Do Phenotypically and Genetically Similar Animals have Similar TNF Genes?

Yvonna Leung

Table of Contents

Background

Tumor Necrosis Factor–alpha (TNFα), is known to regulate a multitude of cellular events, such as cell survival, proliferation, differentiation, and cell death [7]. It is a pro-inflammatory cytokine that was originally identified as a cytokine to suppress tumor cell proliferation [7]. Since its early days of discovery, other roles for TNF have emerged. The primary role of TNF is in its immunological response. TNF is able to induce fever, apoptosis, and inflammation [4]. These characteristics can help reduce tumor growth and viral replication. However, dysregulation of TNF can lead to an excess production of TNF. This phenomenon can be detrimental and be a cause of many diseases, including cancer, diabetes, and autoimmune diseases. Patients with these health conditions often have to be treated with TNF inhibitor drugs to suppress the inflammatory response [5].

Since TNF plays a major role in the host response to many diseases, it would be valuable to further understand the TNF gene by analyzing its evolutionary aspects. Evolutionary studies of genes help uncover what role some evolutionary forces, such as natural selection, have on those genes [1]. While most genes are inherited by the organism from its ancestors, other evolutionary forces can effectively change the composition of the gene over time in order for the organism to adapt to a particular environment [6]. By studying the evolutionary relationships of these genes, we can better understand how genetic changes occur within that organism and also in closely related taxa [1].

This study aims to understand how similar or dissimilar the TNF gene is across humans and primates, and whether other mammals have a similar genetic disposition to humans and primates. A total of 50 TNF sequences will be obtained, 25 of which will be primates, and the other 25 of which will be mammals from other taxa. A multiple sequence alignment (MSA) will be built to infer the presence of ancestral relationships between the sequences. A phylogenetic tree will also be drawn to visually represent the evolutionary relationships among the species analyzed. Furthermore, these sequences will be trained through a Hidden Markov Model (HMM) that then later computes a log-Viterbi score. These analyses will help determine the similarity between the species among the primate clade and mammals from other taxa. Because humans are most similar to primates, it is to be expected that the MSA will align similarly and result in low log-Viterbi scores. Since primates may not be genetically similar with other mammals, it is to be expected that there can be a significant difference between the MSA alignments and could result in much higher log-Viterbi scores. We hypothesize that phenotypically and genetically similar animals will have genotypically similar TNF genes.

II. Methods

Data

The TNF sequences of both the primate set, labeled as the “basic set”, and the mammals set, labeled as the “related set”, were obtained from NCBI Genbank database. A Python script was written to grab these files from GenBank [2] using the accession numbers and start and stop locations from the TNF sequences. The names of each species and their accession number are represented in Table 3 in the Appendix as the basic set and Table 4 in the Appendix as the related set. Each individual sequence was then saved in individual FASTA files and then later combined into one separate file for the “basic set” and another file for the “related set”. The links for these files are located in Table 5 of the Appendix.

Procedure

After saving the sequences into a FASTA file, the next step was to align the sequences using CLUSTALW as an MSA tool. BioPython’s library, AlignIO, is a tool used to align the sequences from both the basic set and the related set. These alignment results were then compared with the results from Clustal Omega [3], a publicly available bioinformatics tool. The alignment from BioPython's library was very similar to Clustal Omega's alignment. Clustal Omega's alignment was used to train the HMM since it had easier readability.

BioPython's library, Phylo, was used to draw phylogenetic trees for both the basic set and the related set to visually inspect the evolutionary relationships between the species. The branch lengths were used to determine which species were more similar to each other. BioPython's trees were then compared with the trees from Clustal Omega. Clustal Omega's cladograms and phylogenetic trees were used in this report since they have better visualization.

We then wrote a custom HMM builder class. This allowed us to easily specify a complex HMM using simple terms. The builder used multi-level dictionaries to store probabilities, which allowed us to easily specify transmissions and emissions. After building the HMM specification, the builder computes appropriate A, B, and pi matrices from the built-up dictionaries. Finally, the builder validates these matrices by checking that all rows sum to 1.

We then wrote a function to build our alignment HMM. Here, we incorporated match states, delete states, and insert states at every column position in the alignment training data. current state. Pseudocounts were used where appropriate to handle situations that didn't occur in the training data. Since the probabilities of a sequence containing amino acids that do not occur in the training set is 0, we have to add pseudocounts to every match and delete state for the transition probabilities, as well as emission probabilities for base pairs that weren't seen in the training data. The transition probability from a match state to an insert state is 0.01, the probability from an insert state to itself is 0.01, and the probability from an insert state back to a match state is 0.99. We added other transition probabilities, such as a match state to another match state, by incorporating them into a builder function that then takes in the previous state, the current state, and the probability of that transition.

We also had to consider instances where there are conserved regions, which means that there are no deletions present in these regions. In this instance, the delete probabilities are 0.01. The probability of going from the previous match state to the current match state will be 0.5, while the probability of going from the previous match state to the current delete state will also be 0.5 because there is equal probability of going to either state.

