Instituto Universitario de Investigación en Ingeniería de Aragón
Universidad Zaragoza

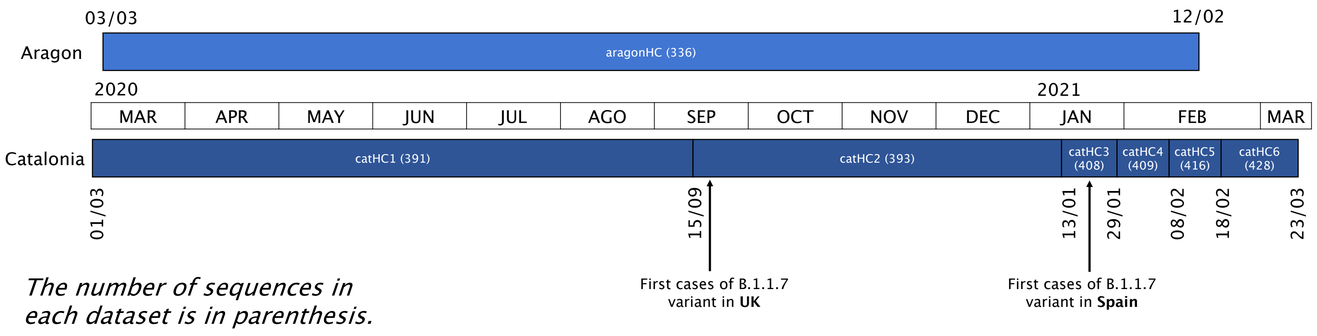
A phylogenetic study of COVID-19 data from Aragon and Catalonia over a year: learning Bioinformatics during a world pandemic

Álvaro García-Díaz, Alejandro Gómez-González, Adrián Martín-Marcos, Fernando Peña[1], Álvaro Romeo, José Manuel Sánchez-Aquilué, Alba Vallés, Elvira Mayordomo

Motivation

- Teaching undergraduate Bioinformatics through a cooperative project in the context of a pandemic (Spring Semester 2021).
- Analyze the variability of SARS-CoV-2 virus in the Aragon area and compare it with Catalonia.
- Track the UK-variant of the virus (B.1.1.7), which was becoming predominant at the time.

Data



The number of sequences in each dataset is in parenthesis.

- The data used consisted of all the high coverage nucleotide sequences available at GISAID coming from Aragon and Catalonia submitted up to March 23rd, 2021.
- Data was split chronologically into 7 datasets of approximately 400 sequences each.
- The sequence of the B.1.1.7 variant was included in each dataset.
- Each student had the task of analyzing one of the datasets.

Methodology

The analysis of each dataset was divided into three tasks:

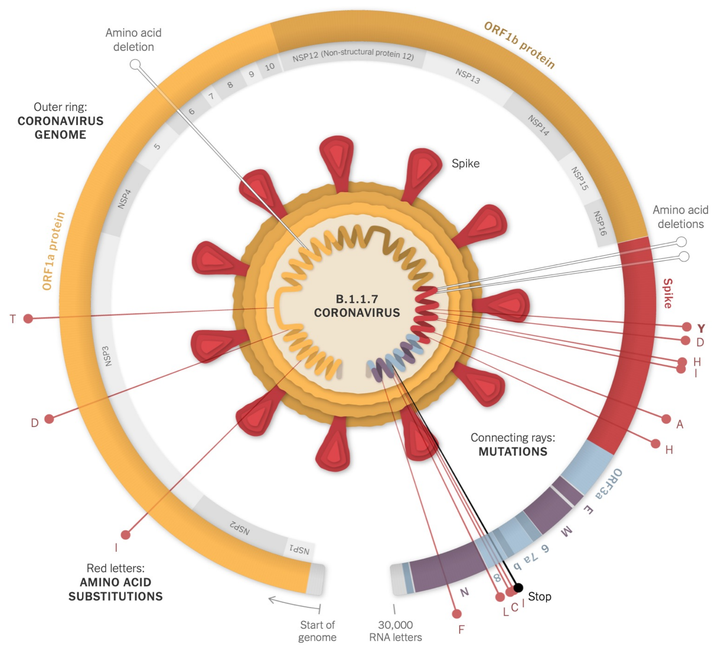
1. **Multialignment:** This task was performed using Clustal Omega and MAFFT. SARS-CoV-2 reference sequence NC_0455512.2 was used for guiding the multialignment. For instance,

```
$ mafft --auto --addfragments gisaid_seqs.fasta RefSeqWuhan.fasta \
   > alignment.fasta
```

