Virus Pop: Expanding viral databases by protein sequence simulation

Julia Kende, Massimiliano Bonomi, Sarah Temmam, Philippe Pérot, Béatrice Regnault, Marc Eloit, Thomas Bigot

Virus Pop: Expanding viral databases by protein sequence simulation

Julia Kende[1], Maximiliano Bonomi[2], Sarah Temmam[3], Philippe Pérot[3], Béatrice Regnault[3], Marc Éloit[3] and Thomas Bigot[1]

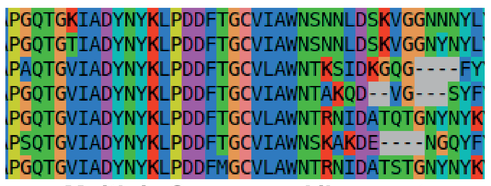
[1] Bioinformatics and Biostatistics Hub, Institut Pasteur, Université Paris Cité, F-75015 Paris, France
[2] Department of Structural Biology and Chemistry, Institut Pasteur, Université Paris Cité, CNRS UMR 3528, F-75015 Paris, France
[3] Pathogen Discovery Laboratory, Institut Pasteur, Université Paris Cité, F-75015 Paris, France

With the advent of NGS and the growing availability of metagenomic data, a new way to ensure the surveillance of the virosphere was born. Metagenomic for virus detection is used in various contexts. It can be used to identify and characterize viruses in medical circumstances (e.g. [7]), to target a virus in a precise research perspective (e.g. [9]) or in the general effort to improve our knowledge on the virosphere. For all those objectives, samples will, at some point, be processed through a bioinformatic pipeline for Next Generation Sequencing reads diagnostic and discovery.

In this context, the **goal of Virus Pop [2] is two-fold**:

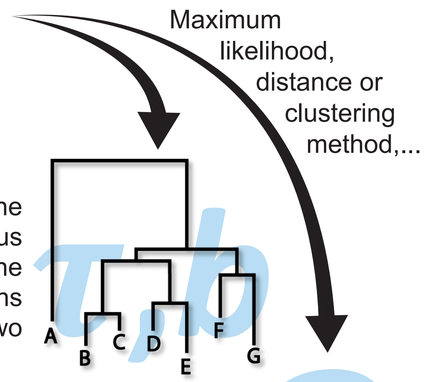
- Evaluate the level of sensibility of a pipeline or of independent tools. That is, being able to know how distant a virus can be from known clades and still be detected.
- Explore whether it could be relevant to extend the databases with simulated sequences to increase our bioinformatic tools sensibility without modifying them directly.

Generalities on protein simulations

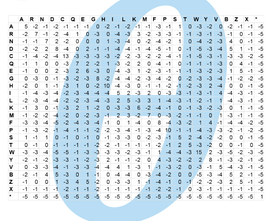


Multiple Sequence Alignment

A phylogenetic tree: it represents the evolution history of the homologous sequences. The tree topology reflects the speciation events and the branch lengths represent the evolutionary time between two nodes.



An evolutionary model: it mainly consists in a substitution matrix. It contains the probabilities for a mutation from one amno-acid to another to happen. The parameters and the type of model they are related to define the process which governs the evolution along the phylogenetic tree branches.



Inferring a tree and an evolutionary model from real data

Simulating with the model, a tree and a root sequence

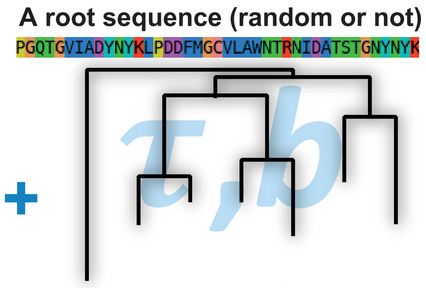
Issues with existing simulators:

- They are filling a new tree with sequences
  → No way to populate an existing tree with new sequences
- There is no convenient way to simulate while respecting variations in the substitution rate along the alignment
  → Preserved areas are lost