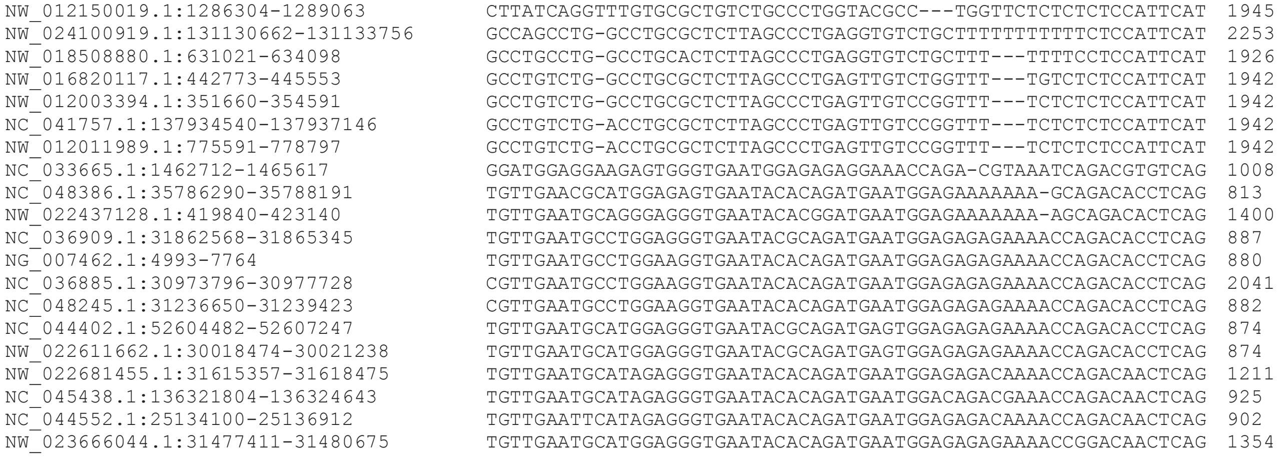
Emission probabilities were calculated for each match state by dividing the total count of each nucleotide in the current column over the length of the column. Pseudocounts were also incorporated for each match and delete state. Insert states do not need pseudocounts since they have balanced emission probabilities, in which we will just add an equal probability to each nucleotide. Finally, all the different types of transition and emission probabilities were added to the HMM builder, including start and end probabilities. Start states and end states are needed because we could start or end with an insertion or a match. Because the start and end states also need to emit something, ‘S’ and ‘E’ emissions were added with emission probability 1.

To calculate the log-Viterbi scores, delta was first calculated. Delta represents a matrix of the best Viterbi scores among any path ending at some state at some timestep. Viterbi scores are probabilities of the most likely state path, though they do not directly reveal the state path on their own. To keep track of the most likely state transitions, we used another matrix called psi. The transition probabilities were calculated from any state to the current state and the emission probabilities were calculated in its current state with the current observation. Delta was then updated to capture the highest score obtained from each timestep. Once delta was fully computed, we called the maximum score among all states at the last timestep the Viterbi score of the whole sequence. If needed, the actual optimal state path could be computed by traversing backwards through psi, starting with the state that maximized Delta at the last timestep.

After both the HMM model and the Viterbi algorithm were setup, training was performed on 80% of the sequences from the basic set, on a region containing both conserved domains and a variable region. This was repeated for the related set. Viterbi scores were calculated against both HMMs for the 20% of sequences left in the basic set and related set. Scores for all HMM-sequence combinations were then captured in the results below.

