Genetic Relationship between SARS-CoV-2 and Other Coronaviruses

Giana Hubbard

December 13, 2020

Abstract

News coverage about the COVID-19 pandemic has often compared SARS-CoV-2, the virus that causes this disease, to other coronaviruses that have made the species jump to humans, such as SARS-CoV of the 2003 SARS outbreak, MERS-CoV of the 2012 MERS outbreak, and the coronaviruses that cause the flu and common cold. Though the comparison of SARS-CoV-2 to the common cold has been weaponized for political reasons, it is important for the study of SARS-CoV-2 and the coronavirus classification, the development of an effective arsenal against this virus and the global pandemic, and the effective implementation of public health measures. In order to tackle these broad areas of research, the genetic makeup of SARS-CoV-2 must be analyzed and compared to that of other human coronaviruses. Research about SARS-CoV-2 and COVID-19 has been conducted extremely rapidly, with the demand for new and groundbreaking information growing more and more each day as we aim to end the global pandemic. The research that has been conducted approaches the virus and the pandemic from many different angles and fields, such as epidemiology, medicine, economics, psychology, and much more; however, there is an enormous information gap in the medicine, biochemistry, and genetics of SARS-CoV-2 and COVID-19. In this research endeavor, I will determine how genetically related SARS-CoV-2, SARS-CoV, and other coronaviruses are, and how these variations effect observed differences in epidemiology, symptoms, and severity, among others.

Introduction

COVID-19 is the infectious respiratory disease that is caused by the coronavirus SARS-CoV-2 that scientists first identified in Wuhan, China in December 2019. SARS-CoV-2 stands for severe acute respiratory syndrome coronavirus 2, and this abbreviation will be used to refer to the virus throughout this document. Other coronaviruses in this same family with the SARS prefix are also severe acute respiratory syndrome diseases (such as SARS-CoV), and other viruses with the CoV abbreviation are coronaviruses (like MERS-CoV).

COVID-19 spread worldwide extremely rapidly, with known cases being identified as early as January 2020, with national lockdowns and the race to find and manufacture treatments, cures, and personal protective equipment taking hold in many nations around the world. Thus, the COVID-19 global pandemic was declared by the World Health Organization (WHO) on March 11, 2020.

Background/Previous Work

With the scramble to assemble as much information about SARS-CoV-2 and COVID-19 as rapidly as possible, much of this research compares SARS-CoV-2, its biochemical characteristics, and transmissibility, to those of other coronaviruses. Several robust studies that explore the codon usage and resulting effects of SARS-CoV-2 and other related coronaviruses have already been published, with the studies I found being cited in the References section. Since codons code for the amino acids that make up the proteins of an organism, which themselves determine the structure, function, and biochemical interactions of the cell/organism, analysis of codon frequencies will allow for important interpretation of the coronaviruses' characteristics. All of these studies use quantitative data and analysis to describe and extrapolate the results of the codon usages in these coronaviruses' genetic sequences, but many of these sources were

published in the early months of the global pandemic, and others lack valuable comparative information to other strains of SARS-CoV-2 as well as other coronaviruses, which could provide useful information about the evolution and functions of these viruses and strains. These studies and their conclusions have proven that analysis of codons in the genetic sequences of these coronaviruses not only reveal the genetic relationships between these viruses, but it also can used to determine the functions of these codons and the amino acids and proteins for which they code. With new information about SARS-CoV-2 and COVID-19 emerging every day, it is useful to systematically analyze and compare the codon usage in SARS-CoV-2 and related coronaviruses and other strains to further explore how these codons affect each virus's characteristics.

Methods

I want to investigate how genetically related SARS-CoV-2 and SARS-CoV are. So, I developed a code in the coding language Python to calculate the codon frequencies (the number of times a set of three nucleotide bases occurs) to determine how comparable the two coronaviruses are at the genomic level. A codon a basic unit of the RNA and DNA molecules/strands that is made up of a sequence of three nucleotide bases. These codons each correspond to an amino acid, which form the proteins that determine the function and structure of these molecules. With this information, I want to analyze the identity and function of these codons to determine how the frequencies and functions of each codon result in the genetic and evolutionary differences as well as the differing effects in humans for these two coronaviruses, and possibly, other coronaviruses.

The Python code that I wrote and developed after conducting research about the importance of codons and codon frequency to genetic makeup and molecular function outputs the codon frequency of any genomic sequence that is input into the program. It also plots the

codon frequency of any set of two or more genomic sequences input into the program on a double bar graph, which can then be analyzed for the questions I seek to investigate: how do the codon frequency and identity of SARS-CoV-2 compare to those of other coronaviruses, what do these results tell us about the genetic relationship between these coronaviruses, and what about these relationships can be used to learn more about SARS-CoV-2 and COVID-19?

