Genetic Algorithm for Phylogenetic Time Tree Construction

Simon Lee, Daniele Venturi

Department of Applied Mathematics, University of California - Santa Cruz

Abstract

The SARS-CoV-2 (COVID-19) outbreak has seen many mutations since it was first reported on December 31, 2019 in Wuhan China. Our primary research objective is to provide a representation of one of its variants, the deltacoronavirus and put it in a gene map known as *Phylogenetic Time Trees*. Our focus is to use a novel method in building distance matrices from a metaheuristic called the *Genetic algorithm*. This algorithm is inspired by Charles Darwin's Natural Selection and it is very strong in navigating through massive search spaces. It finds optimal solutions to combinatorics problems that take a lifetime to compute. Using genome data offered by the National Center of Bioinformatics Information (https://www.ncbi.nlm.nih.gov/datasets/coronavirus/genomes/), we build a time tree of this Coronavirus variant.

Methodology

The Genetic Algorithm (GA), is a metaheuristic algorithm inspired by Charles Darwin's theory of Natural Selection. This machine learning algorithm uses three components (Fitness, Single-Point Crossovers, Mutations) and is primarily used to solve optimization problems. The architecture of our algorithm can be seen in our **Figure 1.** This program produces a distance matrix. Distance-matrix methods are far more rapid compared to popular methods like the *maximum-likelihood*, and it essentially measures the genetic distances between sequences. Once these distances are found, a distance matrix will be constructed, where the distance equates to the approximate evolutionary distance. Now that we have obtained a distance matrix, we can begin to construct the trees. As shown in **Figure 2**, the tree begins in a star formation where there is no tree or lengths of branches shown or combined into internal nodes. But with this distance matrix, the distances represent the dissimilarity of the aligned sequence and can begin subsequent joining between neighbors. This whole process is demonstrated with a small example with 7 individual taxa going through processes (A) - (E). Using this we obtain our phylogenetic time tree

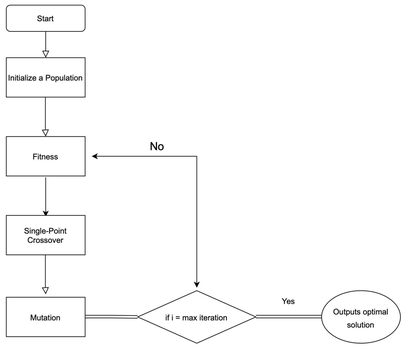
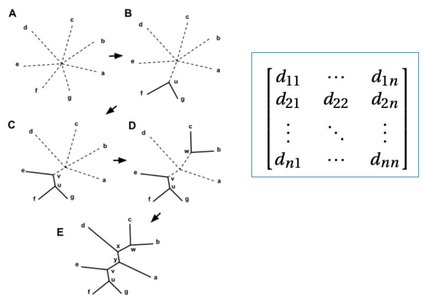


Figure 1: Flowchart of the Genetic Algorithm



$$\begin{bmatrix} d_{11} & \cdots & d_{1n} \\ d_{21} & d_{22} & d_{2n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{bmatrix}$$

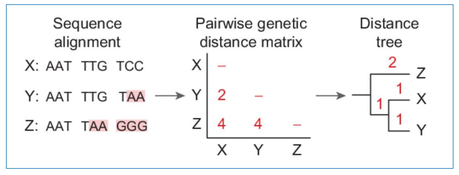


Figure 2: A Distance Matrix based method to construct trees known as The Neighbor Joining Algorithm