III. Results

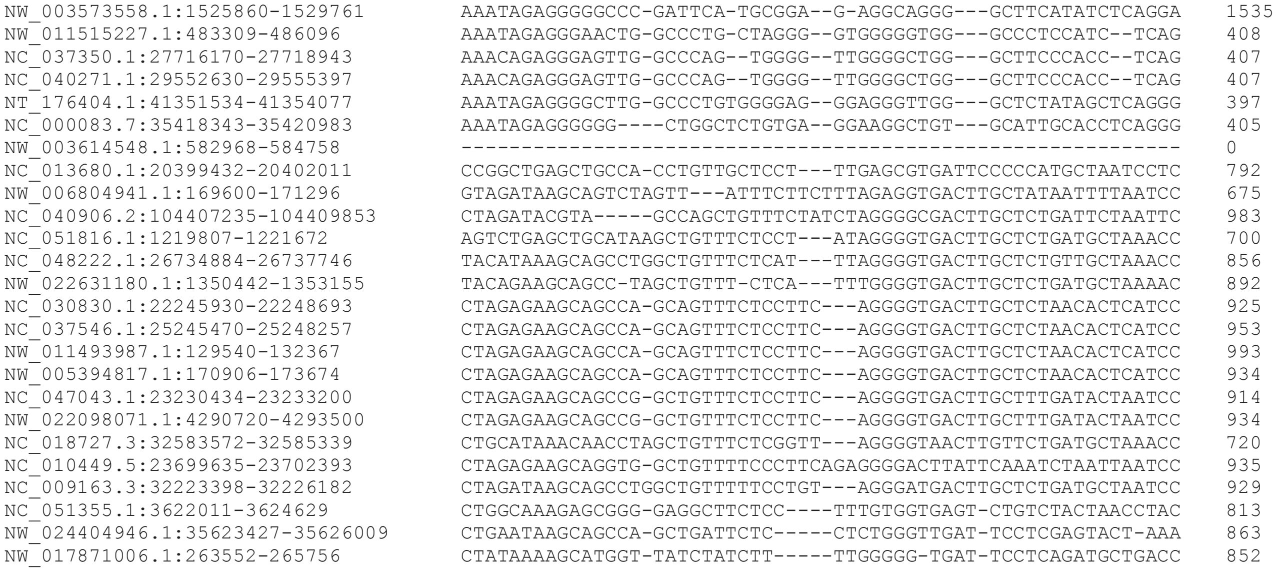
Training was performed on a conserved and variable 250 base pairs region from 2521bp to 2761bp for the basic set. Figure 1 below shows a section of this region. This area was picked because it had a considerable number of conserved regions, yet also contained some variable regions.



```
NW_012150019.1:1286304-1289063        CTTATCAGGTTTGTGCGCTGTCTGCCCTGGTACGCC---TGGTTCTCTCTCTCCATTCAT  1945
NW_024100919.1:131130662-131133756    GCCAGCCTG-GCCTGCGCTCTTAGCCCTGAGGTGTCTGCTTTTTTTTTTTCTCCATTCAT  2253
NW_018508880.1:631021-634098          GCCTGCCTG-GCCTGCACTCTTAGCCCTGAGGTGTCTGCTTT---TTTTCCTCCATTCAT  1926
NW_016820117.1:442773-445553          GCCTGTCTG-GCCTGCGCTCTTAGCCCTGAGTTGTCTGGTTT---TGTCTCTCCATTCAT  1942
NW_012003394.1:351660-354591          GCCTGTCTG-GCCTGCGCTCTTAGCCCTGAGTTGTCCGGTTT---TCTCTCTCCATTCAT  1942
NC_041757.1:137934540-137937146       GCCTGTCTG-ACCTGCGCTCTTAGCCCTGAGTTGTCCGGTTT---TCTCTCTCCATTCAT  1942
NW_012011989.1:775591-778797          GCCTGTCTG-ACCTGCGCTCTTAGCCCTGAGTTGTCCGGTTT---TCTCTCTCCATTCAT  1942
NC_033665.1:1462712-1465617           GGATGGAGGAAGAGTGGGTGAATGGAGAGAGGAAACCAGA-CGTAAATCAGACGTGTCAG  1008
NC_048386.1:35786290-35788191         TGTTGAACGCATGGAGAGTGAATACACAGATGAATGGAGAAAAAAA-GCAGACACCTCAG  813
NW_022437128.1:419840-423140          TGTTGAATGCAGGGAGGGTGAATACACGGATGAATGGAGAAAAAAA-AGCAGACACTCAG  1400
NC_036909.1:31862568-31865345         TGTTGAATGCCTGGAGGGTGAATACGCAGATGAATGGAGAGAGAAAACCAGACACCTCAG  887
NG_007462.1:4993-7764                 TGTTGAATGCCTGGAAGGTGAATACACAGATGAATGGAGAGAGAAAACCAGACACCTCAG  880
NC_036885.1:30973796-30977728         CGTTGAATGCCTGGAAGGTGAATACACAGATGAATGGAGAGAGAAAACCAGACACCTCAG  2041
NC_048245.1:31236650-31239423         CGTTGAATGCCTGGAAGGTGAATACACAGATGAATGGAGAGAGAAAACCAGACACCTCAG  882
NC_044402.1:52604482-52607247         TGTTGAATGCATGGAGGGTGAATACGCAGATGAGTGGAGAGAGAAAACCAGACACCTCAG  874
NW_022611662.1:30018474-30021238      TGTTGAATGCATGGAGGGTGAATACGCAGATGAGTGGAGAGAGAAAACCAGACACCTCAG  874
NW_022681455.1:31615357-31618475      TGTTGAATGCATAGAGGGTGAATACACAGATGAATGGAGAGACAAAACCAGACAACTCAG  1211
NC_045438.1:136321804-136324643       TGTTGAATGCATAGAGGGTGAATACACAGATGAATGGACAGACGAAACCAGACAACTCAG  925
NC_044552.1:25134100-25136912         TGTTGAATTCATAGAGGGTGAATACACAGATGAATGGAGAGACAAAACCAGACAACTCAG  902
NW_023666044.1:31477411-31480675      TGTTGAATGCATGGAGGGTGAATACACAGATGAATGGAGAGAGAAAACCGGACAACTCAG  1354
```

Figure 1: A section of the conserved and variable region from the Basic set MSA alignment from Clustal Omega.

