Intro

We are investigating the SARS-CoV-2 receptor binding domain and its possible adaptations to further facilitate binding to the ACE2 receptor. The ACE2 receptor is proven to mediate viruses like SARS and acute lung injury (ALI). This receptor is targeted for possible therapies to alleviate viral infection. Binding to ACE2 is essential for the virus to infect the host. Therefore, this receptor is the target for preventing viral infection, and the receptor binding domain of sars cov 2 is a target for disease prevention. ACE2 is negatively charged, possibly allowing for a positively charged RBD of the virus in question. there are hydrophobic residues on the ACE2 receptor that could also facilitate in binding to the covid RBD. '...binding site on the SARS-CoVS glycoprotein was localized between amino acid residues 303 and 537'(A model of the ACE2 structure and function as a SARS-CoV receptor).

SARS-CoV-2 contains spike glycoproteins on its surface that allow the virus to attach to the ACE2 receptor in human cells, letting the virus infect the host. Because these spike proteins are on the surface of the virus, they are a target for treatments such as antibodies. The ACE2 receptor has specific points that are vital for the binding of the virus to take place. There are adaptations on the SARS CoV virus that encircle these points and cause COVID to be more virulent. Focusing on these points is vital to understanding how the virus adapts and mutates to increase infectivity.

A database of COVID 19 receptor binding domain protein sequences was compared to a database of COVID 19 surface glycoproteins sequenced from various collection sites. This comparison can infer relatedness and changes to the genome over time. A clustal alignment was performed on

the receptor binding domain database to eliminate redundancy of the query. Many of these proteins share similarities in sequence, possibly indicating the presence of an antibody bound to the protein.

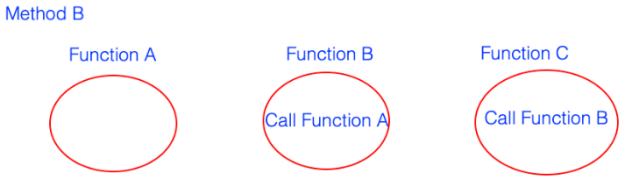
The ACE2 receptor has been recognized as a receptor for the SARS virus receptor binding domain. The specific feature of interest concerning the ACE2 receptor is the catalytic site on the top of the molecule. This is surrounded with ridges that are negatively charged. This could be a binding site for a positively charged receptor binding domain of COVID and focus on this could aid in treatment of the disease.

## Materials and methods

The sequences of the covid receptor binding domain were aligned using clustal omega, a multiple sequence alignment program. The program blastp was used to run an alignment between the query: the clustal alignment of the receptor binding domain of SARS CoV 2, and the database consisting of SARS CoV 2 surface glycoproteins from multiple regions. The pipeline constructed in python used Entrez to first import the SARS CoV 2 RBD data from the NCBI database. This was done with the search term "SARS-CoV-2 Spike protein S1" from the protein database. After these sequences were obtained, I isolated the receptor binding domain by parsing the file and selecting only the sequences containing "Spike protein S1". From here, a clustalo could be run on only the receptor binding domain of SARS CoV 2. After the clustalo, a blastp was run comparing the S1 spike proteins of the receptor binding domain to a large database of SARS CoV 2 surface glycoproteins. This was done with the set parameter of an evalue

<0.1. From here, the hits that were analyzed had a 97% identity. SNP's were investigated for changes in function.

The coding used to obtain the S1 glycoprotein information and run alignments was designed in an analysis pipeline to make this particular analysis repeatable with any input. Functions created for each task make it possible to change in input while keeping the formatting of the rest of the analysis. Function A was designed to retrieve information from NCBI database while allowing for different search terms and out file names to be implemented. The second function, (B), is meant to call the first function and run a clustal alignment on the fasta file obtained through NCBI. This function allows for a different input and output file. The final function, (C), will call the output from function B and run that clustal alignment against a local database. This is the analysis pipeline diagram and description mentioned in lesson 11.




