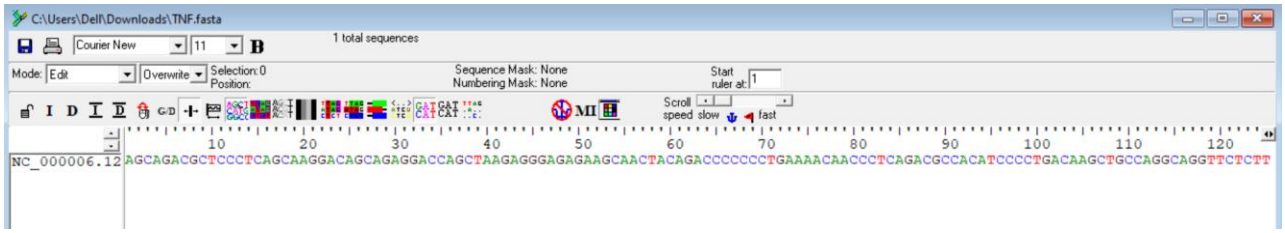
MINI PROJECT

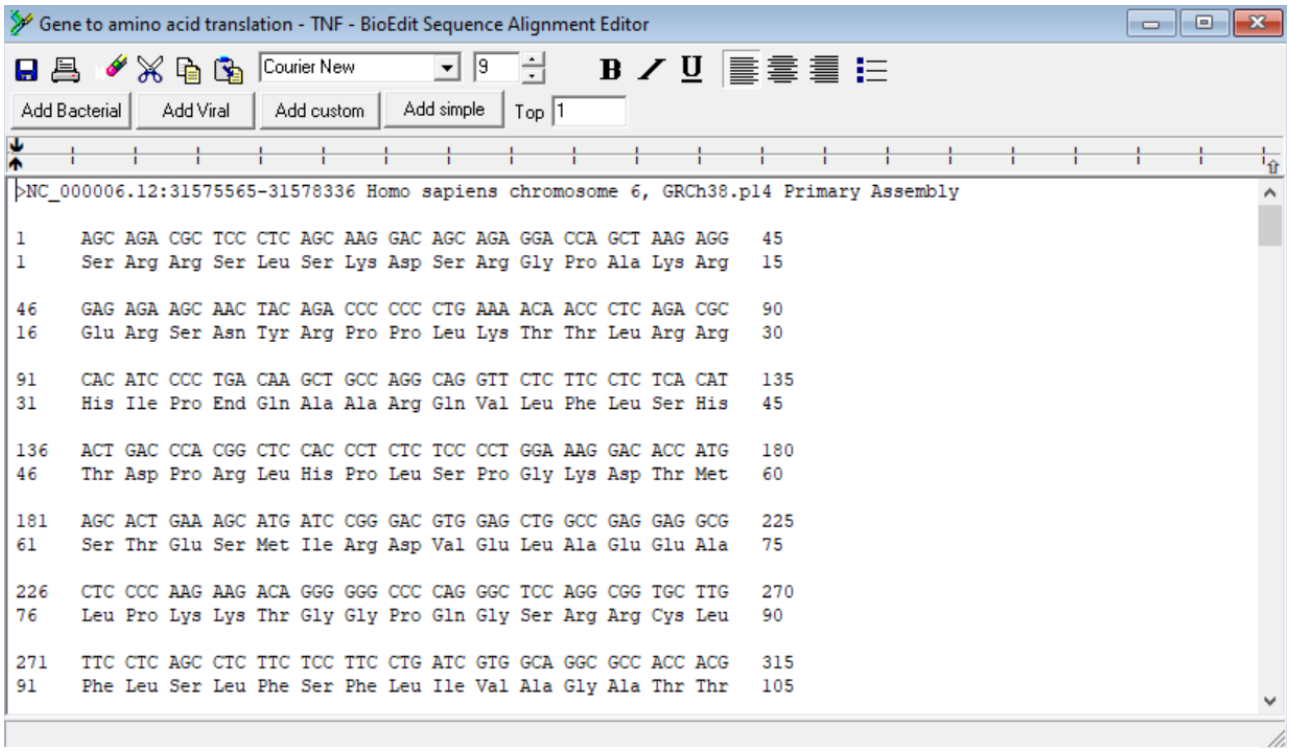
COMPREHENSIVE SEQUENCE ANALYSIS OF THE HUMAN TNF GENE

LAKSHANA B

1. **Obtaining the sequence from NCBI and viewing it in BioEdit:**