2. **Conservation index calculation:** It was computed for each column i of the alignment using an entropy method, paying special attention to the ORF1a, Spike, ORF3a and N genes:

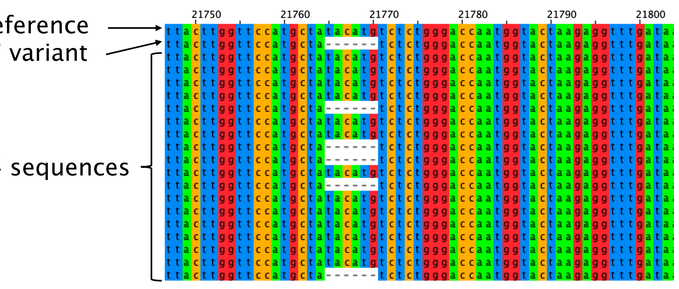
$$C(i) = -\sum_{v\in\{A,G,C,T\}} f_v(i)\ln\big(f_v(i)\big)$$

3. **Phylogeny reconstruction:** From the best multialignment, a phylogeny tree was reconstructed with maximum likelihood methods from FastTree and RAxML. Different evolution models, in particular Jukes-Cantor and Generalized Time-Reversible (GTR), were tested, and the best scoring phylogeny was selected for each dataset.

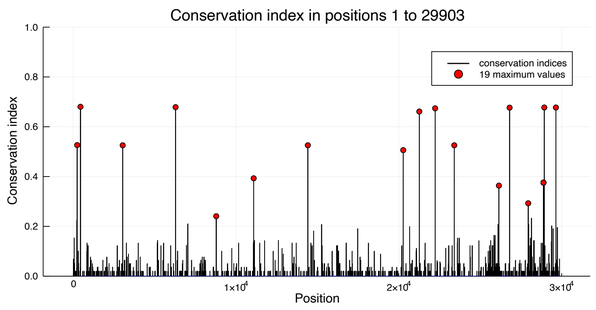


Source: The New York Times

Results



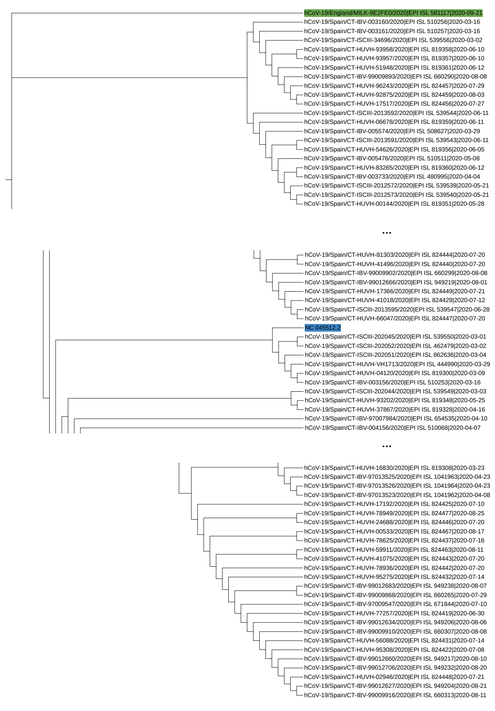
Fragment of a **multialignment** with catHC4 dataset.



