Alexa Anderson

BIOI4870

Dr. Kate Cooper

9 May 2024

SARS-CoV-2 Nucleotide Database

**Abstract**

SARS-CoV-2 Nucleotide Database sets out to deliver users a database of complete SARS-CoV-2 genomic sequences for analysis and reference. This project was initialized with the research question of how NCBI Virus makes its data available and consumable by users, and how to implement such a database on a smaller scale for a course project. Although the creation of the database was successful, the front-end webpage was not. Much educational value has been gained from this project, and a greater appreciation for databases and their management has been gained as well.

**Research Question**

The goal of this project is to create a small scale database that implements some of the mechanisms that NCBI’s Virus database possesses. Thus, the research question would ideally be: how does NCBI Virus manage and make its data available to users, and can it be replicated on a smaller scale?

**Background**

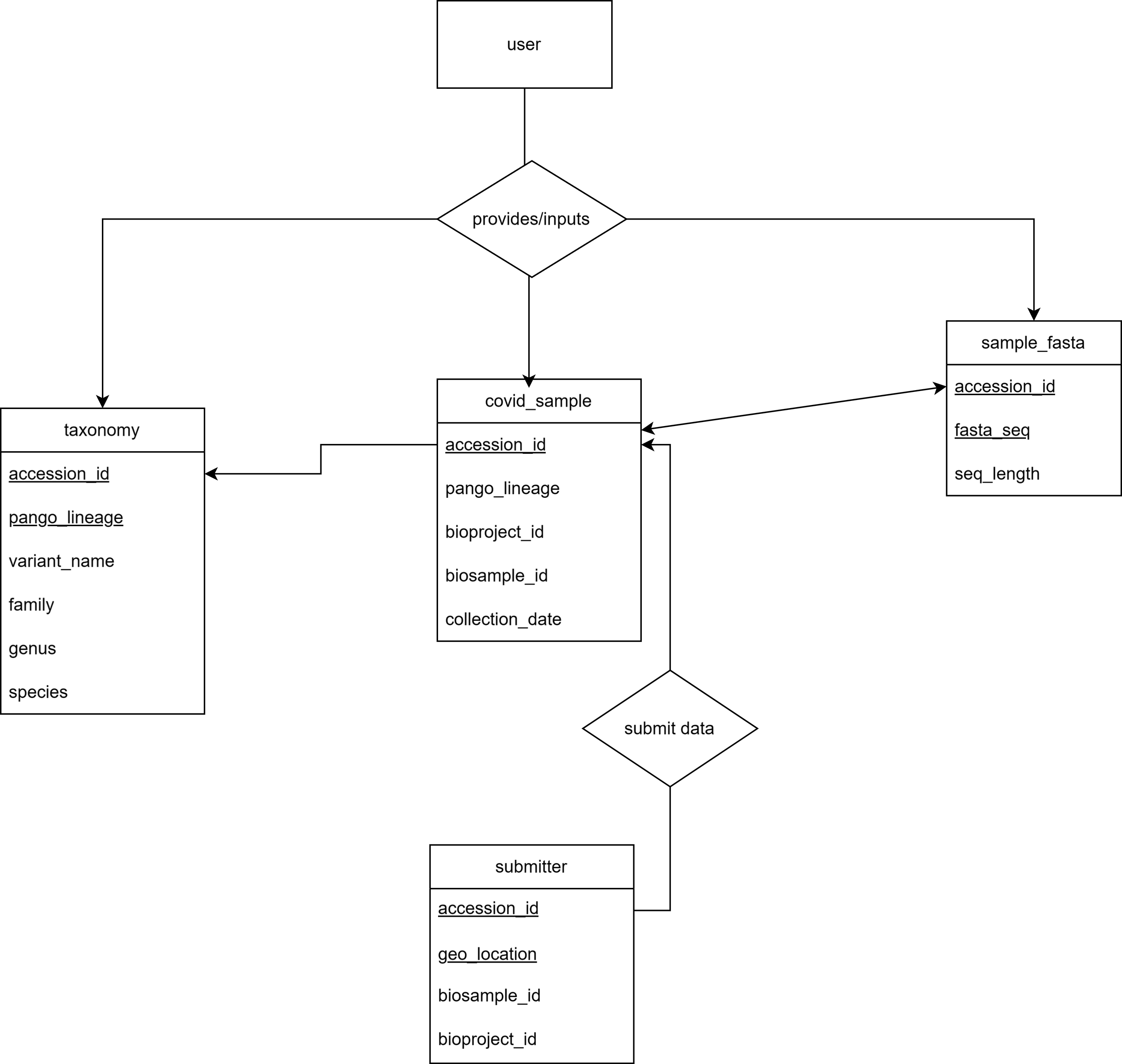
SARS-CoV-2 is a virus that much of the public is familiar with, predominantly for its role in the pandemic. Although they are used interchangeably in colloquial discussions, SARS-CoV-2 is the virus while COVID-19 is the infection it causes. The virus has claimed many lives over the course of the pandemic, due not only to its severe respiratory infection but also its variants. Beyond the death toll, there have been emergent lingering impacts of being infected with COVID-19, referred to as Long COVID. Long COVID is typically understood to be novel symptoms of a secondary infection after being afflicted with COVID-19, and new manifestations of it are being discovered consistently [2,12].

