Creative Education, 2021, 12, 193-202
https://www.scirp.org/journal/ce
ISSN Online: 2151-4771
ISSN Print: 2151-4755

Use of Bioinformatics Technologies and Databases to Teach Analysis of Genetic Sequences to Undergraduate Students in Physics, Biotechnology, and Biology: The Specific Case of the SARS-CoV-2 Spike Protein

Michelle Abigail Fuentes-Acosta, Jorge Mulia-Rodríguez, Daniel Osorio-González*

Molecular Biophysics Laboratory, Faculty of Sciences, Autonomous University of State of Mexico, Toluca, Mexico
Email: *dog@uaemex.mx

How to cite this paper: Fuentes-Acosta, M. A., Mulia-Rodríguez, J., & Osorio-González, D. (2021). Use of Bioinformatics Technologies and Databases to Teach Analysis of Genetic Sequences to Undergraduate Students in Physics, Biotechnology, and Biology: The Specific Case of the SARS-CoV-2 Spike Protein. *Creative Education, 12,* 193-202.
https://doi.org/10.4236/ce.2021.121014

Received: December 20, 2020
Accepted: January 22, 2021
Published: January 25, 2021

Abstract

Worldwide, the implementation of computational tools has allowed the advancement of our understanding, specifically, biotechnological information tools have made possible to tackle global challenges such as the COVID-19 pandemic; therefore, such tools are essential for science students. We propose an activity to teach the use of information and database biotechnologies and their utility for the alignment and comparison of sequences. We use as query sequence the corresponding to SARS-CoV-2 Spike protein in its closed state, and we compare it with 200 sequences obtained from the NCBI databases to identify the mutations and their domain. In the results, we show the frequency of the mutations, domain, and country of the isolated SARS-CoV-2 genome. The activity we propose is aimed at first-year undergraduate students in physics, biology, and biotechnology.

Keywords

Spike, SARS-CoV-2, Bioinformatic Technologies

1. Introduction

SARS-CoV-2, the virus responsible for the COVID19 pandemic, has been the subject of several investigations. It was outbroken in late December 2019 in Wuhan, China, and on March 11, the World Health Organization increased the status to a maximum emergency and declared COVID-19 a Pandemic (World Health Organization, 2020a). SARS-CoV-2 is the causative agent of the severe acute respiratory syndrome and is characterized by an estimated average incubation period of 6.4 days, although it can vary from 2 to 11 days, and in some cases, it extends up to 14 days (Backer, Klinkenberg, & Wallinga, 2020; Lauer et al., 2020; Liu, Gayle, Wilder-Smith, &Rocklöv, 2020).

For the development of the activity, it is necessary to describe the genomic organization and the main characteristics of the virus. In general, CoVs are pleomorphic viruses that contain crown-shaped peplomers with a size of 80 - 160 nM and positive-sense single-stranded RNA (ssRNA+) genome of 27 - 32 kb. The SARS-CoV-2 virus has a 96% similarity with the genome of a bat coronavirus (RaTG13) and similar composition to SARS-CoV-1 (70% - 80%) and MERS-CoV (50%); all of them have emerged from animal reservoirs and are classified within the β-coronavirus (Lu et al., 2020; Zhou et al., 2020; Sun, Yang, Sun, & Su, 2020; Wu, Zhao, Yu, Chen, Wang, & Song, 2020; Taiaroa et al., 2020). The SARS-CoV-2 genome has been completely sequenced and is composed of 29,033 nucleotides that include at least ten open reading frames (ORF), of which the ORF1a/b correspond to 2/3 of the viral RNA and are translated into two large overlapping polyproteins, pp1a and pp1ab, encoding 16 non-structural proteins (NSPs); while the rest of the ORF encodes structural and accessory proteins (Li, Geng, Peng, Meng, & Lu, 2020; Taiaroa et al., 2020). The structural proteins that make up the virus are Spike (Spike or S), nucleocapsid (N), matrix (M), and envelope (E) (Cascella, Rajnik, Cuomo, Dulebohn, & Di Napoli, 2020; Guo et al., 2020). In this activity, we will approach the Spike glycoprotein which consists of a class I fusion trimer in a metastable prefusion conformation and undergoes substantial structural rearrangement and numerous conformational changes to fuse the viral membrane with the host cell membrane. In fact, Spike is the coronavirus's essential surface protein, and some vaccines are made using fragments of its mRNA (World Health Organization, 2020b; Wrapp et al., 2020). The Spike protein comprises two subunits, S1 and S2, where S1 is responsible for binding to the host via the receptor-binding domain (RBD) and S2 for viral and host membrane fusion. In SARS-CoV-2, RBD (~21 kDa) is located in the 332 - 524 region of the Spike protein and exhibits strong binding to its receptor, the angiotensin-converting enzyme 2 (ACE2) (Tai et al., 2020). It should be noted that SARS-CoV-1 and SARS-CoV-2 have an unusually high proportion of evolutionary convergent amino acid sites in the RBD of Spike and its associated protein, ORF3a, which could explain that both viruses are adapted for the same receptor (ACE2) in addition to both having general structural homology (Wrapp et al., 2020; Wu, 2020; Zhou et al., 2020).