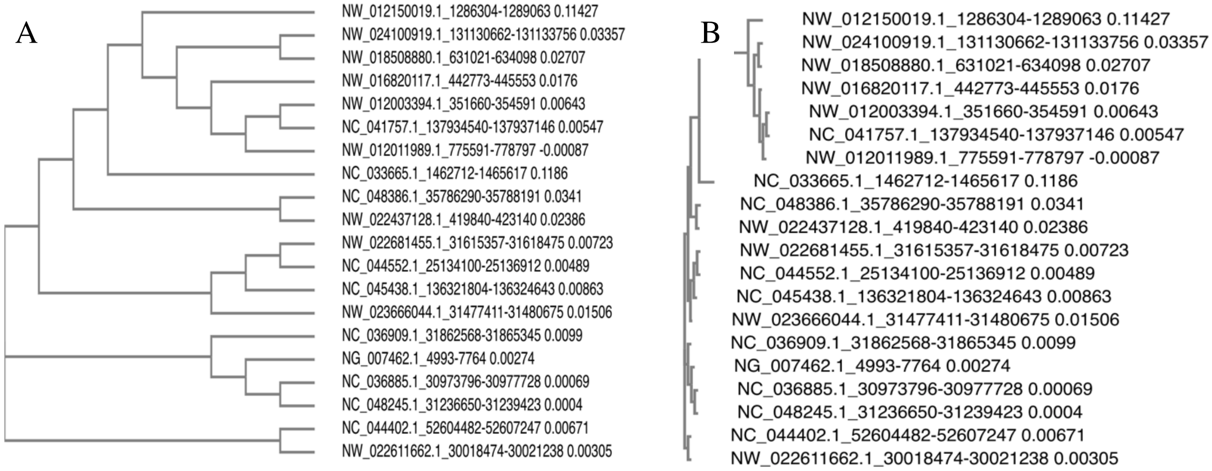
Training was also performed on a conserved and variable 250 base pairs region from 1561bp to 1801bp for the related set as shown in Figure 2.



```
NW_003573558.1:1525860-1529761       AAATAGAGGGGGCCC-GATTCA-TGCGGA--G-AGGCAGGG---GCTTCATATCTCAGGA    1535
NW_011515227.1:483309-486096         AAATAGAGGGAACTG-GCCCTG-CTAGGG--GTGGGGGTGG---GCCCTCCATC--TCAG    408
NC_037350.1:27716170-27718943        AAACAGAGGGAGTTG-GCCCAG--TGGGG--TTGGGGCTGG---GCTTCCCACC--TCAG    407
NC_040271.1:29552630-29555397        AAACAGAGGGAGTTG-GCCCAG--TGGGG--TTGGGGCTGG---GCTTCCCACC--TCAG    407
NT_176404.1:41351534-41354077        AAATAGAGGGGCTTG-GCCCTGTGGGGAG--GGAGGGTTGG---GCTCTATAGCTCAGGG    397
NC_000083.7:35418343-35420983        AAATAGAGGGGGG----CTGGCTCTGTGA--GGAAGGCTGT---GCATTGCACCTCAGGG    405
NW_003614548.1:582968-584758         ------------------------------------------------------------    0
NC_013680.1:20399432-20402011        CCGGCTGAGCTGCCA-CCTGTTGCTCCT---TTGAGCGTGATTCCCCCATGCTAATCCTC    792
NW_006804941.1:169600-171296         GTAGATAAGCAGTCTAGTT---ATTTCTTCTTTAGAGGTGACTTGCTATAATTTTAATCC    675
NC_040906.2:104407235-104409853      CTAGATACGTA-----GCCAGCTGTTTCTATCTAGGGGCGACTTGCTCTGATTCTAATTC    983
NC_051816.1:1219807-1221672          AGTCTGAGCTGCATAAGCTGTTTCTCCT---ATAGGGGTGACTTGCTCTGATGCTAAACC    700
NC_048222.1:26734884-26737746        TACATAAAGCAGCCTGGCTGTTTCTCAT---TTAGGGGTGACTTGCTCTGTTGCTAAACC    856
NW_022631180.1:1350442-1353155       TACAGAAGCAGCC-TAGCTGTTT-CTCA---TTTGGGGTGACTTGCTCTGATGCTAAAAC    892
NC_030830.1:22245930-22248693        CTAGAGAAGCAGCCA-GCAGTTTCTCCTTC---AGGGGTGACTTGCTCTAACACTCATCC    925
NC_037546.1:25245470-25248257        CTAGAGAAGCAGCCA-GCAGTTTCTCCTTC---AGGGGTGACTTGCTCTAACACTCATCC    953
NW_011493987.1:129540-132367         CTAGAGAAGCAGCCA-GCAGTTTCTCCTTC---AGGGGTGACTTGCTCTAACACTCATCC    993
NW_005394817.1:170906-173674         CTAGAGAAGCAGCCA-GCAGTTTCTCCTTC---AGGGGTGACTTGCTCTAACACTCATCC    934
NC_047043.1:23230434-23233200        CTAGAGAAGCAGCCG-GCTGTTTCTCCTTC---AGGGGTGACTTGCTTTGATACTAATCC    914
NW_022098071.1:4290720-4293500       CTAGAGAAGCAGCCG-GCTGTTTCTCCTTC---AGGGGTGACTTGCTTTGATACTAATCC    934
NC_018727.3:32583572-32585339        CTGCATAAACAACCTAGCTGTTTCTCGGTT---AGGGGTAACTTGTTCTGATGCTAAACC    720
NC_010449.5:23699635-23702393        CTAGAGAAGCAGGTG-GCTGTTTTCCCTTCAGAGGGGACTTATTCAAATCTAATTAATCC    935
NC_009163.3:32223398-32226182        CTAGATAAGCAGCCTGGCTGTTTTTCCTGT---AGGGATGACTTGCTCTGATGCTAATCC    929
NC_051355.1:3622011-3624629          CTGGCAAAGAGCGGG-GAGGCTTCTCC----TTTGTGGTGAGT-CTGTCTACTAACCTAC    813
NW_024404946.1:35623427-35626009     CTGAATAAGCAGCCA-GCTGATTCTC-----CTCTGGGTTGAT-TCCTCGAGTACT-AAA    863
NW_017871006.1:263552-265756         CTATAAAAGCATGGT-TATCTATCTT-----TTGGGGG-TGAT-TCCTCAGATGCTGACC    852
```

Figure 2: A section of the conserved and variable region from the Basic set MSA alignment from Clustal Omega.

After the MSA alignment was performed, the cladograms and phylogenetic trees were obtained from our code and compared against Clustal Omega. Figure 3A and 3B show the cladogram and phylogenetic tree for the basic set and Figure 4A and 4B show the cladogram and phylogenetic tree for the related set all from Clustal Omega.