A reason SARS-CoV-2 took the world by storm and is so prolific is there are multiple SARS-CoV-2 variants that have evolved since the beginning of the pandemic in 2019, such as Alpha, Beta, Delta, Omicron, and others [1,3]. Variants come about by mutations in the viral genome, most commonly in the gene that codes for the spike protein. Changes in the spike protein impact the virus’ ability to enter host cells and infect [1]. Thus, although the genomes of SARS-CoV-2 variants are quite similar, they are not identical. Many studies have been conducted to discern the similarities and differences between SARS-CoV-2 variants in order to better understand how to prevent infection, or treat infections once established [1,3,4,7].

Given the relevance of COVID-19, I wanted to create a database that houses representative sequences of each SARS-CoV-2 variant, aggregated from multiple biological projects and data sources. There is definitely a need for resources like this due to the daunting task of keeping track of and comparing different strains of SARS-CoV-2. There are databases that seek to accomplish the goal of documenting and analyzing COVID-19 available, and they are reflected in the literature. There are databases for genomics, proteomics, phylogeny, and variant tracking available to the public that are curated and updated daily [4,7,9,11]. Beyond the databases just mentioned, the community database powerhouses NCBI and EBI have also created COVID-19 specific resources. NCBI Virus and EMBL COVID-19 Data Portal both house SARS-CoV-2 sequences and metadata along with tools to analyze them [6,10].

I took inspiration from NCBI’s Virus database most heavily, as it has a very seamless and user friendly interface with millions of SARS-CoV-2 nucleotide and protein sequences. The COVID-19 Data Portal functions similarly, containing SARS-CoV-2 nucleotide and protein sequences. With excellent models to reference, I was able to create a similar database on a smaller scale. Acknowledging the database I am setting out to create is not novel in its functionality, it is still valuable from an educational standpoint. An important part of programming is understanding existing algorithms, and understanding existing databases is no different. However, if I had to set my database apart, it would be different as it is a smaller, more curated dataset of complete, representative genomes of SARS-CoV-2. Given the dataset I collected, it could be utilized as a reference dataset if one was trying to characterize a new COVID variant and/or relate it to other strains.

**Database Diagram**



**Methods**

*Code*

All code written for this project is available through the first link in the Supplementary Information section of this report.

On the GitHub README document is the license. An MIT License was chosen for the database.

*Data Provenance*

I retrieved my data through three sources: NCBI Virus, EMBL-EBI COVID-19 Data Portal, and GISAID Reference Sequence. 251 samples were collected for this database, totaling roughly 75mb in data. One reason I was able to do a genomic database for COVID-19 is the relatively small genome size, so I can have plenty of representative information while also adhering to odin’s data limit of 1GB. I collected the metadata csv files for most of the samples, and ones I could not do that for I entered the information manually. All this was done in Excel as a form of pre-processing the data. I did this pre-processing step in order to collect data from multiple sources in one place and verify consistency in what data was present for each sample, making the replication of this database as seamless for a potential user as possible. Once the data was collected, I utilized a MySQL data dump method in order to transfer this data into my database. The dump scripts are available for easy replication of the database, along with the preparation python file.

A screenshot of a computer

Description automatically generated

*Webpage*

I used HTML and PHP to create the webpage for this database. HTML was utilized for the appearance of the database and its interactive elements(i.e. buttons, text fields) while PHP was chosen for its ability to connect the front-end to the database for the purpose of querying.

The front-end of this database is available through the second link in the Supplementary Information section of this report.

**Results**

SARS-CoV-2 Nucleotide Database delivers on the data it sought to implement. It aggregated data from multiple sources and organized it based on sample metadata (covid\_sample), taxonomy (taxonomy), sequence availability (sample\_fasta), and submitter/location data (submitter). However, the database failed at being consumable from a user perspective. Unless a user is comfortable using the command line to query the database, it is essentially useless to a consumer. Further, allowing a user to directly query a database in this fashion could be detrimental to the integrity of the database, as a user could unintentionally move or delete data if certain safeguards are not put in place, which they are not currently.

**Discussion/Conclusions**

This project was rewarding but did not come easy as someone who was implementing their first database. I encountered quite a few challenges, such as unsuccessfully coding a webpage for the database. I heavily referenced resources provided in class along with an HTML guide, which were very helpful even with my unsuccessful result. Given more time and discipline, I believe I could get it to work in the manner I wish. Coincidentally, this leads to the next challenge: time management. I did not adequately manage my time for this project and suffered the consequences accordingly. I believe I could have delivered a better project with better time management and belief in myself that I could do it. Finally, A challenge I encountered was the manner in which I put my data into my database. The SQL dump did not behave in the way I expected, with a new line created after the final element was put into the dump statement. I could not get rid of this through coding, so I had to go through line by line (all 251) and correct it for each dump file.

