2021-22

BIO F242
Introduction to Bioinformatics

End-Sem Project

Name : Suchismita Tripathy
ID : 2019A7PS0554P

Gene Name : HUS1 Checkpoint Clamp Component (HUS1) transcript variant X1
Organism : Theropithecus gelada
Accession Number : XM_025380797.1

17 May, 2022

Contents

1 Introduction

The protein encoded by this gene is a component of an evolutionarily conserved, genotoxin-activated checkpoint complex that is involved in the cell cycle arrest in response to DNA damage. This protein forms a heterotrimeric complex with checkpoint proteins RAD9 and RAD1. In response to DNA damage, the trimeric complex interacts with another protein complex consisting of checkpoint protein RAD17 and four small subunits of the replication factor C (RFC), which loads the combined complex onto the chromatin. The DNA damage induced chromatin binding has been shown to depend on the activation of the checkpoint kinase ATM and is thought to be an early checkpoint signaling event. Alternative splicing results in multiple transcript variants.

Here, there are multiple transcript variants as well, with different exon counts, coding sequence length and so on. For the transcript variant X1 with a length of 3031 nucleotides and a coding sequence length of 843, its location is on chromosome 3 and it has 8 exons.

2 Details

2.1 Gene Sequence

>XM_025380797.1 PREDICTED: Theropithecus gelada HUS1 checkpoint clamp component (HUS1), transcript variant X1, mRNA

AACCATGTGCGCGCCGCCGTGTTTCCGGGCGGGGACACTCAGGGCGCGACACTCATCTGTTACCCACAGA
GGCCGTCGCGGCTGCGGCAGGCGCGGCCATGAGGTTTCGGGCCAAGATCGTGGACGGGGCCTGTCTGAAC
CACTTCACACGAATCAGTAACATGATAGCCAAGCTTGCCAAAACCTGCACCCTCCGCATCAGCCCTGATA
AGCTTAACTTCATCCTTTGTGACAAGCTGGCTAATGGAGGGGTGAGCATGTGGTGTGAGCTGGAACAGGA
GAACTTCTTCAACGAATATCAAATGGAGGGTGTCTCTGCAGAAAACAATGAGATTTATTTAGAGCTAACA
TCGGAAAACTTATCTCGAGCCTTGAAAACTGCCCAGAATGCCAGAGCCTTGAAAATCAAACTGACTAATA
AACACTTTCCCTGCCTCACAGTCTCCGTGGAGCTGTTATCTATGTCAAGCAGTAGCCGCATTGTGACACA
TGACATCCCTATAAAGGTGATTCCTAGGAAATTGTGGAAGGACTTGCAAGAACCGGTGGTCCCAGATCCT
GATGTTAGTATTTATTTACCAGTCTTGAAGACTATGAAGAGTGTTGTGGAAAAAATGAAAAACATCAGCA
ATCACCTTGTTATTGAAGCAAACCTAGATGGAGAATTGAATTTGAAAATAGAAACTGAATTAGTATGTGT
TACAACTCATTTTAAAGATCTTGGAAATCCTCCTTTAGCCTCTGAAAGCACCCATCAAGACAGAAACATG
GAACACATGGCTGAAGTGCACATAGATATTAGGAAGCTCCTACAGTTTCTTGCTGGACAACAAGTAAATC
CCACAAAGGCCTTATGCAATATTGTGAATAACAAGATGGTGCATTTTGATCTGCTTCACGAAGACGTCTC
CCTTCAGTATTTCATCCCTGCGCTGTCCTAGCACCCTGTTGCTGGAGTTGGCATGCGGAGACTTTGTCAG
GATGGGAGAGGCCGTAGGCGTTGTGTTCTGATCACTGGTCTGTGTCCTCACAGCACCGCACATCGACACA
CTGTACTTATTTGTCCCTCTCTAACATTTTAACTAAAAGTTGACTTAATGGCACACAGTTGGATAAACAT
ATCACTTCATGTTGCTCATGTCTGTTTTGTTTTGTTTTTAAGACATTGAAAAGAAAAGCTAGAATTTATT
TATTCAGACTTTAAAGAACAATTTCTCATTGACGTTGTGAAAATCCTCATGTATTTAGACTTGGTGTAGT
AGCCAGAATTCGTGCAGCTGTTGCCTGGGAGCTGGGTACTTTCCCTCCGGGCAGAGGCTCTAGCTCAGCA
CGGCCTGTAGCGCACAGTCAGTCTTGCATTTCAGTGTGTTCACCCTGCTTGCTCCTGCCCCTTGGAGCCC
AGTGACAGAAAGTATAGCCTCTGTCACCCCGCTGCCACTGCCTTGGTTACTCAGAGCGCTGTGGGGTGTC
ACAGCTGCAGCATTTGGGGTCTCTCTCTCTCTTGCTGAGTACTCAAGCCCACCTGAGTCCACTCCCCTCT
TGATGCCTGGAGAGCTGGCCCAGCCAACACAGCTTTTCGCTGGGAGCTCCTTCTGCCATTCCAATTAGTT
TCTTCCTGGGGCCAGTTTTGGGTTTAGGTTGTAATTCCTTATATTTCTTTCTTCCACAGTGCATCGGATC
TGTCATTCTGGAAAGAAGACCCTTCTATTTAGAGTAGAAACAAATGAAACTTCTAAGGTATCATCTGTGT
TAGGTGATGAGACCATATTTCTTTGATGTTTCTGAACATCAAAGCTGATTCAGTACTGGTAGATGTGCTC
ATTCTCCCTGAAACATACCTATCATATTTCCTATTATAATTCTATCTCATTGTCCTGTGGAGGTGGACAT
GATCAACAATATCTTTTATTTTCTTGTTTTGTTTTGTTTGAGACAAGGTCTCACTCTGTCACCCAGACTG
GAGTGAAGGGCAACAATCATGGCTCACCGCCTTGACCTCCTTGGCTCAAGGCATCCTCCCACCTCAGCCT
CCTGACTACCTGGGACTACTGGTGTGCGCCACCACATCCAGCTAGTTTTTAATTTTTCATAGAGACAGAG
TCCCACTGTGTTGCCTAGGCTGGTCTTGAACTCCTGGGCTCAAGCCACCCACCCACCTGGGCCTCCCAAA
GTGCTGGGATTACAGGCATTAGCCATCACACCCATCCGTAAACGTTATCTTAATGTCACATTACAAAACT
GAGGCCTAAGTTGCTTAGGACAAGATGAAGAAATCGAAGACTAGCTAACATGAAAATTTATATTTTGGCT
TTTTCATGTTTTTTGGTAAAACCAGTGTATTTGAATAGTTCTTTTGATGTTTCATAATGGTTTTTTGTTT
GTTTGCTTGGTTCAGTTTTTTTTTTTCCTTGAGACAGAGTCTTGCTCTGTCGCCAGGCCAGAGTGCCAGT
GTCACGATCTTGGCTCACTGCAACCTCCACCTCCCAAGTTCAAGCGATTCTCCTGCCTCAGCCCCCCGAG
TAGCTGGGACTATAGGCACGTGCCACCACACCCAGCTAATTTTTGTATTTTTAGTAGAGATGGCATTTCA
CTATGTTGGCCAGGATGGTCTCGATCTCTTGACTTTGTGATCCACCTGCCTTGGCCTCTCAAAAGTGCTG
GGATTACAGGCGTGAGCCACCATGCCTGGCCTGTGTTTAATAATGTTTAAATAGGGTGGAATATTTTGTT
AAATTAACATTTTAAAATTAGAAGACGCCATTTTAATTTTTAAACCCTTTCTCCTCGTTGTAACAAAATT
AATTCCAGCTGTAGTGAGAAAACTTAAAAATCATGATACAAAATGAAACAATATCTGAAAGTAGTTTTAT
AAAACTGAAATTATTGTTAAAGAGAATGGTATTAGTGACTTAACCATTTGCTCTATATGATGTTTATTAT
CAAATACACATAATTTTGAAGATTTTAATGAATGGCTTAAGATTTTATCTTTGTGTAGAATGTGGCTAAA
GAAACCTTAGTTGAGATTCAA