I will collect genomic sequences of SARS-CoV-2 and other coronaviruses like SARS-CoV, while being sure to take note of the date and location of isolation and input them into the code that I wrote. The Python code begins by taking the genomic sequences themselves as inputs and also taking an integer input of three. To split the RNA sequences into codons (three nucleotide bases to one codon), I next included a for loop to divide the length of the RNA strands (genomic sequences) by three and isolate each codon. Each unique codon isolated by the for loop then becomes a key in the dictionary corresponding to that RNA strand. Every time each unique codon is identified in the RNA sequence, one is added to the counter for that unique codon. Those sums become the values associated with those keys in the dictionaries. When each dictionary is printed, it will display each unique codon as a key and its frequency/count in the RNA sequence as its value. Next, a bar graph for each RNA sequence is be created, and for each sequence input into the program, the bars in the bar graph are organized in descending order of the frequency of each codon. For the combined bar graph which displays the frequencies of the codons in all of the coronavirus RNA sequences input into the program, the bars are not ordered. Instead, the bars are organized in no particular order, but the bars corresponding to each codon will overlap, representing and comparing the frequencies of each unique codon in each coronavirus RNA sequence input into the program. I expect the total length of each sequence to differ by some small number of nucleotide bases, so these differences should not greatly skew

the codon frequency comparisons.

An example that I have already conducted is comparing SARS-CoV-2 to SARS-CoV using this code. I found that the SARS-CoV-2 genome that I used has 9,942 nucleotide bases, and the SARS-CoV genome that I used has 10,020 nucleotide bases. As stated previously, I did not expect this difference to be very high; in fact, the difference was only nucleotide bases (0.7815% difference). The following bar graphs (Figures 1, 2, and 3) were the results of this comparison. In Figures 1 and 2, the codon frequencies are organized in decreasing order to clearly demonstrate the comparative magnitudes of the frequencies. On the other hand, the frequencies are not particularly ordered in Figure 3, as this bar graph is just meant to demonstrate which codons have similar frequencies and which have significantly different frequencies.

