Bioinformatics, 37(12), 2021, 1761–1762
doi: 10.1093/bioinformatics/btaa871
Advance Access Publication Date: 12 October 2020
Applications Note

Genome analysis

COVID-Align: accurate online alignment of hCoV-19 genomes using a profile HMM

Frédéric Lemoine [1,2,*], Luc Blassel[1,3], Jakub Voznica[1,4] and Olivier Gascuel[1,5,*]

[1]Unité de Bioinformatique Evolutive, USR 3756 (DBC/C3BI), Institut Pasteur & CNRS, 75015 - Paris, France, [2]Hub de Bioinformatique et Biostatistique, USR 3756 (DBC/C3BI), Institut Pasteur & CNRS, 75015 - Paris, France, [3]ED515, Sorbonne Université, Collège Doctoral, 75006 - Paris, France, [4]Université de Paris, 75006 Paris, France and [5]Académie des Sciences, USR 3756, CNRS, 75015 - Paris, France

*To whom correspondence should be addressed.

Associate Editor: Valencia Alfonso
Received and revised on May 23, 2020; editorial decision on September 11, 2020; accepted on September 24, 2020

Abstract

Motivation: The first cases of the COVID-19 pandemic emerged in December 2019. Until the end of February 2020, the number of available genomes was below 1000 and their multiple alignment was easily achieved using standard approaches. Subsequently, the availability of genomes has grown dramatically. Moreover, some genomes are of low quality with sequencing/assembly errors, making accurate re-alignment of all genomes nearly impossible on a daily basis. A more efficient, yet accurate approach was clearly required to pursue all subsequent bioinformatics analyses of this crucial data.

Results: hCoV-19 genomes are highly conserved, with very few indels and no recombination. This makes the profile HMM approach particularly well suited to align new genomes, add them to an existing alignment and filter problematic ones. Using a core of ~2500 high quality genomes, we estimated a profile using HMMER, and implemented this profile in COVID-Align, a user-friendly interface to be used online or as standalone via Docker. The alignment of 1000 genomes requires ~50 minutes on our cluster. Moreover, COVID-Align provides summary statistics, which can be used to determine the sequencing quality and evolutionary novelty of input genomes (e.g. number of new mutations and indels).

Availability and implementation: https://covalign.pasteur.cloud, hub.docker.com/r/evolbioinfo/covid-align.

Contacts: frederic.lemoine@pasteur.fr or gascuelolivier@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

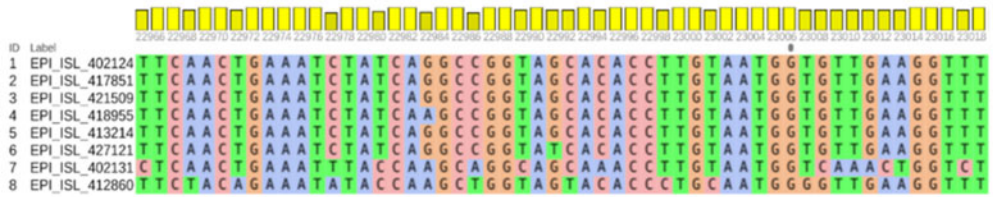
1 Introduction

Since the emergence of the hCoV-19 virus (or SARS-CoV-2) responsible for the COVID-19 pandemic, unprecedented efforts are taking place across the world to sequence genomes of this virus and share the data. As of today (September 9, 2020), the GISAID (Shu *et al.*, 2017) provides access to more than 105 000 full genomes, and ~23 000 for the NCBI and the EBI. The first genomes were sequenced in China by the end of December 2019. Their number first increased slowly and then rapidly when the pandemic appeared on all continents. Submissions of several thousand sequences to GISAID in a single day have become common. Moreover, some genomes may be submitted incomplete, with sequencing and assembly errors. These characteristics pose major challenges to bioinformatics, notably that of multiple sequence alignment (MSA; Chatzou *et al.*, 2016), which is crucial for subsequent analyses (phylogeny, transmission clusters, mutation study, structure, etc.). To solve this difficulty, we use a profile HMM-based approach (Durbin *et al.*, 1998), which is the norm for HIV (www.hiv.lanl.gov), and is particularly well suited to hCoV-19, as its genome is highly conserved, without known recombination in human hosts (De Maio *et al.*, 2020; Xiaolu *et al.*, 2020). Using a profile, the addition of new data to an existing MSA requires linear computing times in the number of input genomes. Moreover, profile-based MSA proved to be very accurate (Earl *et al.*, 2014; Nute and Warnow, 2016). This approach is implemented in COVID-Align, which can be used thanks to a Web service and via Docker.

2 Materials and methods

To estimate our profile HMM, we proceeded in several steps, in order to select an appropriate set of sequences and obtain a clean and reliable MSA to give as input to HMMER (www.hmmer.org):

- We downloaded all hCoV-19 genomes available on GISAID (April 24, 2020) and performed pairwise alignments using MAFFT (Katoh and Standley, 2013) of each of these genomes with the reference strain hCoV-19/Wuhan/WIV04/2019, sequenced in China December 30, 2019. This genome was found perfectly conserved not only in China, but also in Thailand,



	#	High Q.	Low Q.
Genomes	25,298	23,596	1,702
	#	Unique	New
Mutation	11.7	0.25	0.62
Gap	0.59	0.00008	0.00004
Gap Open	0.13	0.00008	0.00004
Ins Open	0.0037	0.0018	0.0011

