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

Afiliación: *Computer* Science for Complex *System Modeling (COS2MOS)*
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, [1]e-mail: *756012@unizar.es*

Abstract

We present here a one year variability analysis of SARS-CoV-2 virus in the Aragon area, comparing it to neighboring Catalonia and focusing on the UK-variant. Research was performed within a Bioinformatics course. Both a detailed assessment on the value of this methodology and novel research conclusions were extracted.

Background

The current world pandemic has been a challenge for education at all levels. We present here a study case of hybrid teaching of a Bioinformatics undergraduate class at the University of Zaragoza, Spain, where the final lab project was fully based on the variability analysis of SARS-CoV-2 virus in the Aragon area, comparing it to the nearby Catalonia area and to the UK-variant which was becoming predominant at the time. We intend to both investigate the teaching value of this methodology and to extract novel research conclusions on the cooperative work performed in the classroom.

The course lasted from February to May 2021 and consisted of online classes and in-person computer laboratory sessions which were the first physical sessions since March 2020 except for exams [2].

According to the National Institute of Statistics [5, 6] Aragon had a population of 1,319,291 in 2019 with an average age 44.81 which was 1.43 years higher than that of Spain. Catalonia had a population of 7,675,217 in 2019 and average age 42,83. According to the Spain Health Ministry, mortality rate per 100,000 inhabitants due to COVID-19 was 263.96 in Aragon, 186.90 in Catalonia, and 168.5 in Spain in May 2021 [3].

Methods and data

Data

GISAID maintains the world's largest repository of SARS-CoV-2 complete genome sequences with over 1,800,000 submissions worldwide until May 2021. This included 30,370 submissions from Spain with 999 from Aragon and 8,050 from Catalonia[1].

Variant of concern B.1.1.7 was first detected in the UK in September 2020 with an estimated increased transmissibility of up to 70% over previous variants [1], In Spain it was first detected in January 2021.

The data used consisted of all the high coverage[2] sequences available at GISAID with geographic origin in Aragon and Catalonia and submitted up to March 23rd 2021. Data was split chronologically into 7 (minimally overlapping) sets of approximately 400 sequences each (see Table 1). Each student was assigned one of them.

Variability analysis

The analysis of the data was divided into three main tasks on each dataset, multialignment, phylogeny reconstruction, and conservation index calculation. Multialignment allows us to compare biological sequences on which evolutionary events and errors may have occurred, the conservation index is computed on the multialignment to measure the conservation of each position, and finally a

[1] A European Commission Recommendation dated 19 January 2021 states that all EU Member States should reach a capacity of sequencing at least 5% - and preferably 10% - of positive test results [EC]. To reach this target, there needs to be a significant increase in sequencing capacity in Member States.

[2] According to GISAID, high coverage means that only entries with less than 1% of undefined bases (NNNs) and no insertions and deletions unless verified by the submitter are tolerated.

phylogeny is a relational hierarchy between a set of data.

The multialignment was performed using the most widely used tools Clustal Omega and MAFFT [8] for which all available different options were tested and that the students have used in previous laboratory sessions. COVID-19 reference sequence NC_0455512.2 was used for the guided multialignment options of. Each student selected the best alignment for their dataset.

The conservation index was computed for each position of the alignment as implemented by each student. The mutations defining variant B.1.1.7 (UK) [1] were particularly observed.

From the best multialignment, a phylogeny was reconstructed with maximum likelihood methods of widely used FastTree and RAxML computational tools [7]. Different evolution models were tested and the best scoring phylogeny was selected for each dataset.

Results

Data from Aragon and Catalonia has provided an interesting test case (thanks to the GISAID initiative, see acknowledgements section). The main objects of our study have been computational phylogeny methodology comparison and observation of SARS-CoV-2 variants of interest and concern. The different computational phylogeny methods have been tested for seven different sets of similar size (around four hundred sequences of size close to 30k bp each) for both accuracy and time complexity, obtaining a detailed comparison for a large real data collection.

The analysis of data from March 2020 (01/03/2020) to March 2021 (23/03/2021) has concluded that the UK-variant appeared earlier in Catalonia than in Aragon, in particular the relevant B.1.1.7 clade of the phylogeny of the Catalan sequences can be observed in the trees corresponding to the second dataset (sequences up to mid January 2021, for which B.1.1.7 constitutes 9%) and later, while it appears only in late January in Aragon and the size of the corresponding clade is smaller (2,4%). Conservation index was used as a second variant corroboration, and the positions defining the UK-variant were consistently less preserved in the datasets where the B.1.1.7 clade was present.

Detailed results can be found in the repository of supplementary materials https://github.com/ferpb/Bioinformatics2021

The students have been able to fully understand the advanced concepts of multiple sequence alignment, phylogeny and conservation index, implementing the later one and using at expert level several available tools for both multialignment and phylogeny reconstruction. The students' involvement in the class was very high and we attribute this fact to the chosen case of use.

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.

Acknowledgements

See GISAID full acknowledgements at https://github.com/ferpb/Bioinformatics2021

Research supported in part by Spanish Ministry of Science and Innovation grants TIN2016-80347-R and PID2019-104358RB-I00 and by the Science dept. of Aragon Government: Group Reference T64_20R (COSMOS research group)

REFERENCES

[1]. ECDC, 2020. 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*. Online. 20 December 2020. [Accessed 6 June 2022]. Retrieved from: https://www.ecdc.europa.eu/en/publications-data/threat-assessment-brief-rapid-increase-sars-cov-2-variant-united-kingdom

[2]. ESCUELA DE INGENIERÍA Y ARQUITECTURA DE ZARAGOZA, 2021. *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,

2021. *Actualización nº 386. Enfermedad por el coronavirus (COVID-19)*. . 31 May 2021.

[4]. EUROPEAN COMMISSION, 2021. *European Commission. Communication from the Commission to the European Parliament, the European Council and the Council, A united front to beat COVID-19*. Online. 19 January 2021. COM(2021)35. [Accessed 6 June 2022]. Retrieved from: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM%3A2021%3A35%3AFIN

[5]. INE, 2021a. Edad Media de la Población por comunidad autónoma, según sexo. *Instituto Nacional de Estadística*. Online. 2021. [Accessed 6 June 2022]. Retrieved from: https://www.ine.es/jaxiT3/Datos.htm?t=3198

[6]. INE, 2021b. Población por comunidades y ciudades autónomas y sexo. *Instituto Nacional de Estadística*. Online. 2021. [Accessed 6 June 2022]. Retrieved from: https://www.ine.es/jaxiT3/Datos.htm?t=2853

[7]. LIU, Kevin, LINDER, C. Randal and WARNOW, Tandy, 2011. 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. DOI 10.1371/journal.pone.0027731.

[8]. PAIS, Fabiano Sviatopolk-Mirsky, RUY, Patrícia de Cássia, OLIVEIRA, Guilherme and COIMBRA, Roney Santos, 2014. Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology*. December 2014. Vol. 9, no. 1, pp. 4. DOI 10.1186/1748-7188-9-4.

Table 1. Chronological datasets

dataset	from (coll.)	to (coll.)	total
aragonHC	03/03/2020	12/02/2021	336
catHC1	01/03/2020	15/09/2020	391
catHC2	15/09/2020	13/01/2021	393
catHC3	13/01/2021	29/01/2021	408
catHC4	29/01/2021	08/02/2021	409
catHC5	08/02/2021	18/02/2021	416
catHC6	19/02/2021	23/03/2021	428