Figure 1: FASTA file as obtained from the NCBI website

2.2 mRNA Sequence

Figure 2: mRNA Sequence obtained using Python code in Experiment-2

2.3 Protein Sequence

MRFRAKIVDGACLNHFTRISNMIAKLAKTCTLRISPDKLNFILCDKLANG
GVSMWCELEQENFFNEYQMEGVSAENNEIYLELTSENLSRALKTAQNARA
LKIKLTNKHFPCLTVSVELLSMSSSSRIVTHDIPIKVIPRKLWKDLQEPV
VPDPDVSIYLPVLKTMKSVVEKMKNISNHLVIEANLDGELNLKIETELVC
VTTHFKDLGNPPLASESTHQDRNMEHMAEVHIDIRKLLQFLAGQQVNPTK
ALCNIVNNKMVHFDLLHEDVSLQYFIPALS

2.4 Analysis

1. Length: 3031 nucleotides
2. Location: chromosome-3
   Sequence : NC_037670.1 -> 65,152,075 ..65,169,087
3. Exon Count: 8
4. Lengths (limits) of exons:
   (a) (1 ..150)
   (b) (151 ..278)
   (c) (279 .. 455)
   (d) (456 ..563)

(e) (564 ..638)

(f) (639 ..738)

(g) (739 ..858)

(h) (859 ..3031)

5. Number of introns: 0 (as there are 8 exons which span the total length of the sequence, the number of introns must be 0 for this variant)
6. Coding sequence length: 99 ..941 i.e. 843 nt
7. Number of A: 792
8. Number of T: 915
9. Number of G: 641
10. Number of C: 683
11. Number of start codons (ATG): 51
12. Number of stop codons (TAG + TGA + TAA): 158
13. GC Content: 43.68195315077532 %
14. Length of mRNA transcript: 3031 nt
15. Number of EcoRI sites: 1
16. Number of BamHI sites: 0
17. Number of HindIII sites: 2
18. Number of Isochores: 2
19. Number of genes:
20. Number of Init = Initial exon (ATG to 5' splice site): 0
21. Number of Intr = Internal exon (3' splice site to 5' splice site): 0
22. Number of Term = Terminal exon (3' splice site to stop codon): 0
23. Number of Sngl = Single-exon gene (ATG to stop): 1
24. Length of longest Prom = Promoter (TATA box / initation site): 0
25. Length of longest PlyA = poly-A signal (consensus: AATAAA): 6

26. Maximum CodRg : coding region score (tenth bit units): 500
27. Maximum P : probability of exon (sum over all parses containing exon): 0.885
28. Location of start codon in protein: 99
29. Location of stop codon in protein: 941
30. Length of protein: 280 amino acids
31. Number of G (Glycine) in protein sequence: 7
32. Number of A (Alanine) in protein sequence: 16
33. Number of V (Valine) in protein sequence: 21
34. Number of C (Cysteine) in protein sequence: 7
35. Number of P (Proline) in protein sequence: 12
36. Number of L (Leucine) in protein sequence: 35
37. Number of I (Isoleucine) in protein sequence: 19
38. Number of M (Methionine) in protein sequence: 10
39. Number of W (Tryptophan) in protein sequence: 2
40. Number of F (Phenylalanine) in protein sequence: 10
41. Number of K (Lysine) in protein sequence: 21
42. Number of R (Arginine) in protein sequence: 10
43. Number of H (Histidine) in protein sequence: 10
44. Number of S (Serine) in protein sequence: 19
45. Number of T (Threonine) in protein sequence: 14
46. Number of Y (Tyrosine) in protein sequence: 4
47. Number of N (Asparigine) in protein sequence: 21
48. Number of Q (Glutamine) in protein sequence: 9
49. Number of D (Aspartic acid) in protein sequence: 13
50. Number of E (Glutamic acid) in protein sequence: 20

