Investigating the K-mer Spectra of Virus Genomes to Uncover Genome Evolution Mechanisms

Indronil Bhattacharjee, Jaspreet Thind

Abstract

The distribution of k-mers in genome sequences can offer insights into their composition and evolution. In this study, we investigate the k-mer spectra of virus sequences ranging from SARS-CoV-2 to HIV, to identify underlying patterns in the k-mer spectra of genome sequences and gain insights into genome evolution mechanisms. To accomplish this, we computed the Hamming distance matrix between a set of DNA sequences, constructed a phylogenetic tree using the UPGMA algorithm, and calculated the most frequent 9-mers for each virus. Additionally, we analyzed the most frequent 9-mers for HIV, Adenovirus, Ebola, and Hepatitis-B. We present a resulting phylogenetic tree of coronaviruses and perform a literature review to validate our hypothesis regarding their evolution. Overall, our findings offer insights into the underlying mechanisms that drive genome evolution and can have implications for understanding the spread and treatment of viral diseases.

Introduction

The k-mers found in DNA sequences carry significant insights about the composition and evolution of the sequence. Our aim is to investigate the k-mer spectra of genome sequences in order to unveil the rules governing sequence evolution. The non-random attribute was commonly employed to anticipate and detect operational areas such as promoter regions (Chan & Kibler, 2005; Hariharan et al., 2013), enhancers (Lee et al., 2011), CpG island sequences (Chae et al., 2013; Hashim & Abdullah, 2015), conservative non-coding sequences, and transcriptional start sites (W. Chen et al., 2014). Carl Woese's work made a significant breakthrough by utilizing the unique features of nucleotide sequences to depict the evolutionary relationships between species (Pace et al., 2012). He utilized conserved (SSU) rRNA sequences to build phylogenetic trees and introduced the Three Domain theory (Woese & Fox, 1977). Researchers have examined the k-mer spectral distributions of genome sequences as a potential area of interest. In this regard, Chen conducted an analysis of the k-mer spectra ($k = 6$) of 9 genome sequences (Y.-H. Chen et al., 2005), followed by Beny Chor's study of the k-mer spectra ($k = 7–11$) of approximately 100 genome sequences (Chor et al., 2009). Their observations indicated that some vertebrates, fungi, and prokaryotes had unimodal k-mer spectra, while four-clawed mammals displayed tri-modal k-mer spectra. The objective of our study was to examine the distribution of 9-mer subsets in virus sequences, ranging from SARS-CoV-2 to HIV, to identify underlying patterns in k-mer spectra of genome sequences and to gain insights into genome evolution mechanisms.

Methods

Task 1: This piece of code is used to compute the Hamming distance matrix between a set of DNA sequences, where the Hamming distance indicates the number of positions at which two sequences have different nucleotides. The hamming_distance function computes the Hamming

distance between two sequences, while pairwise_hamming_distance creates a matrix of pairwise distances by iterating through all possible combinations of sequences and invoking the hamming_distance function for each pair. Subsequently, the distance_matrix obtained is supplied as input to the create_matrix function, which arranges the matrix in a format suitable for building a phylogenetic tree. Finally, the matrix is converted into a Bio.Phylo.TreeConstruction._Matrix object and saved in the variable m.

Task 2: The code provided allows for the construction of a phylogenetic tree with a given method. The build_tree function takes in a distance matrix, method, and an optional outgroup parameter, and constructs a rooted tree using either the UPGMA or Neighbor Joining algorithm based on the method parameter. First, the distance matrix is converted to a Bio.Phylo DistanceMatrix object, and the appropriate algorithm is chosen based on the method parameter. If the method is 'upgma', the UPGMA algorithm is used, otherwise, the Neighbor Joining algorithm is used. Additionally, if an outgroup parameter is provided, the root of the tree is placed at the specified outgroup. The function returns the resulting rooted tree.
Next, the code constructs a UPGMA tree using the build_tree function and the distance matrix m generated earlier. The tree is then made more readable with a ladderize() method call. Finally, the tree is displayed with a graphical representation using Phylo.draw and a text-based representation using Phylo.draw_ascii.

