Comparing Surface ("Spike") Glycoproteins in SARS-CoV-2 Reference and the Omicron Variant

Alex Amari
12/18/2021

Introduction

The evolution of SARS-CoV-2, the coronavirus which causes COVID-19, has posed significant challenges to prevention, diagnosis, and treatment of the condition for health systems around the world. The dissemination of new viral variants, characterized by significant changes to SARS-CoV-2's genome, have rendered established methods for diagnosis less reliable.[1] In addition, experts are now calling into question the course and efficacy of widely-used immunization treatments, such as mRNA-based vaccines, which were developed in large part prior to the inception of novel variants.[2] Thankfully, there is an increasing abundance of sequencing data for SARS-Cov-2 and its variants in publicly available online databases. These data, coupled with open-source bioinformatics approaches allowing for comparison of related viral genomes, are increasing the collective understanding of new variants and their implications for public health.

On November 28 2021, the World Health Organization (WHO) classified the recently discovered SARS-CoV-2 variant B.1.1.529 a Variant of Concern, labeling it Omicron.[3] This designation carries with it several public health recommendations, such as enhanced surveillance and genetic sequencing of cases, as well as a call to action for researchers around the world to help enhance the global community's understanding of Omicron. This report aims to serve as a small contribution to that purpose. With sequencing data from a documented Omicron case in Belgium, as well as data from the SARS-CoV-2

[1] https://www.fda.gov/medical-devices/letters-health-care-providers/genetic-variants-sars-cov-2-may-lead-false-negative-results-molecular-tests-detection-sars-cov-2
[2] https://www.mayoclinic.org/coronavirus-covid-19/covid-variant-vaccine
[3] https://www.who.int/news/item/28-11-2021-update-on-omicron

reference genome, I use bioinformatics methods to compare the surface ("spike") glycoprotein of the new variant with that of the reference virus.

The report proceeds as follows:

1. A short overview of COVID variants and their associated spike proteins
2. Description of data sources and methods for retrieval
3. Exploratory analysis of the Omicron and reference spike proteins (e.g., length and residue compositions)
4. Sequence alignment of the two proteins using dynamic programming with Biopython
5. Identification and mapping of Omicron glycoprotein mutations on the reference genome
6. Comparison of results to existing Omicron studies, and short discussion on implications for public health
7. References

What is Omicron?

The World Health Organization (WHO) monitors the evolution of viruses like SARS-CoV-2 by analyzing viral genomes on an ongoing basis. If a specific mutation or set of mutations is deemed to be significant and consistent enough to cause changes to a virus' behavior, it is characterized as a new viral variant. In the case of SARS-CoV-2, several variants have been classified as Variants of Concern, including the Alpha, Beta, Gamma, and Delta variants.

In November 2021, laboratories in Botswana and South Africa detected a novel COVID variant. Analysis showed that the new variant exhibited a number of mutations in the coding region of its genome associated with its surface glycoprotein. The surface glycoprotein of SARS-CoV-2 is one of the structural proteins which assists the virus' invasion of host cells. Often called "spike" proteins for their spiky appearance, they latch onto proteins on host cells' surfaces and facilitate the transfer of viral material to the host.[4] Due in large part to mutations in its spike protein, Omicron was subsequently classified as a Variant of Concern by the WHO.[5]

[4] https://www.ncbi.nlm.nih.gov/gene/43740568
[5] https://www.who.int/news/item/28-11-2021-update-on-omicron

Data

To investigate the Omicron variant's spike protein, as well as to compare it with that of previously identified strains of SARS-CoV-2, it was necessary to acquire sequencing data of the viral genomes of both the Omicron variant as well as earlier forms of COVID-19. For this report, sequencing data was collected from the NCBI's GenBank sequencing database.[6] The first dataset was the SARS-CoV-2 reference genome (NC_045512.2) generated from a sample collected in Wuhan, China, and posted on GenBank in January 2020.[7] The reference genome represents COVID-19 as it was prior to becoming a global pandemic, and prior to the development of novel variants such as Omicron. As of December 2021 there is no apparent reference genome for the Omicron variant, however there are a number of Omicron sequences that have been made recently available on GenBank. While Omicron was first identified in South Africa and Botswana, it has since been detected in other parts of the world. An Omicron sequence (OL672836.1) generated in Belgium by the Rega Institute for Medical Research and tagged with the designation B.1.1.529 was chosen for this analysis.[8] I chose this data as it was the earliest apparent sequence for the variant posted on GenBank (November 30, 2021) and is also thoroughly annotated and well-documented.

Protein sequences for the spike proteins (annotated as "Spike Glycoprotein" in the Wuhan sample and "Surface Glycoprotein" in the Belgian sample) were included in the FASTA files. For nucleotide sequences corresponding with the two spike proteins, the genomes were indexed using Biopython according to the designated Protein Coding Sequence indices in GenBank.

