An Analysis of SARS-CoV-2 Using ViReport

Miranda Song¹ and Niema Moshiri^{2,*}

¹Department of Bioengineering, UC San Diego, La Jolla, CA, 92123, USA

²Department of Computer Science & Engineering, UC San Diego, La Jolla, CA, 92123, USA

*To whom correspondence should be addressed; a1moshir@ucsd.edu

Abstract

The ongoing outbreak of *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) has resulted in millions of cases and hundreds of thousands of deaths. Given the current lack of treatments or vaccines available, it may be useful to trace the evolution and spread of the virus to better develop methods of preventative intervention. In this study, we analyzed over 4,000 full genome sequences of human SARS-CoV-2 using novel tool ViReport [13], an automated workflow for performing phylogenetic analyses on viral sequences and generating comprehensive molecular epidemiological reports. The complete ViReport output can be found at <https://github.com/mirandajsong/ViReport-SARS-CoV-2>.

1 Introduction

Coronaviruses are a group of RNA viruses that cause various diseases in humans, other mammals, and birds. In the last two decades, there have been three documented strains of coronavirus that have caused major epidemics in humans. These include *severe acute respiratory syndrome coronavirus* (SARS-CoV) in 2002, *Middle East respiratory syndrome coronavirus* (MERS-CoV) in 2012, and most recently, a novel coronavirus referred to as COVID-19 or SARS-CoV-2 that emerged in Wuhan, China in December 2019. As of June 2020, there have been over 8 million confirmed cases of SARS-CoV-2 globally [3].

The rapid spread of SARS-CoV-2 has demonstrated the need to understand how the virus spreads and for measures of epidemic intervention such as stay-at-home orders and social distancing. To gain insight into disease transmission and origins, phylogenetic analyses are often performed using bioinformatics methods to quantify and visualize the relatedness of viral sequences. A typical workflow for phylogenetic analysis and reconstruction involves four main steps: (1) Data Acquisition, (2) Sequence Alignment, (3) Phylogenetic Inference, and (4) Tree Visualization [6].

Acquiring data involves identifying homologous nucleotide or protein sequences that are representative of the point of interest, whether that be a specific gene or species or something else. This step may involve additional preprocessing (e.g. filtering by length or sample type) to ensure that the sequences are valid and that the inferred tree will provide meaningful results. There are many public databases that provide viral (and other) sequences and accompanying metadata, including NCBI (ncbi.nlm.nih.gov), GISAID (gisaid.org), and ViPR (viprbrc.org).

Multiple sequence alignment (MSA) refers to the alignment of related biological sequences of similar length, the output of which may be used to infer more specific evolutionary relationships. Given the classification of MSA computation as an NP-complete problem, there are many existing algorithms that solve this problem [1]. Examples include MAFFT [7], MUSCLE [5], and ViralMSA [15], all of which are implemented in ViReport, along with others.

Phylogenetic inference is the reconstruction of evolutionary trees by grouping individual elements based on shared ancestry. Estimation of phylogenetic trees can be performed through various methods such as Bayesian Inference and Maximum Likelihood [6], which have spawned tools such as BEAST [4] and RAxML [17], respectively. The resulting phylogenies can be input into online tree visualization tools for further examination.

While there exist various tools that seek to automate this process such as MEGA5 [18] and Phylogeny.fr [2], ViReport is robust in its integration of modules that allows for the user to select their desired tool for each stage in the workflow. Furthermore, ViReport includes options for rooting and dating the inferred phylogeny, calculating pairwise distances, ancestral sequence reconstruction, and transmission clustering, all of which provide additional information beyond the tree itself.

Figure 1: Demonstrates the overall ViReport workflow.

2 Methods

ViReport was used to preprocess the data, perform the MSA and phylogenetic inference steps, and produce the input tree files for tree visualization.

2.1 Data Acquisition

For this study, sequences were retrieved from the NCBI Virus database, which hosts data from GenBank. Nucleotide sequences were selected for using the SARS-CoV-2 Virus specification and by filtering out sequences under 26,000 bases as to only include full genome sequences [22]. Sequence metadata was also compiled, specifically that regarding the coun-

try of origin and the collection date.