Task 3: small_parsimony(tree) is a function that takes a phylogenetic tree object as input and returns the same object with inferred ancestral sequences in each node. The function applies the small parsimony algorithm, which is a dynamic programming approach to infer the most likely nucleotide sequence at each node of the tree, given the nucleotide sequences of the terminal nodes (leaves).
The function initializes the score and sequence for each node in the tree, and then iterates over the nodes in a post-order traversal. For each node, it calculates the minimum parsimony score for each possible nucleotide at each position in the sequence, by considering the scores of its child nodes. It then updates the score and sequence of the parent node if a lower score is found. Finally, the function iterates over the nodes again and sets the name attribute of each internal node to the inferred ancestral sequence. The function prints out the inferred ancestral sequences for each internal node.

Task 4: The code downloads fasta sequences of virus genomes including SARS-CoV-2, SARS-CoV, MERS-CoV, Bat-CoV and Pangolin CoV from the NCBI database. The function BetterFrequentWords, developed in Project 1, is used to calculate the kmers of a user-defined length. The most frequent kmers from each virus are then obtained. Next, the pairwise hamming distance is calculated using the pairwise_hamming_distance function. The resulting distance matrix is used as input to generate a phylogenetic tree.

Task 5: A literature review was performed to construct and validate a hypothesis from the resulting phylogenetic tree of coronaviruses.

Task E1: The code downloads fasta sequences of virus genomes including HIV, Human Adenovirus, Ebola and Hepatitis-B from the NCBI database. The function BetterFrequentWords, developed in Project 1, is used to calculate the kmers of a user-defined length. The most frequent kmers from each virus are then obtained. Next, the pairwise hamming distance is calculated using the pairwise_hamming_distance function. The resulting distance matrix is used as input to generate a phylogenetic tree.

Task E2: A literature review was performed to construct and validate a hypothesis from the resulting phylogenetic tree of the given viruses.

Results

Most Frequent 9-mers from Task 4

SARS-CoV-2: 'TAAACGAAC'
MERS-CoV: 'TTAACGAAC'
Bat-CoV: 'TAAACGAAC'
SARS-CoV: 'TAAACGAAC'
Pangolin-CoV: 'TAATGGTAA'

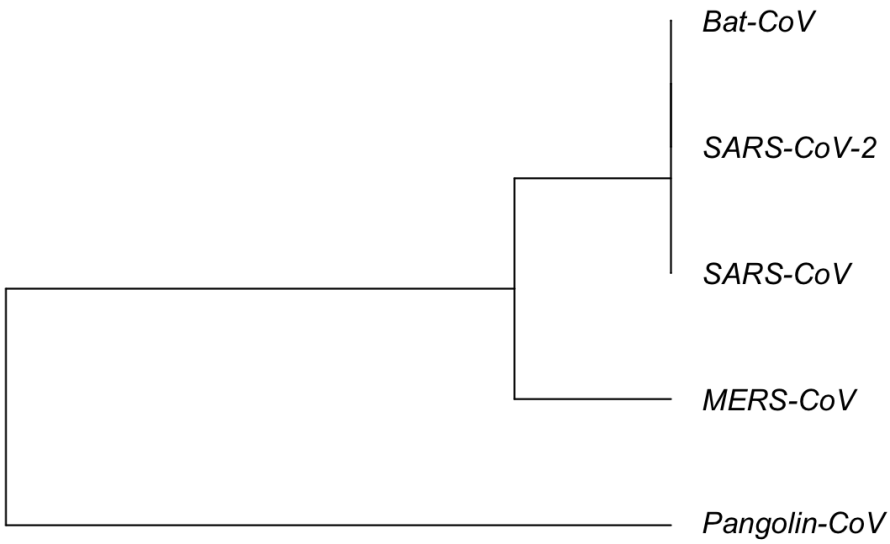

Figure 1: A resulting phylogenetic tree using Phylo.draw function for task 4.