The complementary domains of the Spike protein are the N-terminal domain (NTD), the fusion peptide (FP), the heptads 1 and 2 (HR1, HR2), the central helix (CH), and the connecting domain (CD) (**Figure 1**), whose specific functions are fundamental in the fusion process between the viral envelope and the membrane of the host cell.

The COVID 19 pandemic is a phenomenon that impacted all sectors of society.

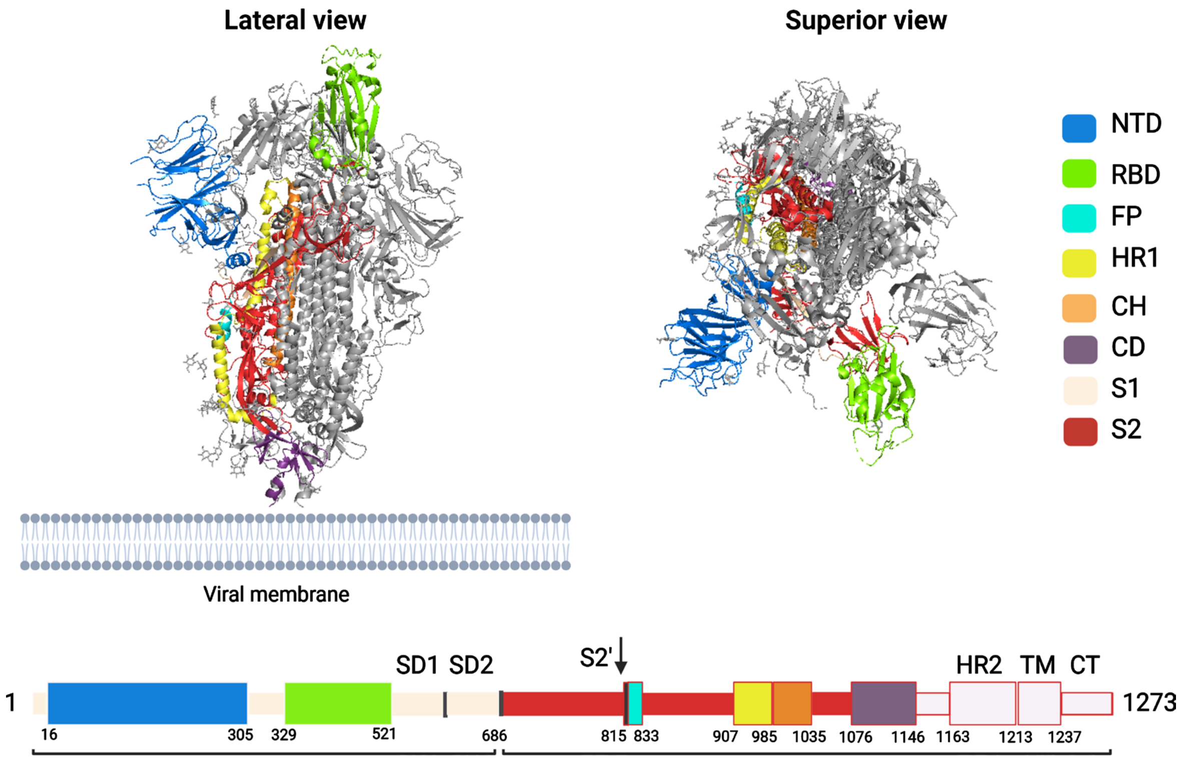


Figure 1. Structure of the SARS-CoV-2 Spike protein. The localization of the domains is based on Walls, Park, Tortorici, Wall, McGuire, & Veesler, 2020.