Below is an example of the clustal alignment output.

```
with open('/Users/jessi/OneDrive/Desktop/clustalo/clustal-omega-1.2.2-win64/clustal.aln')as clustal:
    contents = clustal.read()
print(contents)
```

```
CLUSTAL O(1.2.2) multiple sequence alignment


pdb|7S8U|A      ----------------TNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    35
pdb|7Z0Y|R      ------------QNITNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    38
pdb|7Z0X|R      ------------QNITNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    38
pdb|7U0N|F      --RVVPSGDVVRFPNITNLC-------PFGEVFNATKFPSVYAWERKKISNCVADY--SV    49
pdb|7U0N|E      --RVVPSGDVVRFPNITNLC-------PFGEVFNATKFPSVYAWERKKISNCVADY--SV    49
pdb|7WEF|E      ------------PNITNLC-------PFDEVFNATRFASVYAWNRKRISNCVADY--SV    38
pdb|7UR5|A      --RVQPTESIVRFPNITNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    49
pdb|7URQ|A      GSRVQPTESIVRFPNITNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    51
pdb|7U2E|A      ----------------TNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    35
pdb|7U2D|A      ----------------TNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    35
pdb|7TTY|A      ----------------NLC-------PFGEVFNATTFPSVYAWERKRISNCVADY--SV    34
pdb|7R8U|B      ---------QCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSV-------LHSTQDLFLPF    45
pdb|7TE1|E      --RVQPTESIVRFPNITNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    49
pdb|7TE1|D      --RVQPTESIVRFPNITNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    49
pdb|7R8Y|C      ------------FPNITNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    39
pdb|7R8Y|A      ------------FPNITNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    39
pdb|7WUE|B      ----------------TNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    35
pdb|7WUE|A      ----------------TNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    35
pdb|7TN0|S      -------ETGTRFPNITNLC-------PFDEVFNATRFASVYAWNRKRISNCVADY--SV    44
pdb|7TN0|I      -------ETGTRFPNITNLC-------PFDEVFNATRFASVYAWNRKRISNCVADY--SV    44
pdb|7S0B|F      --RVQPTESIVRFPNITNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    49
pdb|7S0B|E      --RVQPTESIVRFPNITNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    49
pdb|7NAJ|A      ------------NITNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    37
pdb|7CYV|B      ----------AGSPNITNLC-------PFGEVFNAXRFASVYAWNRKRISNCVADY--SV    41
pdb|7FDG|E      ---MFVFLVLLPLVSSTNLC-------PFGEVFNATRFASVYAWNRKRISNCVADY--SV    48
                                :*           :   * *         ... *   .
```

Please note that this analysis pipeline was not put into full use because of the difficulty with installing certain programs on windows. The alignments were run through the command line instead of

through a function in jupyter notebook because it was not possible to run through the jupyter notebook. Below is the output from the blastp. This included only significant evalues and the collection sites of the samples from the database. This blstp was run locally. The query sequences were traced back by parsing the clustal file for the alignment in question.

```
QWT94942.1 |ssRNA(+)|GenBank|Brazil|                                  165     1e-42
QZD99304.1 |ssRNA(+)|GenBank|Iraq|                                    163     1e-42
QWT94947.1 |ssRNA(+)|GenBank|Brazil|                                  165     1e-42
QWT94832.1 |ssRNA(+)|GenBank|Brazil|                                  166     1e-42
QWT94833.1 |ssRNA(+)|GenBank|Brazil|                                  166     1e-42
QWT94839.1 |ssRNA(+)|GenBank|Brazil|                                  166     1e-42
QWT94935.1 |ssRNA(+)|GenBank|Brazil|                                  166     1e-42
QWT94927.1 |ssRNA(+)|GenBank|Brazil|                                  165     1e-42
QWT94835.1 |ssRNA(+)|GenBank|Brazil|                                  165     1e-42


>QSL96982.1 |ssRNA(+)|GenBank|India: Assam|
Length=167

 Score = 166 bits (421),  Expect = 2e-44, Method: Compositional matrix adjust.
 Identities = 90/129 (70%), Positives = 94/129 (73%), Gaps = 15/129 (12%)

Query  3522  PGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSK*PDBSBUAVGGNYNYLYRLFRKSNLKP  3581
             PGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSK        VGGNYNYLYRLFRKSNLKP
Sbjct  2     PGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSK--------VGGNYNYLYRLFRKSNLKP  53

Query  3582  FERDISTEIYQAGSTPCNGVEGFNCYFPDBZYRVGGNYNYLYRLFRKSNLKPFERDISTE  3641
             FERDISTEIYQAGSTPCNGVEGFNCYFP   Y     Y   Y+ +R   L       S E
Sbjct  54    FERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTYGVGYQPYRVVVL-------SFE  106

Query  3642  IYQAGSTPC  3650
             +  A +T C
Sbjct  107   LLHAPATVC  115


 Score = 155 bits (393),  Expect = 9e-41, Method: Compositional matrix adjust.
 Identities = 79/104 (76%), Positives = 83/104 (80%), Gaps = 12/104 (12%)

Query  4815  DBFDGEVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYF*****PDB  4874
             ++ D +VGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYF
Sbjct  29    NNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYF--------  80

Query  4875  SBUAPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPDBZYRPLQS  4918
                 PLQSYGFQPT GVGYQPYRVVVLSFELLHAP     P +S
Sbjct  81    ----PLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKS  120
```