Figure 1

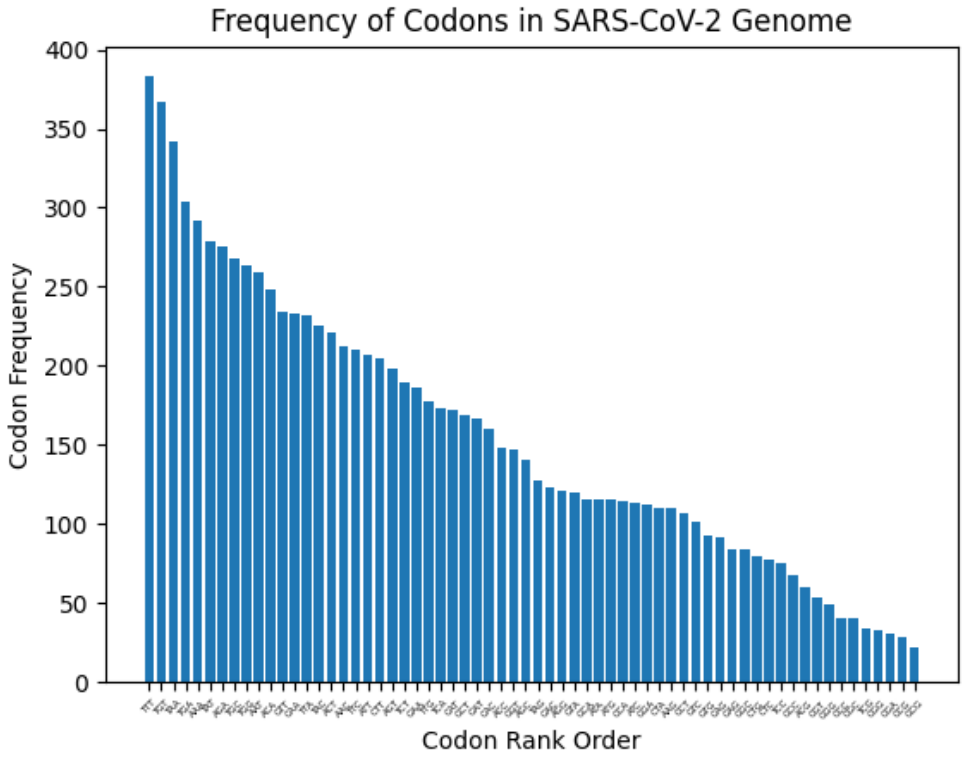


Figure 2

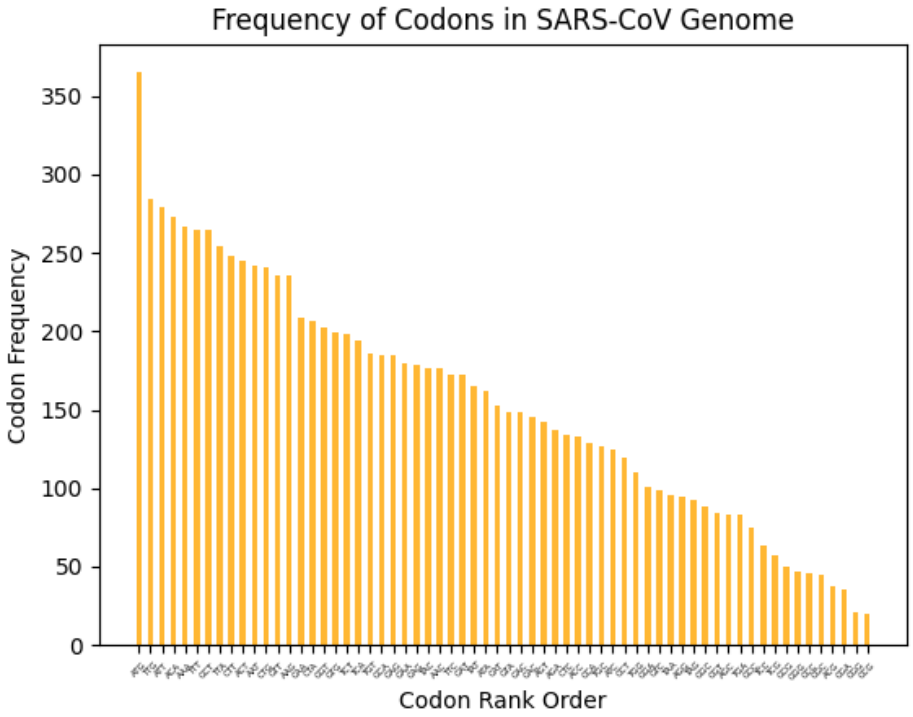
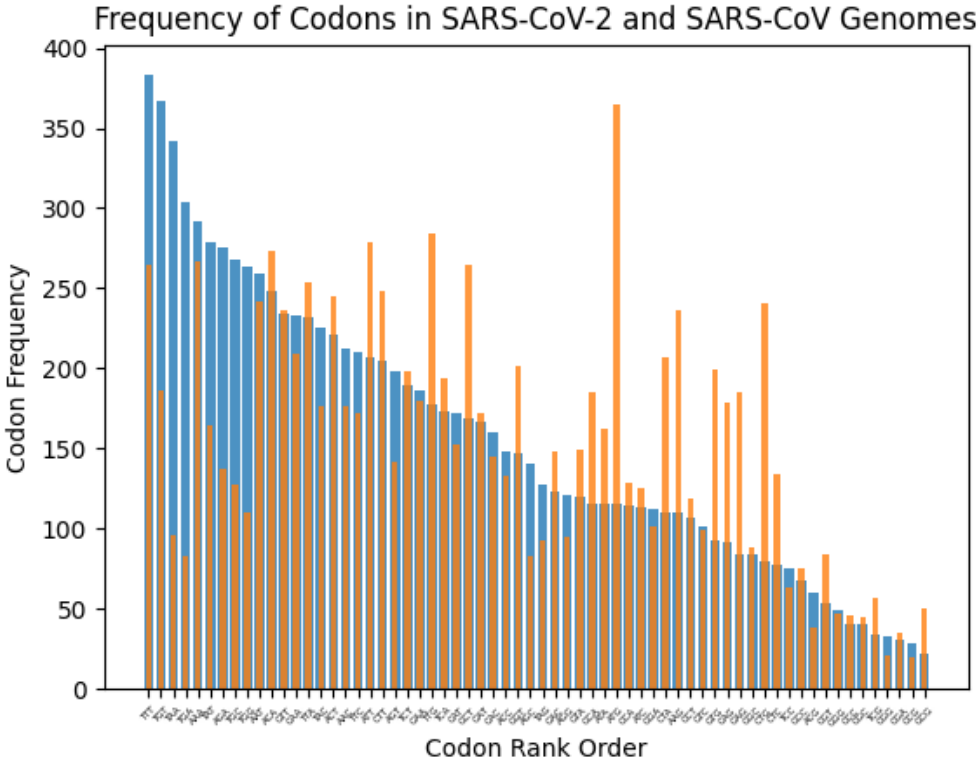