Preliminary Sequence Analysis

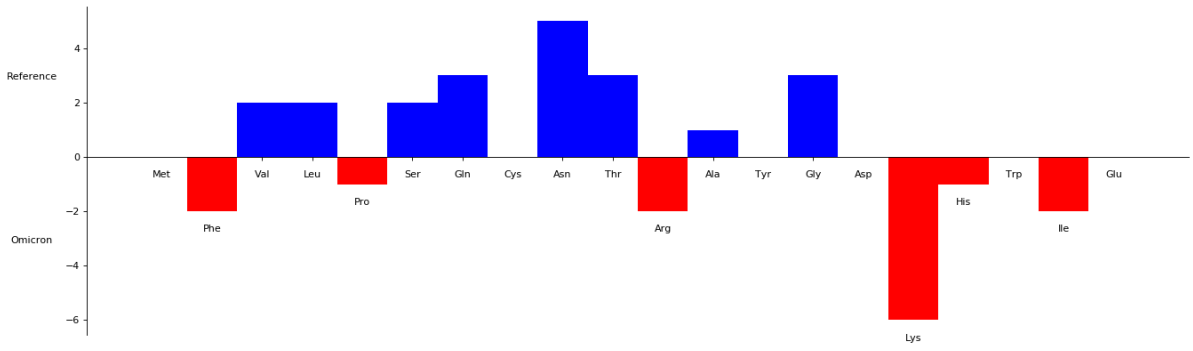
The comparison of the reference and Omicron spike proteins commenced with a surface-level investigation of the two sequences prior to sequence alignment. I wanted to establish that there were

[6] https://www.ncbi.nlm.nih.gov/genbank/
[7] https://www.ncbi.nlm.nih.gov/nuccore/1798174254
[8] https://www.ncbi.nlm.nih.gov/nuccore/OL672836.1/

significant differences between the two glycoproteins, such that it was clear that the Belgian sample could indeed be characteristic of a new variant. First, I used Biopython to count both the nucleotides and residues incorporated in the reference and Omicron spike proteins.[9] The reference COVID-19 surface glycoprotein consisted of 3,822 nucleotides and 1,273 amino acid residues, while the Omicron variant's spike protein contained 3,813 nucleotides and 1,270 residues. This result indicated the presence of several amino acid deletions in the Omicron variant's surface glycoprotein.

However, this result alone tells us little about the nature of the apparent changes to the Omicron genome. For example, if the mutations only affected three residues, they could have little or no impact on the protein's structure or function. Moreover, some of these differences could be the result of errors in the sequencing or data collection processes. To better gauge the extent of the changes to the glycoprotein in Omicron, a next step was to compare the residue frequencies in the sequences. Figure 2 depicts the frequency differences between reference and Omicron for the 20 amino acids that form COVID-19's surface glycoprotein. Blue bars represent more of that amino acid in the reference sequence, while red bars indicate more in the Omicron sequence.

Figure 1: Residue Frequency Differences for Reference and Omicron



While some amino acid frequencies were unchanged, others differed significantly. There were a total of 35 residue differences in Omicron's genome, representing an approximate 2.8% difference in its

[9] http://biopython.org/DIST/docs/tutorial/Tutorial.html

amino acid composition as compared to reference. This result provides strong preliminary evidence of changes to the Omicron spike protein.

Sequence Alignment

To begin making functional comparisons of the reference and Omicron spike proteins, it was important to align the two sequences. Sequence alignment was performed with the Biopython module pairwise2, which utilizes a slightly modified Needleman-Wunsch global alignment strategy for protein sequence alignment.[10] Needleman-Wunsch employs a dynamic programming strategy to find the optimal global alignment between DNA or protein sequences, which in turn allows for the identification and analysis of similar regions and mutations.[11] Needleman-Wunsch relies on several user-inputted parameters that affect the weighting of matches, mismatches, and gaps across the sequences. After some experimentation, I chose alignment parameters match = 1, mismatch = -1, gap penalty = -0.5, gap extension penalty = -0.1, as these parameters were representative of the most consistent alignment result and also resembled results from other analyses of Omicron (described further in the discussion section). The final alignment score for the two sequences using this method was 1211.