3 Dot Plots

3.1 Gene Dot Plot

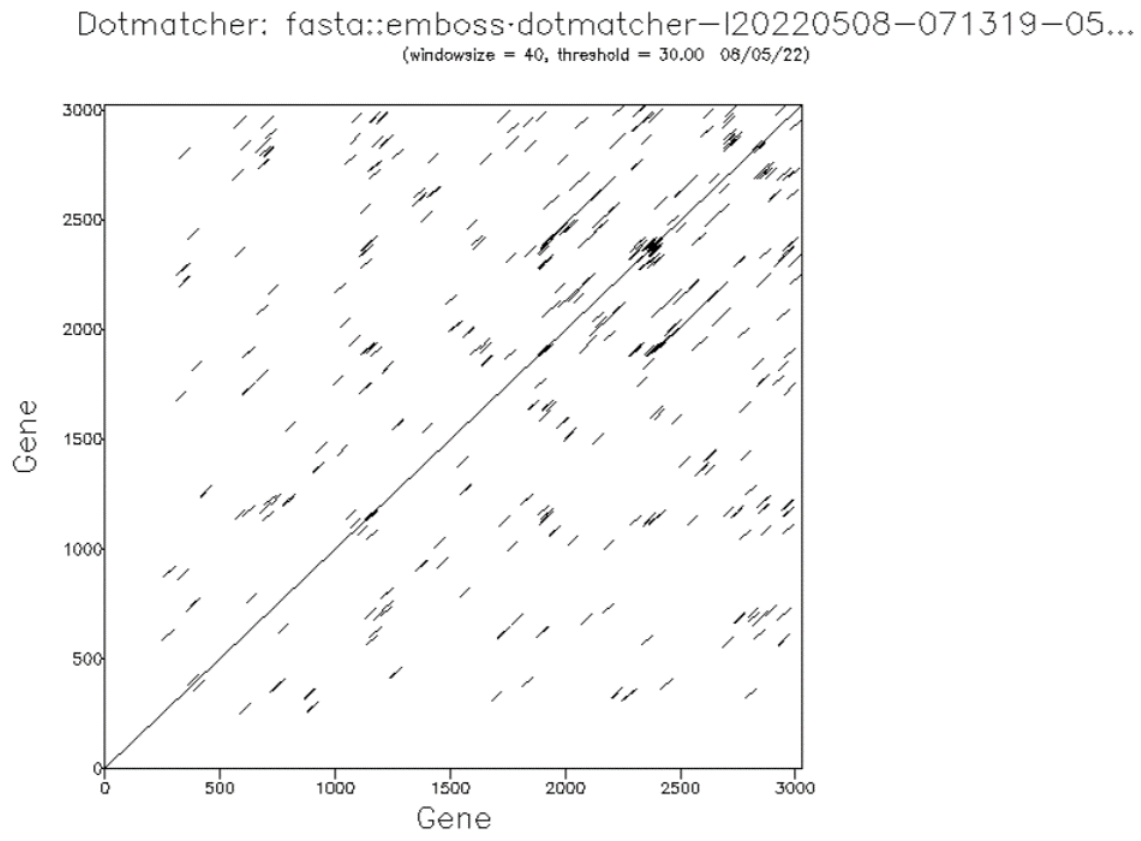

Figure 3: Gene dot plot with DNAFull scoring matrix, window size 40 and threshold 30

