

COVID-19 Protein Comparison Analysis



Author:

Nikhil Malhotra

Global Head of Innovation, Tech Mahindra

THE WORLD OF MICROBES

We are living in a world which got jolted in January by the onset of a brand new enemy, an enemy which cannot be seen but can only be felt, via nausea, headaches, shortness of breath and in some unfortunate cases the quintessential death. The Novel Corona Virus – nCoV19 or a close cousin or a form of the SARS virus family. The virus seems to be virulent and transmissible by touch or on the surfaces that the virus deposits itself onto. Within weeks, the virus infects inhabitants in Wuhan, China from where it started and the number of people affected by the virus quickly reach 1000s and 100000 by the time we are writing this paper.

The entire world primary care community goes to an overdrive where patients keep filtering in and the virus has become virulent to cause 1000s of death. However, the virus shows a very unique behavior, it seems to miss kids of the age 0-9 years and seems to affect people with the age band 65 years and older. This challenge that the world faces prompted Makers Lab, R&D division of TechMahindra to go into an overdrive ourselves to see what can be done round it purely from a computational model perspective, and see if we find something.

The Coronavirus disease 2019 (COVID-19) pandemic is rapidly evolving; it has spread to more than 150 countries. By March 20th, 2020, there are more than 250,000 confirmed cases and at least 10,000 have died. More than two-third of them are outside mainland China where the virus originated. There are no vaccines available and there is little evidence on the effectiveness of potential therapeutic agents. In addition, there is presumably no pre-existing immunity in the population against COVID-19 and everyone in the population is assumed to be susceptible.

OUR DESIGN PRINCIPLE

Our design principle is based on a very common software principle called YAGNI (*You ain't gonna need it*). This principle has been designed to simplify the process and act of writing code, so that it affects the most where needed, without changing the entire underlying premise

We were aware that protein structures and shape do affect the way proteins behave but as a starting point we just wanted to look at peptide chains in Corona and compare them with

peptide chains in other viruses of similar types, the basic idea being if a chemical compound is known to act on one, it might give scientists a head start to act on this virus.

OUR PRE-STUDY

We are computational scientists with little or no knowledge of Biology or the functioning of the body, leave alone the world of viruses and bacteria, so we had to resort to classes online to understand the virus, its structure and also how it functions. We studied for 48 hours straight, and some of the classes I found brilliant which gave me understanding were classes on Virology (Virology Lectures by Vincent Racaniello)

https://www.youtube.com/watch?v=svlKm4S1M3Y&list=PLGhmZX2NKiNlwig68CGPHQI_Pcxri4LDx

I would highly recommend anyone trying to understand viruses to go to this lecture

One thing that made us clear at the lab was the way, a virus behaves in comparison to a bacteria and why mechanisms to fight bacterial infection will never work for viruses.

Bacteria self-replicate and viruses need a host. There has been an endless debate on whether a virus is living or not, a debate I could settle in my head by the fact, if the virus lives outside the host when it is not replicating, it is non-living. The virus only starts living when it enters the host cells and tries to transcribe its mRNA to make copies. Given below is the table of difference between viruses and bacteria

Viruses	Bacteria
Traditionally very very small, although some large viruses have been found	Larger in size compared to viruses
Have no metabolism of their own	Have metabolism of their own
Take no food	Take food by absorption
Do not grow and divide	Grow and divide to produce more
Use the host cells to replicate	Can replicate and reproduce on their own
All viruses are not harmful	All bacteria are not harmful

A virus is defined as ‘An infectious, obligate intercellular parasite, comprising of genetic material (RNA or DNA) surrounded by a protein coat, sometimes a membrane - Vincent Racaneillo

Virologist the world over divide the viral infectious cycle in two phases even though no such boundaries exist. Having said that an infectious cycle is shown below :

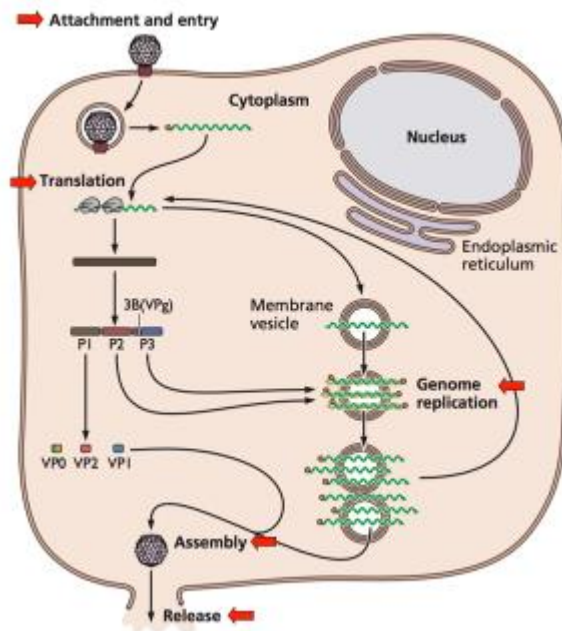


Fig 1: Source: Vincent Racaneillo course

What we see is the way the virus uses the host cell. The viral spike protein after attachment on the cell membrane essentially breaks open in two parts. One of the parts are used for genome replication and the other part is used for translation and creating more viral proteins. The compounds are mixed together to make more viruses which are then released in greater number.

EXPERIMENTAL RESEARCH

Armed with this knowledge, and backed by the fact that protein membrane of the virus plays a major part in attacking a host cell within the human body and then replicates, we wanted to do the following studies, more like questions that we wanted an answer about

- Can we get the viral protein structure and can we make a 3d model of atoms by placement?
- Can we compare the protein amino acid chains of two or more viruses?

In order to do that, we used Bio Python to search from the RCSB database to get PDB files and see if we could construct the viral structures and atom placement. On further finding out, we discovered that BioPython has a PDB parser that enables us to get chains of models, residues and atoms and also gives the x,y and z coordinates of the atom placements. Our job was to find out the structure and plot it. First task was easy and we managed to create a good structure of not just the Corona virus but also other protein structures available to us.

Given below is an image from our system created file showing the atomic placement of corona virus as compared to the structure available on RCSB database.



This close match gave us a lot of confidence in proceeding to the next task for comparing protein nucleotides

There are two mechanisms in which we would have measured the protein peptide links, either through converting them into a vector, or using sequence matching techniques to see which peptide chains match. The structure of pdb is such that it has to be converted into chains. We converted it into chains of atoms like 'N-CA-C-CB-O-CG-CD-CE-NZ' which essentially meant atoms in one single chain and wrote a code to compare only the chains itself between two virus proteins. The idea was to check how many sequences match, and how much does the protein sequence match in total. Even though we knew, proteins do fold at temperatures and it is the protein's shape that gives the protein its characteristic behavior, we wanted to limit our search to only chain values, because we started with the assumption that if chains can be found similar, then similar antidotes might work on them

