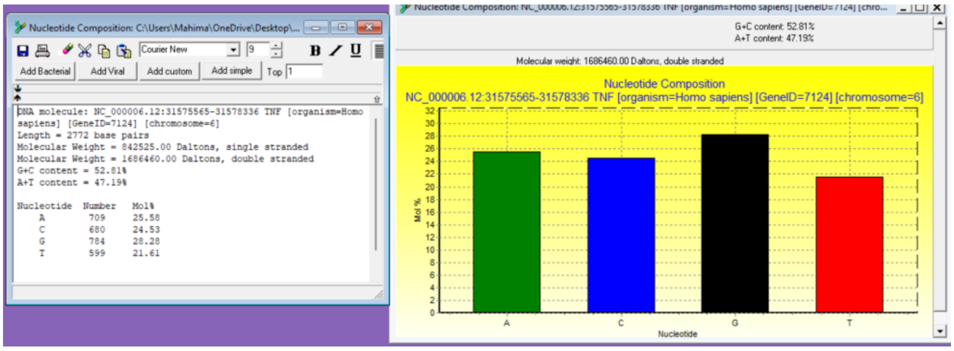
MTDREDRNRMWGGQSSRARMWRVNRHGHTDSPLPLSLPPANPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYSQVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRETPEGAEAKPWYEPIYLGGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL

The phylogenetic tree generated for the protein product is as follows:


tumor necrosis factor a (TNF superfamily, member 2) [Danio rerio]
lymphotoxin-alpha precursor [Mus musculus]
unnamed protein product
tumor necrosis factor [Homo sapiens]
tumor necrosis factor isoform 1 [Mus musculus]
tumor necrosis factor isoform 2 [Mus musculus]

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

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


Nucleotide Composition: C:\Users\Mahima\OneDrive\Desktop\...
Courier New
9
Add Bacterial
Add Viral
Add custom
Add simple
Top
1
Nucleotide Composition: NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chro...
Nucleotide Composition
NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6]
Mol %
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
A
C
G
T
Nucleotide

```
DNA molecule: NC_000006.12:31575565-31578336 TNF [organism=Homo
sapiens] [GeneID=7124] [chromosome=6]
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
```

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

Molecular weight: 1686460.00 Daltons, double stranded

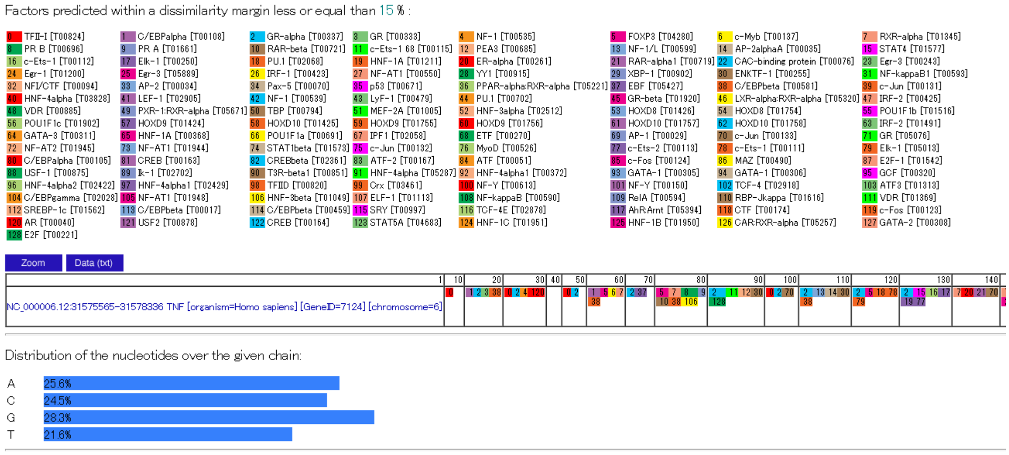
Task 4 output

Interpretation: The sequence has a GC content of 52.81% which is higher than the AT content indicating that it is stable because of stronger H-bonding between G and C.

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

1. Accessed the PROMO tool at https://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3
2. Selected 'Homo sapiens' as the species
3. Input the entire TNF gene sequence
4. Identified potential transcription factor binding sites

PROMO is a virtual laboratory for identifying putative transcription factor binding sites (TFBS) in DNA sequences from a species or groups of species of interest.


