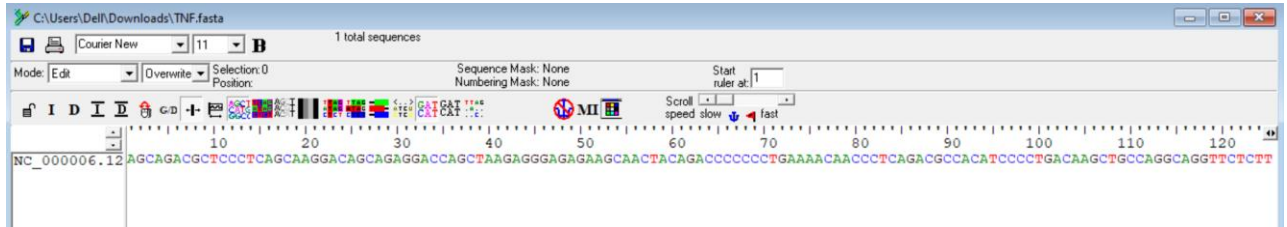


MINI PROJECT

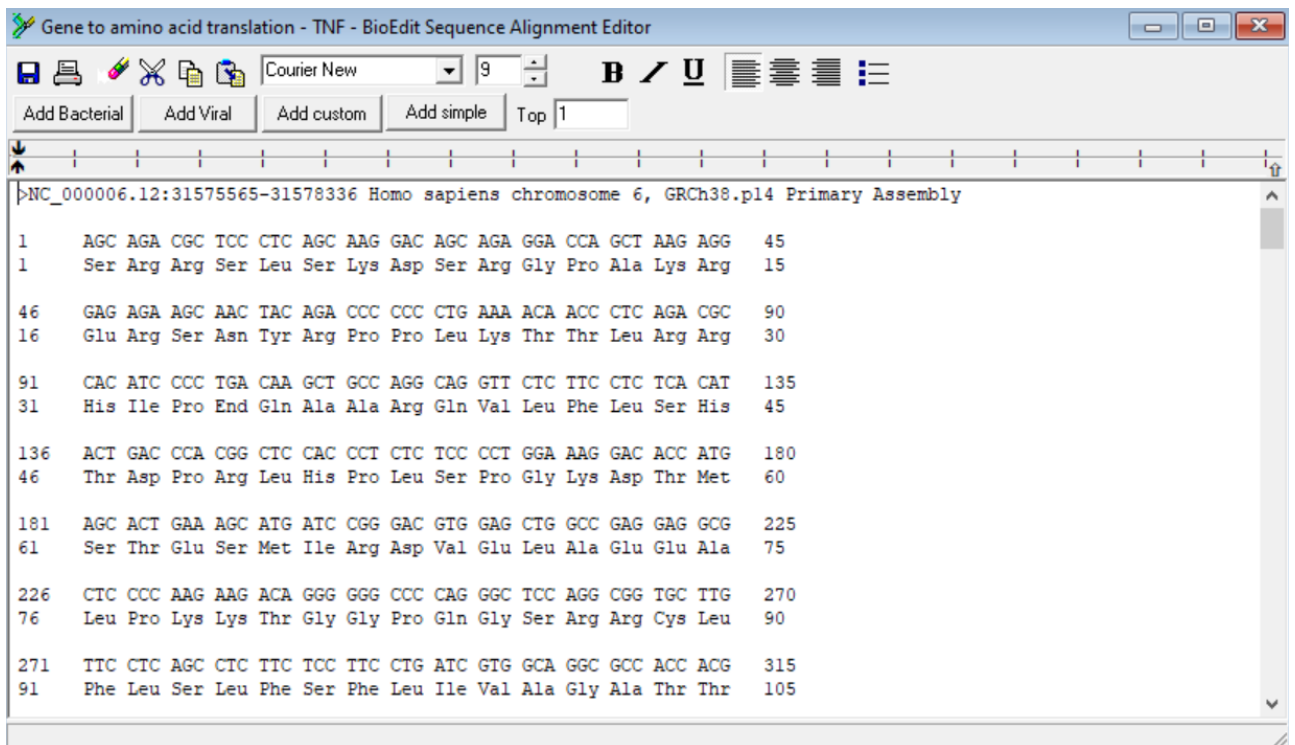
COMPREHENSIVE SEQUENCE ANALYSIS OF THE HUMAN TNF GENE