2.2 Multiple Sequence Alignment

Multiple Sequence Alignment was performed using ViralMSA [15] utilizing Minimap2 [10] and the reference sequence with accession ID NC_045512, defined as "Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1" and dated December 2019. Pairwise sequence distances were calculated using the tn93 tool from HIV-TRACE [8].

2.3 Phylogenetic Inference

FastTree 2 [16] was used to infer a maximum-likelihood phylogeny under the General Time-Reversible model [19] using a Gamma20-based likelihood. The phylogeny was MinVar-rooted using FastRoot [11] and dated using LSD2 [20]. Pairwise phylogenetic distances were calculated using TreeSwift [14].

2.4 Tree Visualization

The phylogenies were visualized using iTOL [9] and IcyTree [21].

3 Results

3.1 Analysis of Different Coronavirus Strains

As a preliminary analysis based on a previous study titled *A Novel Coronavirus from Patients with Pneumonia in China, 2019* [23], 30 full genome sequences of different coronavirus strains (including SARS, MERS, and SARS-CoV-2) were compared using ViReport. After preprocessing, the average sequence length was 30018.133 bases with standard deviation 584.016, with sample dates ranging from April 2003 to March 2020. The average pairwise sequence distance (as calculated using tn93 from HIV-TRACE) was 0.503, with a standard deviation of 0.322.

Figure 2: Rooted phylogenetic tree showing evolutionary relationships between various human and nonhuman coronavirus strains as produced by ViReport and visualized using iTOL; highlighted are SARS genomes in blue, MERS in green, and SARS-CoV-2 in red.

Figure 3: A reproduction of Figure 4B, "Phylogenetic Analysis of 2019-nCoV and Other Betacoronavirus Genomes," from *A Novel Coronavirus from Patients with Pneumonia in China, 2019* [23]

Comparison of the visualization of the ViReport output and the figure from this study [23] shows that the results are very similar, despite the two analyses having used different tools (ViralMSA and IQ-TREE for ViReport, MUSCLE and RAXML for the article [23]), as the overall clades are nearly identical. Both analyses indicate that, while SARS-CoV, MERS, and SARS-CoV-2 fall into distinct clades, SARS-CoV-2 is more closely related to SARS-CoV, as both fall under the subgenus *Sarbecovirus*, whereas MERS falls under a different subgenus *Merbecovirus*. Additionally, SARS-CoV-2 is most closely related to a SARS-like coronavirus found in bats.

3.2 Analysis of SARS-CoV-2 Genomes

This analysis was performed on a dataset of 4635 full genomes sequences of SARS-CoV-2 found in *Homo sapiens*. The average sequence length was 29819.915 bases with a standard deviation of 80.154, with sample dates ranging from December 2019 to May 2020 (sequence data was retrieved on May 30, 2020; additional sequences have been added to the database since).

Figure 4: Distribution of sample dates.

We can see from the histogram above that the peak of the sample distribution falls around late March 2020. Sequences were also grouped by geographical location, as shown below.

Figure 5: Distribution of categories (country of origin).

We can see from the distribution that nearly 40 countries were represented in the dataset, with the five highest sequence counts from the U.S., Australia, Hong Kong, Germany, and China. Incidentally, as of June 2020, the top five countries in terms of total number of confirmed cases are the U.S., Brazil, Russia, India, and the United Kingdom [3]. While there does not seem to be significant correlation between sequence counts and confirmed cases, which could be the result of sampling bias, it is notable that the U.S. has the highest counts for both metrics.

Multiple sequence alignment resulted in 29903 positions (213 invariant) and 4333 unique sequences. The average pairwise sequence distance was 0.000323, with a standard deviation of 0.000181.

Figure 6: Distribution of pairwise sequence distances.