Figure 3



Before running the code and getting the results and bar graphs, my hypothesis was that the majority of codons would have about the same frequency in both genomes of SARS-CoV-2 and SARS-CoV, since both are coronaviruses and thus are closely related, and a small portion of codons will have significantly different frequencies to account for the differences in function/effect of the two coronaviruses. However, the results of the analysis proved my hypothesis wrong. Instead of the majority of codons having similar frequencies in the RNA sequences for SARS-CoV-2 and SARS-CoV, just under half of the codons (about 26 codons) have similar frequencies in the genomes of SARS-CoV-2 and SARS-CoV. Perhaps the two coronaviruses are not as closely related as I previously thought. This is the kind of analysis that I would be completing with this code, and I can take this thinking further by researching the functions of each codon and how these functions result in the observations that scientists have made about SARS-CoV-2 and other coronaviruses. More specifically, I want to investigate whether or not codons that appear in approximately the same frequency in both SARS-CoV-2 and other coronaviruses have similar, if not the same functions, or molecular and/or structural effects on these coronaviruses. To effectively conduct research to answer these questions and fill in these information gaps, I would need to explore transmissibility, symptoms, death rates, and other effects of these coronaviruses compare what is already known about these other coronaviruses to SARS-CoV-2, and I would also need to continue to refine my code and look for more recent, relevant, and reliable information about SARS-CoV-2, COVID-19, and the coronavirus family.

Expected Results

As explained in the Methods section, my initial hypothesis when setting up my first comparison was proven wrong by the results of the Python code. Now that I have that initial

diagnostic, I can modify my hypothesis for the other comparisons that I plan to carry out. As a result, I expect that just under half of codons to have similar frequencies for the different coronaviruses and SARS-CoV-2, but the more closely related the coronaviruses are, the more codons will have similar frequencies between those viruses. In comparing the different strains of SARS-CoV-2, I expect the majority of the codon frequencies to be similar with only a few variations to account for the mutations that lead to the different strains and their differing characteristics. I also expect that the codons with similar frequencies in the different coronaviruses and strains to account for the observed similarities between them, while significantly different codon frequencies will account for the observable differences between the coronaviruses and strains. Lastly, I expect for SARS-CoV-2 to have similar codon frequencies to SARS-CoV than MERS-CoV, for example, supported by the similar naming conventions, which would indicate a closer genetic relationship between SARS-CoV-2 and SARS-CoV than MERS-CoV. Thus, comparing codon frequencies can indicate genetic relationships, which can then support evolutionary similarities and variations.

Conclusion

By conducting this research, I hope to contribute significantly to the sprawling research community that is racing to uncover more information about SARS-CoV-2 and COVID-19. Despite the political divisions that have arisen from this global pandemic and the almost unimaginably immense effects that is has had on the world, I know that this scientific work and research surrounding this virus in general are so important to moving forward with our understanding of our world, scientifically, economically, socially, and otherwise. Comparing SARS-CoV-2 to other coronaviruses that we have more information about and experience with could lead to the coveted revelations that we need to make some significant progress in tackling

this pandemic and better understanding coronaviruses, the spread of diseases, public health, and all of the other realms that are tied to this global pandemic.

References

Center for Disease Control and Prevention. (2017, December 6). *SARS basics fact sheet*. Centers for Disease Control and Prevention. https://www.cdc.gov/sars/about/fs-sars.html

Cilla, G., Montes, M., Pineiro, L., & Marimon, J. M. (2020, June 23). *Severe acute respiratory syndrome coronavirus 2 isolate SARS-Cov-2/hum - Nucleotide - NCBI*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/nuccore/MT655132

Dilucca, M., Forcelloni, S., Georgakilas, A. G., Giansanti, A., & Pavlopoulou, A. (2020, April 30). *Codon usage and phenotypic divergences of SARS-Cov-2 genes*. Multidisciplinary Digital Publishing Institute. https://www.mdpi.com/1999-4915/12/5/498/htm

He, R., Dobie, F., Ballantine, M., Leeson, A., Li, Y., Bastien, N., Cutts, T., Andonov, A., Cao, J., Booth, T. F., Plummer, F. A., Tyler, S., Baker, L., & Li, X. (2020, September 15). *SARS coronavirus Tor2, complete genome - Nucleotide - NCBI*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/nuccore/30271926