Particularly in education, it was an event that generated disruptive thinking among students and teachers due to the urgent need to use information and communication technologies as an indispensable element for the teaching-learning process. The best educational practices require that the contents of the courses are contextualized within an immediate social reality, and therefore, in this paper, we present a proposal for teaching the use of NCBI databases for the alignment and comparison of sequences of the Spike protein from SARS-CoV-2.

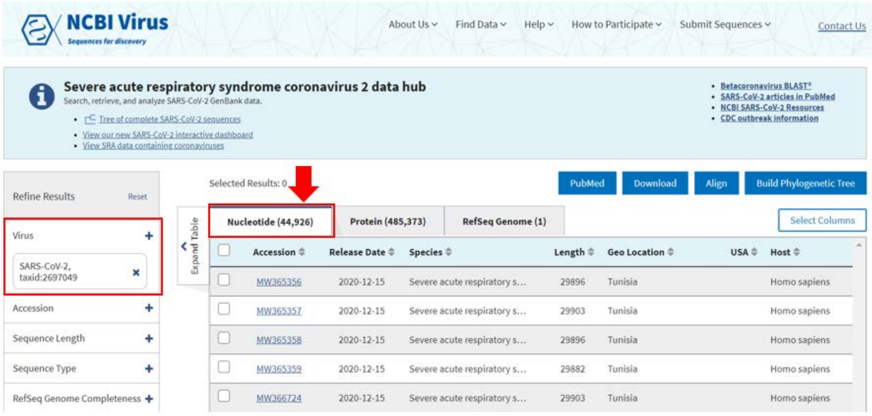
2. Materials and Methods

The activity proposed in this paper uses technological resources and open-source scientific tools to identify SARS-CoV-2 recent genetic mutations from a sample of genomes isolated from patients from different nationalities and regions. To exemplify the activity, we analyzed 200 different sequences of SARS-CoV-2, chosen randomly from more than 44,000 available sequences, to verify the genomic organization and compare them with a reference sequence to find and punctually identify the existing variables of Spike protein.

The database used to access the sequences is available at the National Center for Biotechnology Information (NCBI) (2020a). The reference sequences or query can be obtained by downloading the FASTA file from the Protein Data Bank (PDB) using the identification code 6VYB, while the subject sequences were extracted from the NCBI Gen Bank.

As the starting point of the activity, the professor must instruct the students to enter the NCBI database that has been specially designated for the SARS-CoV-2 sequences (Figure 2), and that is daily updated with new genomes. When entering the database, the students must select the Nucleotides section to quickly consult the accession number, the release date, the species, the length of the sequence, the geographic location, the host, and the date of collection. The results are then filtered by the host so that the chosen sequences correspond to humans (Homo sapiens). The access number link of each sequence redirects to the Gen Bank page where they can consult information regarding the source, authors, title of the work, the journal where it was published, the laboratory where it was submitted for analysis and sequencing, the length, and the amino acid sequence. From this document, the Spike protein's ID is obtained, which will be used as the subject sequence for the alignment.

As a next step, it is recommended that the student understands the importance of finding differences between amino acid sequences to align the chosen Spike sequences later using the National Center for Biotechnology Information (2020b) protein-BLAST tool (Figure 3). As a query sequence, the FASTA file downloaded from the PDB is selected, and in the subject section, the IDs of the Spike protein obtained from the NCBI Gen Bank are captured. For further analysis, students are



Severe acute respiratory syndrome coronavirus 2 data hub

Search, retrieve, and analyze SARS-CoV-2 GenBank data.