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

0 TFII-I [T00824]	1 C/EBPalpha [T00108]	2 GR-alpha [T00337]	3 GR [T00333]	4 NF-1 [T00535]	5 FOXP3 [T04280]	6 c-Myb [T00137]	7 RXR-alpha [T01345]
8 PR B [T00696]	9 PR A [T01661]	10 RAR-beta [T00721]	11 c-Ets-1 68 [T00115]	12 PEA3 [T00685]	13 NF-1/L [T00599]	14 AP-2alphaA [T00035]	15 STAT4 [T01577]
16 c-Ets-1 [T00112]	17 Elk-1 [T00250]	18 PU.1 [T02068]	19 HNF-1A [T01211]	20 ER-alpha [T00261]	21 RAR-alpha1 [T00719]	22 CAC-binding protein [T00076]	23 Egr-3 [T00243]
24 Egr-1 [T01200]	25 Egr-3 [T05889]	26 IRF-1 [T00423]	27 NF-AT1 [T00550]	28 YY1 [T00915]	29 XBP-1 [T00902]	30 ENKTF-1 [T00255]	31 NF-kappaB1 [T00593]
32 NFI/CTF [T00094]	33 AP-2 [T00034]	34 Pax-5 [T00070]	35 p53 [T00671]	36 PPAR-alpha:RXR-alpha [T05221]	37 EBF [T05427]	38 C/EBPbeta [T00581]	39 c-Jun [T00131]
40 HNF-4alpha [T03828]	41 LEF-1 [T02905]	42 NF-1 [T00539]	43 LyF-1 [T00479]	44 PU.1 [T00702]	45 GR-beta [T01920]	46 LXR-alpha:RXR-alpha [T05320]	47 IRF-2 [T00425]
48 VDR [T00885]	49 PXR-1:RXR-alpha [T05671]	50 TBP [T00794]	51 MEF-2A [T01005]	52 HNF-3alpha [T02512]	53 HOXD8 [T01426]	54 HOXD8 [T01754]	55 POU1F1b [T01516]
56 POU1F1c [T01902]	57 HOXD9 [T01424]	58 HOXD10 [T01425]	59 HOXD9 [T01755]	60 HOXD9 [T01756]	61 HOXD10 [T01757]	62 HOXD10 [T01758]	63 IRF-2 [T01491]
64 GATA-3 [T00311]	65 HNF-1A [T00368]	66 POU1F1a [T00691]	67 IPF1 [T02058]	68 ETF [T00270]	69 AP-1 [T00029]	70 c-Jun [T00133]	71 GR [T05076]
72 NF-AT2 [T01945]	73 NF-AT1 [T01944]	74 STAT1beta [T01573]	75 c-Jun [T00132]	76 MyoD [T00526]	77 c-Ets-2 [T00113]	78 c-Ets-1 [T00111]	79 Elk-1 [T05013]
80 C/EBPalpha [T00105]	81 CREB [T00163]	82 CREBbeta [T02361]	83 ATF-2 [T00167]	84 ATF [T00051]	85 c-Fos [T00124]	86 MAZ [T00490]	87 E2F-1 [T01542]
88 USF-1 [T00875]	89 Ik-1 [T02702]	90 T3R-beta1 [T00851]	91 HNF-4alpha [T05287]	92 HNF-4alpha1 [T00372]	93 GATA-1 [T00305]	94 GATA-1 [T00306]	95 GCF [T00320]
96 HNF-4alpha2 [T02422]	97 HNF-4alpha1 [T02429]	98 TFIID [T00820]	99 Crx [T03461]	100 NF-Y [T00613]	101 NF-Y [T00150]	102 TCF-4 [T02918]	103 ATF3 [T01313]
104 C/EBPgamma [T02028]	105 NF-AT1 [T01948]	106 HNF-3beta [T01049]	107 ELF-1 [T01113]	108 NF-kappaB [T00590]	109 RelA [T00594]	110 RBP-Jkappa [T01616]	111 VDR [T01369]
112 SREBP-1c [T01562]	113 C/EBPbeta [T00017]	114 C/EBPbeta [T00459]	115 SRY [T00997]	116 TCF-4E [T02878]	117 AhR:Arnt [T05394]	118 CTF [T00174]	119 c-Fos [T00123]
120 AR [T00040]	121 USF2 [T00878]	122 CREB [T00164]	123 STAT5A [T04683]	124 HNF-1C [T01951]	125 HNF-1B [T01950]	126 CAR:RXR-alpha [T05257]	127 GATA-2 [T00308]
128 E2F [T00221]							

Zoom
Data (txt)
NC_000006.12:31575565-31578336 TNF [organism=Homo sapiens] [GeneID=7124] [chromosome=6]
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140

Distribution of the nucleotides over the given chain:

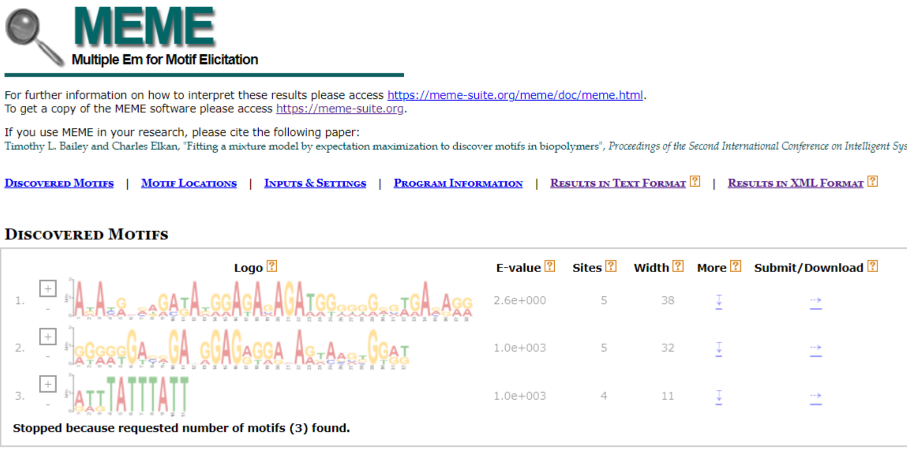
A 25.6%
C 24.5%
G 28.3%
T 21.6%

Task 5 output

Several transcription factors can bind to the promoter region of the Human TNF gene and initiate its transcription at different locations in the gene.