Most Frequent 9-mers from E1:

HIV (3): ‘AAAGAAAAA', 'AAGAAAAAA', 'TTAAAAGAA'
Adenovirus (1): 'GGCGGCGGC,
Ebola (3): ‘GAAGATTAA', 'TTAAGAAAA', 'TAAGAAAAA',
Hepatitis-B (25): 'TTCTTGTTG', 'CAATTTTCT', 'CACCAGCAC',
'TGGGAGTGG','GGGAGTGGG', 'CCCCCACTG', 'AAACAAAAA', 'GGTTGGGGC'
'ATGTCAACG', 'TGTCAACGA', 'GTCTTTTGG', 'GTCTGTGCC', 'TCTGTGCCA',
'ACCCCTCCC', 'GGGGCTCTA', 'CGACCGACC', 'CTGTGCCTT',
'TTCACCTCT', 'TCACCTCTG', 'CACCTCTGC', 'CTTGGACTC', 'GAGTGGGAG',
'CTCCTCCTC', 'TCCTCCTCC', 'AAGGTGGGA']

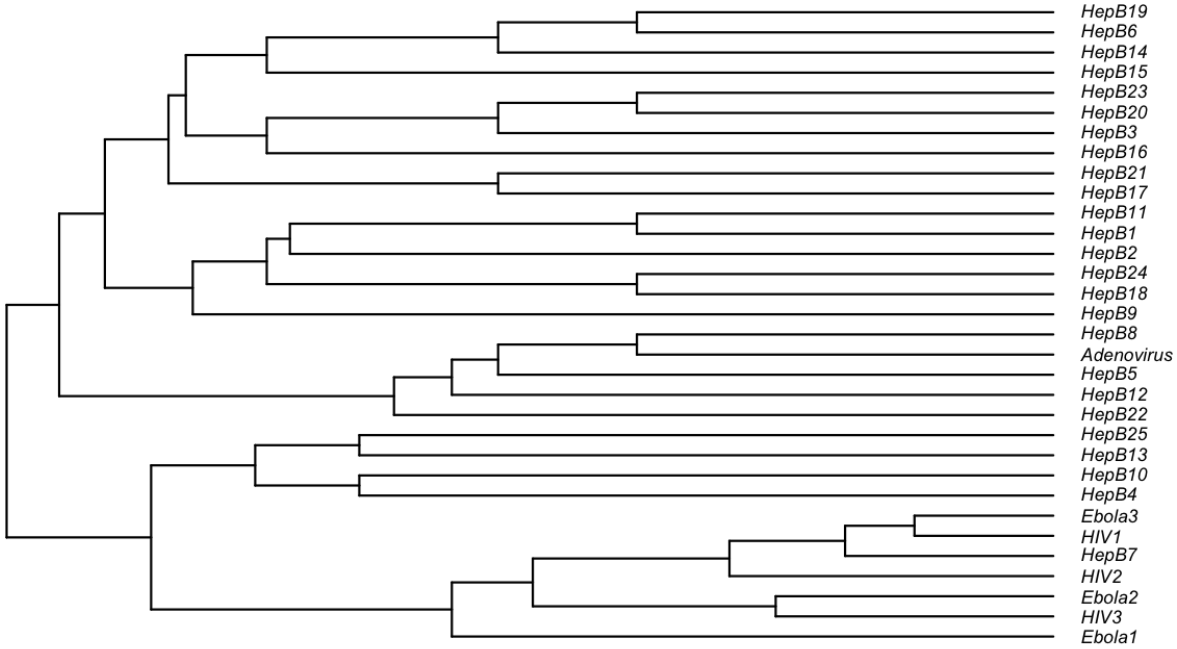

Figure 2: A resulting phylogenetic tree using Phylo.draw function for E1.

Discussion

The tree (Figure 1) is represented in a hierarchical manner, with the Pangolin-CoV virus as the root of the tree. The MERS-CoV virus is the closest relative to the Pangolin-CoV virus, followed by a branch that splits into two: one leading to another clade that consists of SARS-CoV-2, SARS-CoV and Bat-CoV virus. Interestingly, this tree suggests that all three viruses are more closely related to the than to the other viruses included in the analysis. Our hypothesis suggests that there is a strong genetic similarity between Bat-CoV and SARS-CoV-1-2, and that pangolin-CoV is the original common ancestor of all coronaviruses.