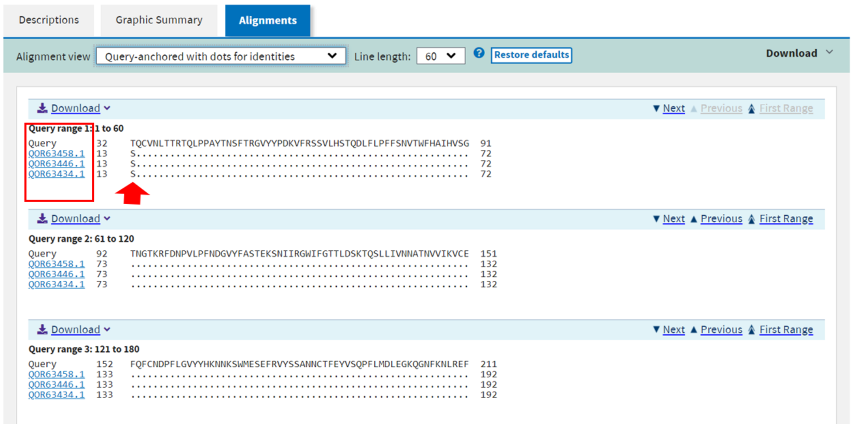
- Tree of complete SARS-CoV-2 sequences
- View our new SARS-CoV-2 interactive dashboard
- View SRA data containing coronaviruses

SARS-CoV-2, taxid:2697049

Selected Results: 0

Accession	Release Date	Species	Length	Geo Location	USA	Host
MW365356	2020-12-15	Severe acute respiratory s...	29896	Tunisia		Homo sapiens
MW365357	2020-12-15	Severe acute respiratory s...	29903	Tunisia		Homo sapiens
MW365358	2020-12-15	Severe acute respiratory s...	29896	Tunisia		Homo sapiens
MW365359	2020-12-15	Severe acute respiratory s...	29882	Tunisia		Homo sapiens
MW366724	2020-12-15	Severe acute respiratory s...	29903	Tunisia		Homo sapiens

Figure 2. SARS-CoV-2 genome database from the NCBI.



Alignment view: Query-anchored with dots for identities
Line length: 60
Restore defaults

```
Query range 1: 1 to 60
Query        32    TQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSG  91
QOR63458.1   13    S...........................................................  72
QOR63446.1   13    S...........................................................  72
QOR63434.1   13    S...........................................................  72

Query range 2: 61 to 120
Query        92    TNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCE  151
QOR63458.1   73    ............................................................  132
QOR63446.1   73    ............................................................  132
QOR63434.1   73    ............................................................  132

Query range 3: 121 to 180
Query        152   FQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREF  211
QOR63458.1   133   ............................................................  192
QOR63446.1   133   ............................................................  192
QOR63434.1   133   ............................................................  192
```

Figure 3. Identification of the Spike-SARS-CoV-2 variables using the NCBI protein-BLAST tool.

required to select the Alignments section and visualized it as pairwise with dots for identities or query-anchored with dots for identities. This type of visualization allows students to identify the variants quickly and to get involved with the analysis. In the event of an offset in the subject sequences with respect to the query, the students must capture it as an observation or a note in the resulting database since this information must be taken into account when analyzing the location of the variants in the protein domains, which is carried out utilizing the Graphics resource of protein-BLAST based on the position of the domains in the query sequence (**Figure 4**).

If an offset has been identified, the student is instructed to make the corresponding adjustment of the variants' positions concerning the query sequence. Subsequently, it is identified whether the variants in question are located in a

