Kristen Steenbergen

May 1, 2021

410.712 Advanced Computer Concepts

Final Project Report

My final project is a web application that displays genetic information about COVID19 Variants of Concern (VOC).

I chose to present this information because I follow the news closely when it comes to the coronavirus pandemic, but I find the evolution of the virus to be particularly fascinating, albeit unsettling. I found myself researching COVID19 genes and specific mutations, but VOCs can be confusing, and difficult to differentiate. This project became a a good opportunity to further explore a personal interest while building my programming skills. I do not expect that this application will have a high impact in the field since this pandemic is an incredibly fast-paced, evolving situation and there are plenty of websites hosting real-time geospatial tracking of COVID19 VOCs alongside information on genetic consequences. However, what my project *will* provide is an informative experience for those, similar to myself, that have an interest in the COVID genome and the most common VOCs.

One of the biggest challenges for me occurred immediately in the beginning of my research as I located source information on variants. I expected there would be well-documented consensus sequences for each of the major variants. However, what I found was that the 5 variants included in this project are only the tip of the variant iceberg and there is an overwhelming amount of information available. One important resource was cov-lineages (cite cov-lineage page(PANGO lineages, 2021)) which has developed the PANGOLIN software that is used to phylogenetically assign COVID19 genomes using the PANGO lineage nomenclature. If a new variant is found, it is submitted to PANGO for consideration. For each of the major VOCs, cov-lineage has an international report which includes a list of distinguishing mutations. Checking these mutation lists with others I deemed it worthy source data and used this information to populate the ‘variant’ table of my database. To be consistent with the information from cov-lineages, I decided to use the PANGO nomenclature in my project alongside country of origin. This nomenclature is easier to distinguish than others and has been referenced by common news sources making it more recognizable to users.

EDIT THIS PARAGRAPH: A new variant emerged while working on the project, and although it was classified as a variant of interest instead of concern, and all the information was not out yet, it had some "concerning" effects and even though it had yet to be defined with an International Lineage Report by PANGO, the distinguishing mutations were found \_\_\_ and this was added to the database. The database could be updated on a regular basis as time allows. The last database update before submission for grading was completed on \_\_\_\_\_\_\_.

The reference genome was fairly easy to track down and well documented. I obtained access to GISAID database (CITE THEM HERE (GISAID - Initiative, 2021)) which had a fully annotated genome with a handy graphic to boot. The COVID reference genome can be downloaded from NCBI (ACC# MN996528) (CITE THEM HERE (Nucleotide -NCBI, 2021)). The ORF10 gene was excluded due to varying information, and also no evidence of expression

I started the project with a list of mutation codes for all of the top VOCs in a text file, and genbank file of the COVID19 official reference sequence. My first step included writing two python programs to parse these files and output the result in the same formatted that I would use in a command to create a table in mySQL database.

To build the SQL database, the schema I designed included one table for the reference genome populated with all the gene info I could parse from the genbank file (except for sequence information). A second table contained the variants and information on their distinguishing mutations. The tables could be joined by gene names. I designed it this way so that if you had a mutation of interest, you could join to the reference table and obtain all the genetic information.

I decided this information would be best displayed in one of two ways: a gene-level display of the mutations, or a comparison of 2 variants. I created two CGI scripts for each task, as well as two HTML templates. The home page allows the user to choose which method to view depending on the form they submit.

When the gene-level display was printed to the screen, it was hard to capture all the information, so I used Jquery to create folding accordion tables allowing the user to click through to view each gene individually. Since all genes appear equally at first, it is an interesting experience to click through the genes and view the number of mutations occurring on each.

Include anything here about CSS.

Reference – CHECK ORDER AND FONT:

* Epicov.org. 2021. *GISAID - Initiative*. [online] Available at: <https://www.epicov.org/> [Accessed 23 April 2021].
* Cov-lineages.org. 2021. *PANGO lineages*. [online] Available at: <https://cov-lineages.org/> [Accessed 23 April 2021].
* Nucleotide [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – . Accession No. MN908947.3, Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome; [cited 2021 Apr 02]. Available from: <https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3>
* Ncbi.nlm.nih.gov. 2021. *Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, co - Nucleotide - NCBI*. [online] Available at: <https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3> [Accessed 23 April 2021].

Intext: (Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, co - Nucleotide - NCBI, 2021) OR: (Nucleotide -NCBI, 2021)

NCBI, 2021