MINI DISCOVERY

We first started by blasting open the pdb structure obtained from BioPython PDB parser as shown below

```
if(extension=='pdb'):

    data = parser.get_structure('pdb',main_virus_needed+'\\'+filename)

    initial_models = data.get_models()

    models = list(initial_models)

    type(models[0])

    for model in models:

        chains = list(models[0].get_chains())

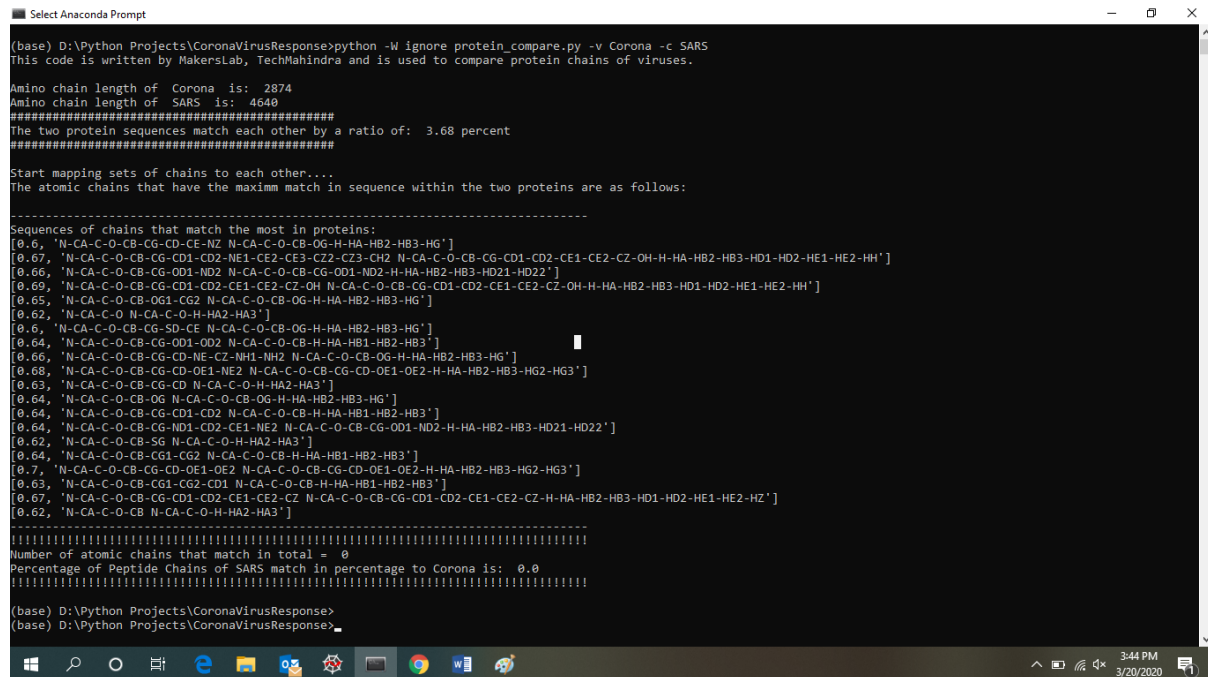
        for chain in chains:
```

```
residues = list(chains[0].get_residues())  
  
for residue in residues:  
  
    atoms = list(residue.get_atoms())  
  
    atoms_li = []  
  
    #Getting atom chains here to append  
  
    for atom in atoms:  
  
        atoms_li.append((atom.get_name()))  
  
    residue_atoms.append(atoms_li)
```

PDB structures are maintained in sequences of residues containing atoms at a specific position on x,y and z axis so we combined the atoms in one residue to form a chain like the one explained above

```
str_atoms_c.append(''.join(map(str,i)))
```

Once these chains were obtained we started comparing peptide chains of one protein to another, given below are some results we obtained from the system which completely astonished us



```

Select Anaconda Prompt

(base) D:\Python Projects\CoronaVirusResponse>python -W ignore protein_compare.py -v Corona -c SARS
This code is written by MakersLab, TechMahindra and is used to compare protein chains of viruses.

Amino chain length of Corona is: 2874
Amino chain length of SARS is: 4648
#####
The two protein sequences match each other by a ratio of: 3.68 percent
#####

Start mapping sets of chains to each other....
The atomic chains that have the maximm match in sequence within the two proteins are as follows:
-----
Sequences of chains that match the most in proteins:
[0.6, 'N-CA-C-O-CB-CG-CD-CE-NZ N-CA-C-O-CB-CG-H-HA-HB2-HB3-HG']
[0.67, 'N-CA-C-O-CB-CG-CD1-CD2-NE1-CE2-CE3-CZ2-CZ3-CH2 N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ-OH-H-HA-HB2-HB3-HD1-HD2-HE1-HE2-HH']
[0.66, 'N-CA-C-O-CB-CG-OD1-ND2 N-CA-C-O-CB-CG-OD1-ND2-H-HA-HB2-HB3-HD21-HD22']
[0.69, 'N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ-OH N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ-OH-H-HA-HB2-HB3-HD1-HD2-HE1-HE2-HH']
[0.65, 'N-CA-C-O-CB-OG1-CG2 N-CA-C-O-CB-OG-H-HA-HB2-HB3-HG']
[0.62, 'N-CA-C-O N-CA-C-O-H-HA2-HA3']
[0.6, 'N-CA-C-O-CB-CG-SD-CE N-CA-C-O-CB-CG-H-HA-HB2-HB3-HG']
[0.64, 'N-CA-C-O-CB-CG-OD1-OD2 N-CA-C-O-CB-H-HA-HB1-HB2-HB3']
[0.66, 'N-CA-C-O-CB-CG-CD-NE-CZ-NH1-NH2 N-CA-C-O-CB-CG-H-HA-HB2-HB3-HG']
[0.68, 'N-CA-C-O-CB-CG-CD-NE1-NE2 N-CA-C-O-CB-CG-CD-CE1-CE2-H-HA-HB2-HB3-HG2-HG3']
[0.63, 'N-CA-C-O-CB-CG-CD N-CA-C-O-H-HA2-HA3']
[0.64, 'N-CA-C-O-CB-OG N-CA-C-O-CB-OG-H-HA-HB2-HB3-HG']
[0.64, 'N-CA-C-O-CB-CG-CD1-CD2 N-CA-C-O-CB-H-HA-HB1-HB2-HB3']
[0.64, 'N-CA-C-O-CB-CG-ND1-CD2-CE1-NE2 N-CA-C-O-CB-CG-OD1-ND2-H-HA-HB2-HB3-HD21-HD22']
[0.62, 'N-CA-C-O-CB-SG N-CA-C-O-H-HA2-HA3']
[0.64, 'N-CA-C-O-CB-CG1-CG2 N-CA-C-O-CB-H-HA-HB1-HB2-HB3']
[0.7, 'N-CA-C-O-CB-CG-CD-CE1-CE2 N-CA-C-O-CB-CG-CD-CE1-CE2-H-HA-HB2-HB3-HG2-HG3']
[0.63, 'N-CA-C-O-CB-CG1-CG2-CD1 N-CA-C-O-CB-H-HA-HB1-HB2-HB3']
[0.67, 'N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ-H-HA-HB2-HB3-HD1-HD2-HE1-HE2-HZ']
[0.62, 'N-CA-C-O-CB N-CA-C-O-H-HA2-HA3']
-----
Number of atomic chains that match in total = 0
Percentage of Peptide Chains of SARS match in percentage to Corona is: 0.0
-----

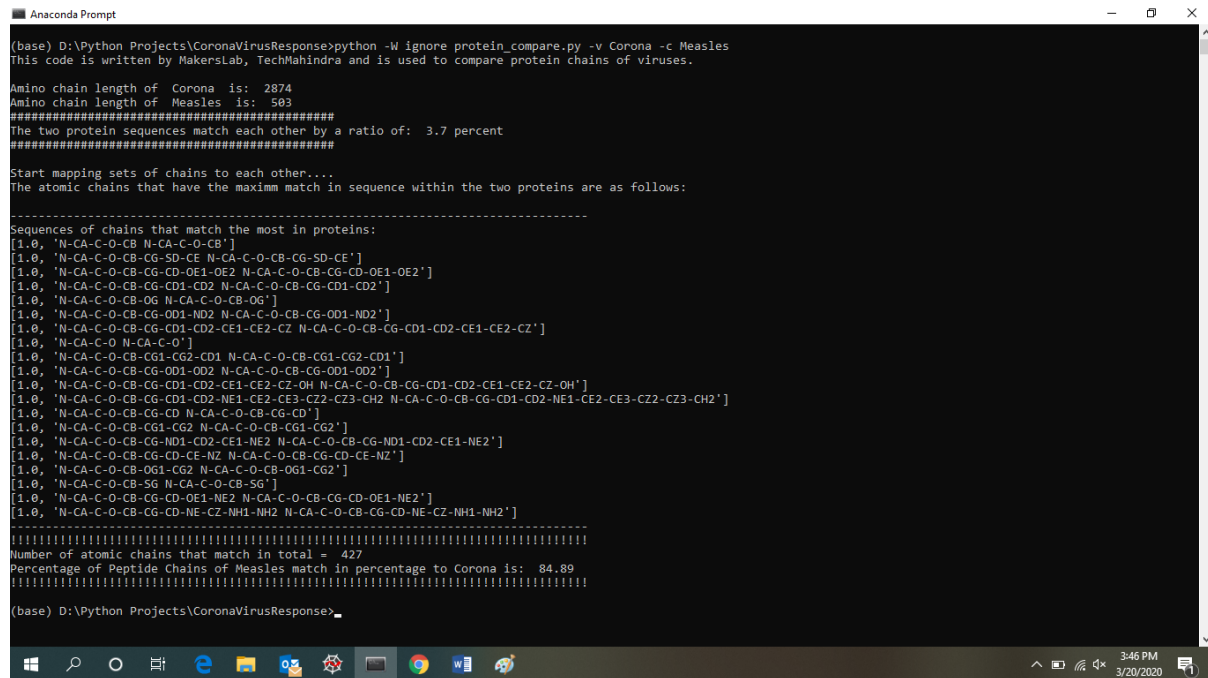
(base) D:\Python Projects\CoronaVirusResponse>
(base) D:\Python Projects\CoronaVirusResponse>