3.2 mRNA Dot Plot

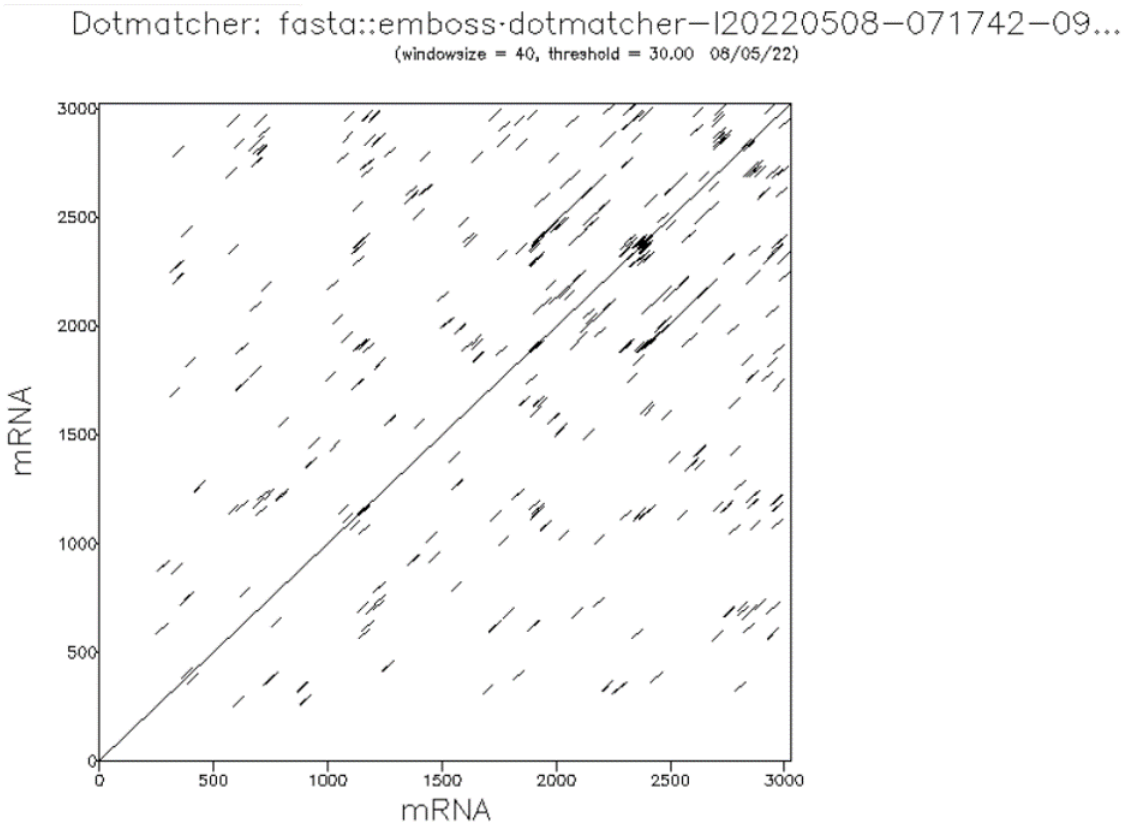

Figure 4: mRNA dot plot with DNAFull scoring matrix, window size 40 and threshold 30

3.3 Protein Dot Plot

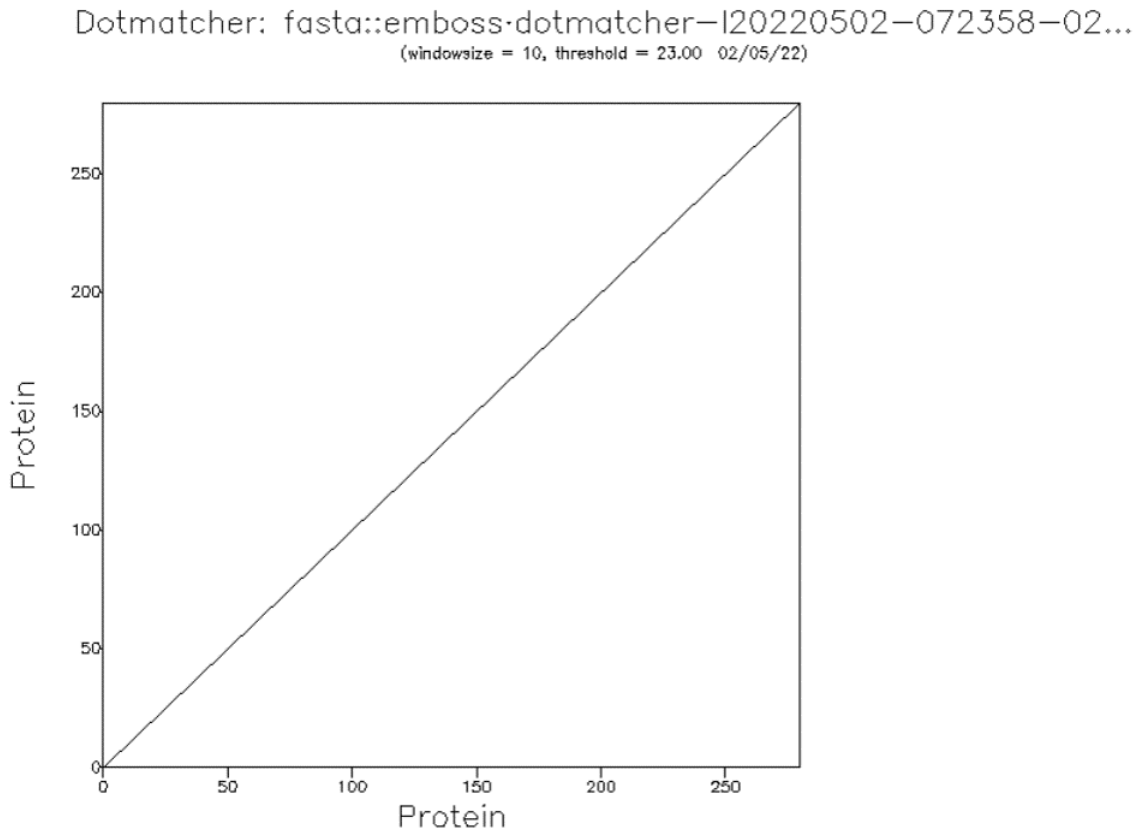

Figure 5: Protein dot plot with BLOSUM60 scoring matrix, window size 10 and threshold 23

3.4 Analysis

Both gene and mRNA dot plots look the same. They have one main diagonal (since they are self dot plots) and some smaller dispersed diagonals. These represent the noise and indicate the presence of short repeats.

The protein dot plot only has the main diagonal and hence has no short repeats within the sequence. The protein dot plot results were obtained with less noise at a lower window size and threshold value since proteins involve 20 different characters unlike nucleotide sequences where only 4 characters are involved at a time, greatly increasing the chance of noise.

4 BLAST

4.1 Gene

4.1.1 Top 5 Eukaryotic

Species	Gene	Max Score	Total Score	Query Cover	E Value	% Identity	Acc. Length
Papio anubis	PKD1L1	1576	1660	31%	0.0	99.77	10787
Cercocebus atys	HUS1	1291	1836	33%	0.0	99.58	1393
Macaca mulatta	HUS1	1772	1772	32%	0.0	99.19	1482
Macaca nemestrina	HUS1	1772	1772	32%	0.0	99.19	1462
Colobus angolensis palliatus	HUS1	1755	1904	35%	0.0	98.88	1126

4.1.2 Top 5 Prokaryotic

No prokaryote matches were found.