A
NW_012150019.1_1286304-1289063 0.11427
NW_024100919.1_131130662-131133756 0.03357
NW_018508880.1_631021-634098 0.02707
NW_016820117.1_442773-445553 0.0176
NW_012003394.1_351660-354591 0.00643
NC_041757.1_137934540-137937146 0.00547
NW_012011989.1_775591-778797 -0.00087
NC_033665.1_1462712-1465617 0.1186
NC_048386.1_35786290-35788191 0.0341
NW_022437128.1_419840-423140 0.02386
NW_022681455.1_31615357-31618475 0.00723
NC_044552.1_25134100-25136912 0.00489
NC_045438.1_136321804-136324643 0.00863
NW_023666044.1_31477411-31480675 0.01506
NC_036909.1_31862568-31865345 0.0099
NG_007462.1_4993-7764 0.00274
NC_036885.1_30973796-30977728 0.00069
NC_048245.1_31236650-31239423 0.0004
NC_044402.1_52604482-52607247 0.00671
NW_022611662.1_30018474-30021238 0.00305
B
NW_012150019.1_1286304-1289063 0.11427
NW_024100919.1_131130662-131133756 0.03357
NW_018508880.1_631021-634098 0.02707
NW_016820117.1_442773-445553 0.0176
NW_012003394.1_351660-354591 0.00643
NC_041757.1_137934540-137937146 0.00547
NW_012011989.1_775591-778797 -0.00087
NC_033665.1_1462712-1465617 0.1186
NC_048386.1_35786290-35788191 0.0341
NW_022437128.1_419840-423140 0.02386
NW_022681455.1_31615357-31618475 0.00723
NC_044552.1_25134100-25136912 0.00489
NC_045438.1_136321804-136324643 0.00863
NW_023666044.1_31477411-31480675 0.01506
NC_036909.1_31862568-31865345 0.0099
NG_007462.1_4993-7764 0.00274
NC_036885.1_30973796-30977728 0.00069
NC_048245.1_31236650-31239423 0.0004
NC_044402.1_52604482-52607247 0.00671
NW_022611662.1_30018474-30021238 0.00305

Figure 3: A. Cladogram of the Basic Set displayed from Clustal Omega. B. Phylogenetic tree of the Basic Set from Clustal Omega.

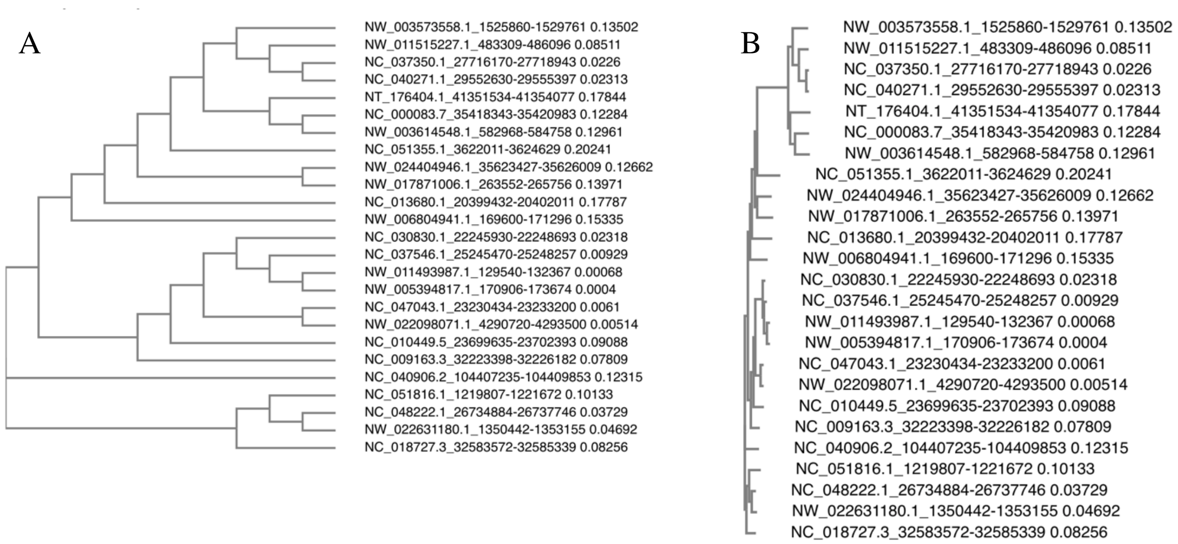

A
NW_003573558.1_1525860-1529761 0.13502
NW_011515227.1_483309-486096 0.08511
NC_037350.1_27716170-27718943 0.0226
NC_040271.1_29552630-29555397 0.02313
NT_176404.1_41351534-41354077 0.17844
NC_000083.7_35418343-35420983 0.12284
NW_003614548.1_582968-584758 0.12961
NC_051355.1_3622011-3624629 0.20241
NW_024404946.1_35623427-35626009 0.12662
NW_017871006.1_263552-265756 0.13971
NC_013680.1_20399432-20402011 0.17787
NW_006804941.1_169600-171296 0.15335
NC_030830.1_22245930-22248693 0.02318
NC_037546.1_25245470-25248257 0.00929
NW_011493987.1_129540-132367 0.00068
NW_005394817.1_170906-173674 0.0004
NC_047043.1_23230434-23233200 0.0061
NW_022098071.1_4290720-4293500 0.00514
NC_010449.5_23699635-23702393 0.09088
NC_009163.3_32223398-32226182 0.07809
NC_040906.2_104407235-104409853 0.12315
NC_051816.1_1219807-1221672 0.10133
NC_048222.1_26734884-26737746 0.03729
NW_022631180.1_1350442-1353155 0.04692
NC_018727.3_32583572-32585339 0.08256
B
NW_003573558.1_1525860-1529761 0.13502
NW_011515227.1_483309-486096 0.08511
NC_037350.1_27716170-27718943 0.0226
NC_040271.1_29552630-29555397 0.02313
NT_176404.1_41351534-41354077 0.17844
NC_000083.7_35418343-35420983 0.12284
NW_003614548.1_582968-584758 0.12961
NC_051355.1_3622011-3624629 0.20241
NW_024404946.1_35623427-35626009 0.12662
NW_017871006.1_263552-265756 0.13971
NC_013680.1_20399432-20402011 0.17787
NW_006804941.1_169600-171296 0.15335
NC_030830.1_22245930-22248693 0.02318
NC_037546.1_25245470-25248257 0.00929
NW_011493987.1_129540-132367 0.00068
NW_005394817.1_170906-173674 0.0004
NC_047043.1_23230434-23233200 0.0061
NW_022098071.1_4290720-4293500 0.00514
NC_010449.5_23699635-23702393 0.09088
NC_009163.3_32223398-32226182 0.07809
NC_040906.2_104407235-104409853 0.12315
NC_051816.1_1219807-1221672 0.10133
NC_048222.1_26734884-26737746 0.03729
NW_022631180.1_1350442-1353155 0.04692
NC_018727.3_32583572-32585339 0.08256

Figure 4: A. Cladogram of the Related Set displayed from Clustal Omega. B. Phylogenetic tree of the Related Set from Clustal Omega.