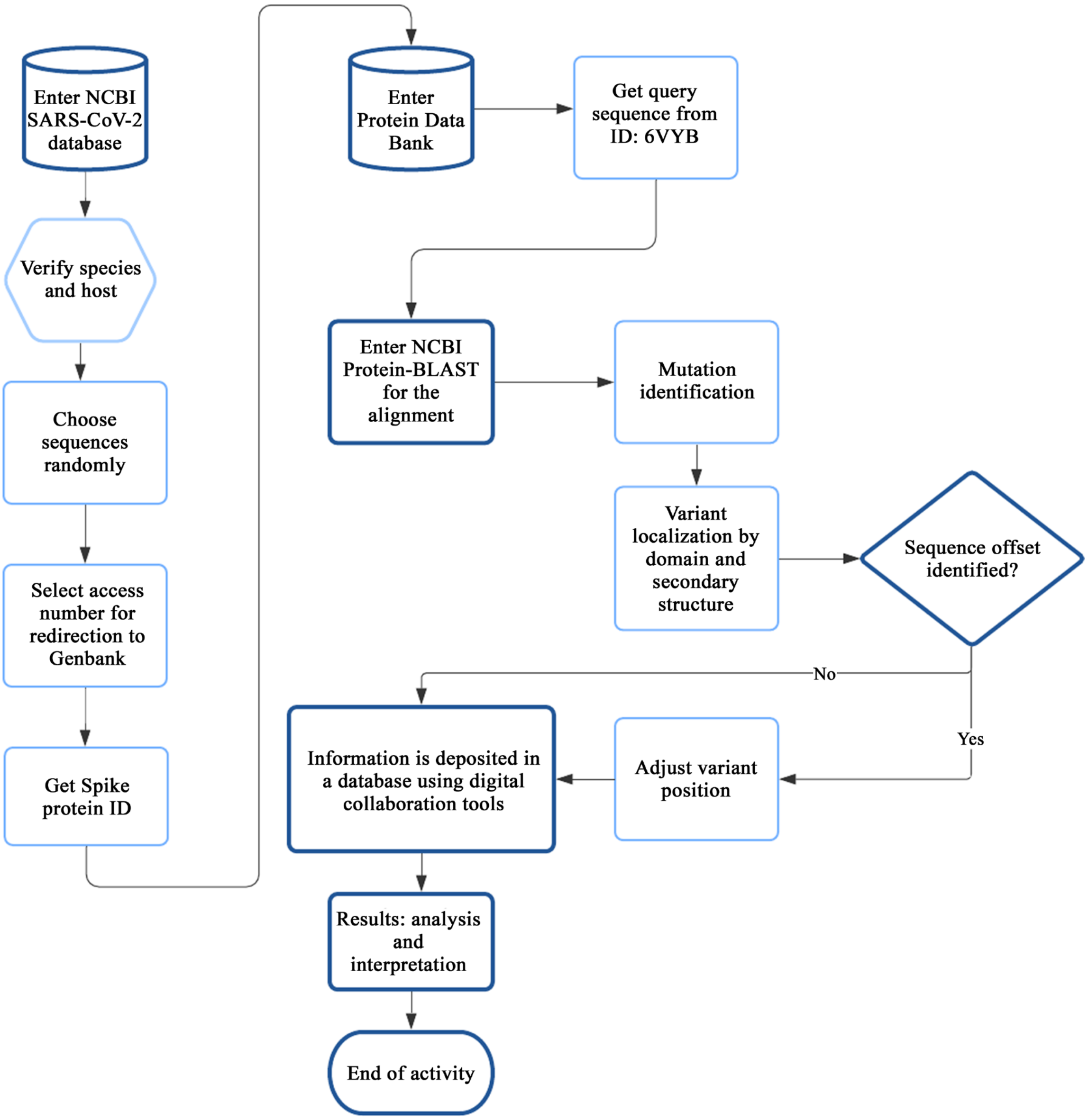

Enter NCBI SARS-CoV-2 database
Verify species and host
Choose sequences randomly
Select access number for redirection to Genbank
Get Spike protein ID
Enter Protein Data Bank
Get query sequence from ID: 6VYB
Enter NCBI Protein-BLAST for the alignment
Mutation identification
Variant localization by domain and secondary structure
Sequence offset identified?
No
Yes
Adjust variant position
Information is deposited in a database using digital collaboration tools
Results: analysis and interpretation
End of activity

Figure 4. Activity diagram.

structured region using the Sequence section of the PDB where the Spike protein is described in terms of a symbology associated with the secondary structure.

Finally, the information is deposited in a database that includes sequence ID, geographic region, Spike protein ID, list of the identified variants, the domain in which each variant is located, type of secondary structure, and observations. This database is made available to all participants through a digital platform as a shared resource.

3. Results

We propose an activity-oriented to the teaching of bioinformatics; in this example, we analyzed 200 randomly selected sequences of the SARS-CoV-2 Spike protein, which were obtained from 34 regions of the world, being the most frequent the United States of America, which corresponds to 17.5% (**Figure 5**), followed by India and Germany. Defining the geographic distribution of the genomes, especially of the specific variants, is of great importance for developing strategies to contain the virus at regional level. Koyama and collaborators (2020) carried out a large-scale analysis similar to this activity, in which 15,755 sequences from 68 countries were analyzed, and they found that most of the genomes came from the United States of America. Therefore, with this activity, the students can extrapolate the information obtained to a real situation of global importance.

