ViralMSA: Massively scalable reference-guided multiple sequence alignment of viral genomes

Niema Moshiri[1,*]

[1]Department of Computer Science & Engineering, UC San Diego, La Jolla, 92093, USA

*To whom correspondence should be addressed.

Abstract

Motivation: In molecular epidemiology, the identification of clusters of transmissions typically requires the alignment of viral genomic sequence data. However, existing methods of multiple sequence alignment scale poorly with respect to the number of sequences.

Results: ViralMSA is a user-friendly reference-guided multiple sequence alignment tool that was built to enable the alignment of ultra-large viral genome datasets. It scales linearly with the number of sequences, and it is able to align tens of thousands of full viral genomes in seconds.

Availability: ViralMSA is freely available at https://github.com/niemasd/ViralMSA as an open-source software project.

Contact: a1moshir@ucsd.edu

Supplementary information: Supplementary data are available at online.

1 Introduction

Real-time or near real-time surveillance of the spread of a pathogen can provide actionable information for public health response (Poon *et al*., 2016). Though there is currently no consensus in the world of molecular epidemiology regarding a formal definition of what exactly constitutes a "transmission cluster" (Novitsky *et al*., 2017), all current methods of inferring transmission clusters require a multiple sequence alignment (MSA) of the viral genomes: distance-based methods of transmission clustering require knowledge of homology for accurate distance measurement (Pond *et al*., 2018), and phylogenetic methods of transmission clustering require the MSA as a precursor to phylogenetic inference (Balaban *et al*., 2019; Rose *et al*., 2017; Ragonnet-Cronin *et al*., 2013; Prosperi *et al*., 2011).

The standard tools for performing MSA such as MAFFT, MUSCLE, and Clustal Omega typically utilize progressive-iterative approaches due to their broad applicability (Katoh & Standley, 2013; Edgar, 2004; Sievers & Higgins, 2014). However, these methods become prohibitively slow for real-time pathogen surveillance as the number of viral genomes grows. For example, during the COVID-19 pandemic, the number of viral genome assemblies available from around the world grew exponentially in the initial months of the pandemic, but MAFFT, the fastest of the aforementioned MSA tools, scales quadratically with respect to the number of sequences.

In the case of closely-related viral sequences for which a high-confidence reference genome exists, MSA can be accelerated by comparing each viral genome in the dataset against the reference genome and then using the reference as an anchor with which the individual alignments can be merged into a single MSA.

Here, we introduce ViralMSA, a user-friendly open-source MSA tool that utilizes read mappers such as Minimap2 (Li, 2018) to enable the reference-guided alignment of ultra-large viral genome datasets.

2 Related work

VIRULIGN is another reference-guided MSA tool designed for viral genomes (Libin *et al*., 2019). While VIRULIGN also aims to support MSA of large sequence datasets, its primary objective is to produce codon-correct alignments, whereas ViralMSA's objective is to produce alignments for use in transmission clustering as fast as possible. Thus, while ViralMSA is not guaranteed to yield codon-aware alignments, it is orders of magnitude faster than VIRULIGN, which is critical for rapidly-growing epidemics. Further, VIRULIGN is codon-aware, making it appropriate for coding regions, whereas ViralMSA is appropriate for the alignment of whole viral genomes. Lastly, VIRULIGN requires a thorough annotation of the reference genome, which may be difficult to obtain (especially towards the beginning of a novel outbreak) and does not provide the user to easily utilize different reference genomes for different viral strains. ViralMSA, on the other

hand, only requires the reference genome assembly's GenBank accession number and can build any required index files on-the-fly

3 Results and discussion

ViralMSA is written in Python 3 and is thus cross-platform. ViralMSA depends on BioPython (Cock *et al*., 2009) as well as on whichever read mapper the user chooses to utilize, which is Minimap2 by default (Li, 2018). In addition to Minimap2, ViralMSA supports STAR (Dobin *et al*., 2013), Bowtie 2 (Langmead & Salzberg, 2012), and HISAT2 (Kim *et al*., 2019), though the default of Minimap2 is strongly recommended for both performance and accuracy (Fig. 1).

ViralMSA takes the following as input: (1) a FASTA file containing the viral genomes to align, (2) the GenBank accession number of the reference genome to use, and (3) the mapper to utilize (Minimap2 by default). ViralMSA will pull the reference genome from GenBank and generate an index using the selected mapper, both of which will be cached for future alignments of the same viral strain, and will then execute the mapping. ViralMSA will then process the results and output a multiple sequence alignment in the FASTA format. For commonly-studied viruses (e.g. HIV, SARS-CoV-2), the user can simply provide the name of the virus instead of an accession number, and ViralMSA will automatically select what we believe would be an appropriate reference genome.

Because it uses the positions of the reference genome as anchors with which to merge the individual pairwise alignments, ViralMSA only keeps matches, mismatches, and deletions with respect to the reference genome: it discards all insertions with respect to the reference genome. For closely-related viral strains, insertions with respect to the reference genome are typically unique and thus lack usable phylogenetic or transmission clustering information, so their removal results in little to no impact on downstream analyses (Fig. 1b).