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

1. Accessed the MEME Suite at https://meme-suite.org/meme/
2. Uploaded the TNF gene sequence in FASTA format
3. Used the default settings to search for motifs
4. Interpreted and saved the results of the motif search



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 Sys*

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.		2.6e+000	5	38	↧	⇢
2.		1.0e+003	5	32	↧	⇢
3.		1.0e+003	4	11	↧	⇢

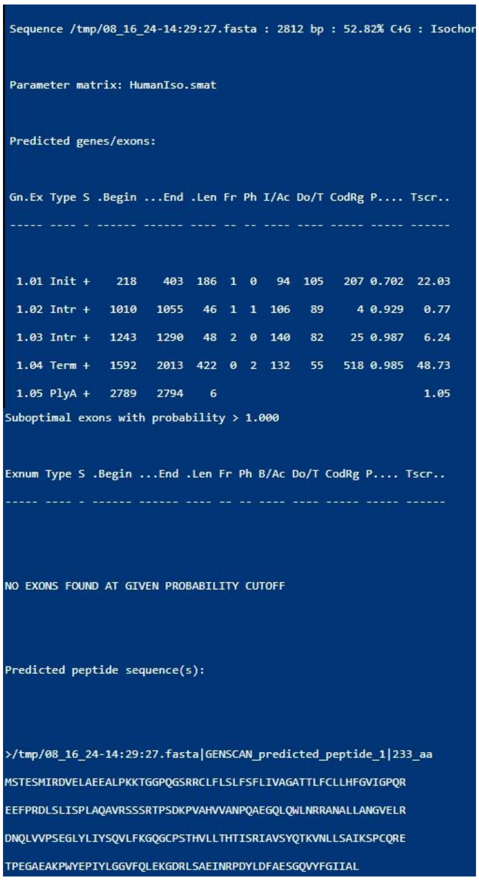
Stopped because requested number of motifs (3) found.

Task 6 output

Interpretation: MEME (Multiple Em for Motif Elicitation) found the lowest E-value motifs with the highest statistical significance as shown above. Motif 1 has a width of 38 and is found at 5 different sites in the sequence with an E-value of 2.6e+000 which is the lowest, suggesting its significance.

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

1. Accessed the GENSCAN tool at http://hollywood.mit.edu/GENSCAN.html
2. Input the TNF gene sequence in the appropriate format
3. Ran the analysis to predict coding and non-coding regions
4. Saved and interpreted the results.



```
Sequence /tmp/08_16_24-14:29:27.fasta : 2812 bp : 52.82% C+G : Isochor

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 +    218    403  186  1  0   94  105   207 0.702  22.03
 1.02 Intr +   1010   1055   46  1  1  106   89     4 0.929   0.77
 1.03 Intr +   1243   1290   48  2  0  140   82    25 0.987   6.24
 1.04 Term +   1592   2013  422  0  2  132   55   518 0.985  48.73
 1.05 PlyA +   2789   2794    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

Predicted peptide sequence(s):

>/tmp/08_16_24-14:29:27.fasta|GENSCAN_predicted_peptide_1|233_aa
MSTESMIRDVELAEEALPKKTGGPQGSRRCLFLSLFSFLIVAGATTLFCLLHFGVIGPQR
EEFPRDLSLISPLAQAVRSSSRTPSDKPVAHVVANPQAEGQLQWLNRRANALLANGVELR
DNQLVVPSEGLYLIYSQVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRE
TPEGAEAKPWYEPIYLGGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL
```

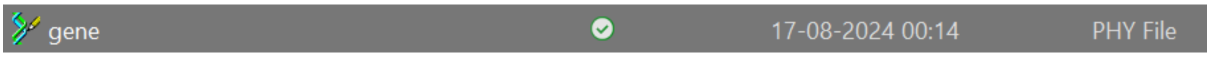
Task 7 output

Interpretation: GENSCAN predicted 5 exons/coding regions in the positive strand of the gene sequence.

- Exon 1 is an initiator exon that marks the beginning of the coding sequence with the start codon. It starts from residue 218 and ends at residue 403 with a length of 186 residues.
- Exon 2 and Exon 3 are internal exons
- Exon 4 is a terminal exon that includes the stop codon
- Exon 5 is the poly-A signal. It is a sequence in the 3' untranslated region (UTR) of the mRNA that signals for the addition of the poly-A tail to the mRNA

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

1. Opened the TNF 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



gene ✓ 17-08-2024 00:14 PHY File

Task 8 output