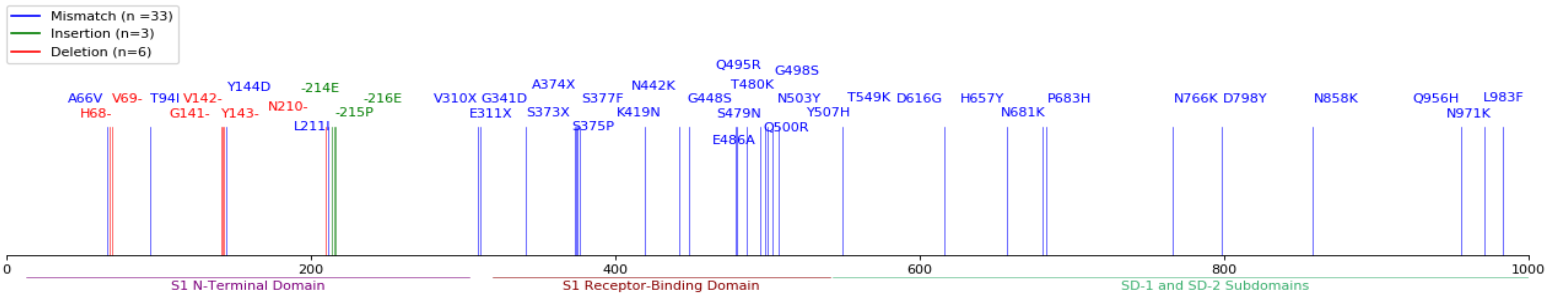
Mutation Mapping

Having aligned the two sequences, it became possible to investigate individual mutations in the Omicron variant. The alignment yielded 33 residue mismatches and 9 (potentially attached) indels. To visualize the distribution of these mutations, I next mapped them with the Python library Matplotlib onto an axis representing the 1,273 reference glycoprotein residues, annotated with regions including the S1 N-Terminal Domain, the S1 Receptor-Binding Domain, and the SD-1 and SD-2 Subdomains, which were identified according to previously annotated COVID spike proteins associated with sequencing data on GenBank.[12] Indels were color-coded red and green according to the location of a gap (representing an

[10] https://biopython.org/docs/1.75/api/Bio.pairwise2.html
[11] BIOT E104 Lecture 2
[12] https://www.ncbi.nlm.nih.gov/protein/YP_009724390.1?feature=any

apparent insertion or deletion in Omicron) while mismatches were coded blue. Each mutation was labelled on the plot according to the nature of the residue change (e.g., A374X representing an A in reference becoming an X in Omicron at residue 374).

Figure 1: Mutation Mapping of Omicron Surface Glycoprotein vs. Reference



Visualization showed three apparent deletions in the Omicron variant (residues 68-69, 141-143, and 210) and one apparent insertion (residues 214-216), all of which were robust to changes in the alignment parameters. Interestingly, these deletions and insertions all appeared in the S1 N-Terminal Domain of the glycoprotein, a region located at the outer edge of the protein's "spike" which is associated with receptor binding.[13] There was also a strong clustering of residue mismatches in the S1 Receptor-Binding Domain, a related region which is associated with the transmissibility of SARS-CoV-2.[14] 15 of the total 33 mismatches were concentrated in this region.

Discussion and Results Comparison

Taken together, these results suggest that there are significant changes to the surface glycoprotein of the Omicron variant as compared to the reference SARS-CoV-2 sample. The mutations in the variant appear to be concentrated within the first 500 residues of the protein, which represent the S1 N-Terminal and Receptor-Binding Domains. Given that these domains have been associated with the transmissibility

[13] https://www.ebi.ac.uk/interpro/entry/InterPro/IPR032500/
[14] Ibid.,

of COVID variants in the past, it is likely that Omicron exhibits substantively different transmissibility behavior than previous COVID strains. While these analyses are not enough to conclude that infectivity has increased with Omicron, mounting evidence from case reporting and laboratory testing seems to indicate that this is unfortunately the case.[15]

To gauge their consistency, I compared these results with existing studies and reports on Omicron produced by various global health agencies. The European Centre for Disease Prevention and Control (ECDC) recently published results of its own comparative analysis of Omicron and reference genomes using Omicron sequencing data hosted on GISAID.[16]

Figure 3: Mutations Identified in This Report vs. ECDC Report

This Report	ECDC Nov. 2021
A66V, Δ68-69, T94I, Δ141-143, Y144D, Δ210, L211I, ins214EPE, V310X, G311X, G341D, S373X, A374X, S375P, S377F, K419N, N442K, G448S, S479N, T480K, E486A, Q495R, G498S, Q500R, N503Y, Y507H, T549K, D616G, H657Y, N681K, P683H, N766K, D798Y, N858K, Q956H, N969K, L983F Total = 37	A67V, Δ69-70, T95I, G142D, Δ143-145, Δ211, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493K, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F Total = 34