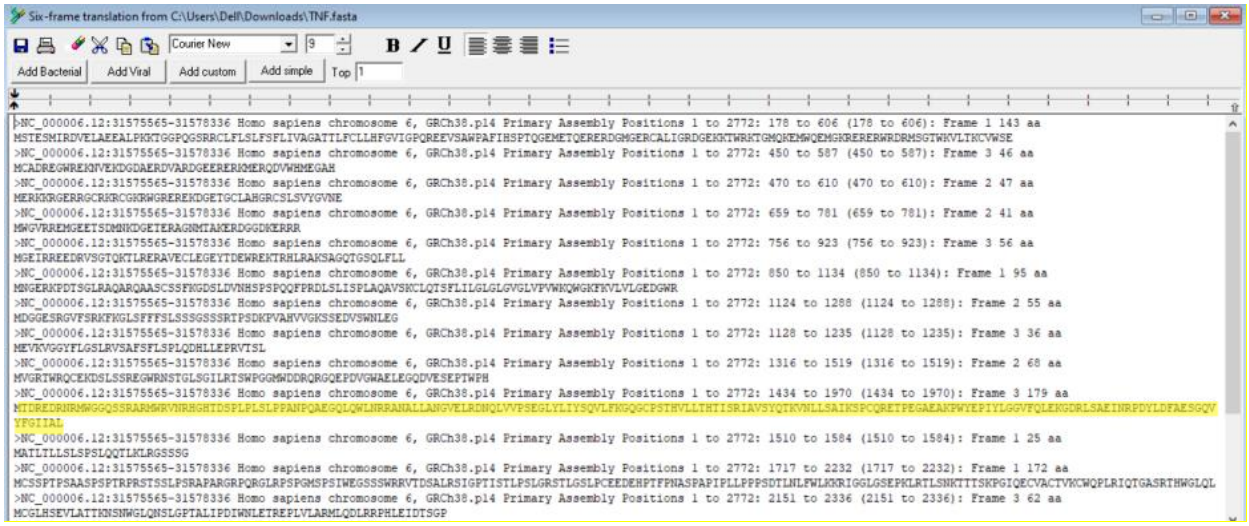


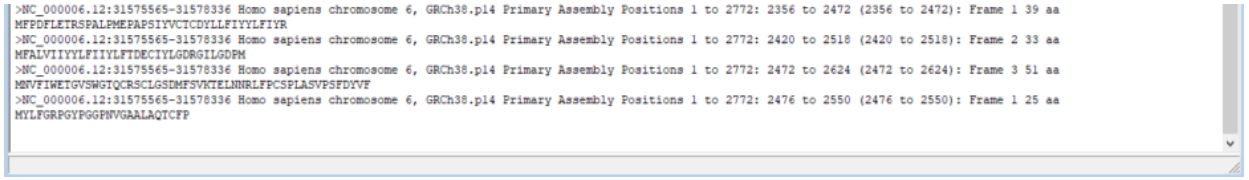
2. **Translation of DNA sequence to amino acid:**



The translated amino acid sequence from position 1-315 is displayed.