Hou, W. (2020, September 14). *Characterization of codon usage pattern in SARS-CoV-2*. Virology Journal. https://virologyj.biomedcentral.com/articles/10.1186/s12985-020-01395-x

Hussain, S., Shinu, P., Islam, M. M., Chohan, M. S., & Rasool, S. T. (2020, May 4). *Analysis of codon usage and nucleotide bias in Middle East respiratory syndrome coronavirus genes*. SAGE Journals. https://journals.sagepub.com/doi/full/10.1177/1176934320918861

Kandeel, M., Ibrahim, A., Fayez, M., & Al-Nazawi, M. (2020, March 11). *From SARS and MERS CoVs to SARS-Cov-2: Moving toward more biased codon usage in viral structural and nonstructural genes*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/full/10.1002/jmv.25754

Koyama, T., Platt, D., & Parida, L. (2020, June 2). *Variant analysis of SARS-CoV-2 genomes*. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7375210/

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., … Tan, W. (2020, January 30). *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding*. The Lancet. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30251-8/fulltext#seccestitle170

Ludwig, S., & Zarbock, A. (2020, March 31). *Coronaviruses and SARS-CoV-2: A brief overview*. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7173023/

Mayo Clinic. (2019, September 24). *Severe acute respiratory syndrome (SARS) - Symptoms and causes*. https://www.mayoclinic.org/diseases-conditions/sars/symptoms-causes/syc-20351765

Mayo Clinic. (2020, October 15). *Coronavirus disease 2019 (COVID-19) - Symptoms and causes*. https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963

Petersen, E., Koopmans, M., Go, U., Hammer, D. H., Petrosillo, N., Castelli, F., Storgaard, M., Al Khalili, S., & Simonsen, L. (2020, July 3). *Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics*. The Lancet: Infectious Diseases. https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30484-9/fulltext

Python Software Foundation. (2020, October 20). *5. Data structures — Python 3.9.0 documentation*. The Python Tutorial. https://docs.python.org/3/tutorial/datastructures.html

Seladi-Schulman, J. (2020, April 2). *COVID-19 vs. SARS: How do they differ?* Healthline. https://www.healthline.com/health/coronavirus-vs-sars#symptoms

The Jupyter Book Community. (2019, September 30). *Visualization*. Inferential Thinking. https://www.inferentialthinking.com/chapters/07/Visualization.html

Tort, F. L., Castells, M., & Cristina, J. (2020, April 12). *A comprehensive analysis of genome composition and codon usage patterns of emerging coronaviruses*. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7152894/

Spiteri, G., Fielding, J., Diercke, M., Campese, C., Enouf, V., Gaymard, A., Bella, A., Sognamiglio, P., Sierra Moros, M. J., Riutort, A. N., Demina, Y. V., Mahieu, R., Broas, M., Bengnér, M., Buda, S., Schilling, J., Filleul, L., Lepoutre, A., Saura, C., … Ciancio, B. C. (2020, March 5). *First cases of coronavirus disease 2019 (COVID-19) in the WHO European region, 24 January to 21 February 2020*. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7068164/

Timeline, Required Materials, and Budget

I will continue my research first by refining my already written Python code to be more efficient, which should only take one to two weeks. Next, for about one week, I will compile the genetic sequences from online databases and systematically analyze with the Python code, which immediately calculates the codon frequencies for the coronaviruses' genetic sequences and plots them on the bar graphs in only a couple seconds. This will bring me to the beginning of January 2021, when the next steps will be to use information that is widely available online about RNA sequences, codons, coronaviruses, genetic mutations, etc. to research what the codon frequencies could mean for the coronaviruses' structures, functions, and genetic relationships to each other. This research and compilation of information is the core of the research project, and I expect it to take about eight months to assemble, interpret, and extrapolate that information to the codon frequencies calculated by my Python code, which will bring me to the end of August 2021. During these months, I will continue to look for and incorporate new genetic sequences and published research into my analysis. Thus, I ask for a stipend to compensate myself for my time and work on this project and funds to travel to present my findings at three conferences (if the COVID-19 pandemic has ended by the time I am ready to publish and present my results) within the six months following the projected August end date for the research itself. I want to spend two to three days in the location of each conference. I plan to work on this project ten hours per week for six to eight months with a salary of $20 per hour. Thus, I propose a maximum $7000 stipend, with an additional $2500 allotted for travel (flights, hotel rooms, car rental/transportation, meals) for research conferences through February 2022.