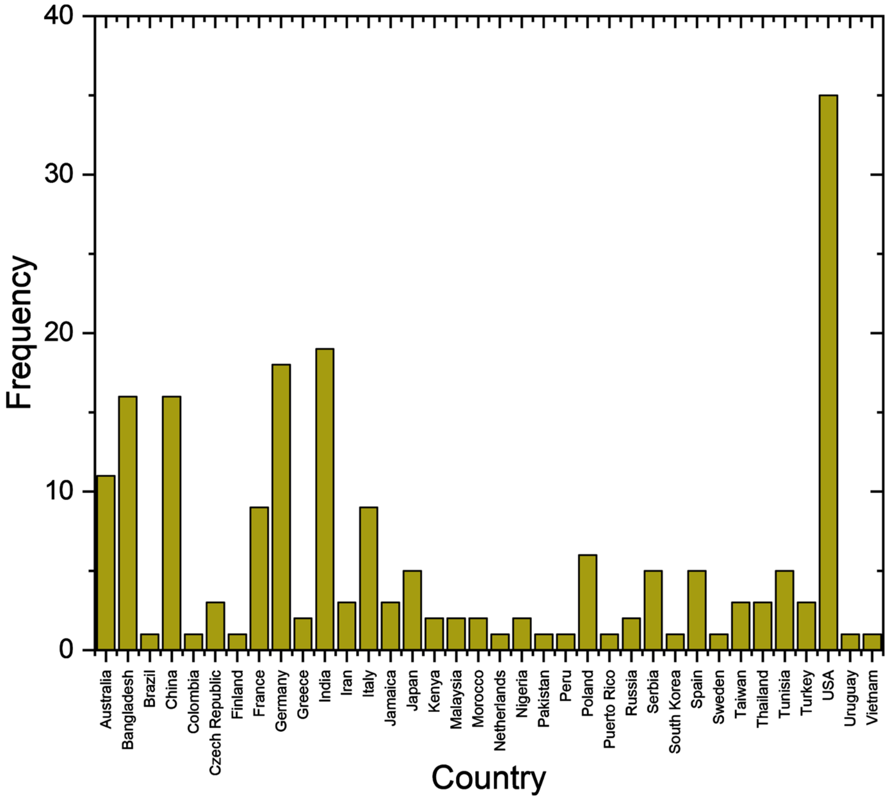


Figure 5. Sequences frequency by country.

In this analysis, 1512 mutation events were found, corresponding to 82 different variants of which 57.4% are located in the SD domain of the Spike protein (**Figure 6**), the most frequent being E607Q and P986K (**Figure 7**). The D614G variant should be highlighted since it has been identified in different large-scale studies and constitutes one of the most common clades of Spike variants that amino acid is responsible for the initial interaction of the virus with the human host cell (Mercatelli & Giorgi., 2020; Koyama, Platt & Parida, 2020).

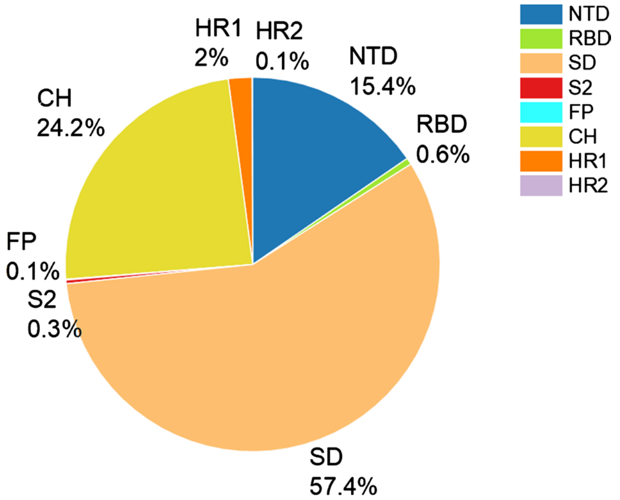


Figure 6. Percentage of mutations by location in the SARS-CoV-2 Spike protein domains.