Pairwise sequence distances indicate how closely related the sequences are, and in comparison to the average pairwise sequence distance from the various coronavirus strains (Section 3.1) of 0.503, the average for the SARS-CoV-2 genomes of 0.000323 is multiple orders of magnitude smaller, indicating that the SARS-CoV-2 genomes are much more closely related to each other, as expected.

Across the multiple sequence alignment positions, the average coverage was 0.983 with a standard deviation of 0.0416, indicating good coverage. The average Shannon entropy as 0.00993 with a standard deviation of 0.0431; these relatively small values indicated that there was not much sequence variability.

Figure 7: Coverage (proportion of non-gap characters) across MSA positions.

Figure 8: Shannon entropy across MSA positions. Due to the abundance of zero-entropy positions, all non-zero entropies were deemed significant. The significance threshold is shown as a red dashed line, and significant points are shown in red.

The following maximum-likelihood phylogeny was inferred, with an average pairwise phylogenetic distance of 0.000381 and a standard deviation of 0.000208. The maximum pairwise phylogenetic distance (i.e. tree diameter) was 0.00505, which describes the distance between the two least closely related sequences.

Figure 9: Phylogenetic tree as visualized using IcyTree, path to reference sequence from root highlighted in red.

Figure 10: Distribution of pairwise phylogenetic distances.

Given the large number of sequences included, it is evidently difficult to visualize the specific relationships between sequences in the tree. These individual relationships can be visualized by inputting the rooted tree file to iTOL, but the tree is too large to include in this report.

The pairwise phylogenetic distances were also used for transmission clustering using TreeN93 [12], which resulted in 736 clusters and 2535 singletons. The average cluster size was 2.853 with a standard deviation of 2.0284, and cluster sizes ranged from 2-30.

Figure 11: Distribution of cluster sizes.

4 Discussion

Some considerations of the limitations of this analysis include the dataset used; given that there are well over 8 million confirmed cases of SARS-CoV-2 in the world (as of June 2020) but the sample only included around 4500 sequences, the sample is evidently a very small subset that may not be representative of current events both in geographical distribution and time. For example, the UK has the fifth most cases in the world [3] but is not listed as one of the categories represented by the sample. As the viral sequence database updates with new sequences or sequences from other sources are pooled together, the analysis may become more representative.

Future directions for this work include additional analysis of the generated statistics; a topic of interest is how the sequences are distributed by date for each country, as this may be indicative of how successful policies in place are in terms of managing the ongoing epidemic. Another topic is exploring the distribution of pairwise distances over time, which may lead to conclusions regarding the mutation rate of the SARS-CoV-2 virus. With regard to ViReport, additional work is being done to analyze potential differences in results based on which tools are selected at each step.

5 Data Availability

All genome sequences are available at **NCBI Virus**. The complete ViReport output can be found at <https://github.com/mirandajsong/ViReport-SARS-CoV-2>.

References

- [1] M. Chatzou, C. Magis, J.-M. Chang, C. Kemena, G. Bussotti, I. Erb, and C. Notredame. Multiple sequence alignment modeling: methods and applications. *Briefings in Bioinformatics*, 17(6):1009–1023, nov 2015. ISSN 1467-5463. doi: 10.1093/bib/bbv099. URL <https://doi.org/10.1093/bib/bbv099>.
- [2] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.-F. Dufayard, S. Guindon, V. Lefort, M. Lescot, J.-M. Claverie, and O. Gascuel. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids research*, 36(Web Server issue):W465–W469, jul 2008. ISSN 1362-4962. doi: 10.1093/nar/gkn180. URL <https://pubmed.ncbi.nlm.nih.gov/18424797><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2447785/>.
- [3] G. L. Dong E, Du H. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis*, 2020. URL [https://doi.org/10.1016/S1473-3099\(20\)30120-1](https://doi.org/10.1016/S1473-3099(20)30120-1).
- [4] A. J. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214, 2007. ISSN 1471-2148. doi: 10.1186/1471-2148-7-214. URL <https://doi.org/10.1186/1471-2148-7-214>.
- [5] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, mar 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh340. URL <https://doi.org/10.1093/nar/gkh340>.
- [6] B. G. Hall. Building Phylogenetic Trees from Molecular Data with MEGA. *Molecular Biology and Evolution*, 30(5):1229–1235, mar 2013. ISSN 0737-4038. doi: 10.1093/molbev/mst012. URL <https://doi.org/10.1093/molbev/mst012>.
- [7] K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, apr 2013. ISSN 1537-1719. doi: 10.1093/molbev/mst010. URL <https://pubmed.ncbi.nlm.nih.gov/23329690><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3603318/>.
- [8] S. L. Kosakovsky Pond, S. Weaver, A. J. Leigh Brown, and J. O. Wertheim. HIV-TRACE (TRANsmiission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. *Molecular Biology and Evolution*, 35(7):1812–1819, 2018. ISSN 0737-4038. doi: 10.1093/molbev/msy016. URL <https://academic.oup.com/mbe/article/35/7/1812/4833215>.
- [9] I. Letunic and P. Bork. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*, 47(W1):W256–W259, apr 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz239. URL <https://doi.org/10.1093/nar/gkz239>.