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:

```
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 178 to 606 (178 to 606): Frame 1 143 aa
MSTESMIRDELAEALPKMTGGPQSSRRCLFLSLFSLIVAGATILFCLLHFGVIGPQREVSANFAFIHSPQTQGMETQERERDGMGERCALIGRDGKNTWRKTGMQKEMWQEMGHRERERWRDRMSGTWVLIKCVSE
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 450 to 587 (450 to 587): Frame 3 46 aa
MCADREGWREINVEKDGAERDVARDEERERERQVWMEGAG
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 470 to 610 (470 to 610): Frame 2 47 aa
MERKRGERRGCRKRCGRWRGREREKDGTCGLAHRCSLSVYGVNE
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 659 to 781 (659 to 781): Frame 2 41 aa
MMGVVREMGEEISNNNDGETERAGNMTAMERDGGKERRR
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 756 to 923 (756 to 923): Frame 3 56 aa
MGEIRREEDRVSGTUTLRLRAVECEGETTDEKREKTRRLRAKSAGQTGSQLELL
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 850 to 1134 (850 to 1134): Frame 1 95 aa
MNGERFVTSGLAQAQAASCSFSGSLDYNMSPSQFFRDLSLISPLAQVSHCLQTSFLILGLGLGVGVFVWQWGWKFKVLVLGEDGWR
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 1124 to 1288 (1124 to 1288): Frame 2 55 aa
MDGGSRGVTSKAFKGLSFFFLSSSGSSSTPSDKFVAVVWGSSEVDVWNLG
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 1128 to 1235 (1128 to 1235): Frame 3 36 aa
MEVVGGVGLGSLRVSAFSLPLQDLLEFVTVSL
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 1316 to 1519 (1316 to 1519): Frame 2 68 aa
MVGRTWQCEKDLSSREGWRNSTGLSLRTSWFGGNDQRGRQGPVGVNALEGGQDVESEPTWPH
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 1434 to 1970 (1434 to 1970): Frame 3 179 aa
MTTREDRIRNMGQSSSRARWVVRHGHSTDSPLFLSLFPANFQAEGQLWLRRAIALLANGVLRDQQLVPSGLYLYISQVLFNGQGCPSRVLTLTISRIVSYQTKVNLLSAIKSPQRETFEGAAKWPYEPYILGGVFLQKHGRLSAETINRPOYDFAESGVV
YFGIILAL
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 1510 to 1584 (1510 to 1584): Frame 1 25 aa
MATILLLSLSPSLQQTILKRGSSSG
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 1717 to 2232 (1717 to 2232): Frame 1 172 aa
MCSSPTPSAASPSPTRFRTSSLSFSAFARGRPQGLAPSPGMSPSINWSSSWRRVDSALRSIGTISTLPSLGRSTLGLSLPCEEDENPTFPHASPAFIPFLPPSDTLNLFWLGRIGGLGSEPKRLTSLNHTTTSKPGIQECVACTVWQWQPLRIQTGASRTHWGLQL
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 2151 to 2336 (2151 to 2336): Frame 3 62 aa
MGLHSEVLATTNSNMGQLNSLOPTALPDWNLTEFLVLRMLQDLRRPHLEIDTSGF

>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 2356 to 2472 (2356 to 2472): Frame 1 39 aa
MFFDFLETRSPALMEPAPSIYVCTDYLLFIYLYFIYR
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 2420 to 2518 (2420 to 2518): Frame 2 33 aa
MFALVLIYLYFIYLYFTDECIYLGDRGILGDPH
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 2472 to 2624 (2472 to 2624): Frame 3 51 aa
MNVFIWETGVSWGTQCRSLGSDMFVKTLEINRLFPCLASVPSFDYVF
>NC_000006.12:31575565-31578336 Homo sapiens chromosome 6, GRCh38.p14 Primary Assembly Positions 1 to 2772: 2476 to 2550 (2476 to 2550): Frame 1 25 aa
MYLFGPGYGGPNVGAALAQTCFP
```