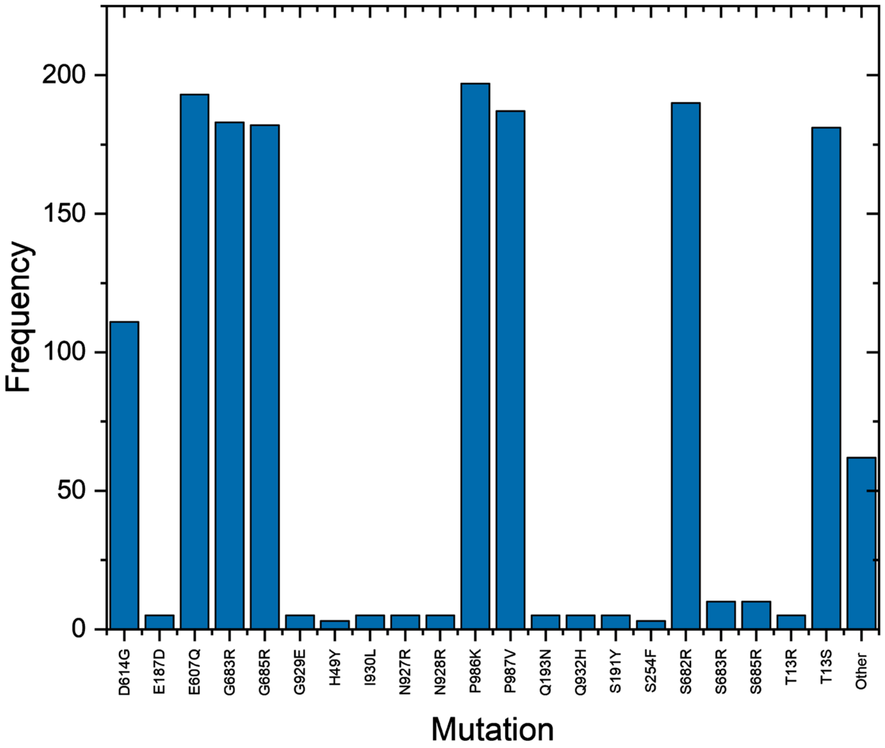


Figure 7. Frequency of mutations. The most representative mutations obtained from the random analysis are shown.

The activity allows students to obtain preliminary data from the sequences and generate hypotheses about the biological significance of the variants and their implications; it should be taken into account that the databases are updated continuously and that the virus's intrinsic characteristics can lead to new mutations.

We achieved the main goal of this activity, which is that the students can understand the importance of biotechnological information tools and apply them to tackle real problems, this is particularly significant for first graders since they are beginning to shape their path in science and need a vision as broad and complete as possible of the tools available according to their level. As for the learning process results, all students performed the activity satisfactorily, enthusiastically expressed what they learned from the activity, and showed interest in pursuing an advanced course with a focus on bioinformatics.

3. Conclusion

The implementation of computational tools for the analysis of biological information is of great relevance worldwide. In this activity, these tools focus on the analysis of sequences whose interest, particularly during the SARS-CoV-2 outbreak, is such that several large-scale studies have been carried out to understand the viral genome's variability since this influences the strategies and policies implemented to contain and understand the outbreak. This activity allows students to apply their knowledge to a relevant topic and expand their global vision of bioinformatics impact on science. The activity proposed in this work is a clear example that digital literacy is essential for teaching processes during times of pandemic, and its use will transcend beyond it. We also show that bioinformatics is a necessary complement in sciences such as Physics, Biology, and Biotechnology because it enables efficient experimental methods.

The nature of SARS-COV-2 allows it to be a more stable virus, with fewer mutations than others such as HIV or Hepatitis C, however, the specific identification of its mutations is of utmost importance since they could become resistant or unrecognizable by the currently approved vaccines. Of course, this work's activity is an academic exercise whose purpose is only to introduce students to the management of biotechnological information tools and motivate them to take advanced courses that allow them to develop skills inherent to their professional training.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Backer, J. A., Klinkenberg, D., & Wallinga, J. (2020). Incubation Period of 2019 Novel Coronavirus (2019-nCoV) Infections among Travellers from Wuhan, China, 20-28 January 2020. *Eurosurveillance, 25,* Article ID: 2000062. https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Di Napoli, R. (2020). *Features, Evaluation and Treatment Coronavirus (COVID-19).* Treasure Island, FL: Stat Pearls Publishing.