Viterbi Scores were calculated from our Viterbi algorithm. The log-Viterbi scores were computed for the remaining 20% sequences for both the basic set and related set and compared against both the trained 80%, or 20, sequences of the basic set and the trained 80% of sequences of the related set. Their values are displayed below in Table 1 and Table 2.

Test Sequence from Basic Set	Accession Number	Log-Viterbi Scores: From Trained Basic Set on 20 sequences	Log-Viterbi Scores: From Trained Related Set on 20 sequences
Gelada	NC_037671	-756.85	-609.25
Small-Earged Galago	NW_003852441	-798.31	-616.43
Angola Colobus	NW_012118154	-780.07	-563.01
Drill	NW_012104920	-812.35	-596.18
Crab-Eating Macaque	NC_022287	-799.43	-505.25

Table 1: Log-Viterbi scores computed from the basic test set trained from 2521bp to 2761bp

Test Sequence from Related Set	Accession Number	Log-Viterbi Scores: From Trained Basic Set on 20 sequences	Log-Viterbi Scores: From Trained Related Set on 20 sequences
Giant Panda	NC_048222	-799.94	-563.95
N.American River Otter	NW_022631180	-inf	-575.65
Angola Colobus	NW_022098071	-759.77	-579.29
Beluga Whale	NW_005394817	-758.69	-602.94
W. European Hedgehog	NW_006804941	-inf	-inf

Table 2: Log-Viterbi scores computed from the related test set trained from 1561bp to 1801bp.

IV. Discussion

After aligning the 25 human and primate TNF sequences for the basic set using multiple sequence alignment, it was found that there were many indel regions in the beginning of the alignment. Only after the alignment reached 60 base pairs did we finally begin to see some alignment between the first 7 species. However, some of the other species in the basic set continued to not show alignment for hundreds of base pairs. This observation could be explained by the fact that not all primates are that closely related genetically. One example of species that do show close alignment are humans and chimpanzees. This would make sense as it had been found that humans are most closely related to chimpanzees and pygmy chimpanzees. The high volume of indels could also be attributed to the observation that not all the sequence lengths are the same. As we traverse through the base pairs, we do begin to see more conserved domains that contain some variable regions. These regions would be good candidates for the HMM to train on.

Similarly, we also saw that there were many regions of indels in the beginning of the alignment in the related set. The more conserved regions occurred later in the alignment, indicating that these conservations could be related to the functional property of TNF. The related set's alignment results were not surprising since having so many indels could be due to how these mammals could be vastly different genetically.

From the phylogenetic tree results of the basic set, it was found that there are many closely related primates. For instance, the branch length between chimpanzees and pygmy chimpanzees was among one of the shorter branch lengths. As shown in Figure 5A, the next branch length was that of human, indicating that humans are closely related to both these chimpanzees. In the phylogenetic tree results of the related set, it was found that mammals of similar taxa were closely related to each other. As shown in Figure 5B, the branch length between cattle and sheep was among one of the shorter branch lengths. These results correlate to our current understanding of the cattle family, which contains sheep.