4.1.3 Analysis

Papio anubis polycystin 1 like 1, transient receptor potential channel interacting protein was found to have the greatest percentage identity (99.77%) with the input sequence. These results all had an e-value of 0.0, indicating that they are very significant.

4.1.4 Conclusions

The Paio anubis gene has the highest identity with the query sequence while Colobus angolensis palliatus has the lowest identity value.

4.2 Protein

4.2.1 Top 5 Eukaryotic

Species	Gene	Max Score	Total Score	Query Cover	E Value	% Identity	Acc. Length
Cercocebus atys	HUS1	553	553	100%	0.0	100	280
Chlorocebus sabaeus	HUS1	551	551	100%	0.0	99.64	280
Macaca nemestrina	HUS1	551	551	100%	0.0	99.29	280
Macaca mulatta	HUS1	550	550	100%	0.0	99.29	280
Pilicolobus tephrosceles	HUS1	548	548	100%	0.0	98.93	280

4.2.2 Top 5 Prokaryotic

No prokaryote matches were found.

4.2.3 Analysis

Cercocebus atys checkpoint protein HUS1 isoform 2 was found to have 100% identity with the given protein. The top 5 results all had the same sequence length as the input and hence there was 100% query coverage for all. These results all had an e-value of 0.0, indicating that they are very significant.

4.2.4 Conclusions

All the top 5 results returned by BLAST have 100% identity with the query sequence and all the results are significant as indicated by the 0.0 E-value.

5 Needleman-Wunsch Global Alignments

5.1 Gene

5.1.1 Matrix

Species	Gene	Score	Length	Identity	Similarity	Gaps
Papio anubis	PKD1L1	6371.0	11041	21.9%	21.9%	74.8%
Cercocebus atys	HUS1	5709.5	3267	33.9%	33.9%	64.6%
Macaca mulatta	HUS1	6025.0	3379	32.1%	32.1%	66.4%
Macaca nemestrina	HUS1	6014.5	3359	32.3%	32.3%	66.2%
Colobus angolensis palliatus	HUS1	6090.5	3089	34.2%	34.2%	65.4%

5.1.2 Analysis

Papio anubis, that had the highest percentage identity according to the BLAST results has the lowest identity and similarity score according to the Needleman-Wunsch algorithm. Colobus angolensis palliatus on the other hand had the highest identity and similarity score. All results have non-zero gap percentage as their lengths are not the same as the length of the input sequence.

5.1.3 Conclusions

Here it is shown that the Colobus angolensis HUS1 gene has the highest similarity and identity percentage with the query sequence while the Papio anubis HUS1 gene has the lowest similarity and identity percentage.

5.2 Protein

5.2.1 Matrix

Species	Gene	Score	Length	Identity	Similarity	Gaps
Cercocebus atys	HUS1	1443.0	280	100.0%	100.0%	0.0%
Chlorocebus sabaeus	HUS1	1439.0	280	99.6%	100.0%	0.0%
Macaca nemestrina	HUS1	1437.0	280	99.3%	100.0%	0.0%
Macaca mulatta	HUS1	1435.0	280	99.3%	100.0%	0.0%
Pilicolobus tephrosceles	HUS1	1432.0	280	98.9%	100.0%	0.0%

5.2.2 Analysis

Cercocebus atys, that had the highest percentage identity according to BLAST also has the highest indentity score according to the Needleman-Wunsch algorithm. In addition, all these sequences have 100% similarity with the input. As all these sequences have length 280 aa, the same as the input sequence, there are no gaps in the alignment and the gap percentage is 0 for all.

5.2.3 Conclusions

All top 5 returned sequences have 100% similarity with the query sequence, with the Cercocebus atys HUS1 protein sequence also having 100% identity.

6 Multiple Sequence Alignment

6.1 Gene

6.1.1 Consensus Sequence

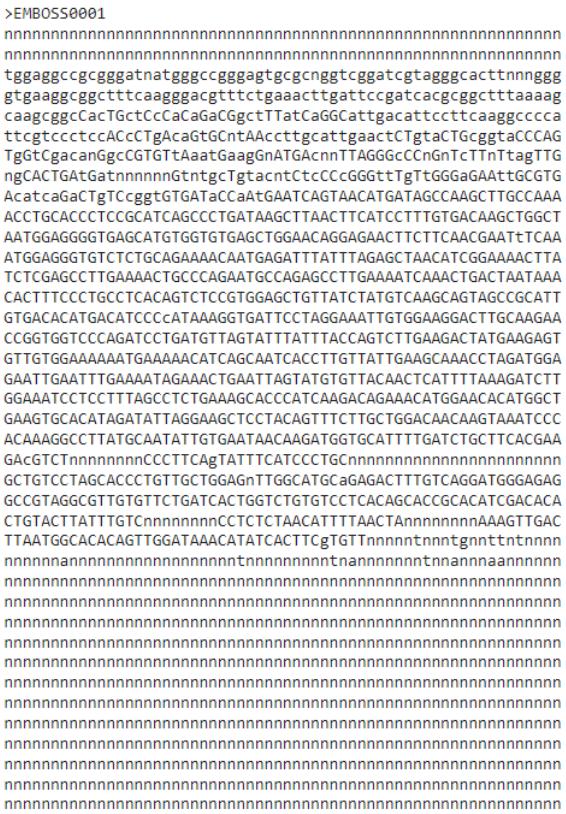

Figure 6: Partial sequence shown, due to pausity of space

6.1.2 Analysis

The 'n's in the sequence represent lack of information i.e. in the case of gaps. The EMBOSS Cons online software uses a threshold value equal to half the total weight of all the sequences to judge the matches. For positive matches for which the score is above the threshold, the consensus uses upper case characters and lower case characters otherwise. Here, in the consensus sequence obtained, a large section is in upper case characters, indicating that a large part of the genes is mostly common, and is only broken in a few places by lower case letters. The rest of the sequence has lower case characters, representing unaligned residues and long chains on 'n's at the

beginning and end of the consensus sequence i.e. one gene sequence (Papio anubis) is longer than the others.