3. **Finding ORF (Open-Reading Frames) of the sequence:**

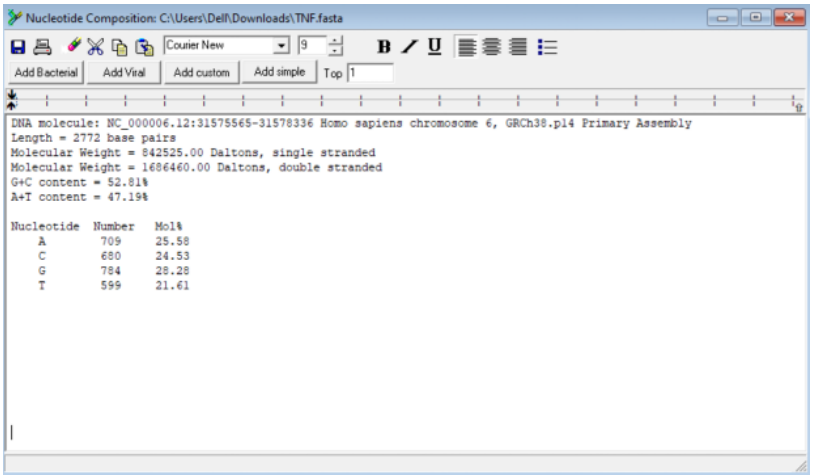


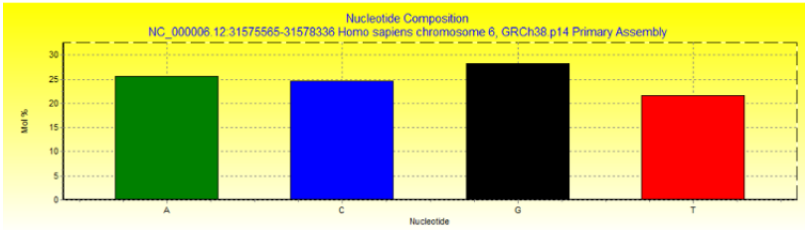


Inference:

A total of 17 open reading frames were found in the entire sequence of TNF gene. The largest open reading frame is made of 179 amino acids, found at position 1434-1970. The protein sequence of the largest ORF is highlighted. A double helix DNA molecule has three distinct reading frames, which produces six possible frame translations. In this sequence frame 1 has 6 ORF's and frame 2 and 3 has 5 ORF's each.

4. **Sequence composition analysis:**

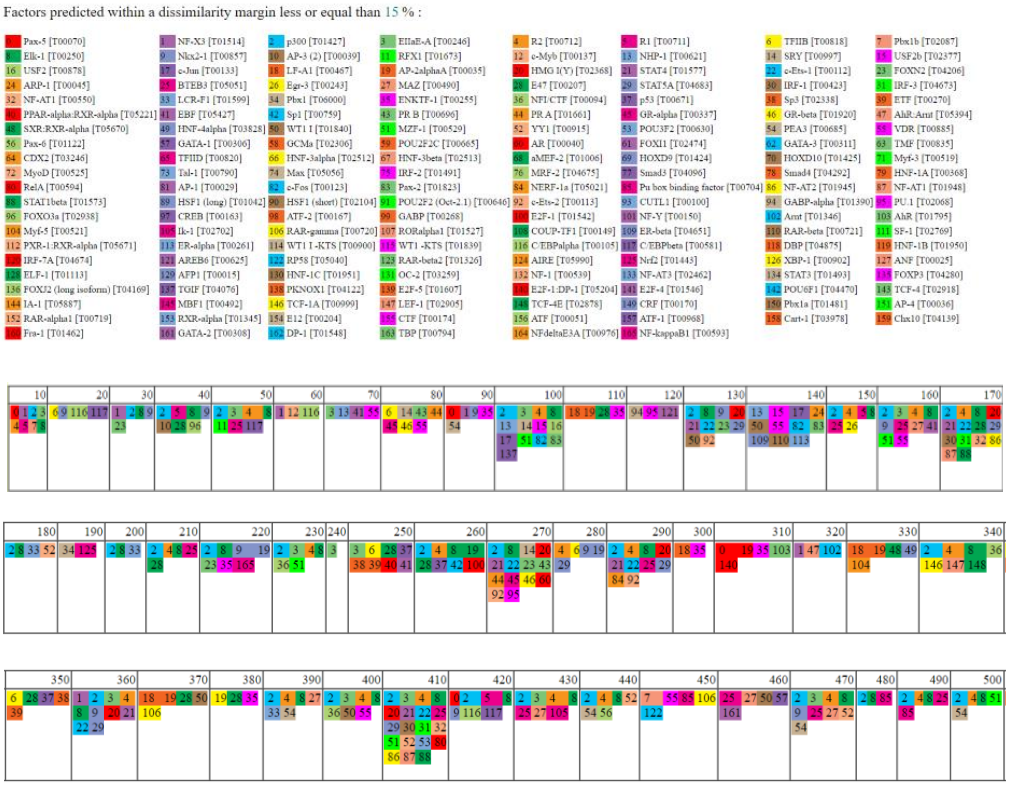




Inference:

