Phylogeography: Analysis of genetic and climatic data of SARS-CoV-2

Wanlin Li, Aleksandr Koshkarov, My-Linh Luu, Nadia Tahiri.
Department of Computer Science, Université de Sherbrooke, Canada

ABSTRACT

- This research discusses phylogeographic approaches to examine how patterns of divergence within SARS-CoV-2 coincide with geographic features, such as climatic features.
- We propose a python-based bioinformatic pipeline called ***aPhylogeo*** for **phylogeographic** analysis written in Python 3.9 that help researchers better understand the distribution of the virus in specific regions, and then run all the analysis operations in a single run.
- In particular, the *aPhylogeo* tool determines which parts of the **genetic sequence** undergo a high mutation rate depending on **geographic conditions**, using a *sliding window* that moves along the genetic sequence alignment in user-defined steps and a window size.
- *aPhylogeo* runs on Windows®, MacOS X® and GNU/Linux, and the code is freely available to researchers and collaborators on GitHub.

DATASET

Our study focuses on SARS-CoV-2 lineages that were **first identified** and **widely disseminated** in a particular country during a certain period (O'Toole et al., 2021).

38 lineages with regional characteristics were selected for further study (Rambaut et al., 2020).

Based on location information, complete nucleotide sequencing data for these 38 lineages was collected from the NCBI Virus website.

Our analysis focus on sliding window that moves along the genetic sequence alignment in user-defined steps and a window size (see Figure 1).

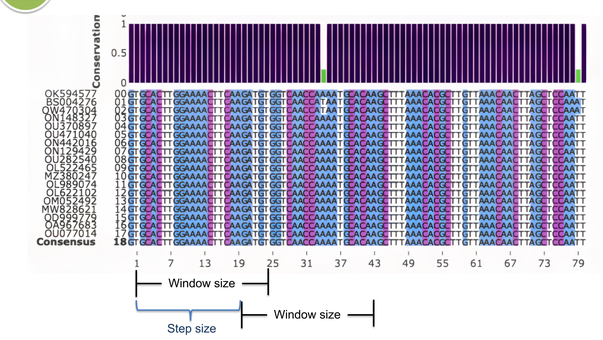

Figure 1: Alignment of partial nucleotide sequencing data

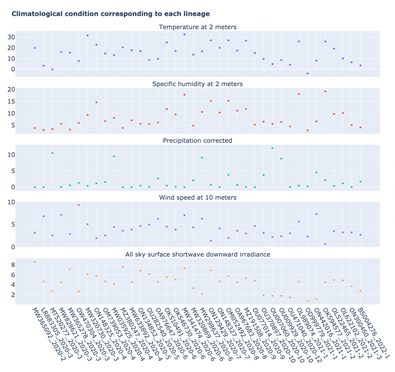

Figure 2: Climatic conditions of each lineage in most common country at the time of first detection

The meteorological conditions of each lineage in most common country at the time of first detection were considered (see Figure 2).

PIPELINE

We focus on developing a new algorithm to find ***relationships*** between a **reference tree** (i.e., a temperature tree, a habitat precipitation tree, or others) with their **genetic compositions**. This new algorithm (Li et al., 2022) can help find which genes or which subparts of a gene are sensitive or favorable to a given environment. This algorithm is inspired by Tahiri 2012.

The algorithm includes **four main steps** (see Figure 3).

- Input parameter validation
- Creation of trees based on the climate data (Dissimilarity analysis between each pair of variants)
- Creation of phylogenetic trees from Multiple Sequence Alignment (MSA) using a sliding window approach
- Comparison of phylogenetic trees and climatic trees using the Robinson and Foulds (RF) topological distance

Figure 3: Workflow of the algorithm based on our new pipeline aPhylogeo (Li et al., 2022).