## Discussion

Upon analyzing the clustal data, I noticed there were many sequences that had big chunks of amino acids in common with each other. These sequences are neutralizing antibodies bound to the S1 protein. This complex was usually the COVID spike protein bound with a potent antibody, J08 Fab. These neutralizing antibodies could then be ignored following further analysis. The blastp results analyzed were focused on single amino acid substitutions of the alignments with a 97% identity and evalue of <0.1. This allowed me to infer function and adaptation.

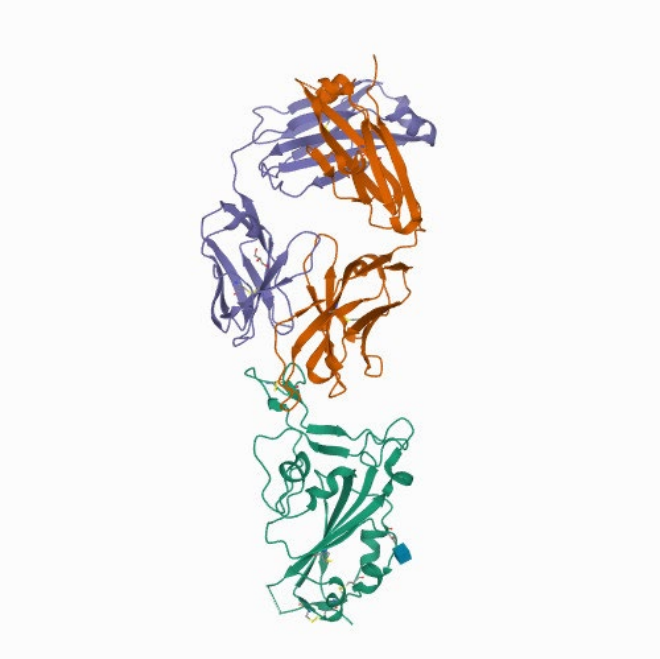
The following sequence meets the above parameters:

QSL96982.1 India: Assam, 97% identity. This query is pdb|7SBU|A from the clustal alignment.

| | | |
|---|---|---|
| Query 3126 | CPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSK | 3159 |
| | PGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSK | |
| Sbjct 1 | APGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSK | 34 |

Subject contains Alanine at the beginning of this sequence, while the query contains a cysteine. Alanine is hydrophobic, while cysteine is hydrophilic. This may indicate that the query protein (pdb|7SBU|A), is more hydrophilic than the subject. This may give it higher binding affinity to the ACE2 receptor than the subject.

This query is 7SBU Crystal structure of SARS-CoV-2 spike protein receptor-binding domain in complex with a highly potent antibody (J08 Fab).This is the structure of the query receptor binding domain bound with an antibody.

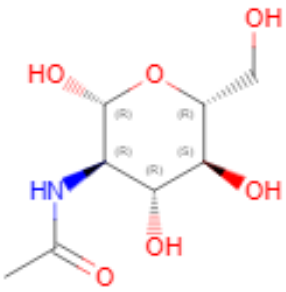


https://cdn.rcsb.org/images/structures/sb/7sbu/7sbu_assembly-1.jpeg

The similarity of the RBD of SARS CoV 2 to other well documented surface glycoproteins of other coronaviruses is lower than 20% similarity. ('A Model of the ACE2 structure and function as a SARS-CoV receptor'). Therefor, treatments for SARS-CoV-2 will be novel and address the S1 glycoproteins. These S1 glycoproteins are known to have an overall positive charge on the surface of the virus and a loop that is electronegative. The hydrophobicity of the viruses surface surrounds the positively charged areas.