```

[Figure 2: Comparison between Corona and Sars virus proteins Source: Makers Lab Tech Mahindra]

Amino peptide chains when compared between Corona and SARS was an astonishment as even though the novel coronavirus is considered to be a form of SARs, the amino peptide chains hardly matched. The sequence similarity was 3.68 percent and none of the chains actually matched within a set.

Similarly comparisons were drawn with Measles and Influenza virus proteins and the results are shown below



```

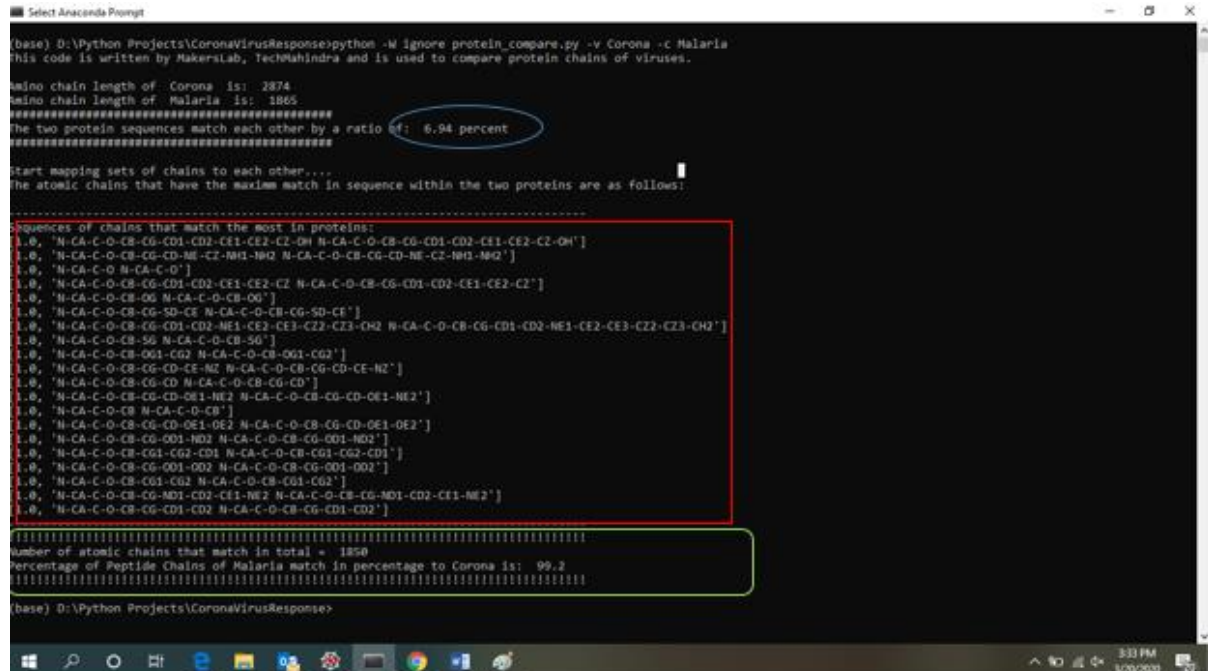
Anaconda Prompt
(base) D:\Python Projects\CoronaVirusResponse>python -W ignore protein_compare.py -v Corona -c Measles
This code is written by MakersLab, TechMahindra and is used to compare protein chains of viruses.

Amino chain length of Corona is: 2874
Amino chain length of Measles is: 503
=====
The two protein sequences match each other by a ratio of: 3.7 percent
=====
Start mapping sets of chains to each other...
The atomic chains that have the maximm match in sequence within the two proteins are as follows:
-----
Sequences of chains that match the most in proteins:
[1.0, 'N-CA-C-O-CB-N-CA-C-O-CB']
[1.0, 'N-CA-C-O-CB-CG-SD-CE N-CA-C-O-CB-CG-SD-CE']
[1.0, 'N-CA-C-O-CB-CG-CD-0E1-0E2 N-CA-C-O-CB-CG-CD-0E1-0E2']
[1.0, 'N-CA-C-O-CB-CG-CD1-CD2 N-CA-C-O-CB-CG-CD1-CD2']
[1.0, 'N-CA-C-O-CB-OG N-CA-C-O-CB-OG']
[1.0, 'N-CA-C-O-CB-CG-OD1-ND2 N-CA-C-O-CB-CG-OD1-ND2']
[1.0, 'N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ']
[1.0, 'N-CA-C-O-N-CA-C-O']
[1.0, 'N-CA-C-O-CB-CG1-CG2-CD1 N-CA-C-O-CB-CG1-CG2-CD1']
[1.0, 'N-CA-C-O-CB-CG-OD1-OD2 N-CA-C-O-CB-CG-OD1-OD2']
[1.0, 'N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ-OH N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ-OH']
[1.0, 'N-CA-C-O-CB-CG-CD1-CD2-NE1-CE2-CE3-CZ2-CZ3-CH2 N-CA-C-O-CB-CG-CD1-CD2-NE1-CE2-CE3-CZ2-CZ3-CH2']
[1.0, 'N-CA-C-O-CB-CG-CD N-CA-C-O-CB-CG-CD']
[1.0, 'N-CA-C-O-CB-CG1-CG2 N-CA-C-O-CB-CG1-CG2']
[1.0, 'N-CA-C-O-CB-CG-ND1-CD2-CE1-NE2 N-CA-C-O-CB-CG-ND1-CD2-CE1-NE2']
[1.0, 'N-CA-C-O-CB-CG-CE-NZ N-CA-C-O-CB-CG-CD-CE-NZ']
[1.0, 'N-CA-C-O-CB-OG1-CG2 N-CA-C-O-CB-OG1-CG2']
[1.0, 'N-CA-C-O-CB-S6 N-CA-C-O-CB-S6']
[1.0, 'N-CA-C-O-CB-CG-CD-0E1-NE2 N-CA-C-O-CB-CG-CD-0E1-NE2']
[1.0, 'N-CA-C-O-CB-CG-CD-NE-CZ-NH1-NH2 N-CA-C-O-CB-CG-CD-NE-CZ-NH1-NH2']
-----
Number of atomic chains that match in total = 427
Percentage of Peptide Chains of Measles match in percentage to Corona is: 84.89
-----
(base) D:\Python Projects\CoronaVirusResponse>

