Modified Edit Distance on Global SARS-CoV-2 Analysis

Jenny Fan, Princeton Day School

Introduction

Edit distance is a measure of how far two strings are from each other, typically using three types of edits: insertion, deletion, and substitution. Typically, each edit has a weight of 1, but the weight can also be changed to be applied more generally. The most common edit distance algorithm finds the distance by iterating through the strings letter by letter to find the best alignment and return smallest distance at the end. [1]

Edit distance is like sequence comparison in bioinformatics. The algorithms behind bioinformatics tools like BLAST are based off these algorithms that align sequences [2]. Research has been done to apply Edit Distance algorithm to bioinformatics. A common method is to implement substitution matrices, or scoring matrices, which change the weight of substituting between different amino acids. Research has been done to find suitable matrices to find similarities between sequences [3].

Random errors often happen in DNA replication. Thus, finding a single character difference is not as significant as longer chains of differences. A substitution matrix does not take into account this factor, so the idea behind this research was to modify the Edit Distance algorithm from this angle instead. This research focuses on finding a mathematical function where given the number of consecutive edits, the function would give a factor to multiply the cost by, and that distance is what is added to the total distance between two sequences. This function is termed as a "cost function," and this research will evaluate whether this method is applicable by comparing with results from bioinformatics tools.

Research Question

Are cost functions a viable alternative way to modify the edit distance algorithm for the purpose of sequence comparison?

Results

$$d_f = d_0\left(1 - \frac{1}{(x+75)^{0.5}}\right)$$

Figure 1: Final Cost Function

Comparison	Percent Identities (%)		
	Original	BLASTP	Modified
CHN and SARS	86.39	86.12	87.94
CHN and bat	39.64	46.81	46.41

Figure 2: Percent identities between coronavirus sequences given by the original edit distance algorithm, BLASTP, and the modified edit distance algorithm

	CHN	ITA	CA	NY	TX	AUS	RUS	BRA
CHN		0.885	1.771	1.771	1.771	3.541	0.885	1.771
ITA	0.885		2.656	0.885	0.885	4.426	0	2.656
CA	1.771	2.656		3.541	3.541	5.312	2.656	3.541
NY	1.771	0.885	3.541		0	5.312	0.885	3.541
TX	1.771	0.885	3.541	0		5.312	0.885	3.541
AUS	3.541	4.426	5.312	5.312	5.312		4.426	5.312
RUS	0.885	0	2.656	0.885	0.885	4.426		2.656
BRA	1.771	2.656	3.541	3.541	3.541	5.312	2.656	

Figure 3: Table of distances given by the modified edit distance algorithm between SARS-CoV-2 sequences around the world

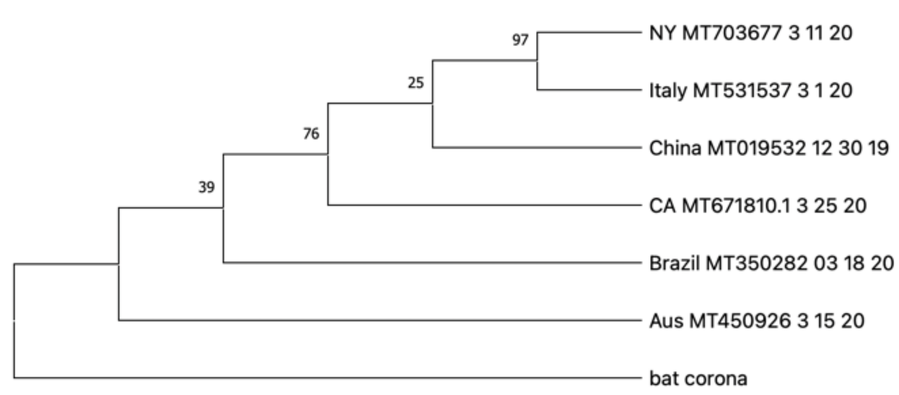

Figure 4.A: Phylogeny Tree drawn by MEGAX on SARS-CoV-2 sequences using Maximum Likelihood method