An evolutionary model

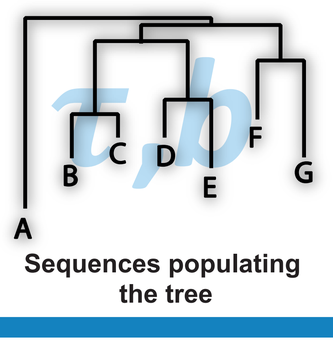
IQ-TREE [5], Seq-Gen [6], INDELible [1]

A root sequence (random or not)



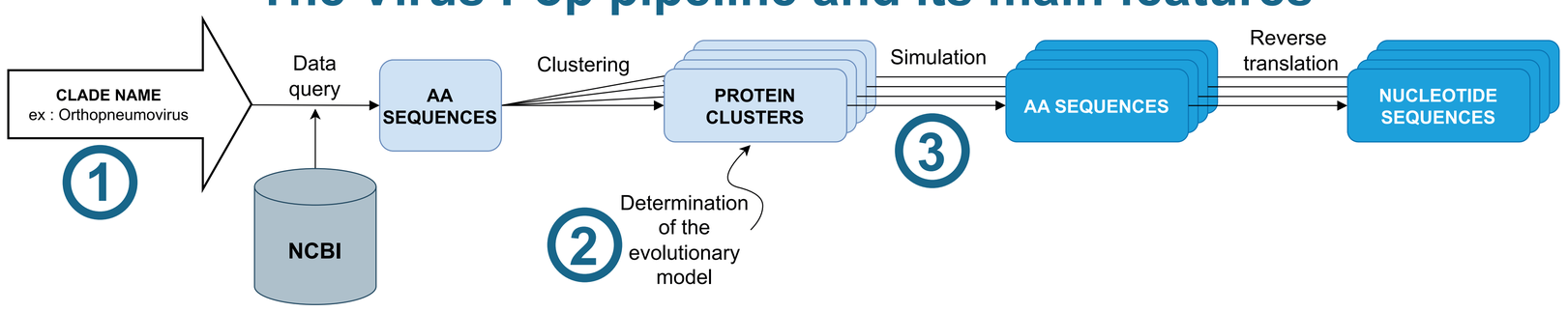
A phylogenetic tree (random or not)

Starting from the root sequence **mutations** are generated (and potentially also **insertion** and **deletion**) respecting the model parameters. The number of mutations is determined by the branch lengths. At a branching point, the sequence is duplicated and the copies evolve independantly. This represents **speciation events**.



Sequences populating the tree

The Virus Pop pipeline and its main features



1 A user-friendly program with many automated step

To generate simulations, the user only needs to enter a taxonomic group name and an ouput folder name. Two parameters are however recommanded (see below).

Virus Pop will automatically determine the species within the group and fetch the data online in the NCBI database.

```
python3 ./run_virus_pop.py group_name Orthopneumovirus Project_1 -t 15 -s 20
```

`-t --target_n_genome` Target number of genomes to be downloaded per species found (default = 5)

`-s --min_cluster_size` minimum size of clusters that will be processed (default = 25)

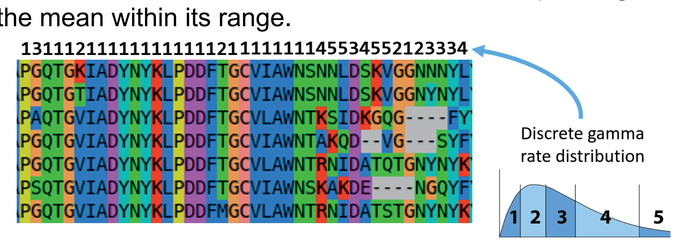
Other important parameters:

`--all_ancestral_nodes` : run simulation on each internal node of the phylogenetic tree

`--n_simu` : number of ancestral sequences per node (default = 10). Then the default number of simulation will be this number times 3.