All these challenges considered; SARS-CoV-2 Nucleotide Database is not a complete failure. There is quality data housed in an organized fashion and it can be queried from the conceptual level on the command line. I think this database has potential to be improved, and with more time and funding I would like to see it house protein sequences and have the option to match a user-submitted fasta to similar sequences in the database.

**References**

1. Akkız H. (2022). The Biological Functions and Clinical Significance of SARS-CoV-2 Variants of Corcern. Frontiers in medicine, 9, 849217. https://doi.org/10.3389/fmed.2022.849217.
2. Al-Aly, Z., Bowe, B. & Xie, Y. Long COVID after breakthrough SARS-CoV-2 infection. Nat Med 28, 1461–1467 (2022). https://doi.org/10.1038/s41591-022-01840-0.
3. Aleem, A., Akbar Samad, A. B., & Vaqar, S. (2023). Emerging Variants of SARS-CoV-2 and Novel Therapeutics Against Coronavirus (COVID-19). In StatPearls. StatPearls Publishing.
4. Cheng, Yexiao, Chengyang Ji, Hang-Yu Zhou, Heng Zheng, and Aiping Wu. 2023. "Web Resources for SARS-CoV-2 Genomic Database, Annotation, Analysis and Variant Tracking" Viruses 15, no. 5: 1158. https://doi.org/10.3390/v15051158.
5. Glenet, M., Lebreil, A. L., N'Guyen, Y., Meyer, I., Baud, S., & Andreoletti, L. (2024). Structural impact of a new spike Y170W mutation detected in early emerging SARS-CoV-2 Omicron variants in France. Virus research, 343, 199354. https://doi.org/10.1016/j.virusres.2024.199354.
6. Harrison PW, Lopez R, Rahman N, Allen SG, Aslam R, Buso N, Cummins C, Fathy Y, Felix E, Glont M, Jayathilaka S, Kadam S, Kumar M, Lauer KB, Malhotra G, Mosaku A, Edbali O, Park YM, Parton A, Pearce M, Estrada Pena JF, Rossetto J, Russell C, Selvakumar S, Sitjà XP, Sokolov A, Thorne R, Ventouratou M, Walter P, Yordanova G, Zadissa A, Cochrane G, Blomberg N, Apweiler R. The COVID-19 Data Portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. Nucleic Acids Res. 2021 Jul 2;49(W1):W619-W623. doi: 10.1093/nar/gkab417. PMID: 34048576; PMCID: PMC8218199.
7. Jakob McBroome, Bryan Thornlow, Angie S Hinrichs, Alexander Kramer, Nicola De Maio, Nick Goldman, David Haussler, Russell Corbett-Detig, Yatish Turakhia, A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees, Molecular Biology and Evolution, Volume 38, Issue 12, December 2021, Pages 5819–5824, https://doi.org/10.1093/molbev/msab264.
8. Liotti, F. M., De Maio, F., Ippoliti, C., Santarelli, G., Monzo, F. R., Sali, M., Santangelo, R., Ceccherini-Silberstein, F., Sanguinetti, M., & Posteraro, B. (2022). Two-Period Study Results from a Large Italian Hospital Laboratory Attesting SARS-CoV-2 Variant PCR Assay Evolution. Microbiology spectrum, 10(6), e0292222. https://doi.org/10.1128/spectrum.02922-22.
9. Mariona Torrens-Fontanals, Alejandro Peralta-García, Carmine Talarico, Ramon Guixà-González, Toni Giorgino, Jana Selent, SCoV2-MD: a database for the dynamics of the SARS-CoV-2 proteome and variant impact predictions, Nucleic Acids Research, Volume 50, Issue D1, 7 January 2022, Pages D858–D866, https://doi.org/10.1093/nar/gkab977.
10. NCBI Virus [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2024/05/06]. Available from: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/.
11. Shuyi Fang, Kailing Li, Jikui Shen, Sheng Liu, Juli Liu, Lei Yang, Chang-Deng Hu, Jun Wan, GESS: a database of global evaluation of SARS-CoV-2/hCoV-19 sequences, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D706–D714, https://doi.org/10.1093/nar/gkaa808.
12. Virhammar, J., Nääs, A., Fällmar, D., Cunningham, J.L., Klang, A., Ashton, N.J., Jackmann, S., Westman, G., Frithiof, R., Blennow, K., Zetterberg, H., Kumlien, E. and Rostami, E. (2021), Biomarkers for central nervous system injury in cerebrospinal fluid are elevated in COVID-19 and associated with neurological symptoms and disease severity. Eur J Neurol, 28: 3324-3331. https://doi.org/10.1111/ene.14703.

**Supplementary Information**

GitHub Repository: <https://github.com/aand117/BIOI4870-COVID-Database>

Odin Webpage: <http://odin.unomaha.edu/~aanderson117/hello.php>