52.81% of total G+C content and 47.19% of total A+T content is found in the sequence. Higher G+C content represents high stability of the DNA sequence. The molecular weight of both single and double strand is calculated in terms of 'Dalton' which is highlighted. Additionally, the graphical representation of molecule % of each nucleotide is also obtained, which shows highest % of 'G', followed by 'A', 'C' and 'T' respectively.

5. Identifying transcription factors binding sites:



Inference:

Using the PROMO tool, information about various transcription factors and the number of times they are bound in the entire sequence is obtained. Additionally, detailed representation of binding of different factors at positions ranging from 1-500 is obtained. Based on the obtained results, position 410 has a greater number of factors bound to it.

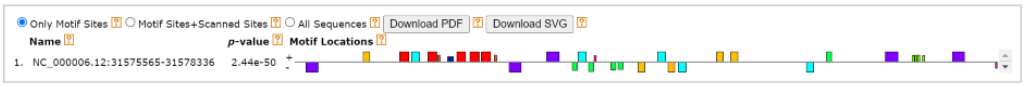
6. Functional motifs in a Genome:

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		
4.		7.4e+003	5	50		
5.		6.0e+003	5	29		
6.		1.0e+005	5	21		
7.		1.2e+005	3	8		
8.		7.0e+004	2	10		
9.		2.6e+005	2	11		

Stopped because requested number of motifs (9) found.

MOTIF LOCATIONS



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

	Name	p-value	Motif Locations
1.	NC_000006.12:31575565-31578336	2.44e-50	

INPUTS & SETTINGS

Sequences

Role	Source	Alphabet	Sequence Count	Total Size
Primary Sequences	TNF.fasta	DNA	1	2772

Background Model

Source: built from the (primary) sequences

Order: 0

Name	Freq.	Bg.			Bg.	Freq.	Name
Adenine	0.236	0.236	A	~	T	0.236	0.236	Thymine
Cytosine	0.264	0.264	C	~	G	0.264	0.264	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	9
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

Inference:

Using the MEME Suite tool, 9 functional motifs were requested and obtained. Based on the obtained results, 50 is the width of the largest motif and it was found at 5 sites. The nucleic acid 'thymine' was mostly conserved in the DNA sequence. A detailed view of the motif's location was also obtained.

7. Predicting coding and non-coding regions:

```
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
```

```
Predicted peptide sequence(s):



>/tmp/08_22_24-08:13:38.fasta|GENSCAN_predicted_peptide_1|233_aa
MSTESMIRDVELAEEALPKKTGGPQGSRRCLFLSLFSFLIVAGATTLFCLLHFGVIGPQR
EEFPRDLSLISPLAQAVRSSSRTPSDKPVAHVVANPQAEGQLQWLNRRANALLANGVELR
DNQLVVPSEGLYLIYSQVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRE
TPEGAEAKPWYEPIYLGGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL
```

Inference:

Using GENSCAN tool, the exons of the gene was predicted. Almost all the predicted sequence has probability > 70% which confirms that they are coding regions/ exons. The predicted peptide sequence was also obtained. The length of the predicted sequence is 233 amino acids.

8. FASTA to PHYLIP:

```
 1 2772
NC_000006.    AGCAGACGCT CCCTCAGCAA GGACAGCAGA GGACCAGCTA AGAGGGAGAG

              AAGCAACTAC AGACCCCCCC TGAAAACAAC CCTCAGACGC CACATCCCCT

              GACAAGCTGC CAGGCAGGTT CTCTTCCTCT CACATACTGA CCCACGGCTC
```

The fasts format of sequence was converted into PHYLIP format using Bio Edit software.