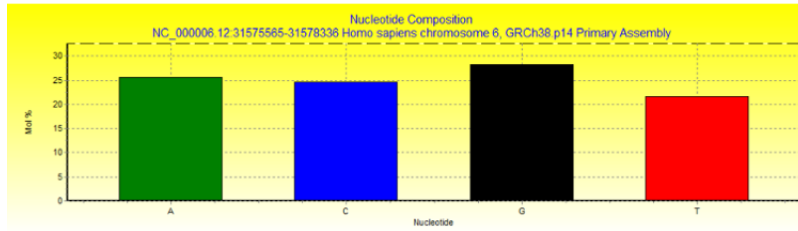
Inference:

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:

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



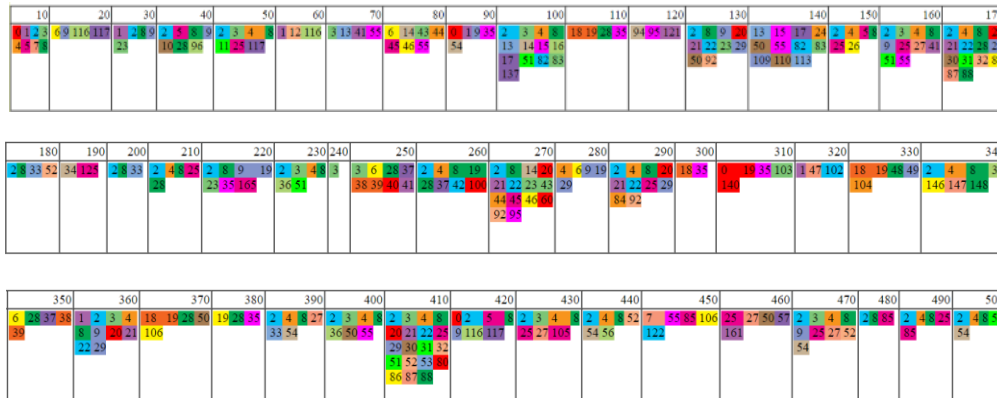
Inference:

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:

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

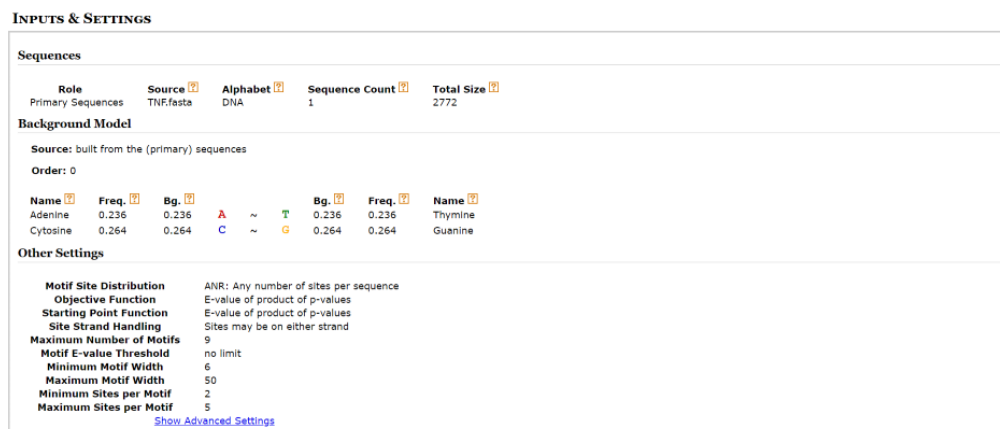
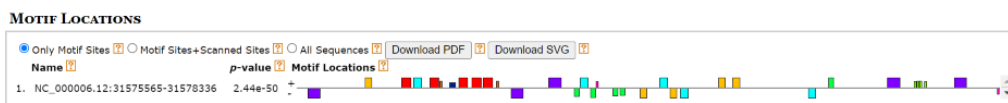
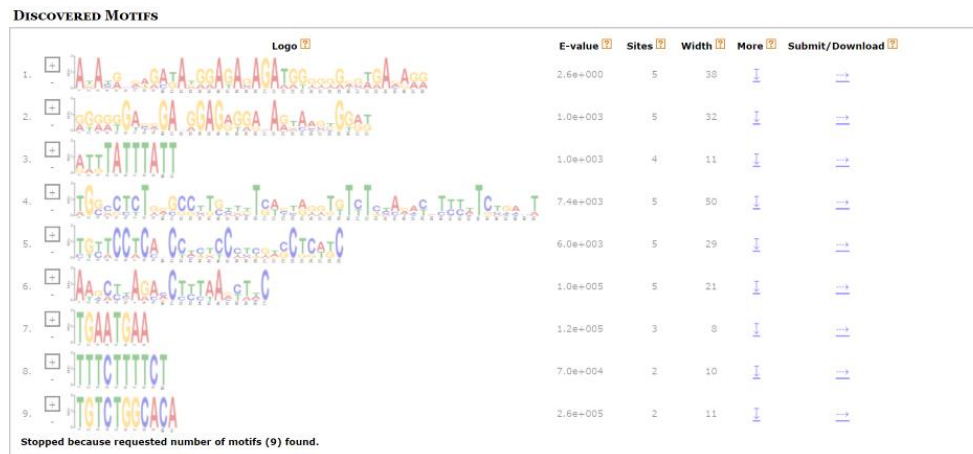
1 Pax-5 [T00070]	101 NF-κB [T01514]	203 p300 [T01427]	325 EtsA [T00246]	447 R2 [T00712]	569 R1 [T00711]	691 TFIIIB [T00818]	813 Ptx1b [T02087]
2 Ets-1 [T00250]	102 Nkx2-1 [T00857]	204 AP-3 (2) [T00039]	326 RFX1 [T01675]	448 c-Myc [T01317]	570 NHP-1 [T00621]	692 SRY [T00997]	814 USF2b [T02377]
3 USF2 [T00878]	103 c-Jun [T00133]	205 LP-A1 [T00487]	327 AP-2alpha [T00035]	449 HMO1 (Y) [T02368]	571 STAT4 [T01577]	693 c-Ets-1 [T00112]	815 FOXN2 [T04206]
4 ARF-1 [T00845]	104 BTEB1 [T00691]	206 Egr-3 [T00245]	328 MAZ [T00499]	450 E17 [T00207]	572 STAT5A [T04683]	694 IRF-3 [T04675]	816 IRF-3 [T04675]
5 NF-AT1 [T00550]	105 LCR-β1 [T01599]	207 Pbx1 [T00600]	329 ENKTF-1 [T00255]	451 NF-CTF [T00694]	573 p53 [T00671]	695 Sp1 [T02335]	817 E17 [T00207]
6 PPAR-alpha [T05221]	106 EBF [T05427]	208 Sp1 [T00759]	330 PR-9 [T00699]	452 PR-A [T01661]	574 GR-alpha [T00337]	696 GR-beta [T01920]	818 AHR-Ant [T05394]
7 RXR-alpha [T05670]	107 HNF-4alpha [T03828]	209 WT-1 [T01840]	331 MZF-4 [T06529]	453 VV1 [T00915]	575 POU3F2 [T00630]	697 PEA3 [T00685]	819 VDR [T00855]
8 Pax-6 [T01122]	108 GATA-1 [T00300]	210 GCMA [T00230]	332 POU2F2C [T00665]	454 AR [T00940]	576 FOXJ1 [T02474]	698 GATA-3 [T00511]	820 TAF [T00835]
9 CDX2 [T01246]	109 TFIIID [T00820]	211 HNF-3alpha [T02512]	333 HNF-3beta [T02513]	455 AMEP-2 [T01006]	577 HONDD9 [T01424]	699 HONDD10 [T01425]	821 Myf-3 [T00519]
10 MyoD [T00525]	110 Tbx-1 [T00790]	212 Max [T00506]	334 IRF-2 [T01491]	456 MRF-2 [T04675]	578 SmaD3 [T04096]	699 SmaD4 [T04292]	822 HNF-1A [T00568]
11 RelA [T00594]	111 AP-1 [T00029]	213 c-Fos [T00123]	335 Pex-2 [T01823]	457 NERF-1a [T05021]	579 Pa box binding factor [T00704]	699 NF-AT2 [T01945]	823 NF-AT1 [T01948]
12 STAT1beta [T01573]	112 HSF1 (long) [T01042]	214 HSF1 (short) [T02104]	336 POU2F2 (Oct-2.1) [T00646]	458 c-Ets-2 [T00113]	580 C/EBP-1 [T00100]	699 GABP-alpha [T01390]	824 PU-1 [T02068]