SARS-CoV-2 shares 96% (Zhou et al., 2020) of the genome sequence with bat CoV and SARS-CoV share 79.6% (Cheng & Shan, 2020) of genomic similarity with SARS-CoV-2 which corroborates the fact the most frequent 9-mers in these three viruses are conserved. It has been reported that pangolin CoV is an intermediate ancestor between bats and humans (Lam et al., 2020) which is opposite to what we got in our phylogenetic tree. It is interesting to note that amino acid sequences of the pangolin CoV receptor-binding domain (RBD) shows a very high binding affinity to the ACE2 receptors of humans, slightly less than the SARS-CoV-2 to human ACE2 (Guo et al., 2021) which means the most frequent 9mers that we located on pangolin CoV and SARS-CoV-2 has nothing to do with the transmissibility of viral RNA to host human cells because in the phylogenetic tree, both of the viruses are far away from each other.

In Figure 2, there are two different clades. Interestingly, one clade only has the 9mers from hepatitis B while the other one has both ebola and HIV 9mers.

Distribution of Work

Jaspreet: Coded for TASK 1, TASK 2 and TASK 4. Completed the TASK 5, E1, and E2 and wrote a report.
Time Spent: 6-7 Hours
Indronil: Coded for TASK 1, TASK 2, TASK 3, TASK 4, TASK E1 and wrote the report.
Time Spent: 40 Hours Approx.

References:

[1] Chae, H., Park, J., Lee, S.-W., Nephew, K. P., & Kim, S. (2013). Comparative analysis using K-mer and K-flank patterns provides evidence for CpG island sequence evolution in mammalian genomes. *Nucleic Acids Research*, *41*(9), 4783–4791.

[2] Chan, B. Y., & Kibler, D. (2005). Using hexamers to predict cis-regulatory motifs in Drosophila. *Bmc Bioinformatics*, *6*(1), 1–9.

[3] Chen, W., Feng, P.-M., Deng, E.-Z., Lin, H., & Chou, K.-C. (2014). ITIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Analytical Biochemistry*, *462*, 76–83.

[4] Chen, Y.-H., Nyeo, S.-L., & Yeh, C.-Y. (2005). Model for the distributions of k-mers in DNA sequences. *Physical Review E*, *72*(1), 011908.

[5] Cheng, Z. J., & Shan, J. (2020). 2019 Novel coronavirus: Where we are and what we know. *Infection*, *48*, 155–163.

[6] Chor, B., Horn, D., Goldman, N., Levy, Y., & Massingham, T. (2009). Genomic DNA k-mer spectra: Models and modalities. *Genome Biology*, *10*, 1–10.

[7] Guo, H., Hu, B., Si, H.-R., Zhu, Y., Zhang, W., Li, B., Li, A., Geng, R., Lin, H.-F., & Yang, X.-L. (2021). Identification of a novel lineage bat SARS-related coronaviruses that use bat ACE2 receptor. *Emerging Microbes & Infections*, *10*(1), 1507–1514.

[8] Hariharan, R., Simon, R., Pillai, M. R., & Taylor, T. D. (2013). Comparative analysis of DNA word abundances in four yeast genomes using a novel statistical background model. *PloS One*, *8*(3), e58038.

[9] Hashim, E. K. M., & Abdullah, R. (2015). Rare k-mer DNA: Identification of sequence motifs and prediction of CpG island and promoter. *Journal of Theoretical Biology*, *387*, 88–100.

[10] Lam, T. T.-Y., Jia, N., Zhang, Y.-W., Shum, M. H.-H., Jiang, J.-F., Zhu, H.-C., Tong, Y.-G., Shi, Y.-X., Ni, X.-B., & Liao, Y.-S. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature*, *583*(7815), 282–285.

[11] Lee, D., Karchin, R., & Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Research*, *21*(12), 2167–2180.

[12] Pace, N. R., Sapp, J., & Goldenfeld, N. (2012). Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. *Proceedings of the National Academy of Sciences*, *109*(4), 1011–1018.

[14] Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, *74*(11), 5088–5090.

[15] Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., & Huang, C.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*(7798), 270–273.