March, 2022

IA325

Algorithmic Information & Artificial Intelligence

Micro-study

teaching.dessalles.fr/FCI

Name: Aloysio GALVÃO LOPES

A SARS-CoV-2 Phylogenetic tree based on the Normalized Compression Distance

Abstract

This project aims at creating a phylogenetic tree based on concepts of algorithmic information. More precisely, the normalized compression distance (NCD) is used as a base to compute a distance matrix that's later used to compute a phylogenetic tree for SARS-CoV-2. The code for the project can be found in https://github.com/aloysiogl/phylo-covid.

Problem

Building phylogenetic trees such as [3] often requires sequence alignment as well lots of domain-specific knowledge. Using AIT concepts, it's possible to build such trees as shown in [5] with a very general framework. In this sense, AIT methods are also much simpler and faster to develop, which might be ideal for fast prototypes and to gain further intuition about the problem. With that in mind, this project aims at reproducing the genealogy of the SARS-CoV-2 virus.

Method

In a first moment, to obtain the genetic information, 52.17 GB of genetic data from SARS-CoV-2 from within the whole 2020 COVID-19 pandemic period to date has been downloaded. The source was the National Center for Biotechnology Information (NCBI) [2].

The dataset consisted of 883,020 strands of RNA from which I sampled 1247 strands by limiting to only 2 sequences per day. After that, a distance matrix D was computed, containing every pairwise distance on the dataset. The distance used was a slight modification (to make it symmetrical) of the Normalized Compression Distance. This distance is shown in Equation 1, where $Z(x)$ is the length of the compressed string x using python's `zlib`.

$$NCD(x,y) = \frac{\frac{Z(xy)+Z(yx)}{2} - min(Z(x), Z(y))}{max(Z(x), Z(y))} \tag{1}$$

After that, I clustered the dataset using D with hierarchical clustering [6] with complete linkage. The reason for that is that I wanted to work with a small quantity of elements, but I wanted all elements to represent single or almost single COVID-19 lineages. Using complete linkage ensures that all elements in a cluster are close to each other.

I took 100 clusters and filtered out all small clusters (with less than 8 elements). Then, I computed a second distance matrix D_c which contains the distance between each pair of clusters, defined as the average NCD between each pair of points on each cluster.

Finally, I used D_c alongside the package `biopython`[1] to compute the phylogenetic tree using the neighbor joining method. I also computed an embedding for the clusters in 2D using scikit-learn's MDS [6].

Results

The three main results are: the phylogenetic tree shown in Figure 1, the clusters embedded in 2D via MDS shown in Figure 2 and Table 1 which compares the onset dates of each variant with the ones obtained by clustering. It's important to notice that the clusters were mostly pure and represented either a WHO named variant or a variant without WHO name, which is represented using an abbreviation of the Pango [4] nomenclature in the images.

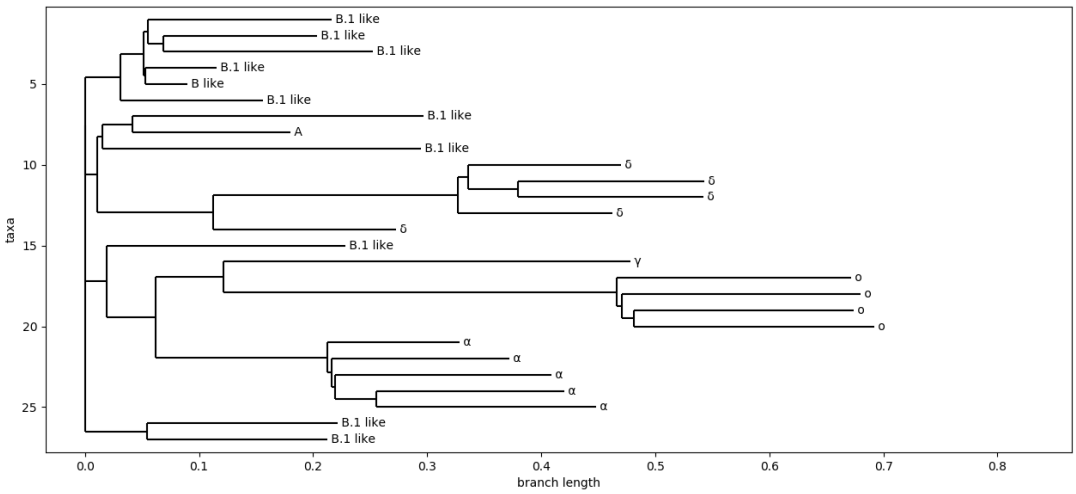

Figure 1: Phylogenetic tree built using neighbor joining from `biopython`[1].

Table 1: Date information (YYYY/MM/DD) for clusters with WHO variant names

Variant	Cluster start date	Cluster end date	Approximate onset date
α	2021/01/05	2021/11/24	2020/11/XX
δ	2021/07/14	2022/03/04	2021/05/XX
γ	2021/05/05	2022/03/18	2021/02/XX
o	2021/12/31	2022/03/21	2021/11/XX

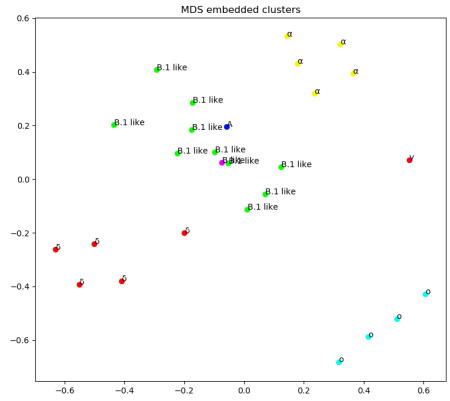

Figure 2: RNA clusters embedded in 2D using MDS.

Discussion

Foremost, we notice that having pure clusters means that the NCD distance was successfully able to bring together RNA strands of the same variants. We see the same effect as well in the phylogenetic tree. It's also interesting to notice that the MDS embedding shows the A strain (original one found in Wuhan) in the center suggesting that mutations irradiated in all directions from the origin as we would expect.

Using the tool nextstrain[3] it's possible to find very accurate representations of SARS-CoV-2 phylogeny. By comparing that to my results, it's noticeable that my tree has successfully captured the separation between δ and o. We also see that o, α and γ are close, which is indeed true, the only mistake is that γ should be closer to α than o.

As a small subsample has been taken, it's hard to determine precise start dates as, near those dates, very few individuals are infected with the concerned variants. I however, tried to work with as many individuals as I could, just linearly scanning the entire dataset took me some minutes and computing the pairwise distances for the reduced dataset took about 3 hours in a 12 core CPU using a parallelized implementation.

In conclusion, the relations shown in the phylogenetic tree built in this project are meaningful and provide insights about the evolution of the pandemic. The great advantage of the AIT is the generality and the lack of need for alignment or any kind of preprocessing.

Bibliography

[1] *Biopython · Biopython*. `https://biopython.org/`.

[2] *National Center for Biotechnology Information*. `https://www.ncbi.nlm.nih.gov/`.

[3] *Nextstrain*. `https://nextstrain.org/`.

[4] *Pango Network – Helping track the transmission and spread of SARS-CoV-2*. `https://www.pango.network/`.

[5] *Phylogeny of the COVID-19 Virus SARS-CoV-2 by Compression*. `https://www.biorxiv.org/content/biorxiv/early/2020/07/23/2020.07.22.216242.full.pdf`.

[6] *scikit-learn: machine learning in Python — scikit-learn 1.0.2 documentation*. `https://scikit-learn.org/stable/`.