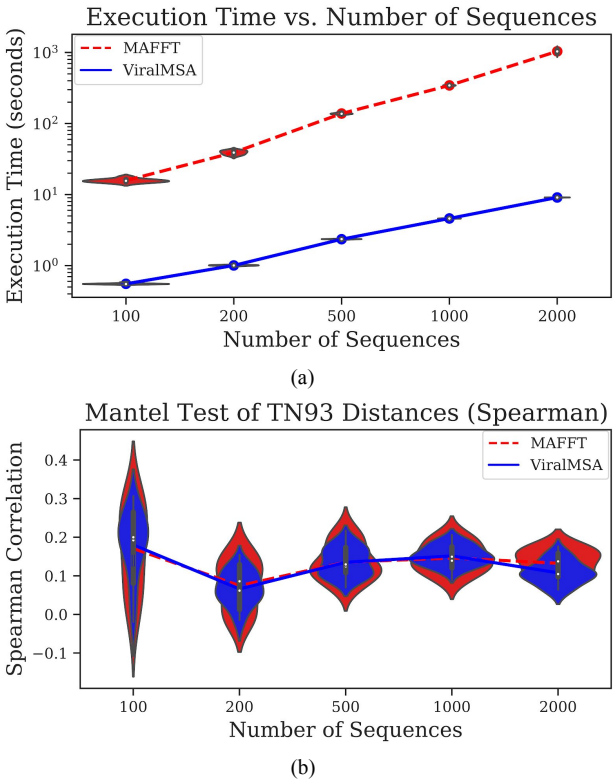

Fig. 1. Performance and accuracy of ViralMSA. Execution time (seconds) and Mantel test Spearman correlations of TN93 distance matrices are shown for MSAs estimated by MAFFT and ViralMSA for various dataset sizes.

4 Results and discussion

In order to test ViralMSA's accuracy, we obtained a curated HIV-1 full genome MSA from the Los Alamos National Laboratory (LANL) HIV Sequence Database, which we used as our ground truth. We then used both MAFFT and ViralMSA (using the HXB2 reference genome) to recompute MSAs. No other MSA tools were included in the comparison due to being orders of magnitude slower than MAFFT (which would render the full experiment infeasible).

Because ViralMSA's objective is to be utilized in transmission clustering workflows, which typically rely on pairwise distances between samples, we computed pairwise sequence distances between every pair of samples in each MSA (true, MAFFT, and ViralMSA) under the TN93 model of sequence evolution (Tamura & Nei, 1993) using pairwise distance calculator implemented in HIV-TRACE (Pond et al., 2018). Then, for each estimated MSA, we measured alignment accuracy by computing the Mantel correlation test (using Spearman correlation) between the true and estimated MSAs.

In order to measure performance, we subsampled the full HIV dataset into datasets of n = 100, 200, 500, 1,000, and 2,000 sequences, with 10 replicates for each value of n. We then computed MSAs of each subsampled alignment replicate.

The results of the accuracy and performance experiments can be found in Figure 1. As can be seen, ViralMSA in its default mode is consistently orders of magnitude faster than MAFFT (Fig. 1a), yet it produces MSAs with the same level of accuracy (Fig. 1b).

Acknowledgements

We would like to thank Heng Li and his exceptional work in developing Minimap2, which has enabled significant speedup in traditional viral analysis workflows. His work is absolutely integral to ViralMSA's improvement on existing approaches.

Funding

This work has been supported by NSF grant NSF-2028040 to N.M. as well as the Google Cloud Platform (GCP) Research Credits Program.

Conflict of Interest: none declared.

References

Balaban, M. *et al*. (2019) TreeCluster: Clustering biological sequences using phylogenetic trees. *PLoS One*, **14**(8), e0221068.
Cock, P.J.A. *et al*. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.
Dobin, A. *et al*. (2009) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1), 15–21.
Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.*, **5**, 113.
Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
Kim, D. *et al*. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.
Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18), 3094–3100.
Libin, P.J.K. (2019) VIRULIGN: fast codon-correct alignment and annotation of viral genomes. *Bioinformatics*, **35**(10), 1763–1765.
Katoh, K. and Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.*, **30**(4), 772–780.
Novitsky, V. *et al*. (2017) Phylogenetic Inference of HIV Transmission Clusters. *Infect. Dis. Transl. Med.*, **3**(2), 51–59.
Pond, S.L.K. *et al*. (2018) HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. *Mol. Biol. Evol.*, **35**(7), 1812–1819.
Poon, A.F.Y. *et al*. (2016) Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *Lancet HIV*, **3**(5), e231–e238.
Prosperi, M.C.F. *et al*. (2011) A novel methodology for large-scale phylogeny partition. *Nat. Commun.*, **2**, 321.
Ragonnet-Cronin, M. *et al*. (2013) Automated analysis of phylogenetic clusters. *BMC Bioinform.*, **14**,317.
Rose, R. *et al*. (2017) Identifying Transmission Clusters with Cluster Picker and HIV-TRACE. *AIDS Res. Hum. Retroviruses*, **33**(3), 211–218.
Sievers, F. and Higgins, D.G. (2014) Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.*, **1079**, 105–116.
Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, **10**(3), 512–526.