Guo, Y. R., Cao, Q. D., Hong, Z. S., Tan, Y. Y., Chen, S. D., Jin, H. J. et al. (2020). The Origin, Transmission and Clinical Therapies on Coronavirus Disease 2019 (COVID-19) Outbreak—An Update on the Status. *Military Medical Research, 7,* Article No. 11. https://doi.org/10.1186/s40779-020-00240-0

Koyama, T., Platt, D., & Parida, L. (2020). Variant Analysis of SARS-CoV-2 Genomes. *Bulletin of the World Health Organization, 98,* 495-504. https://doi.org/10.2471/BLT.20.253591

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R. et al. (2020). The Incubation Period of Coronavirus Disease 2019 (COVID-19) from Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine, 172,* 577-582. https://doi.org/10.7326/M20-0504

Li, X. W., Geng, M. M., Peng, Y. Z., Meng, L. S., & Lu, S. M. (2020). Molecular Immune Pathogenesis and Diagnosis of COVID-19. *Journal of Pharmaceutical Analysis, 10,* 102-108. https://doi.org/10.1016/j.jpha.2020.03.001

Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The Reproductive Number of COVID-19 Is Higher Compared to SARS Coronavirus. *Journal of Travel Medicine, 27,* taaa021. https://doi.org/10.1093/jtm/taaa021

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W. et al. (2020). Genomic Characterisation and Epidemiology of 2019 Novel Coronavirus: Implications for Virus Origins and Receptor Binding. *The Lancet, 395,* 565-574. https://doi.org/10.1016/S0140-6736(20)30251-8

Mercatelli, D., & Giorgi, F. M. (2020). Geographic and Genomic Distribution of SARS-CoV-2 Mutations. *Frontiers in Microbiology, 11,* 1800. https://doi.org/10.3389/fmicb.2020.01800

National Center for Biotechnology Information (2020a). *Severe Acute Respiratory Syndrome Coronavirus 2 Data Hub.* https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049

National Center for Biotechnology Information (2020b). *BLAST.* https://blast.ncbi.nlm.nih.gov/Blast.cgi

Sun, M. L., Yang, J. M., Sun, Y. P., & Su, G. H. (2020). Inhibitors of RAS Might Be a Good Choice for the Therapy of COVID-19 Pneumonia. *Chinese Journal of Tuberculosis and Respiratory Diseases, 43,* 219-222.

Tai, W., He, L., Zhang, X., Pu, J., Voronin, D., Jiang, S. et al. (2020). Characterization of the Receptor-Binding Domain (RBD) of 2019 Novel Coronavirus: Implication for Development of RBD Protein as a Viral Attachment Inhibitor and Vaccine. *Cellular & Molecular Immunology, 17,* 613-620. https://doi.org/10.1038/s41423-020-0400-4

Taiaroa, G., Rawlinson, D., Featherstone, L., Pitt, M., Caly, L., Druce, J. et al. (2020). Direct RNA Sequencing and Early Evolution of SARS-CoV-2. *bioRxiv.* https://doi.org/10.1101/2020.03.05.976167

Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell, 181,* 281-292.E6. https://doi.org/10.1016/j.cell.2020.02.058

World Health Organization (2020a). *WHO Director-General's Opening Remarks at the Media Briefing on COVID-19—11 MARCH 2020.*

https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

World Health World Health Organization (2020b). *Infection Prevention and Control Guidance for Long-Term Care Facilities in the Context of COVID-19.* https://apps.who.int/iris/bitstream/handle/10665/331508/WHO-2019-nCoV-IPC_long_term_care-2020.1-eng.pdf

Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C. L., Abiona, O. et al. (2020). Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *Science, 367,* 1260-1263. https://doi.org/10.1126/science.abb2507

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G. et al. (2020). A New Coronavirus Associated with Human Respiratory Disease in China. *Nature, 579,* 265-269. https://doi.org/10.1038/s41586-020-2008-3

Wu, Y. (2020). Strong Evolutionary Convergence of Receptor-Binding Protein Spike between COVID-19 and SARS-Related Coronaviruses. *bioRxiv.* https://doi.org/10.1101/2020.03.04.975995

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W. et al. (2020). A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. *Nature, 579,* 270-273. https://doi.org/10.1038/s41586-020-2012-7