```

[Figure 3: Comparison between Corona and Measles proteins Source: Makers Lab Tech Mahindra]

84.89 % of the peptide chains in Measles were found in Corona and about 78.5% of chains matched with Influenza. It was a by a chance that we also measured this against Malaria...



```

Select Anaconda Prompt
(base) D:\Python Projects\CoronaVirusResponse>python -W ignore protein_compare.py -v Corona -c Malaria
This code is written by MakersLab, TechMahindra and is used to compare protein chains of viruses.

Amino chain length of Corona is: 2874
Amino chain length of Malaria is: 1865
=====
The two protein sequences match each other by a ratio of: 6.94 percent
=====
Start mapping sets of chains to each other...
The atomic chains that have the maximm match in sequence within the two proteins are as follows:
-----
Sequences of chains that match the most in proteins:
[1.0, 'N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ-OH N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ-OH']
[1.0, 'N-CA-C-O-CB-CG-CD-NE-CZ-NH1-NH2 N-CA-C-O-CB-CG-CD-NE-CZ-NH1-NH2']
[1.0, 'N-CA-C-O-N-CA-C-O']
[1.0, 'N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ']
[1.0, 'N-CA-C-O-CB-OG N-CA-C-O-CB-OG']
[1.0, 'N-CA-C-O-CB-CG-SD-CE N-CA-C-O-CB-CG-SD-CE']
[1.0, 'N-CA-C-O-CB-CG-CD1-CD2-NE1-CE2-CE3-CZ2-CZ3-CH2 N-CA-C-O-CB-CG-CD1-CD2-NE1-CE2-CE3-CZ2-CZ3-CH2']
[1.0, 'N-CA-C-O-CB-S6 N-CA-C-O-CB-S6']
[1.0, 'N-CA-C-O-CB-OG1-CG2 N-CA-C-O-CB-OG1-CG2']
[1.0, 'N-CA-C-O-CB-CG-CD-CE-NZ N-CA-C-O-CB-CG-CD-CE-NZ']
[1.0, 'N-CA-C-O-CB-CG-CD N-CA-C-O-CB-CG-CD']
[1.0, 'N-CA-C-O-CB-CG-CD-0E1-NE2 N-CA-C-O-CB-CG-CD-0E1-NE2']
[1.0, 'N-CA-C-O-CB-N-CA-C-O-CB']
[1.0, 'N-CA-C-O-CB-CG-CD-0E1-0E2 N-CA-C-O-CB-CG-CD-0E1-0E2']
[1.0, 'N-CA-C-O-CB-CG-OD1-ND2 N-CA-C-O-CB-CG-OD1-ND2']
[1.0, 'N-CA-C-O-CB-CG1-CG2-CD1 N-CA-C-O-CB-CG1-CG2-CD1']
[1.0, 'N-CA-C-O-CB-CG-OD1-OD2 N-CA-C-O-CB-CG-OD1-OD2']
[1.0, 'N-CA-C-O-CB-CG-ND1-CD2-CE1-NE2 N-CA-C-O-CB-CG-ND1-CD2-CE1-NE2']
[1.0, 'N-CA-C-O-CB-CG-CD1-CD2 N-CA-C-O-CB-CG-CD1-CD2']
-----
Number of atomic chains that match in total = 1850
Percentage of Peptide Chains of Malaria match in percentage to Corona is: 99.2
-----
(base) D:\Python Projects\CoronaVirusResponse>

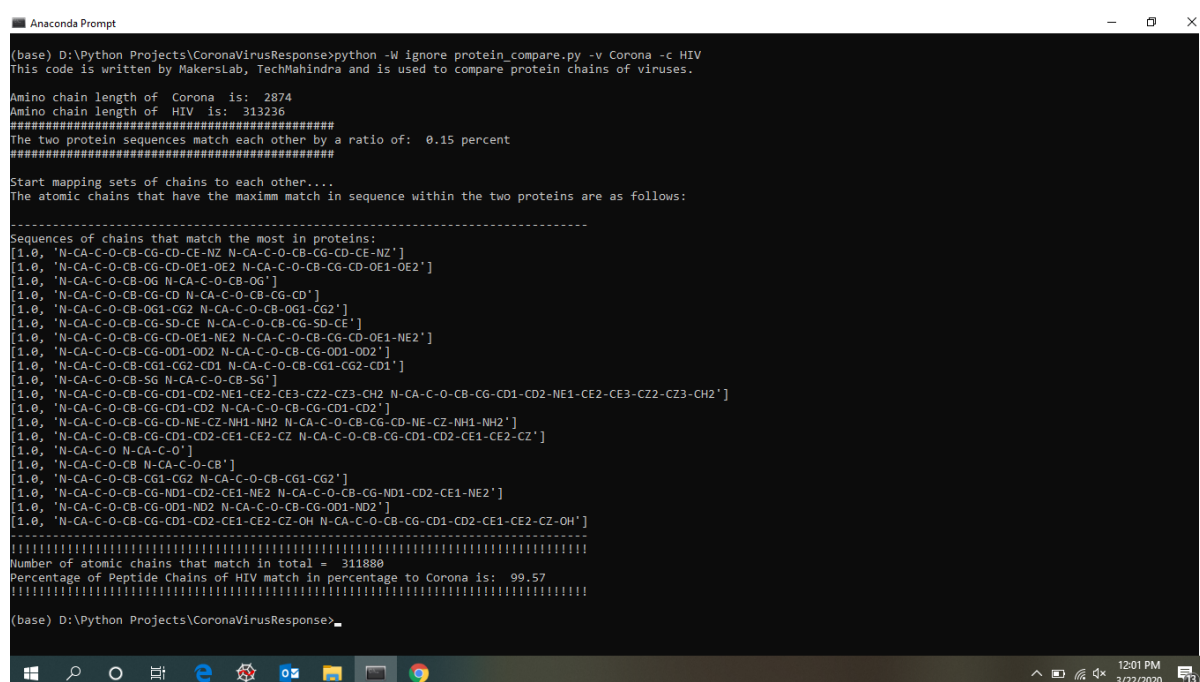
```

[Figure 4: Comparison between Malaria and Corona virus Source: Makers Lab Tech Mahindra]

Amino peptide chains when compared between Corona and Malaria give us the result as given above in figure. As shown above, following things become clear

- The sequence match of proteins in both viral proteins is only 6.94 % . This is not a surprise as there are structural dis-similarities
- The red box explains what all sets of protein chains match each other and what percentage. The ones shown give the maximum match amongst a given set in comparison
- The astonishment happened with the green box. It is seen, that 99% of Malarial protein sequence is found in Corona virus as well!!