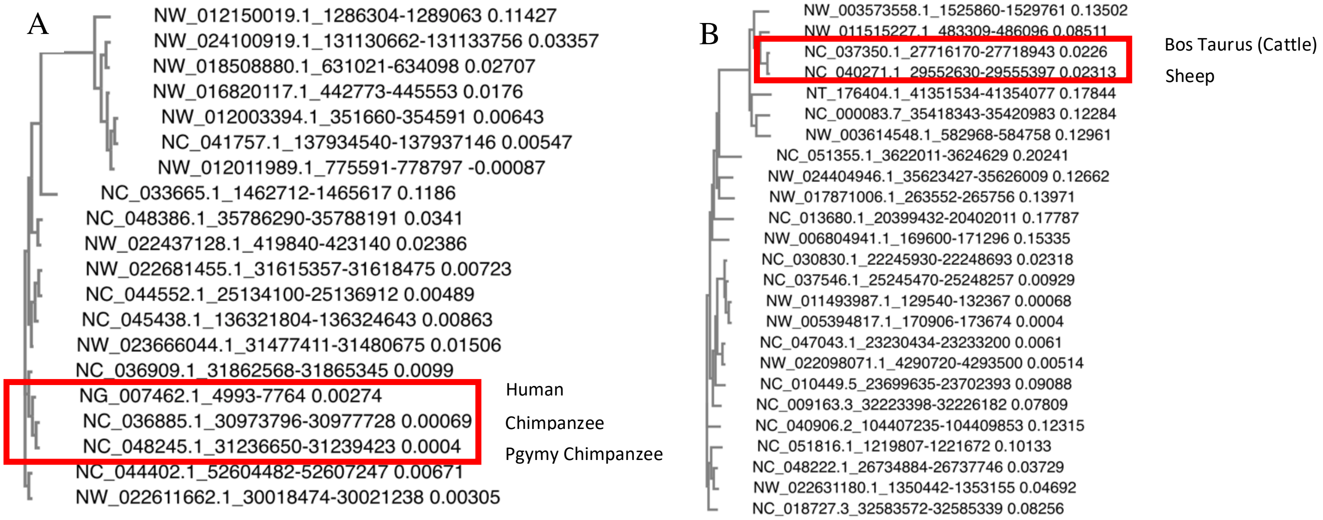

A
NW_012150019.1_1286304-1289063 0.11427
NW_024100919.1_131130662-131133756 0.03357
NW_018508880.1_631021-634098 0.02707
NW_016820117.1_442773-445553 0.0176
NW_012003394.1_351660-354591 0.00643
NC_041757.1_137934540-137937146 0.00547
NW_012011989.1_775591-778797 -0.00087
NC_033665.1_1462712-1465617 0.1186
NC_048386.1_35786290-35788191 0.0341
NW_022437128.1_419840-423140 0.02386
NW_022681455.1_31615357-31618475 0.00723
NC_044552.1_25134100-25136912 0.00489
NC_045438.1_136321804-136324643 0.00863
NW_023666044.1_31477411-31480675 0.01506
NC_036909.1_31862568-31865345 0.0099
NG_007462.1_4993-7764 0.00274
NC_036885.1_30973796-30977728 0.00069
NC_048245.1_31236650-31239423 0.0004
NC_044402.1_52604482-52607247 0.00671
NW_022611662.1_30018474-30021238 0.00305
Human
Chimpanzee
Pgymy Chimpanzee
B
NW_003573558.1_1525860-1529761 0.13502
NW_011515227.1_483309-486096 0.08511
NC_037350.1_27716170-27718943 0.0226
NC_040271.1_29552630-29555397 0.02313
NT_176404.1_41351534-41354077 0.17844
NC_000083.7_35418343-35420983 0.12284
NW_003614548.1_582968-584758 0.12961
NC_051355.1_3622011-3624629 0.20241
NW_024404946.1_35623427-35626009 0.12662
NW_017871006.1_263552-265756 0.13971
NC_013680.1_20399432-20402011 0.17787
NW_006804941.1_169600-171296 0.15335
NC_030830.1_22245930-22248693 0.02318
NC_037546.1_25245470-25248257 0.00929
NW_011493987.1_129540-132367 0.00068
NW_005394817.1_170906-173674 0.0004
NC_047043.1_23230434-23233200 0.0061
NW_022098071.1_4290720-4293500 0.00514
NC_010449.5_23699635-23702393 0.09088
NC_009163.3_32223398-32226182 0.07809
NC_040906.2_104407235-104409853 0.12315
NC_051816.1_1219807-1221672 0.10133
NC_048222.1_26734884-26737746 0.03729
NW_022631180.1_1350442-1353155 0.04692
NC_018727.3_32583572-32585339 0.08256
Bos Taurus (Cattle)
Sheep

Figure 5: Comparison of phylogenetic trees. A: Phylogenetic tree of the basic set. B: Phylogenetic tree of the related set.

After analyzing the MSA and the phylogenetic tree results, the sequences were then used to train an HMM for scoring other sequences. Comparing the log-Viterbi scores could help us determine the similarity between the species among the primate clade and mammals from the other taxa. The higher the log-Viterbi score, the more similar the sequences in the test set are to the training set. Likewise, the lower the log-Viterbi score, the less similar the sequences in the test set are to the training set. Because the primates are in the same clade and therefore are phenotypically and genetically similar, it was expected that the basic training set would produce higher log-Viterbi scores to the basic test set. Since primates are not genetically similar to other mammals in the related set, it was expected that the basic training set would produce lower log-Viterbi scores to the related test set. However, the opposite result was observed in Table 1, in that the trained basic set produced lower log-Viterbi scores for the basic test set compared with the related test set. This result could be due to the regions selected for training and testing.

Because the subset chosen to train was somewhat arbitrarily chosen, the quality of the alignment could be highly dependent on the regions used. Although the chosen region appeared to have contained a good combination of conserved domains and variable region, perhaps it still wasn't conserved enough. There could be higher variability in the sequences than expected, and thus led to higher log-Viterbi scores.

We would expect that a TNF sequence in the related set would be more closely aligned with sequences in the related set than in the primate set. Hence, the related test set would produce higher log-Viterbi scores to the related training set. Likewise, the related test set would produce lower log-Viterbi scores to the basic training set. The results observed in Table 2 follow the hypothesis. The related set that was used to train an HMM on the related test set show better alignment compared with the basic set. However, there were a couple of scores that had a result of "-inf". This could be due to the sequence having a zero probability of occurring, either because of a length mismatch or possibly an oversight in our code. These results indicated that mammals in the related set are still more closely aligned with each other than the primates in the basic set.

The results of this research were somewhat inconclusive, neither strongly supporting nor refuting the hypothesis. This is because the MSA supports the hypothesis, but the HMM does not. More analysis would need to be done. If there was more time to improve the findings of this project, multiple regions of the alignment could be tested to determine which region is the best for training the HMM model. Overall, the log-Viterbi scores were still very low, which could be a result of the chosen region's poor alignment quality. Since the TNF sequences ranged upward of 2000 bp, it was also difficult to train the HMM model well. Training on the whole gene, rather than a single section could also produce interesting results. Perhaps, it would also have been more fruitful to find a gene that had a shorter sequence first before testing it on a much longer sequence or look at multiple genes to find any similarity between them.

References

[1] Awasthi G, Singh S, Dash AP, Das A. (2008). Genetic characterization and evolutionary inference of TNF-α through computional analysis. Braz J Infect Dis 12(5): 1678-4391.

