Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Applications Note

Genome analysis

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.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

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 leverages the algorithmic techniques of read mappers to enable the multiple sequence 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. However, alignments produced by ViralMSA omit insertions with respect to the reference genome.
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 *Bioinformatics* 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; Ragonnet-Cronin *et al*., 2013; Prosperi *et al*., 2011).

The standard tools for performing MSA such as MAFFT (Katoh & Standley, 2013), MUSCLE (Edgar, 2004), and Clustal Omega (Sievers & Higgins, 2014) are 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 independently comparing each viral genome in the dataset against the reference genome and then using the reference as an anchor to merge the individual alignments into a single MSA (Pond *et al*., 2018).

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 whole-genome datasets.

2 Related work

VIRULIGN is another reference-guided MSA tool designed for viruses (Libin *et al*., 2019). While VIRULIGN also aims to support MSA of large sequence datasets, its primary objective is to produce codon-correct MSAs (i.e., avoiding frameshifts), making it appropriate for aligning coding regions, whereas ViralMSA's primary objective is to align whole viral genomes in real-time. Further, ViralMSA is orders of magnitude faster than VIRULIGN (Fig. 1) and uses a fraction of the memory.

3 Results and discussion

ViralMSA is written in Python 3 and is thus cross-platform. ViralMSA depends on BioPython (Cock *et al*., 2009) and whichever read mapper the user chooses, 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: Minimap2 is much faster than the others (Li, 2018) and is the only mapper that consistently

succeeds to align all genome assemblies against an appropriate reference across multiple viruses. ViralMSA's support for read mappers other than Minimap2 is primarily to demonstrate that ViralMSA is flexible, meaning it will be simple to incorporate new read mappers in the future.

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, 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 an MSA in the FASTA format. For commonly-studied viruses, the user can simply provide the name of the virus instead of an accession number, and ViralMSA will select an appropriate reference genome. The user can also choose to provide a local FASTA file containing a reference genome, which may be useful if the desired reference does not exist on GenBank or if the user wishes to conduct the analysis offline.

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 (Tab. 1).

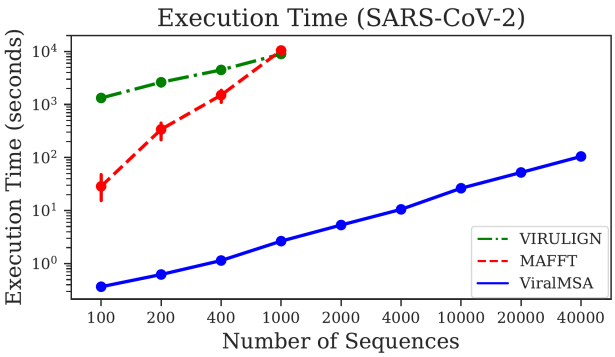

Fig. 1. Execution time. Execution time for SARS-CoV-2 MSAs (genome length 29kb) estimated by VIRULIGN, MAFFT, and ViralMSA for various dataset sizes. All runs were executed sequentially on an 8-core 2.0 GHz Intel Xeon CPU with 30 GB of memory.

Table 1. Multiple sequence alignment accuracy

Virus	MAFFT (*S*)	ViralMSA (*S*)	MAFFT (*P*)	ViralMSA (*P*)
Ebola	0.9957	0.9873	0.9998	0.9816
HCV	0.9995	0.9506	0.9999	0.9678
HIV-1	0.9786	0.9705	0.9957	0.9941

Correlation coefficients are shown for Mantel tests between curated "ground truth" MSAs and those estimated by MAFFT and ViralMSA. *S* and *P* denote Spearman and Pearson Correlation, respectively. 1 indicates perfect correlation, -1 indicates perfect anticorrelation, and 0 indicates no correlation.

In order to assess MSA estimation accuracy, we obtained curated Ebola, HCV, and HIV-1 full-genome MSAs from the Los Alamos National Laboratory (LANL) Sequence Databases, which we used as our ground truth. In order to benchmark MSA runtime, we obtained a large collection of SARS-CoV-2 complete genomes from the Global Initiative on Sharing All Influenza Data (GISAID) database. VIRULIGN crashed when run on all datasets aside from the SARS-CoV-2 dataset.

To measure performance, we subsampled the full SARS-CoV-2 dataset, with 10 replicates for each dataset size, and then computed MSAs of each replicate. ViralMSA is consistently orders of magnitude faster than both MAFFT and VIRULIGN (Figs. 1, S1). Further, for all SARS-CoV-2 datasets, both ViralMSA and MAFFT required less than 1 GB of memory, but VIRULIGN required ~10 GB of memory.

Because ViralMSA's objective is to be utilized in real-time applications such as transmission clustering workflows, which typically rely on pairwise distances between samples, we computed pairwise distance matrices from each MSA under the TN93 model of sequence evolution (Tamura & Nei, 1993) using the pairwise distance calculator implemented in HIV-TRACE (Pond et al., 2018). Then, we measured alignment accuracy by computing the Mantel correlation test between the distance matrix of the curated ("true") MSA against that of each estimated MSA. ViralMSA seems to produce MSAs with just slightly lower accuracy than those produced by MAFFT across different viruses (Tab. 1) and different levels of subsampling (Figs. S2-S3). Furter, both MAFFT and ViralMSA seem to produce MSAs that tend to underestimate TN93 distance, with MSAs produced by ViralMSA underestimating slightly more significantly (Fig. S4).

To obtain a phylogenetic metric of accuracy as well, we inferred phylogenies from each MSA under the General Time Reversible (GTR) model (Tavaré, 1986) with the "Gamma20" model of site-rate heterogeneity using VeryFastTree (Piñeiro *et al.*, 2020). We then compared the phylogenies inferred from the curated ("true") MSA against those inferred from each estimated MSA by computing the normalized Robinson-Foulds (RF) distance (Robinson & Foulds, 1981), a metric ranging from 0 to 1, with 0 indicating identical tree topology. The MSAs estimated by ViralMSA yielded phylogenies with far better topological accuracy (RF 0.531, 0.411, and 0.599 for Ebola, HCV, and HIV-1) than those estimated by MAFFT (0.956, 0.965, and 0.994).

Note that ViralMSA's speed and accuracy stem from the algorithmic innovations of the selected read mapper (not from ViralMSA itself), meaning ViralMSA can natively improve as read mapping tools evolve.

Acknowledgements

We would like to thank Heng Li and his exceptional work in developing Minimap2. His work is integral to ViralMSA's speed and accuracy.

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.
Piñeiro, C. *et al.* (2020) VeryFastTree: speeding up the estimation of phylogenies for large alignments through parallelization and vectorization strategies. *Bioinformatics*, btaa582.
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.
Robinson, D.F. and Foulds, L.R. (2017) Comparison of phylogenetic trees. *Math. Biosci.*, **53**(1–2), 131–147.
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.
Tavaré, S. (1986) Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures Math. Life Sci.*, **17**, 57–86.