A similar run was made against the HIV virus and the following results were obtained



```

Anaconda Prompt
(base) D:\Python Projects\CoronaVirusResponse>python -W ignore protein_compare.py -v Corona -c HIV
This code is written by MakersLab, TechMahindra and is used to compare protein chains of viruses.

Amino chain length of Corona is: 2874
Amino chain length of HIV is: 313236
#####
The two protein sequences match each other by a ratio of: 0.15 percent
#####

Start mapping sets of chains to each other....
The atomic chains that have the maxim match in sequence within the two proteins are as follows:

-----
Sequences of chains that match the most in proteins:
[1.0, 'N-CA-C-O-CB-CG-CD-CE-NZ N-CA-C-O-CB-CG-CD-CE-NZ']
[1.0, 'N-CA-C-O-CB-CG-CD-0E1-0E2 N-CA-C-O-CB-CG-CD-0E1-0E2']
[1.0, 'N-CA-C-O-CB-06 N-CA-C-O-CB-06']
[1.0, 'N-CA-C-O-CB-CG-CD N-CA-C-O-CB-CG-CD']
[1.0, 'N-CA-C-O-CB-0G1-CG2 N-CA-C-O-CB-0G1-CG2']
[1.0, 'N-CA-C-O-CB-CG-SD-CE N-CA-C-O-CB-CG-SD-CE']
[1.0, 'N-CA-C-O-CB-CG-CD-0E1-NE2 N-CA-C-O-CB-CG-CD-0E1-NE2']
[1.0, 'N-CA-C-O-CB-CG-0D1-0D2 N-CA-C-O-CB-CG-0D1-0D2']
[1.0, 'N-CA-C-O-CB-CG1-CG2-CD1 N-CA-C-O-CB-CG1-CG2-CD1']
[1.0, 'N-CA-C-O-CB-S6 N-CA-C-O-CB-S6']
[1.0, 'N-CA-C-O-CB-CG-CD1-CD2-NE1-CE2-CE3-CZ2-CZ3-CH2 N-CA-C-O-CB-CG-CD1-CD2-NE1-CE2-CE3-CZ2-CZ3-CH2']
[1.0, 'N-CA-C-O-CB-CG-CD1-CD2 N-CA-C-O-CB-CG-CD1-CD2']
[1.0, 'N-CA-C-O-CB-CG-CD-NE-CZ-NH1-NH2 N-CA-C-O-CB-CG-CD-NE-CZ-NH1-NH2']
[1.0, 'N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ']
[1.0, 'N-CA-C-O N-CA-C-O']
[1.0, 'N-CA-C-O-CB N-CA-C-O-CB']
[1.0, 'N-CA-C-O-CB-CG1-CG2 N-CA-C-O-CB-CG1-CG2']
[1.0, 'N-CA-C-O-CB-CG-ND1-CD2-CE1-NE2 N-CA-C-O-CB-CG-ND1-CD2-CE1-NE2']
[1.0, 'N-CA-C-O-CB-CG-0D1-ND2 N-CA-C-O-CB-CG-0D1-ND2']
[1.0, 'N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ-OH N-CA-C-O-CB-CG-CD1-CD2-CE1-CE2-CZ-OH']
-----
Number of atomic chains that match in total = 311880
Percentage of Peptide Chains of HIV match in percentage to Corona is: 99.57
-----

(base) D:\Python Projects\CoronaVirusResponse>

```

[Figure 5: Comparison between HIV and Corona virus Source: Makers Lab Tech Mahindra]

CONCLUSIONS

Disclaimer: These are experimental conclusions drawn numerically. These need to be validated by labs and other practitioners to see if they work

- 99.2% of malarial peptide chains are similar to Corona peptide chains which also is corroborated with some study around the world where HydroxyChloroquine , a non-toxic version of Chloroquine is known to work against the virus
- The paper and experiments are intended to give lab scientists and practitioners an easy reference guide to kick start treatment , lest an epidemic arises due to microbes and viruses

- 3) Protein shapes do account for the way a protein functions but to kickstart a vaccination trial for a new epidemic viral protein, just comparing peptide chains could be a starting point. We plan to release this research for the scientific community as a web page
 - 4) Similarity of measles peptide chains and Influenza also suggest treatments followed for measles and influenza can be tested with the virus
 - 5) A theoretical directional move could be why kids are not getting affected badly by the virus is because kids from 0-9 get healthy doses of MMR and Malaria vaccines. Kids are administered malaria vaccine from 5 to 17 months as per WHO recommendation, which might have antibodies working against the virus protein.*[Facts to be ascertained by clinical trial]*
-

Further Update: 19th April 2020

A rapid rational drug repurposing approach to stop COVID-19 virus entry into human cells as therapy

SARS-CoV-2 (COVID-19) emerged in late 2019, spread rapidly across the globe as pandemic claiming lives at unprecedented rate. Scientists are endeavoring to find various approaches to contain the spread of COVID-19 such as finding new drugs, drug repurposing of older drugs approved for other diseases, vaccines and use of antibodies.

The objective of our work is to antagonize the entry of virus into human host cells such as lung airway epithelial cells. The choice of this strategy is important because the high transmission rates of COVID-19 is attributed partly to the high affinity binding and entry of the virus into host cells. The virus uses its spike protein to bind to a receptor on human cells ACE-2 or angiotensin converting enzyme 2 to enter the cells. Once the virus cannot enter the host cell, it is harmless.

The uniqueness of our work stems from identifying a peptide sequence specific to COVID-19 spike protein which is used for virus entry. This peptide sequence was identified through bioinformatics tool approaches. The classical new drug discovery route for antagonizing the activity against SARS-CoV-2 is long and high risk. Therefore, rational computational and rapid approaches for drug-repurposing such as ours are needed. The present work from our team will elucidate the repurposing of FDA approved anti-viral drugs, drugs approved for various diseases, and even some of the natural products/GRAS compounds approved by FDA.

Our peptide sequence will be used as the guiding principle to find only those drugs that are selective for COVID-19. The repurposed drugs will be computationally screened, prioritized, and rank-ordered - followed by in vitro evaluation of the short-listed drugs. Our structure-based approaches will use molecular docking and will not only give a greater visual understanding of a protein/peptide interacting with a small molecule ligand/drugs but also provide rank-ordering of the FDA approved drugs.

We differ from other approaches for finding COVID-19 drugs in three ways:

1. Targeting the entry of virus into host cells as a therapy
2. Our focus on modeling around the unique peptide to find repurposed drugs rationally, and
3. Using the drug-repurposing strategy – rather than focusing on new drug discovery (NCE approach)

We believe that our targeted and rational approach may accelerate the timeline of bringing a drug to the patient.

This research has been divided into three steps as follows:

1) Modelling the atomic structure of the virus proteins:

We understand that the proteins work based on the structure but instead of using a template based modelling, we decided to undergo a study of the protein cloud.

2) Protein atomic structure comparison of various viruses:

This comparison of structural similarity is done by obtaining PDB format structures and comparing the structure of atoms. This comparison is basically a structural formatting. This protein cloud comparison can be found at our site. The protein cloud is formed using the PDB X,Y,Z Cartesian coordinates and taking the same structure to be displayed in 3d using tools like plotly and matplotlib lib.

This comparison of atomic structures can be found on our website:

https://entellio3.techmahindra.com/covid19_research/

While doing a sequence alignment of Corona with other proteins, we found that the alignment scores are highest for POLIO virus followed by Malaria. These results also match with a well renowned scientist Dr. Robert Gallow - a world renowned virologist who also helped discover the HIV virus has a similar thought.

<https://www.pbs.org/wnet/amanpour-and-company/video/can-an-oral-polio-vaccine-help-stop-the-coronavirus/>

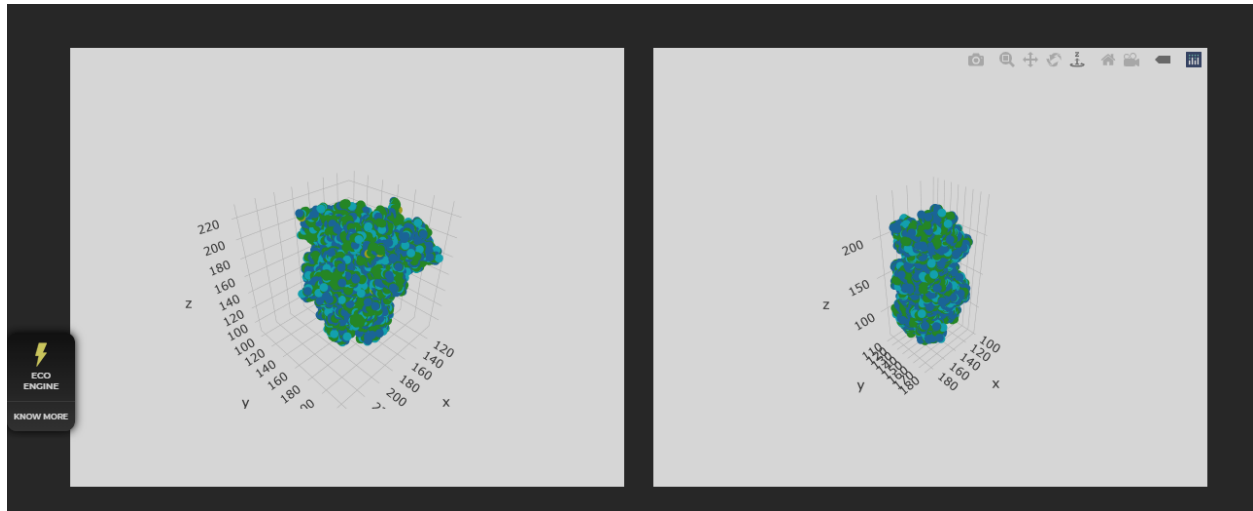
3) Therapeutic drug research and predictive virulence:

We are in the process of using convolutional neural network techniques to find the ligand match for therapeutic drug. Initially we plan to test this with the existing FDA approved drugs and would like to extend to indigenous/traditional drugs for immunity boosting as well. This application is planned to be live by end of May 2020.