6.1.3 Conclusions

The consensus sequence shows that there is a large section that has most residues common across all sequences used for MSA, with surrounding regions of low matches as indicated by the lower case characters.

6.2 Protein

6.2.1 Consensus Sequence

```
>EMBOSS0001
MRFRAKIVDGACLNHFTRISNMIAKLAKTCTLRISPDKLNFILCDKLANGGVSMWCELEQ
ENFFNEFQMEGVSAENNEIYLELTSENLSRALKTAQNARALKIKLTNKHFPCLTVSVELL
SMSSSSRIVTHDIPIKVIPRKLWKDLQEPVVPDPDVSIYLPVLKTMKSVVEKMKNISNHL
VIEANLDGELNLKIETELVCVTTHFKDLGNPPLASESTHQDRNMEHMAEVHIDIRKLLQF
LAGQQVNPTKALCNIVNNKMVHFDLLHEDVSLQYFIPALS
```

Figure 7: Full protein consensus sequence

6.2.2 Analysis

Since all top 5 sequences as obtained from BLAST showed 100% similarity with the query sequence, the consensus sequence also has the same characters in the same order as the query sequence itself. Therefore, there are only upper case letters in the consensus sequence obtained from EMBOSS Cons and no lower case letters as all residues are aligned. As all the top 5 sequences returned by BLAST also had the same length as the query sequence (280 amino acids), there are no 'n's in the consensus sequence.

6.2.3 Conclusions

The consensus sequence is the query protein sequence itself as the sequences used for MSA had 100% similarity with the query seqeunce.