RESULTS

- **Robinson and Foulds baseline and bootstrap threshold:** the phylogenetic trees constructed in each sliding window are compared to the climatic trees using the Robinson and Foulds topological distance (the RF distance). We defined the value of the RF distance obtained for regions without any mutations as a baseline, and the bootstrap values corresponding to this baseline as a bootstrap threshold.
- **Sliding window:** the implementation of a sliding window technique with a bootstrap threshold provides a more accurate identification of regions with high gene mutation rates. For comparison, we applied five combinations of parameters (window size and step size) to the same dataset (see Figure 4). These combinations of window sizes and steps provide an opportunity to have three different movement strategies (overlapping, non-overlapping, with gaps).

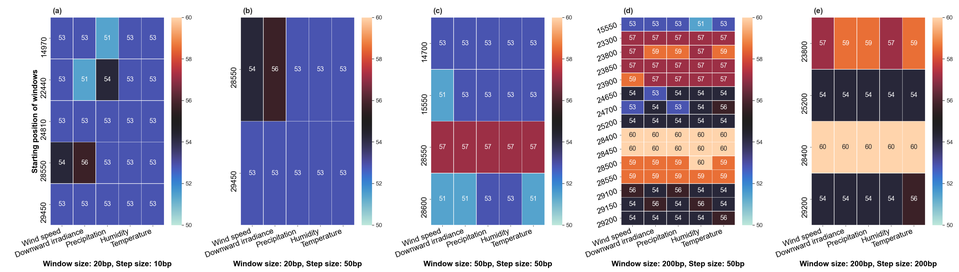

Figure 4: Heatmap of Robinson and Foulds topological distance over alignment windows

- **Comparaison between genetic trees and climatic trees:** Relatively low RF distance values represent relatively more similarity between the phylogenetic tree and the climatic tree. With our algorithm based on the sliding window technique, regions with high mutation rates can be identified (see Fig 5).

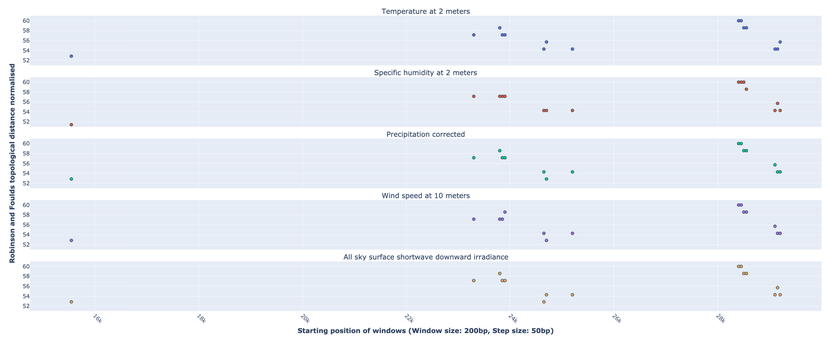

Figure 5: Robinson and Foulds topological distance normalized changes over the alignment windows

CONCLUSIONS

- ***aPhylogeo*** allows the user to quickly and intuitively create trees from genetic and climate data.
- Using a sliding window approach, *aPhylogeo* finds specific subparts of viral genetic sequences that can be sensitive to the climatic conditions of the region.
- *aPhylogeo* aims to help the scientific community by facilitating research in the field of phylogeography.

SOURCE

GitHub Link
https://github.com/tahiri-lab/aPhylogeo

Tahiri Lab website
https://tahirinadia.github.io/

ACKNOWLEDGEMENTS

NSERC CRSNG
Université de Sherbrooke
Centre de recherche en ECOLOGIE DE L'UdeS

REFERENCES

Li, W., Luu, M.-L., Koshkarov, A. and Tahiri, N. aPhylogeo (version 1.0), July 2022. URL: https://github.com/tahiri-lab/aPhylogeo, doi:doi.org/10.5281/ zenodo.6773603.

O'Toole, et al. (2021). Tracking the international spread of SARS-CoV-2 lineages B. 1.1. 7 and B. 1.351/501Y-V2 with grinch. Wellcome open research, 6.

Rambaut, A et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature microbiology, 5(11), 1403-1407.

Tahiri (2012). Un nouvel algorithme pour retrouver les relations phylogénétiques entre la distribution géographique des espèces et leurs compositions génétiques.