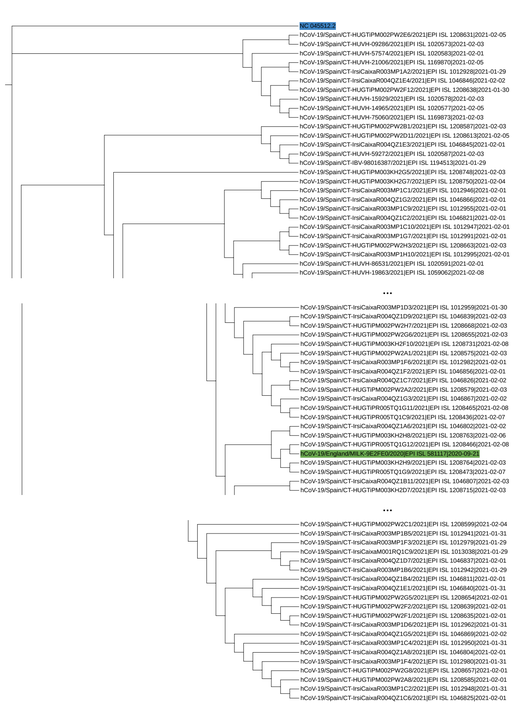
Conservation indices from aragonHC multialignment.

Reconstructed **phylogenies** from different multialignments.

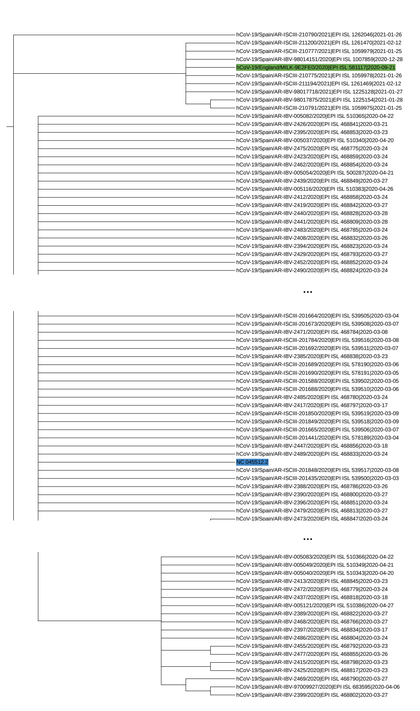
*SARS-CoV-2 reference sequence is marked in **blue**, and the B.1.1.7 variant in **green**.*



catHC1 (before first cases of B.1.1.7)



catHC4 (after first cases of B.1.1.7)



aragonHC

Conclusions

- SARS-CoV-2 virus can be used to teach the main genomic bioinformatics tools to Computer Science students.
- FastTree and RAxML are both able to obtain meaningful phylogenies for datasets of size 400 and sequence size 30k bp for the test case considered.
- The results obtained from the phylogenies and conservation index analysis are consistent with the national and European periodical reports. The lack of detailed local reports prevents further comparisons.

Escuela de Ingeniería y Arquitectura
Universidad Zaragoza

References

1. ECDC. Threat Assessment Brief: Rapid increase of a SARS-CoV-2 variant with multiple spike protein mutations observed in the United Kingdom. *European Centre for Disease Prevention and Control*.
2. ESCUELA DE INGENIERÍA Y ARQUITECTURA DE ZARAGOZA. *Directrices y recomendaciones para la impartición de la docencia en el segundo semestre del curso 2020-2021*. 11 January 2021.
3. ESPAÑA, CENTRO DE COORDINACIÓN DE ALERTAS Y EMERGENCIAS SANITARIAS. *Actualización nº 386. Enfermedad por el coronavirus (COVID-19)*. 31 May 2021.
4. EUROPEAN COMMISSION. *European Commission. Communication from the Commission to the European Parliament, the European Council and the Council, A united front to beat COVID-19*.
5. INE. Edad Media de la Población por comunidad autónoma, según sexo. *Instituto Nacional de Estadística*.
6. INE. Población por comunidades y ciudades autónomas y sexo. *Instituto Nacional de Estadística*.
7. LIU, Kevin, LINDER, C. Randal and WARNOW, Tandy. RAxML and FastTree: Comparing Two Methods for Large-Scale Maximum Likelihood Phylogeny Estimation. WU, Rongling (ed.), *PLoS ONE*. 21 November 2011. Vol. 6, no. 11, pp. e27731.
8. PAIS, Fabiano Sviatopolk-Mirsky, RUY, Patrícia de Cássia, OLIVEIRA, Guilherme and COIMBRA, Roney Santos. Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology*. December 2014. Vol. 9, no. 1, pp. 4.

[1]**Contact:** 756012@unizar.es

Scan the code to download the paper, supplementary material, and GISAID acknowledgements →