[2] Blast: Basic local alignment search tool. (n.d.). Retrieved April 07, 2021, from https://blast.ncbi.nlm.nih.gov/

[3] Clustal Omega. Retrieved April 07, 2021, from https://www.ebi.ac.uk/Tools/msa/clustalo/

[4] Josephs SF, Ichim TE, Prince SM, Kesari S, Marincola, FM, Escobedo AR, Jafri A. (2018). Unleashing endogenous TNF-alpha as a cancer immunotherapeutic. Journal of Translational Medicine 16(242): 1-8.

[5] Kryzysztof L, Kuzawinska O, Balkowiec-Iskra E. (2014). Tumor necrosis factor inhibitors – state of knowledge. Arch Med Sci. 10(6): 1175-1185.

[6] Peaston AE, Whitelaw E. (2006). Epigenetics and phenotypic variation in mammals. Mamm Genome 17(5): 365-374.

[7] Wang X, Lin Y. (2008). Tumor necrosis factor and cancer, buddies or foes? Acta Pharmacol Sin 29(11): 1275-1288.

Appendix

Species Name	Length of TNF Sequence	Accession Number	Start Seq, End Seq
Human	2772 bp	NG_007462	4993, 7764
Rhesus Monkey	2607 bp	NC_041757	137934540, 137937146
Chimpanzee	3933 bp	NC_036885	30973796, 30977728
Sumatran Orangutan	2778 bp	NC_036909	31862568, 31865345
White-Tufted-Ear Marmoset	1902 bp	NC_048386	35786290, 35788191
Northern White-Cheeked Gibbon	2766 bp	NC_044402	52604482, 52607247
Sooty Mangabey	2932 bp	NW_012003394	351660, 354591
Francois's Langur	3119 bp	NW_022681455	31615357, 31618475
Tufted Capuchin	3301 bp	NW_022437128	419840, 423140
Blank Snub-Nosed Monkey	2781 bp	NW_016820117	442773, 445553
Gray Mouse Lemur	2906 bp	NC_033665	1462712, 1465617
Coquerel's Sifaka	2760 bp	NW_012150019	1286304, 1289063
Ma's Night Monkey	3078 bp	NW_018508880	631021, 634098
Pig-Tailed Macaque	3207 bp	NW_012011989	775591, 778797
Golden Snub-Nosed Monkey	2813 bp	NC_044552	25134100, 25136912
Bolivian Squirrel Monkey	3095 bp	NW_024100919	131130662, 131133756
Silvery Gibbon	2765 bp	NW_022611662	30018474, 30021238
Pgymy Chimpanzee (Bonobo)	2774 bp	NC_048245	31236650, 31239423
Green Monkey	3265 bp	NW_023666044	31477411, 31480675
Ugandan Red Colobus	2840 bp	NC_045438	136321804, 136324643
Gelada	2767 bp	NC_037671	34037747, 34040513
Small-Earged Galago	3509 bp	NW_003852441	11768637, 11772145
Angola Colobus	3434 bp	NW_012118154	3616485, 3619918
Drill	3436 bp	NW_012104920	1288916, 1292351
Crab-Eating Macaque	3669 bp	NC_022287	7627176, 7630844

Table 3: Primate species names for the basic set and their accession numbers

Species Name	**Length of TNF Sequence**	**Accession Number**	**Start Seq, End Seq**
Mouse	2641 bp	NC_000083	35418343, 35420983
Norway Rat	2619 bp	NC_051355	3622011, 3624629
Pale Spear-Nosed Bat	2619 bp	NC_040906	104407235, 104409853
Bos Taurus (Cattle)	2774 bp	NC_037350	27716170, 27718943
Sus scrofa (pig)	2759 bp	NC_010449	23699635, 23702393
Canis Lupis (dog)	1866 bp	NC_051816	1219807, 1221672
Rabbit	2580 bp	NC_013680	20399432, 20402011
Sheep	2768 bp	NC_040271	29552630, 29555397
Horse	2785 bp	NC_009163	32223398, 32226182
Bison	2828 bp	NW_011493987	129540, 132367
Felis catus (Cat)	1768 bp	NC_018727	32583572, 32585339
Water Buffalo	2788 bp	NC_037546	25245470, 25248257
Common Bottlenose Dolphin	2767 bp	C_047043	23230434, 23233200
Guinea Pig	2544 bp	NT_176404	41351534, 41354077
Bactrian Camel	2788 bp	NW_011515227	483309, 486096
Goat	2764 bp	NC_030830	22245930, 22248693
Chinese Hamster	1791 bp	NW_003614548	582968, 584758
African Savanna Elephant	3902 bp	NW_003573558	1525860, 1529761
Squirrel	2583 bp	NW_024404946	35623427, 35626009
Beaver	2205 bp	NW_017871006	263552, 265756
Giant Panda	2863 bp	NC_048222	26734884, 26737746
N.A. River Otter	2714 bp	NW_022631180	1350442, 1353155
Beluga Whale	2781 bp	NW_022098071	4290720, 4293500
Wild Yak	2769 bp	NW_005394817	170906, 173674
W. Euro. Hedgehog	1697 bp	NW_006804941	169600, 171296

Table 4: Mammals species names for the related set and their accession numbers

Set Name	Fasta File Link Shared on Google Drive
Basic Set	https://drive.google.com/file/d/1AYFgjGRFfyuf6LcVzfrEkhxot3I5H7Ta/view?usp=sharing
Related Set	https://drive.google.com/file/d/1Iww5MEPz1jmd0IC0Fu_zeWY2_exH_YsL/view?usp=sharing

Table 5: Fasta files containing MSA alignment for both the basic set and the related set.