7 Phylogenetic Trees

7.1 Gene

7.1.1 Maximum Parsimony Method

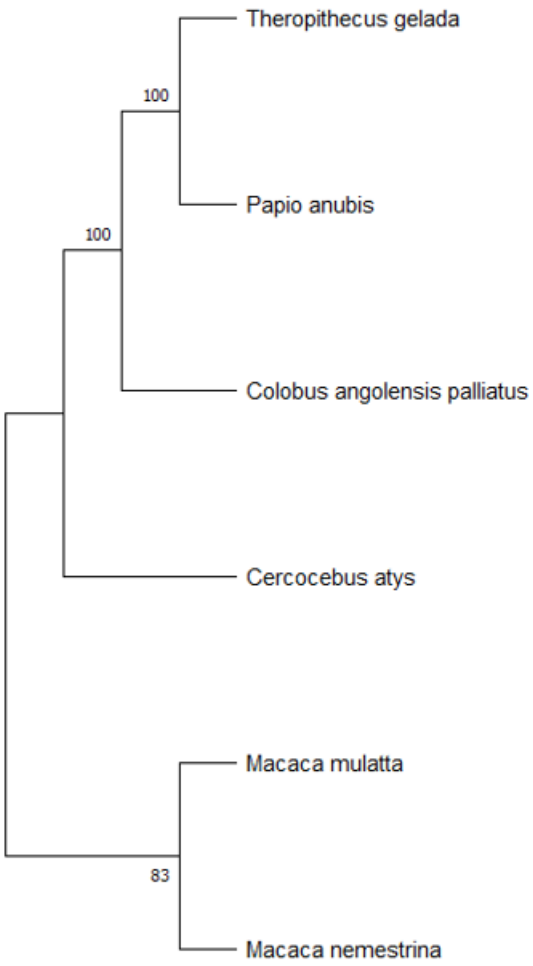

Figure 8: Maximum Parsimony Tree generated for the gene sequences

7.1.2 Neighbour Joining Method

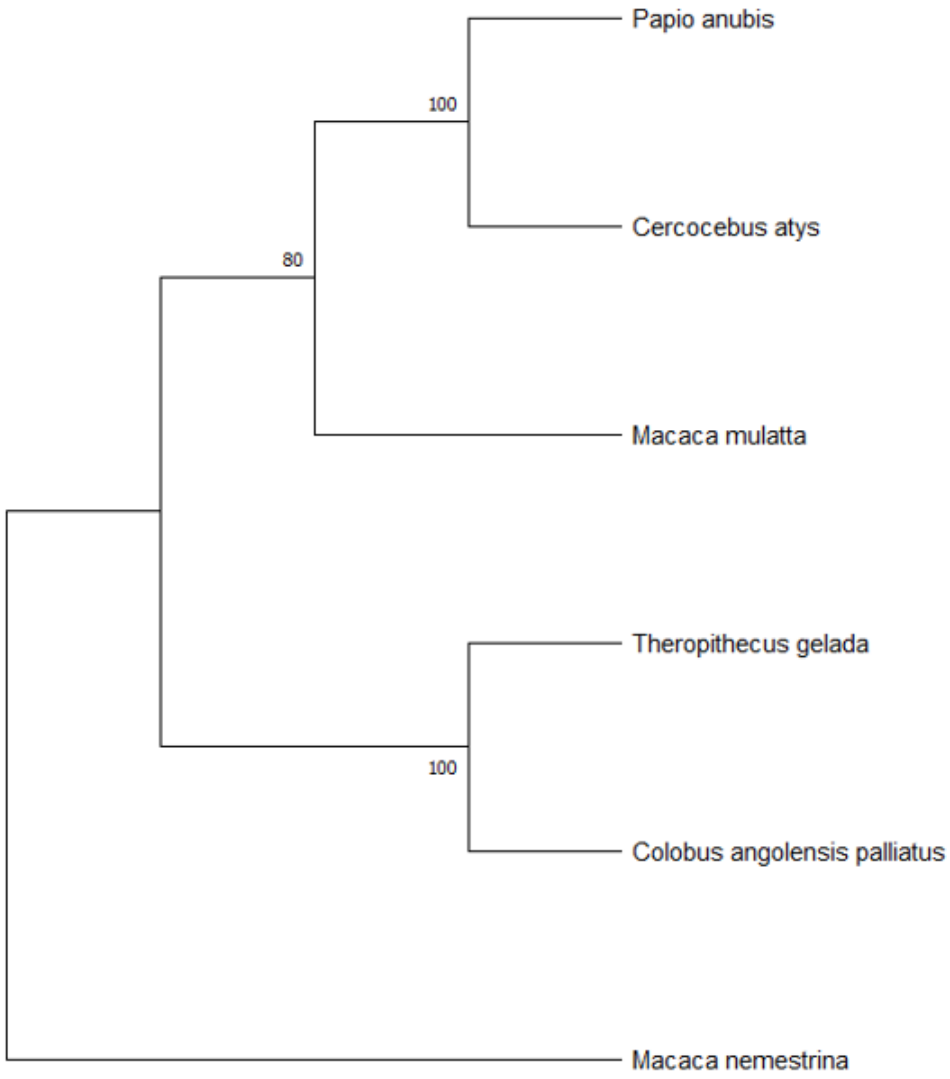

Figure 9: Neighbour Joining Tree generated for the gene sequences

7.1.3 Maximum Likelihood Method

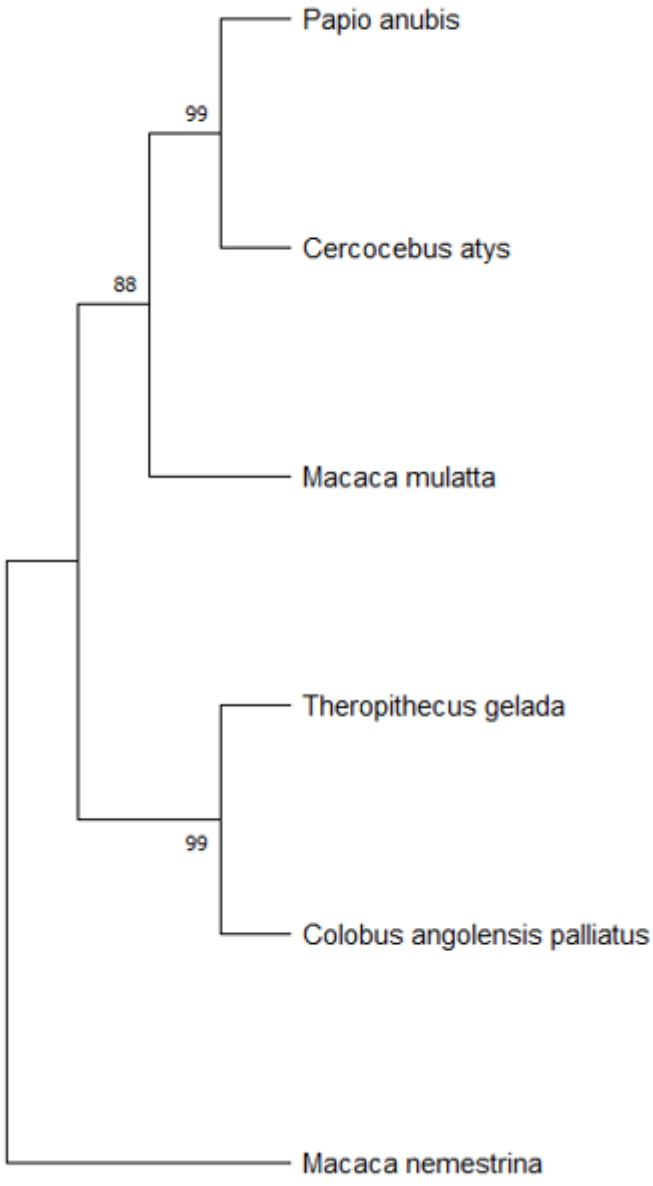

Figure 10: Maximum Likelihood Tree generated for the gene sequences

7.1.4 Analysis

According to the NJ and ML, Papio anubis - Cercocebus atys and Theropithecus gelada - Colobus angolensis palliatus are the closest pairs (both equally close). Here, Macaca nemestrina is the farthest from all other sequences. In the MP tree, Theropithecus gelada - Papio anubis are the 2 most closely related sequences, with Macaca mulatta and Macaca nemestrina being the most distantly related and equally far from all other sequences.

7.1.5 Conclusions

The most closely related to the query sequence, Theropithecus gelada is Papio anubis according to the MP method and Cercocebus atys according to the NJ and ML methods.