13 FOXO3a [T02938]	113 CREB [T01643]	215 ATF-2 [T00167]	337 GABP [T00268]	459 E2F-1 [T01542]	581 NF-Y [T00150]	699 AHR [T01346]	825 AHR [T01346]
14 Myc-1 [T00521]	114 IκB-1 [T00702]	216 RAR-gamma [T00720]	338 RORalpha1 [T01527]	460 COUP-7β1 [T00149]	582 E2F-beta [T04651]	699 RAR-beta [T00721]	826 SF-1 [T02769]
15 PXR-1 [T00571]	115 ER-alpha [T00261]	217 WT-1-K78 [T00906]	339 WT-1-K78 [T01839]	461 C/EBPalpha [T00105]	583 C/EBPbeta [T00581]	699 DRP [T04875]	827 HNF-1B [T01950]
16 DRP-7A [T04674]	116 ARB [T00625]	218 RFX5 [T00640]	340 RAR-beta2 [T01326]	462 AHR [T00990]	584 N4G [T01443]	699 XBP-1 [T00902]	828 ANF [T00025]
17 EL-1 [T01113]	117 ATF1 [T00015]	219 HNF-1C [T01951]	341 OC-2 [T01299]	463 NF-1 [T00539]	585 NF-AT3 [T02462]	699 STAT3 [T01401]	829 POU3 [T04200]
18 FOXD2 (long isoform) [T04169]	118 TGIF [T04076]	220 POU3 [T04122]	342 E2F-5 [T01607]	464 E2F-1-DP-1 [T05204]	586 E2F-4 [T01546]	699 POU3 [T04170]	830 TCF-4 [T02918]
19 IκA-1 [T05857]	119 MBF1 [T00492]	221 TCF-1A [T00999]	343 LEP-1 [T02005]	465 TCF-4E [T02878]	587 CRE [T00170]	699 Ptx1a [T01481]	831 AP-4 [T00036]
20 RAR-alpha1 [T00719]	120 RXR-alpha [T01345]	222 E12 [T00204]	344 CTF [T00174]	466 ATF [T00051]	588 ATF-1 [T00988]	699 Cux-1 [T03978]	832 Cux10 [T04139]
21 Fts-1 [T01462]	121 GATA-2 [T00308]	223 DP-1 [T01548]	345 TBP [T00794]	467 NFdeltaB3A [T00975]	589 NF-kappaB1 [T00593]		



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:



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
MSTESMIRDVELAEELPKKTGGPQGSRRCLFSLFSFLIVAGATTLFCLLHFGVIGPQR
EEFPRDLSLISPLAQAVRSSRTPSDKPVAVHVPANPQAEGLQWLNRANALLANGVELR
DNQLVVPSEGLYLIYSQVLFKGQGPCSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRE
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 AGACCCCCC TGAAAACAAC CCTCAGACGC CACATCCCCT
GACAAGCTGC CAGGCAGGTT CTCTTCCTCT CACATACTGA CCCACGGCTC
```

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