2 Handling heterogeneous substitution rates

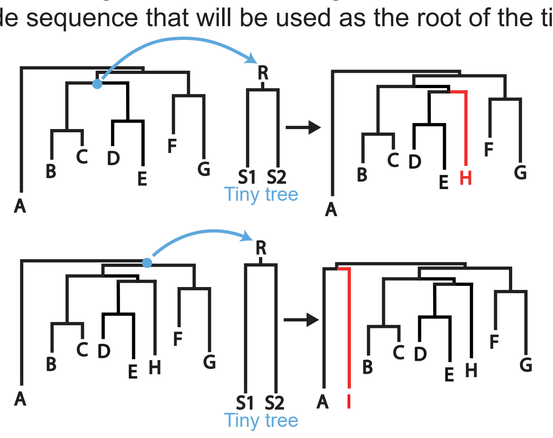
With IQ-TREE, Virus Pop determines the Gamma discretized profile of a protein alignment [10]. The law is discretized in five bins of equal area. Each site is associated to a bin. The rate corresponding to each bin is the mean within its range.



Then, Virus Pop uses mixture models [8] to tinker with the possibilities of IQ-TREE [5] and model with constrains on each site rate. In mixture models, a sequence can be partitioned in any number of parts and each part will have a distinct evolutionary model. In our case, we will use the same substitution model and equilibrium amino acid frequencies, but use variable substitution rates.

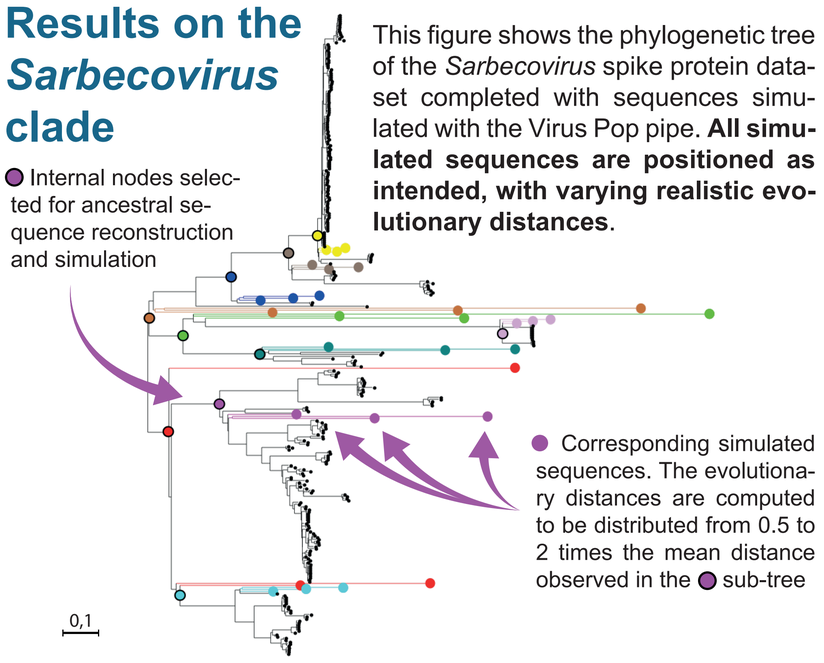
3 Giving a perfect control on the simulations

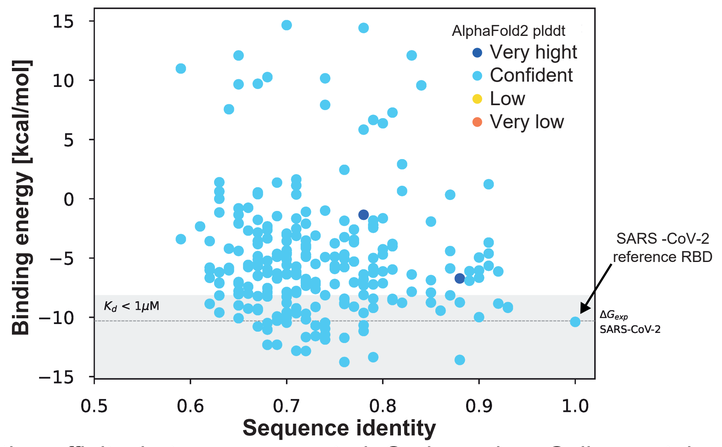
The simulations themselves are performed with Alisim [4] on tiny trees. Their Newich expression is always the same: *(S1:d, S2:d).* With *d* the varying branch length and, for each targetted simulation, we infer an internal node sequence that will be used as the root of the tiny tree.