- [10] H. Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, may 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty191. URL <https://doi.org/10.1093/bioinformatics/bty191>.
- [11] U. Mai, E. Sayyari, and S. Mirarab. Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction. *PLOS ONE*, 12(8):e0182238, aug 2017. URL <https://doi.org/10.1371/journal.pone.0182238>.
- [12] N. Moshiri. TreeN93: a non-parametric distance-based method for inferring viral transmission clusters. *bioRxiv*, page 383190, jan 2018. doi: 10.1101/383190. URL <http://biorxiv.org/content/early/2018/08/02/383190.1.abstract>.
- [13] N. Moshiri. ViReport. 2020. URL <https://github.com/niemasd/ViReport>.
- [14] N. Moshiri. TreeSwift: A massively scalable Python tree package. *SoftwareX*, 11:100436, 2020. ISSN 2352-7110. doi: <https://doi.org/10.1016/j.softx.2020.100436>. URL <http://www.sciencedirect.com/science/article/pii/S2352711019300767>.
- [15] N. Moshiri. ViralMSA: Massively scalable reference-guided multiple sequence alignment of viral genomes. *bioRxiv*, page 2020.04.20.052068, jan 2020. doi: 10.1101/2020.04.20.052068. URL <http://biorxiv.org/content/early/2020/04/21/2020.04.20.052068.1.abstract>.
- [16] M. N. Price, P. S. Dehal, and A. P. Arkin. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE*, 5(3):e9490, mar 2010. URL <https://doi.org/10.1371/journal.pone.0009490>.
- [17] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, jan 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu033. URL <https://doi.org/10.1093/bioinformatics/btu033>.
- [18] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution*, 28(10):2731–2739, oct 2011. ISSN 0737-4038. doi: 10.1093/molbev/msr121. URL <https://doi.org/10.1093/molbev/msr121>.
- [19] S. Tavaré. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.
- [20] T.-H. To, M. Jung, S. Lycett, and O. Gascuel. Fast Dating Using Least-Squares Criteria and Algorithms. *Systematic Biology*, 65(1):82–97, sep 2015. ISSN 1063-5157. doi: 10.1093/sysbio/syv068. URL <https://doi.org/10.1093/sysbio/syv068>.

- [21] T. G. Vaughan. IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*, 33(15):2392–2394, apr 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx155. URL <https://doi.org/10.1093/bioinformatics/btx155>.
- [22] P. C. Y. Woo, Y. Huang, S. K. P. Lau, and K.-Y. Yuen. Coronavirus genomics and bioinformatics analysis. *Viruses*, 2(8):1804–1820, aug 2010. ISSN 1999-4915. doi: 10.3390/v2081803. URL <https://pubmed.ncbi.nlm.nih.gov/21994708><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3185738/>.
- [23] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, and W. Tan. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine*, 382(8):727–733, jan 2020. ISSN 0028-4793. doi: 10.1056/NEJMoa2001017. URL <https://doi.org/10.1056/NEJMoa2001017>.