Results

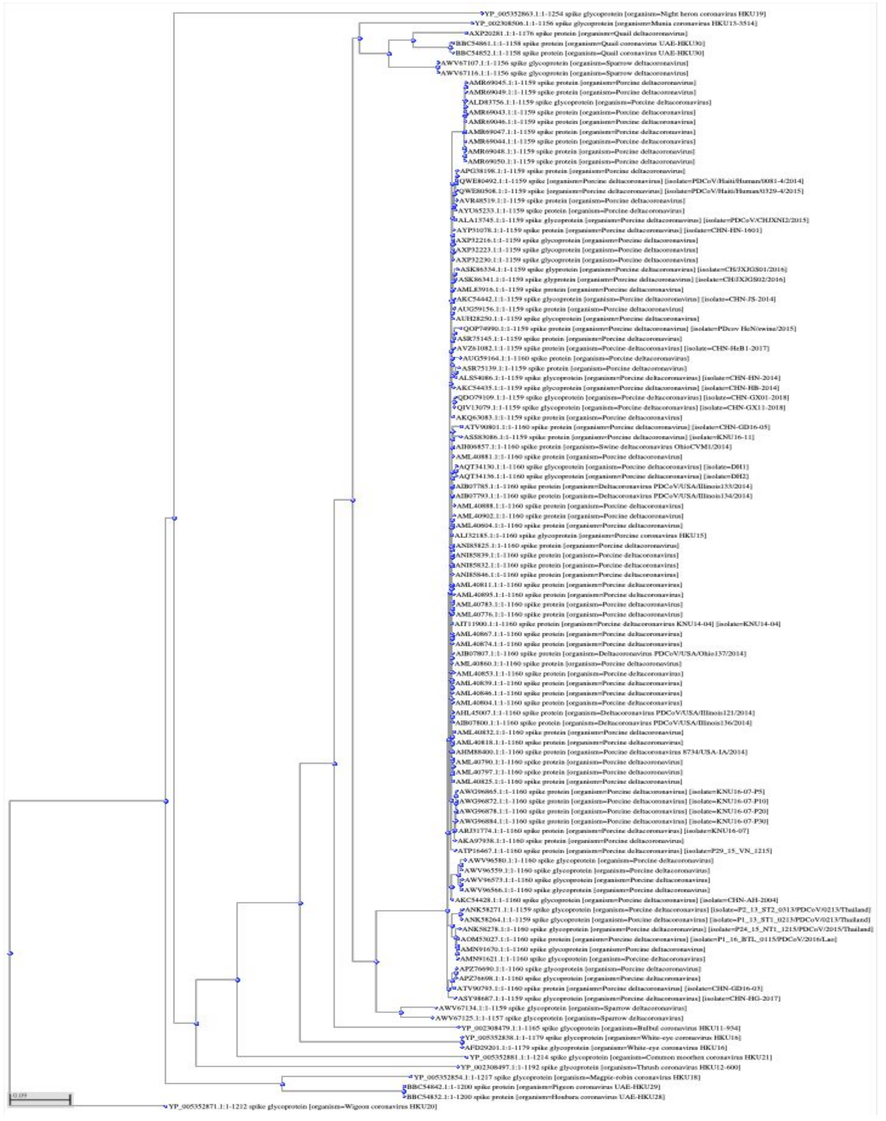


Figure 3: Our Phylogenetic Time tree

After countless of hours of figuring out the best approaches to this problem, we have finally been able to produce a result. The measure at the bottom is equivalent to 0.09 which is around 10.8 months. For every 0.1 in length it is equivalent to 1 year. To our surprise, the Delta variant has been around for much longer than we expected. Originating from the Middle East we were able to observe some of our early findings of this long spread disease. While we did have some surprises, we can also see on the right hand side of the time tree all the recent chaos of the Delta variant spreading across species.

Analysis

With some of our genomes beginning at 2014, it makes intuitive sense as to why the graph is divided up as it is. With the majority of the tree falling between the 2019 to 2021 interval, it aligns with the media and how this Coronavirus variant has spread within our "pandemic years". However within the trees there are also a few incorrect approximations.There could be many reasons for these occurrences. One of the reasons I could think of first off is our small sample size of genome data. The Spike protein coming from the SARS Delta Coronavirus is roughly 1,160 amino acids long. In retrospect the DNA sequence for one of these species is roughly 26,865 nucleotides long. Obviously if we were to run this experiment with DNA, our machine learning algorithm could pick up more insight based off the data it is presented. But as described one of of our main goals in this project was efficiency and computing time.Another reasons is due to its stochastic behavior as well

References

[1] Fliesler, N., *Sturdier spikes may explain SARS-CoV-2 variants' faster spread*, March 2021, Boston's Children Hospital, https://answers.childrenshospital.org/sars-cov-2-variants-spike, Accessed October 7.

[2]Ahn,C.W.,Ramakrishna,R.S.,*Elitism-based compact genetic algorithms*,IEEETransactionsonEvolutionary Computation, Volume: 7, Issue: 4, Aug. 2003, Pages 367-385.

[3] Gammaitoni L., Hänggi P., Jung P., and Marchesoni F., *Stochastic Resonance* Rev. Mod. Phys. 70, 223–288 (1998)

[4] 3Blue1Brown (n.d.) Neural Networks [What is backpropagation really doing? | Chapter 3, Deep learning]. Retrieved from https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi

[5] Hardesty, L., *Explained: P vs. NP, The most notorious problem in theoretical computer science remains open, but the attempts to solve it have led to profound insights.*, October 2009, https://news.mit.edu/2009/explainer-pnp, Accessed October 11.

[6] *Bifurcating Trees*, Ecology Center, https://www.ecologycenter.us/genetic-diversity/bifurcating-trees.html, Accessed October 14.

[7] *Understanding Evolution: your one-stop source for information on evolution*, The Tree Room, https://evolution.berkeley.edu/evolibrary/article/0_0_0/evotrees_intro, Accessed October 14.

Acknowledgements

I want to thank the UCSC Applied Mathematics Department for supporting this project. The UCSC AM department have definitely supported my research interests and they have allowed me to conduct some of this research as I took on the AM 114 class at UCSC. Dynamical systems have wide-applications in biological sciences and I thought it would be a fun idea to try doing a problem with a set of data that was very relevant to the current pandemic. For these reasons I am happy that I was able to use some of this knowledge on dynamical systems and present at the UCSC SURU 2021 Conference.