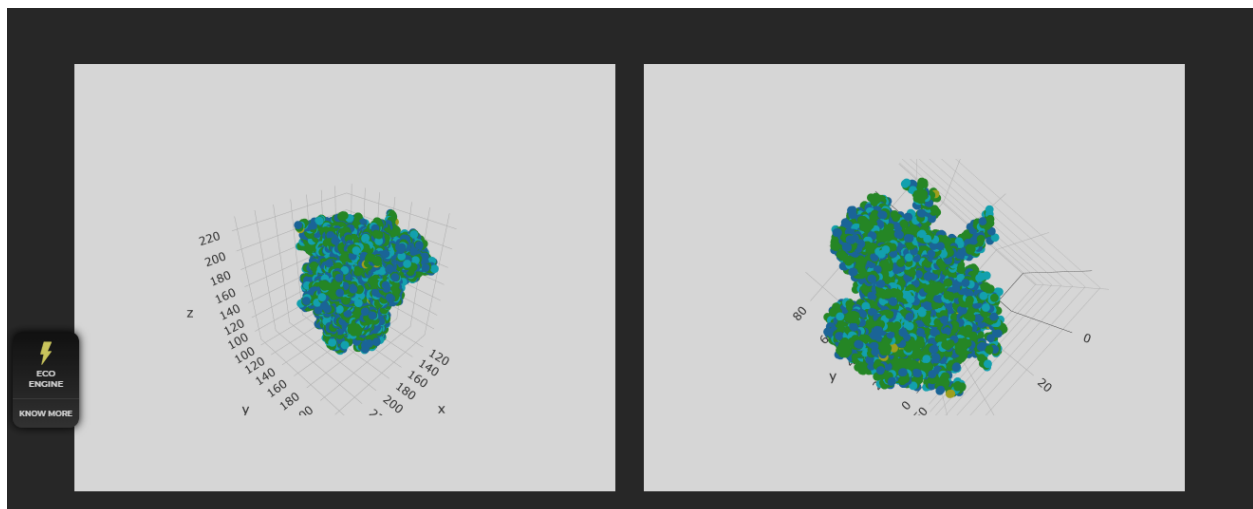
WORK DONE SO FAR

We have covered the first two steps of the research so far. Here are the findings.

Here I have shown a simple protein match structure and atomic structure comparison of two proteins Corona and Malaria



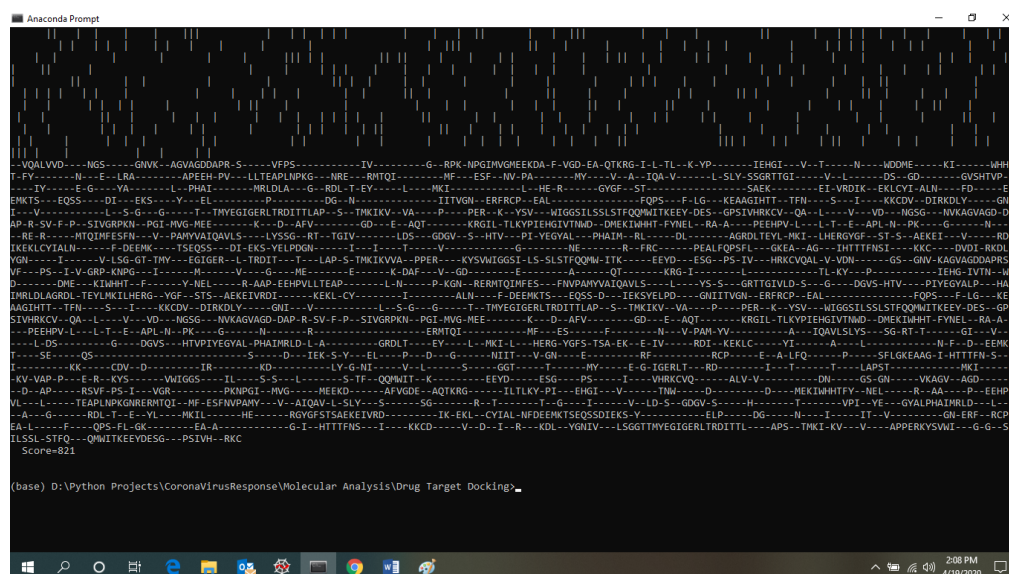
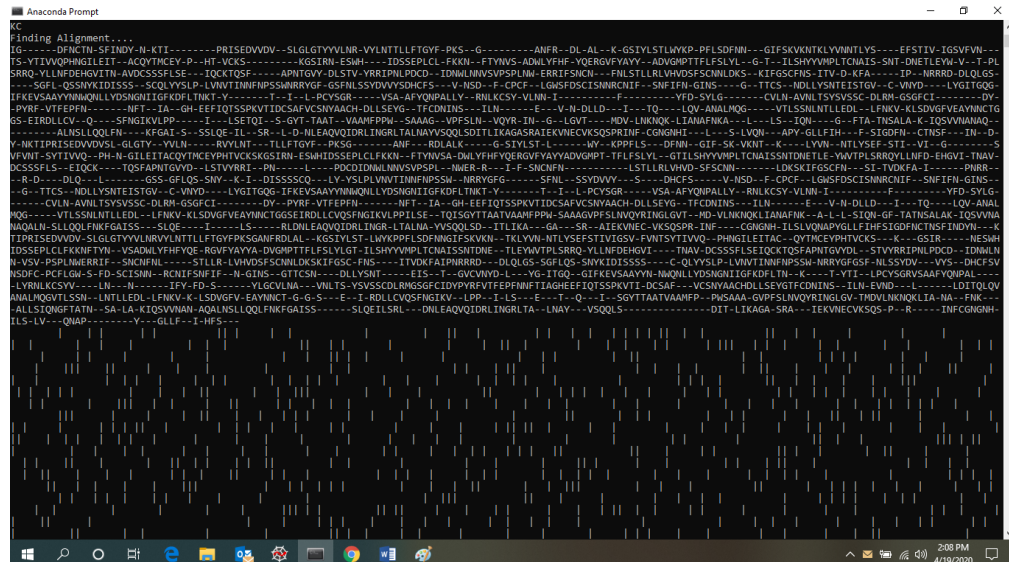
The program also compares the protein atomic sequences and matches them. Given below is also a structure match of Corona virus with Polio



Some interesting idea begins to take shape as we see how the Polio virus structure resembles the Corona virus in a significant way. The other approach we took was how we compare polypeptide sequences with different protein structures together and not just the atomic comparison of sorts

Given below are scores comparison of the peptide chains and residues between various proteins of different viruses. This alignment is the sequence alignment match between various virus proteins