Fig. 1. Visualization and statistics summary. Left: MSAviewer visualization of the Receptor Binding Domain (RBD) of the Spike gene, with reference genome (top), recently sequenced ones and the Bat and Pangolin genomes (bottom). The site numbering corresponds to that of the reference, to be used to recover the ORFs and genes. In RBD region the Pangolin virus genome is closer to Human's than is Bat's, suggesting a possible recombination. On the opposite, Human viruses are highly conserved. Right: Statistics summary, displaying the number of High and Low Quality genomes, and the number of evolutionary events (mutations, gaps, gap openings, insertions, insertion openings). We distinguish the number of unique events (not seen yet and present only once in submitted genomes, possibly due to errors) and the number of new events (seen at least twice, likely corresponding to evolutionary novelties). This table was filled with GISAID sequences deposited between August 10 and September 21 2020, with unique and new statistics with respect to the database as of August 9 (Supplementary Material)

Japan, USA, UK, etc. and is considered as the origin of the virus (Li *et al.*, 2020; www.gisaid.org).
- Then, using loose thresholds, we removed the genomes that were excessively divergent from the reference and had too many unknown (N) characters. We edited the remaining ones (e.g. removing the first gappy positions and the poly-A tail) and aligned them with MAFFT.
- The MSA so obtained was further filtered by removing the genomes having too many unique (i.e. not shared by any other genome) mutations and indels. We used more stringent thresholds than in the previous stage. This resulted in an MSA of 2426 genomes, where the 12 first and 22 last positions of the reference genome were removed due poor alignment and low signal, but all other reference positions were preserved and showed high conservation. We used HMMER to estimate our profile from this curated MSA. All details and program options are available in Supplementary Information.

The resulting profile was implemented in a Nextflow (Di Tommaso *et al.*, 2017) and Galaxy workflow combining hmmalign from HMMER to align the input genomes to the profile, GoAlign to format the input/output files (https://github.com/evolbioinfo/goalign), and Python to compute summary statistics. These statistics help users evaluate the sequencing quality and potential evolutionary novelties of input genomes; for example: number of unique mutations and indels, number of mutations compared to the reference genome... A user-friendly interface, implemented in GO (similar to Lemoine *et al.*, 2019) allows users to launch their analyses without having to know how to use the Galaxy system. For advanced users, COVID-Align can be installed locally via Docker (https://www.docker.com). COVID-Align is also available on www.gisaid.org, using GISAID sequence identifiers.

3 Results

All results are given in a zipped file containing:

- The MSA of the input genomes plus the reference one that is displayed first, but cutting the first 12 and last 22 positions. With small datasets, this MSA can be visualized using MSAviewer (Fig. 1; Yachdav *et al.*, 2016).
- The hmmalign output in FASTA format, for each of the input genomes. This can be used to recover the insertions, deletions and match positions (to be reported to the reference genome).
- A CSV file with all statistics computed for each of the input genomes. Unique mutations and indels are possibly due to errors (sequencing, assembly etc.), while new ones (seen at least twice in submitted genomes, for the first time) likely correspond to evolutionary novelties (see Supplementary Information for details).
- A table in CSV format, summarizing the main average statistics and features of submitted genomes (Fig. 1).

Our Web service processes 1000 genomes and return the MSA in ~50 minutes (average over 10 trials, max = 80 minutes), thanks to parallelization that is easy to set up with profiles. Comparison with MAFFT-based GISAID MSA shows that our MSA: (i) can be used as is, while MAFFT's cannot due to ~10 000 highly gappy columns resulting from sequencing and assembly errors; (ii) helps to detect and filter these errors; (iii) is similar for most sequences to a properly trimmed version of MAFFT's MSA, and more accurate for the few others (Supplementary Information). Importantly, our profile and statistics will be regularly updated to account for user needs and the evolutionary novelties (mutations, indels...) of the emerging genomes to come.

Acknowledgements

Sincere thanks to Amandine Perrin and Fabien Mareuil (Institut Pasteur) for help, and the GISAID Team and all its Data Contributors for sharing their genome data.

Funding

L.B. PhD Grant: PRAIRIE (ANR-19-P3IA-0001); JV: PhD grant by Ecole Normale Supérieure, Paris-Saclay and financial support by Ecole Doctorale Frontières de l'Innovation en Recherche et Education - Programme Bettencourt.

Conflict of Interest: none declared.

References

Chatzou,M. *et al.* (2016) Multiple sequence alignment modeling: methods and applications. *Brief. Bioinform.*, **17**, 1009–1023.

De Maio,N. *et al.* (2020). Issues with SARS-CoV-2 sequencing data, virological.org.

Di Tommaso,P. *et al.* (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–-319.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Earl,D. *et al.* (2014) Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res.*, **24**, 2077–2089.

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Lemoine,F. *et al.* (2019) NGPhylogeny.fr: new generation phylogenetic services for non-specialists. *Nuc. Acids Res.*, **47**, W260–W265.

Li,Y. *et al.* (2020) COVID-19 Evolves in Human Hosts. In Proceedings of KDD Health Day, 2020. arXiv:2003.05580v6

Nute,M. and Warnow,T. (2016) Scaling statistical multiple sequence alignment to large datasets. *BMC Genomics*, **17**, 764.

Shu,Y. *et al.* (2017) GISAID: global initiative on sharing all influenza data – from vision to reality. *EuroSurveillance*, **22**,

Xiaolu,T. *et al.* (2020) On the origin and continuing evolution of SARS-CoV-2. *Natl. Sci. Rev.*, **7**, 1012–1023.

Yachdav,G. *et al.* (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, **32**, 3501–3503.