Aman Patel

December 12, 2020

CSCI-B 363

Evaluating the Phylogeny of SARS-CoV-2 Envelope Protein

**Background**

Phylogeny is the study of the evolutionary history of a virus or organism. It is commonly used to show the evolutionary relationships between different organisms in the form of phylogenetic trees. Phylogenetic trees are important in bioinformatics and epidemiology, as they are insightful about the history of a pathogen or organism and they can be used to predict the spread of disease. They can also provide information on the method of transmission as well as the rate and direction of evolution of the pathogen. Also, phylogenetic trees can give insight into when and how a mutation in the pathogen occurred, which can help epidemiologists discover the source of an outbreak.

Envelope proteins are integral proteins of the membrane that form the outermost layer of some viruses, referred to as the envelope. Most viruses have envelopes surrounding their capsid, which is the protein shell that contains the virus’s genetic material. The envelope is responsible for protecting the virus from the host’s immune system as well as allowing the contents of the virus to enter the host’s cells by fusing with the host cell’s membrane. Understanding envelope proteins can be significant when attempting to create a vaccine or medication to combat the virus.

**Problem Statement**

Evaluating the phylogeny of viral proteins can shed some light on the history and mutations of the virus, and it can be helpful to predict the further transmission of the virus. Analyzing the phylogeny of SARS-CoV-2 envelope proteins can be useful in determining the mutability of the virus and creating vaccines and medications to slow down the pandemic. The problem statement for this project is, “How have SARS-CoV-2 envelope protein residue compositions changed throughout the pandemic?”

**Algorithm**

Constructing a phylogenetic tree consists of three steps:

1. Multiple Sequence Alignment
2. Distance Matrix Creation
3. UPGMA

Multiple sequence alignment is the first step in building a phylogenetic tree. A multiple sequence alignment is an arrangement of three or more sequences such that the overall score of the arrangement is maximized (or minimized depending on the scoring system). There are a wide variety of potential scoring systems, and the system I chose goes as follows: 0 for matches, +1 for mismatches, +1 for gap insertion, and +1 for residue deletion. Multiple alignments are used in phylogenetic tree construction as opposed to pairwise alignments because multiple pairwise alignments are not guaranteed to be compatible with each other. For example, consider the three sequences and . The pairwise alignments are and When combining the first and second alignments, the issues begin. The first alignment assumes comes before , while the second assumes it comes after. Also, they cannot be matched directly on top of one another because the will line up with the , creating a suboptimal alignment. For this project, I imported the multiple sequence alignment from the NCBI website and converted the alignment into a text file for use in Python. Using a dynamic programming approach, this algorithm has a runtime of , where is the number of sequences and is the length of each sequence.

The second step of phylogenetic tree construction is the creation of the distance matrix.

This is a straightforward algorithm: each sequence in the multiple sequence alignment is compared to every other sequence in the alignment and the pairwise score is calculated based on the selected scoring system. These scores are stored in the distance matrix, which is converted into a lower triangular form to remove redundant information and simplify the array when printed. The runtime of this algorithm is , where is the length of the sequences. Each sequence is compared to all others, and for each comparison, there are iterations.

The final step of phylogenetic tree construction is the UPGMA algorithm, also known as the Unweighted Pair Group Method with Arithmetic Mean. UPGMA is an iterative algorithm, repeatedly reducing the distance matrix until the phylogenetic tree is fully built. To begin, the indices (x, y) of the minimum value in the array are found. Then, the values in the xth row and column are averaged with the yth row and column. The averaged values replace the values in the xth row and column, and the yth row and column are deleted. The labels for x and y are joined together, and the process repeats with the union of x and y acting as a single entity. The runtime of this algorithm is , as the algorithm runs times and the distance matrix reduction step has time complexity.

**Results**

The result of this algorithm is an unrooted tree that shows the relationships between each sequence. 50 primary protein structures of SARS-CoV-2 envelope proteins were analyzed. While looking at the multiple alignment produced by NCBI, it was evident that the protein structure had not changed significantly since the onset of the pandemic, with only three sequences having any differences in structure, each differing in one residue from the common sequence. Running the algorithm on this set of sequences yielded the following string: (((1,((((((((((((((((((((((((((((((((((((((((((((((2, 3), 4), 6), 7), 8), 9), 10), 11), 12), 13), 14), 15), 16), 17), 18), 19), 20), 21), 22), 23), 24), 25), 26), 27), 28), 29), 30), 32), 33), 34), 35), 36), 37), 38), 39), 40), 41), 42), 43), 44), 45), 46), 47), 48), 49), 50)), 5), 31)

Sequences 2 through 50 were identical excluding sequences 5 and 31. The identical sequences were all direct neighbors with one another, having a score of 0 for their pairwise alignments. The remaining sequences – 1, 5, and 31 – had a score of 2 amongst themselves and a score of 1 when compared to any other sequences. Due to the high number of identical sequences, it can be assumed that the differences in the remaining sequences were due to random mutations and not a significant change in the composition of the protein.

In the future this experiment can be performed on a protein that is more susceptible to mutation. This would lead to more significant results, and the phylogenetic tree could be used to determine the source, timing, and cause of the mutation(s).

**Works Cited**

1. Ciccozzi, M., Lai, A., Zehender, G., Borsetti, A., Cella, E., Ciotti, M., . . . Angeletti, S. (2019, July 12). The phylogenetic approach for viral infectious disease evolution and epidemiology: An updating review. Retrieved December 9, 2020, from https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.25526
2. Construction of a distance tree using clustering with the Unweighted Pair Group Method with Arithmatic Mean (UPGMA). (n.d.). Retrieved December 9, 2020, from <https://www.icp.ucl.ac.be/~opperd/private/upgma.html>
3. Embl-Ebi. (n.d.). Multiple Sequence Alignment. Retrieved December 10, 2020, from <https://www.ebi.ac.uk/Tools/msa/>
4. Figure - Phylogeny of Zika Virus in Western Hemisphere, 2015 - Volume 22, Number 5-May 2016 - Emerging Infectious Diseases journal - CDC. (n.d.). Retrieved December 9, 2020, from https://wwwnc.cdc.gov/eid/article/22/5/16-0065-f1
5. Phylogenetic Tree: Vertebrates. (n.d.). Retrieved December 9, 2020, from <https://kids.britannica.com/students/assembly/view/235364>
6. Severe acute respiratory syndrome coronavirus 2 data hub. (n.d.). Retrieved December 9, 2020, from <https://www.ncbi.nlm.nih.gov/labs/virus/vssi/>