COMPARISON OF CORONA WITH MALARIA: SCORE ALIGNMENT MATCH : 821



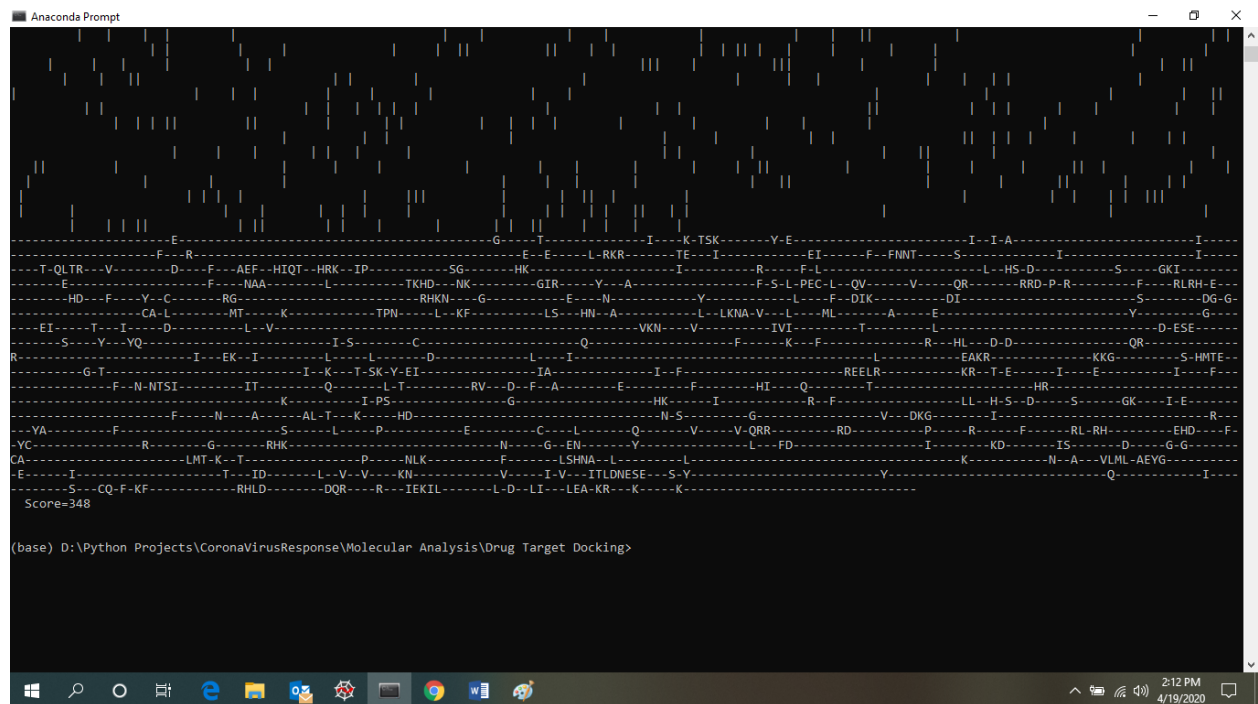
```

Anconda Prompt
MRCIG-----I--SN-----R--D-FV-EGVG-G-G-SWV-----DIV-----L--EHG-----S-----CVTTMA-K-----NK-P-L-DF-----ELI-----ETEAK-----QPATLRKY-CIE-AKLT-----N--T-----
T--T--D-SRSC--P-T-----QGEPS--LNE-----E-----QDK-RF-----VCKHS-MVD-----RG-----WG-----NGCGLF-----G-KGGI-----V--TC-A-----MFT-CKKNMKGKVQPE-NLEY-TIIVTP-HS-----GEE
H--K--L--HMF-----TGKHGKEI--KITPOS-SITEA--ELTG-Y-----GTV-----TMEC-----SP--R-----T--GL--D-F--NEMVL-----LQMEN-----K-A-----WLVRHQ-----WFL-----
D-----LPL-----P--W-----L-----P--G--AD-----TGQS--N-WI--QKETLVT--FKNPHAKKQD-----V-V--VL--GSQ-----E--GAMHTA-----
L--L--T-GATEI--QMS-----GN--LL--FTGHLKC--RL-----RMD-----KL--QL-KGMSY-----S--M--C-----TGKFKVVKVE--IA--E--TQHTG--IV-T--RV--QY-----E-G
D--GSPCKPIPEIMDL-EKRHV--LGRLL-IT--V-N-----PIV-----T--E-----K-D-----S-----PV--N-----T--E-----A-EP-----PFGDS--Y-T--ITG--V--EPGQ-L
K--K--L--HMF-----LITG--V--T--TIT-----V--WIG-MW-----S--RST-----LS-V--SL-V--L-----V-G-V-V-T-LY--LGVWVQASVALVPH-VG-MGLE-TA--T
FS-G-----VS-NIT--MK--T-----LITG--V--T--TIT-----V--WIG-MW-----S--RST-----LS-V--SL-V--L-----V-G-V-V-T-LY--LGVWVQASVALVPH-VG-MGLE-TA--T
E--T--T--MMS--S--E--GAMKHAQRI--E-----T--WIL-----RHPG-F-TIMA--A--ILAYTIG--TTHFQALIF-----IL-----LGT-A-----V-V--AP--MR-----CIG-I-SN--RD--R--F--VE--
E-----GV--S--GGSW-----DI-----V--L--E--HGS-C--VTTMAKN--KPT--L--D-F--EL-----TET-----EA-----KQ-----PATL-----RKYCIEAKLTNTTID--SRC--P--
T--QGEPS-S-LN-EQDKR-----F-----V--C--KHSW--D--RGM-----G-N-G-C-----G-----L-F-G--KG-GI--V--TCA-----MF-TCKKNMK
G-KV-----VQ-P--ENL--EYTV--I--TPHS--G--EEHA-----V-----G-----N-DT--GKH-GKE-I--K-I--TPQS-----S--ITEA-----ELTGYGT-VTMEC--RV--PRTGLDFNEMV--L
L--LQ--MENKAWL--V-----HRQWFL-DLPL--PWL--PG--A--DT-QGS-NWI--QKETL-V--TFKN--PHAKKQDV--VVLGS--Q--EG--AMHT-ALTGATEIQM--S--SG--NLFL-TGHL--KCR--L--RM
D--K-LQ-L--K--GMS--YSMCTGKF--KVKKEIAETQ-HGTIVTRV--Q--YEGDGS--PCK--I--PFEEI--MD-LE--KRHVL--GRLLT--V--NPIVT-----EKDSPVNIE--A
EP--PF--G--DS-YIIGV--EP-GQLKL--NMFKKGSSTG--QMI--ET--TMRGAKRMAI--LGD--TAWD-FGSLG-G--V--FT--SIG--KAL-HQVGF-AI--Y-----G-AAFS-GV--
SWIMKL-IGV-IITWIGMNSRSTLS-S-VS--LV-LVG-V-V-T-LY--LGVWVQASVALVPH-VG-MGLE-TA--T--E--T--MMS--E--GAMKHAQRI--E-----T--WIL-----RHPG-F-TIMA--A--ILAYTIG--TT
HFORALIF-IL-L-T--AV-AP--MRC--I-TGISN-RD--FVE--GV--S--GGSWVDIVLHSGSCV-T-TMA-KNKTPLD--F--ELT--ET--EAKQD-AT--L--R-----KYC-I--EAKL-N-TTDS-RCP-TQGEPS
SLN-E-EQDKR-----F-----VCKHSMD--RGWNGCGL--FG-KGIVTICAF--T--CK--KN-----MKG-----V-V--Q-----ENL--EY--TIV--ITPH--SGEEHA
VH-D-----TG-----KHKKET--MT-TPOS--SI--TEALITG--VGVITMECS--PRTGL--DFH-E--MW--L--LQ--ME-----NKAILVHRQ-WFLD-D-PHLPDQ--TQSSW-VI--QK--ETLVTGKNPHAK--K
QDVVVL--G--SOGAMHTA--LGRATEQMSGSL--L--FTGHLKC--RL--RMD--KL--QL-KGMSY--E--M--C-----TGKFKVVKVE--IA--E--TQHTG--IV-T--RV--QY-----E-G
D--GSPCKPIPEIMDL-EKRHV--LGTI-IT--V-N-----PIV-----T--E-----K-D--D--SPVNIEA-----E--PPF-GDSY-IIGV--EP-GQLKLNMFKKGS--SIG-QMIETTMRGAKRMAIGDT--A--W--D--
D-FGSL--G--GVFT-----STGKALHQVGFALYGA-AF-SGVSWMK--IL-I--G--VITTIQWMS-RSTLS--S-V--SLVL--VGVVTL-----YLVGMVQA--SVAL-----VPHVGMGLTATETWMS-----SEGA--MKAHQ--
R-I--ETWILRHGPGFIIMAAI-AY-----T-I--G-----TTHFQ--RALI-F-----IL-LTAV--AP-----
Score=786
(base) D:\Python Projects\CoronaVirusResponse\Molecular Analysis\Drug Target Docking

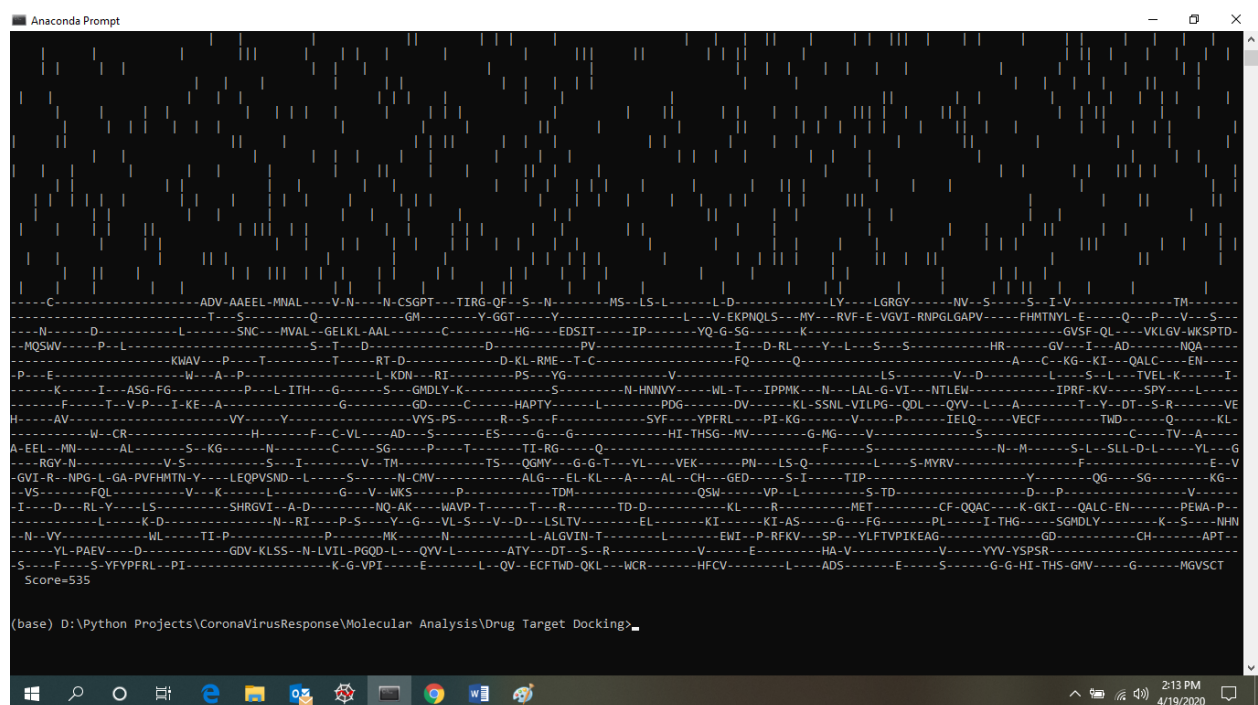
```