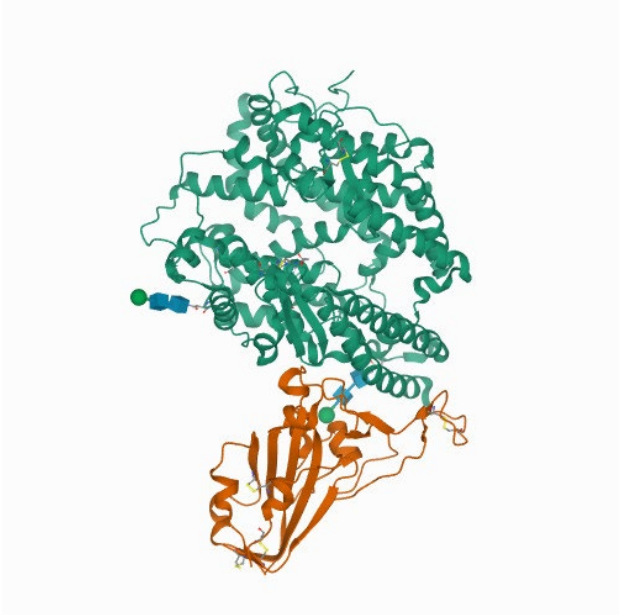
There is a sequence that is conserved: CPFGEVFN where the N is bound to NAG 601(NAG: 2-acetamido-2-deoxy-beta-D-glucopyranose), Pictured above in blue. This structure is also of interest because NAG 601 is a known epitope and may also be targeted for treatment development. Below is the structure of NAG 601. (2) The omicron variant contains more of these

NAG structures, possibly contributing to its increased virulence. These are bound to an Asparagine.

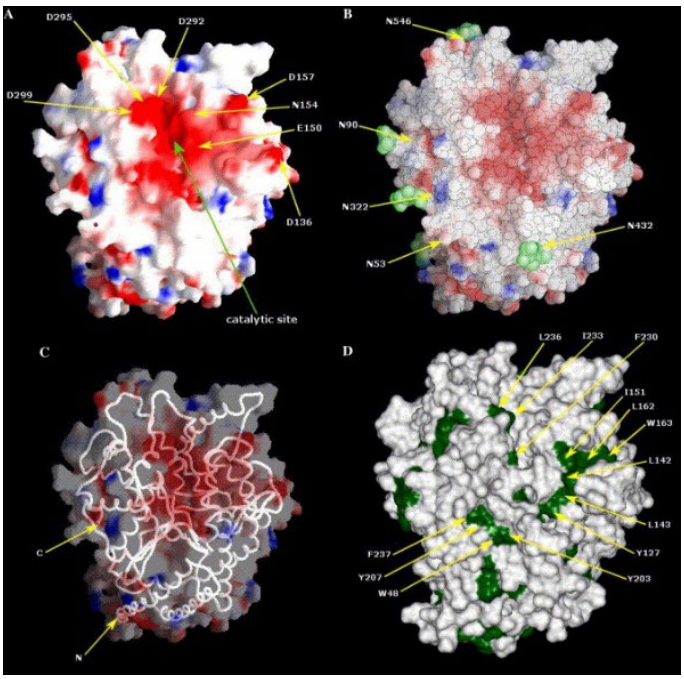



https://cdn.rcsb.org/images/ccd/unlabeled/N/NAG.svg

This is an image of 7U0N Crystal structure of chimeric omicron RBD (strain BA.1) complexed with human ACE2. The omicron variant is bound to ACE2 with a more enhanced affinity than other strains of COVID. NAG: 2-acetamido-2-deoxy-beta-D-glucopyranose, pictured in blue.



(Source: A model of the ACE2 structure and function as a SARS-CoV receptor)




Shown here is a model of the ACE2 receptor showing the potential of negative and positive charges on the surface of the molecule. The catalytic site is overall negatively charged. The last model pictured shows the hydrophobic regions of the molecule in green. Binding of viruses to protein receptors such as ACE2 are established with complementing charge, geometric configuration, and hydrophobicity. The negatively charged catalytic site at the top of the ACE2 molecule is surrounded with hydrophobic points that are compromised of Phe, Trp, and Tyr. These cause binding site to be charged and are likely the main molecules that allow for virus attachment. The highly negative ridges surrounding the catalytic site of ACE2 likely give the enzyme higher binding affinity. Conservation of the amino acids Asn and Asp in ACE2 could also contribute to higher binding affinity to surface glycoproteins, such as those on SARS CoV 2.

Another 97% hit from the query pdb|7SBU|A (mentioned above) is:

| | | |
|---|---|---|
| Query 5097 | PLQSYGFQPTNGVGYQPYRVVVLSFELLHAP | 5127 |
| | PLQSYGFQPT GVGYQPYRVVVLSFELLHAP | |
| Sbjct 81 | PLQSYGFQPTYGVGYQPYRVVVLSFELLHAP | 111 |

There is a SNP from a Y to an N. Y(Tyrosine): polar, neutral. N(Asparagine): polar uncharged, hydrophilic. Considering that there are hydrophobic patches near the catalytic site of the ACE2 receptor, this change from neutral to hydrophilic may aid in binding the COVID virus to ACE2.