ECDC reported 34 total mutations to the spike protein in Omicron, as compared to the 37 found in these analyses. Identified mutations were broadly similar, apart from slight offsets of between 1 and 3 indices in the residue counts (e.g., A66V -> A67V) which could have been the result of different alignment approaches and parameters, errors or discrepancies in sequencing data, or actual residue differences in the Omicron samples on GenBank as compared to GISAID. Three mismatch residue substitutions appeared in my analyses which did not correspond to any in the ECDC report (highlighted red in Figure 3) while two identified mismatches corresponded with different residue substitutions

[15] https://www.nature.com/articles/d41586-021-03614-z

[16] https://www.ecdc.europa.eu/sites/default/files/documents/Implications-emergence-spread-SARS-CoV-2%20B.1.1.529-variant-concern-Omicron-for-the-EU-EEA-Nov2021.pdf

(highlighted blue in Figure 3). These discrepancies could have been the result of the same factors described above. Importantly, ECDC reported the same 3 deletions and 1 insertion in the S1 N-Terminal Domain that these analysis identified, in addition to the large concentration of mutations in the S1 Receptor-Binding Domain, indicating a broad consistency across the Omicron samples. Overall, these results suggest that Omicron does exhibit consistent mutations of the kinds described in these analyses, and that healthcare policymakers and providers should take measures commensurate with the spread of a new, potentially more transmissible form of COVID-19 going into 2022.

References

"Bio.pairwise2 Module." *Bio.pairwise2 Module - Biopython 1.75 Documentation*, https://biopython.org/docs/1.75/api/Bio.pairwise2.html.

"Biopython Documentation." *Biopython Tutorial and Cookbook*, http://biopython.org/DIST/docs/tutorial/Tutorial.html.

Callaway, Ewen, and Heidi Ledford. "How Bad Is Omicron? What Scientists Know so Far." *Nature News*, Nature Publishing Group, 2 Dec. 2021, https://www.nature.com/articles/d41586-021-03614-z.

"Do Covid-19 Vaccines Protect against the Variants?" *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 17 Dec. 2021, https://www.mayoclinic.org/coronavirus-covid-19/covid-variant-vaccine.

"GenBank Overview." *National Center for Biotechnology Information*, U.S. National Library of Medicine, https://www.ncbi.nlm.nih.gov/genbank/.

"Genetic Variants May Lead to False Negatives with SARS-COV-2 Molecular." *U.S. Food and Drug Administration*, FDA, https://www.fda.gov/medical-devices/letters-health-care-providers/genetic-variants-sars-cov-2-may-lead-false-negative-results-molecular-tests-detection-sars-cov-2.

"Implications of the Emergence and Spread of the SARS-COV-2 ..." *European Centre for Disease Prevention and Control*, 26 Nov. 2021, https://www.ecdc.europa.eu/sites/default/files/documents/Implications-emergence-spread-SARS-CoV-2%20B.1.1.529-variant-concern-Omicron-for-the-EU-EEA-Nov2021.pdf.

Lee, Soohyun. “Lecture 2: Pairwise Sequence Alignment.” *Introductory Bioinformatics*, Harvard Extension School, 13 Sept. 2021, https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=edb361bb-df3d-2779-9c68-b3d23bcab476.

Logist, A.-S., Vanmechelen, B., Cuypers, L., Wawina-Bokalanga, T., Verlinden, J., Marti-Carerras, J., Dellicour, S., Andre, E., Baele, G. and Maes, P. “Identification of a SARS-CoV-2 Lineage B.1.1.529 Virus in Belgium.” *NCBI GenBank*, NCBI, 30 Nov. 2021, https://www.ncbi.nlm.nih.gov/nuccore/OL672836.1/.

“S Surface Glycoprotein [Severe Acute Respiratory Syndrome Coronavirus 2] - Gene - NCBI.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, 11 Dec. 2021, https://www.ncbi.nlm.nih.gov/gene/43740568.

“Spike Glycoprotein S1, N-Terminal Domain, Betacoronavirus-Like.” *InterPro*, https://www.ebi.ac.uk/interpro/entry/InterPro/IPR032500/.

“Surface Glycoprotein [Severe Acute Respiratory Syndrome Coronavirus 2] - Protein - NCBI.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, https://www.ncbi.nlm.nih.gov/protein/YP_009724390.1?feature=any.

“Update on Omicron.” *World Health Organization*, World Health Organization, 28 Nov. 2021, https://www.who.int/news/item/28-11-2021-update-on-omicron.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C. and Zhang, Y.Z. “A New Coronavirus Associated with Human Respiratory Disease in China.” *NCBI GenBank*, NCBI, 17 Jan. 2020, https://www.ncbi.nlm.nih.gov/nuccore/1798174254.

Code Repository: https://github.com/aamari94/bioinformatics_final/blob/main/Bioinformaticsfinal.ipynb