Results on the *Sarbecovirus* clade

This figure shows the phylogenetic tree of the *Sarbecovirus* spike protein dataset completed with sequences simulated with the Virus Pop pipe. **All simulated sequences are positioned as intended, with varying realistic evolutionary distances**.





Test of the affinity between generated *Sarbecovirus* Spike proteins (RBD zone) and the hACE2 receptor. A low binding energy suggests the likely formation of the complex. This demonstrates how the evolutionary model based on protein mutation probabilities with varying substitution rate succeeds at generating **realistic sequences in terms of structure** and, thus maybe, in terms of function. It's important to note that, given its probabilistic nature, not all simulations retain the structure.

Conclusion

Virus Pop has proven to be able to generate various sequences that are homologous to the input proteins and that are precisely positioned within its phylogenetic tree. Furthermore, tests on the structure preservation show that some of the generated sequences could be functional, even amongst generated sequences with a low percent of identity.

The Virus Pop command line program is accessible from GitHub and easy to install. Furthermore, we created a database with ready-to-use simulations of 89 virus genera.

Check out the Virus Pop database [3]!

- 53 genera infecting humans
- 36 genera infecting other vertebrates
- 995 homologous protein group found
- 300 simulations at each internal nodes

→ **24,138,277 sequences!**

[1] Fletcher,W. and Z. Yang (Aug. 2009). "INDELible: A Flexible Simulator of Biological Sequence Evolution." en. In: Molecular Biology and Evolution 26.8, pp. 1879–1888. issn: 0737-4038, 1537-1719. doi: 10.1093/molbev/msp098.
[2] Kende J, Bonomi M, Temmam S, Regnault B, Pérot P, Eloit M, Bigot T. Virus Pop - Expanding Viral Databases by Protein Sequence Simulation. Viruses. 2023; 15(6):1227. doi: 10.3390/v15061227
[3] Kende, J., & Bigot, T. (2023). Virus Pop Database V1 (Version 1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7867259
[4] Ly-Trong, Nhan, Suha Naser-Khdour, Robert Lanfear, and Bui Quang Minh (Dec. 2021). AliSim: A Fast and Versatile Phylogenetic Sequence Simulator For the Genomic Era. en. preprint. Bioinformatics. doi: 10.1101/2021.12.16.472905.
[5] Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear (May 2020). "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era." en. In: Molecular Biology and Evolution 37.5. Ed. by Emma Teeling, pp. 1530–1534. issn: 0737-4038, 1537-1719. doi: 10 . 1093 / molbev / msaa015.
[6] Rambaut, Andrew and Nicholas C. Grass (1997). "Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees." en. In: Bioinformatics 13.3, pp. 235–238. issn: 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/13.3.235.
[7] Regnault, Béatrice et al. (May 2021). "First Case of Lethal Encephalitis in Western Europe Due to European Bat Lyssavirus Type 1." In: Clinical Infectious Diseases 74.3, pp. 461–466. doi: 10 . 1093 / cid /ciab443.
[8] Si Quang, Le, Olivier Gascuel, and Nicolas Lartillot (Oct. 2008). "Empirical profile mixture models for phylogenetic reconstruction." en. In: Bioinformatics 24.20, pp. 2317–2323. issn: 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btn445.
[9] Temmam, Sarah et al. (Apr. 2022). "Bat coronaviruses related to SARSCoV-2 and infectious for human cells." en. In: Nature 604.7905, pp. 330–336. issn: 0028-0836, 1476-4687. doi: 10.1038/s41586-022-04532-4.

INSTITUT PASTEUR