This subject is UCF18145.1 |ssRNA(+)|GenBank|Canada|. Here this subject is aligned with pdb|7TN0|S from the clustal alignment.

| | | |
|---|---|---|
| Query 912 | CPFDEVFNATRFASVYAWNRKRISNCVADYSV | 943 |
| | CPF EVFNATRFASVYAWNRKRISNCVADYSV | |
| Sbjct 1 | CPFGEVFNATRFASVYAWNRKRISNCVADYSV | 32 |

This query sequence is the omicron variant of SARS CoV 2 in complex with ACE2 and two antibodies. There is a shift here from glycine to aspartic acid. This is another SNP that is an example of a shift from neutral to hydrophilic. This may be another indication of a polar molecule exhibiting higher binding affinity to ACE2.

## Summary

In summation, many of the SNP's recognized from the blast were a shift from a neutral or hydrophobic amino acid to a hydrophilic amino acid. This could indicate adaptations of the virus toward higher binding affinity to the ACE2 receptor. Further investigation would involve

examining the properties of the ACE2 catalytic site and further analyzing the NAG structures and how they are bound to the Omicron variant in comparison to other variants of COVID 19. Further analysis of 3D models of COVID in complex with the ACE2 receptor would also be a next step. This could help shed some light on how different variations of COVID bind to ACE2 and if their binding affinity has correlation with virulence. This would ultimately be in efforts to design a treatment for the disease and to see how SARS viruses adapt over time.

References

Liu, Mengyuan, et al. "Potential role of ACE2 in coronavirus disease 2019 (COVID-19) prevention and management." *Journal of translational internal medicine* 8.1 (2020): 9-19.

Lan, J., Ge, J., Yu, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581,** 215–220 (2020). https://doi.org/10.1038/s41586-020-2180-5

Ponraj Prabakaran, Xiaodong Xiao, Dimiter, S Dimitrov A model of the ACE2 structure and function as a SARS-CoV receptor. Biochem Biophys Res Commun 314(1): 235–241 (2004)

Python:(https://www.datacamp.com/learn-python-with-anaconda?utm_source=adwords_ppc&utm_medium=cpc&utm_campaignid=1565610360&utm_adgroupid=63334253950&utm_device=c&utm_keyword=python%20programming&utm_matchtype=e&utm_network=g&utm_adpostion=&utm_creative=295213502597&utm_targetid=aud-392016246653:kwd-59293786&utm_loc_interest_ms=&utm_loc_physical_ms=9031300&gclid=Cj0KCQjw1ZeUBhDyARIsAOzAqQKpmoqEpo0zO5BofCwd1yLzgPQyH5eMswEoXIeIy6hJE6SQyViX1d4aAgnSEALw_wcB)

Biopython: (https://biopython.org/)

Bioservices: (https://bioservices.readthedocs.io/en/master/)

RCSB Protein data bank: (https://www.rcsb.org/)

The SARS-CoV-2 Spike Glycoprotein Biosynthesis, Structure, Function, and Antigenicity: Implications for the Design of Spike-Based Vaccine Immunogens

Liangwei Duan1,2, Qianqian Zheng1,2, Hongxia Zhang1,2, Yuna Niu1,2, Yunwei Lou1,2 and Hui Wang1,2*

Structural basis of receptor recognition by SARS-CoV-2

Jian Shang, Gang Ye, Ke Shi, Yushun Wan, Chuming Luo, Hideki Aihara, Qibin Geng, Ashley Auerbach & Fang Li

BLAST® Command Line Applications User Manual [Internet]. Bethesda (MD): National Center for Biotechnology. Information (US); 2008-.

Stephen F. AltschuP, Warren Gish ~, Webb Miller 2, Eugene W. Myers 3 and David J. Lipman ~

“BLAST+: architecture and applications. “Basic Local Alignment Search Tool”. ~Nalional Center for Biotechnology Information National Library of Medicine, National lnstitules of Health Belhesda, MD 20894, U.S.A. 2Department of Computer Science, The Pennsylvania State University, Universily Park, PA 16802, U.S.A. 3Department of Computer Science, University of Arizona, Tucson, AZ 85721, U.S.A. (Received 26 February 1990; accepted 15 May 1990).