7.2 Protein

7.2.1 Maximum Parsimony Method

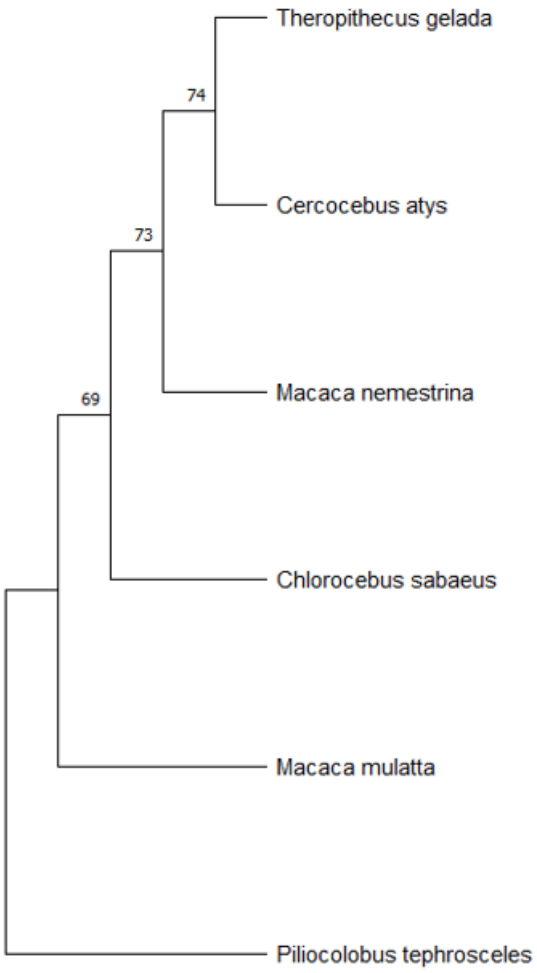

Figure 11: Maximum Parsimony Tree generated for the protein sequences

7.2.2 Neighbour Joining Method

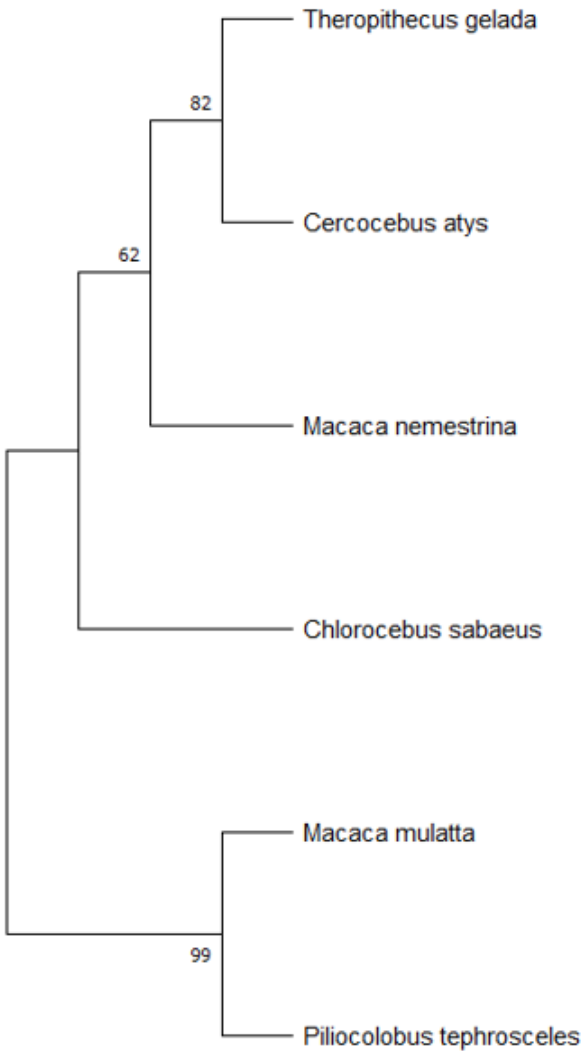

Figure 12: Neighbour Joining Tree generated for the protein sequences

7.2.3 Maximum Likelihood Method

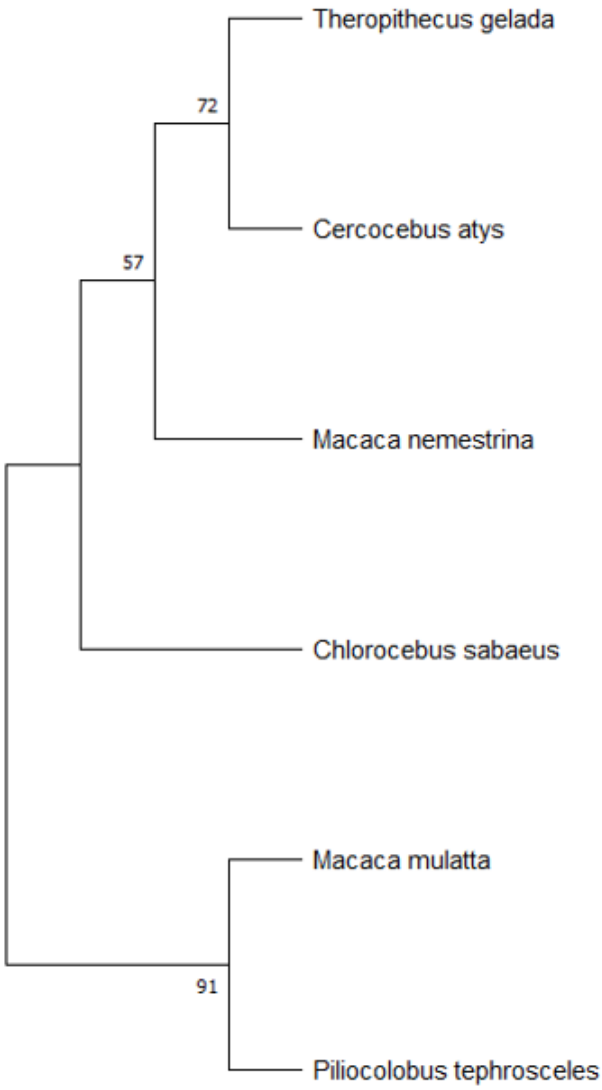

Figure 13: Maximum Likelihood Tree generated for the protein sequences

7.2.4 Analysis

According to the NJ and ML trees, Macaca mulatta and Piliocolobus tephrosceles are the most closely related sequences and are also the most distantly related and equally far from all other sequences. In the MP tree, Theropithecus gelada and Cercocebus atys are the most closely related pair, with Piliocolobus tephrosceles being the most distantly related from all other sequences.

7.2.5 Conclusions

The most closely related to the query sequence, Theropithecus gelada is Cercocebus atys according to the MP, NJ and ML methods.