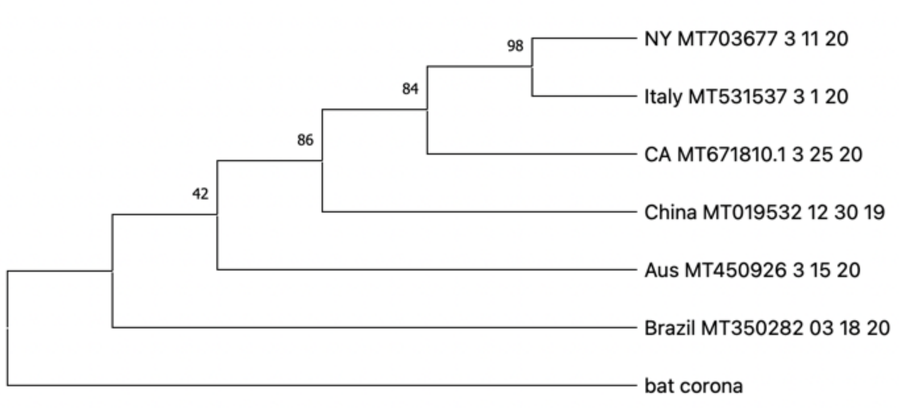

Figure 4.B: Phylogeny Tree drawn by MEGAX on SARS-CoV-2 sequences using Maximum Parsimony method

Note: the numbers on the nodes of the trees are the confidence values of the branches

Methods

- A SARS-CoV-2 sequence from China, a SARS sequence, and a sequence of a bat coronavirus were obtained from NCBI databases
- The original edit distance algorithm was run comparing China with SARS and China with bat. The percent identities produced are compared with those given by BLASTP
- Backtracking was added to the algorithm so it could count consecutive edits and use cost function to produce new distance
- Numbers and shape of cost function was altered numerous times until results produced percent identities close to those by BLASTP
- More SARS-CoV-2 sequences were obtained from around the world and compared against each other using modified edit distance algorithm
- Sequences (minus duplicates) were run on MEGAX to produce phylogeny trees and results compared to see if they matched

Conclusions

Figure 2 demonstrates that adding the cost function found in Figure 1 to the edit distance algorithm produces results that are closer to the percent identities given by BLASTP than the original algorithm. This showed that the idea of cost functions as a modification was plausible.

The ordering of distances in increasing order from New York in Figure 3 exactly match the ordering of sequences from closest to furthers on the Maximum Likelihood tree (Fig. 4.A). The low confidence of the split between California and Brazil parallel the equal distances of the two from all other sequences. All of this showed that the algorithm worked and is also like the Maximum Likelihood method of determining sequence distance and relationships.

Thus, it was concluded that cost functions are a viable alternative approach to modifying the edit distance algorithm for bioinformatics sequence comparison. The results don't exactly match those of BLASTP and MEGAX and are very simplistic considering that the algorithm here only accounts for one factor, but it certainly provides a new lens that when finetuned and incorporated with other factors can improve the accuracy of the edit distance algorithm as applied to bioinformatics.

References

[1] Kleinberg, Jon, and Eva Tardos. Algorithm Design. Pearson Education, 2014.

[2] Berger, Bonnie & Waterman, Michael & Yu, Yun. "Levenshtein Distance, Sequence Comparison and Biological Database Search." IEEE Transactions on Information Theory (2020): PP. 1-1. doi:10.1109/TIT.2020.2996543.

[3] Pearson, William R. "Selecting the Right Similarity-Scoring Matrix." Current protocols in bioinformatics vol. 43 (2013): 3.5.1-3.5.9. doi:10.1002/0471250953.bi0305s43

Acknowledgements

Thank you to Dr. Andrew Vershon and Dr. Janet Mead at the Waksman Institute, Dr. Matthew Nielpielko at Kean University, and Dr. Jeffrey Heinz at Stony Brook University for giving the background knowledge and tools that allowed this research project to happen. Thank you to my advisor, Mr. Brian Mayer at Princeton Day School, for his support in all my scientific endeavors.