[illegible]

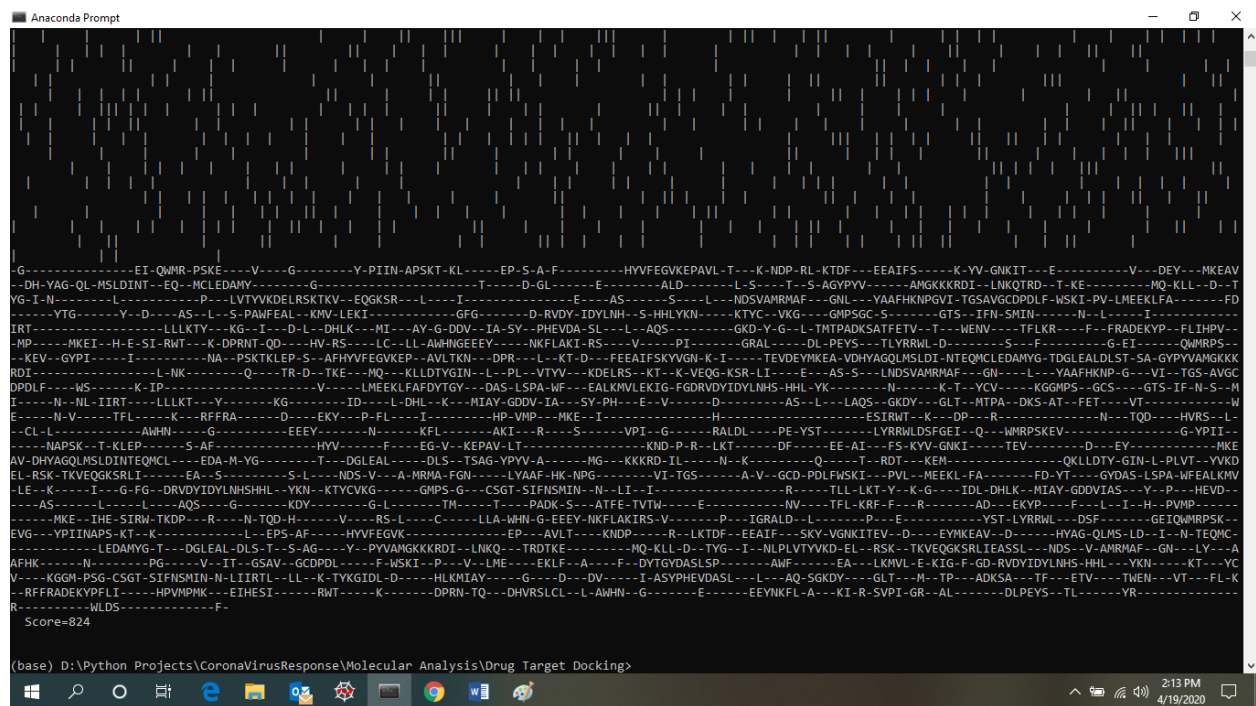
COMPARISON OF CORONA WITH PENUMONIA: SCORE ALIGNMENT MATCH: 348



COMPARISON OF CORONA WITH MEASLES: SCORE ALIGNMENT MATCH: **535**



COMPARISON OF CORONA WITH POLIO: SCORE ALIGNMENT MATCH: 824



This could be an interesting find. The peptide chains and their sequence alignment match with malaria (821) and the maximum with Polio (824) which we wish to explore to find possible Ligands for the attack.

EXTENSION OF WORK

We plan to extend this work by using convolutional 3d networks to see protein ligand interaction with active sites as an attack plan.

We also have a provisional patent approved for Rapid Rational Identification of Repurposed FDA Approved Drugs AND NCE'S as therapeutics for COVID-19 using a specific amino acid sequence of VIRUS surface spike protein (No. TEMP/E1/18551/2020CH)

Watch this space as we publish further results of our research