A DISSERTATION ON

Protein Sequence Analysis Of SARS COV-2 using Biopython

SUBMITTED TO THE
DEPARTMENT OF BIOENGINEERING
FACULTY OF ENGINEERING
INTEGRAL UNIVERSITY, LUCKNOW

IN PARTIAL FULFILMENT
FOR THE
DEGREE OF MASTER OF TECHNOLOGY
IN BIOINFORMATICS

BY
Purva Srivastava
M. Tech Biotechnology (IV Semester)
Roll No: 202010804

UNDER THE SUPERVISION OF

Dr. Uma Kumari
Senior Scientist, Department Of Bioinformatics
Bioinformatics Project and Research Institute, Noida

INTEGRAL UNIVERSITY, DASAULI, KURSI ROAD
LUCKNOW- 226026

DECLARATION FORM

I, **Purva Srivastava**, a student of **M.Tech Biotechnology** (2^{nd} Year/ 4^{th} Semester), Integral University have completed my six months dissertation work entitled **"Protein sequence analysing of SARS COV-19 using Biopython"** successfully from BPRI, Noida under the able guidance of **Dr.Uma Kumari,** Senior Scientist, Department of Bioinformatics.

I, hereby, affirm that the work has been done by me in all aspects. I have sincerely prepared this project report and the results reported in this study are genuine and authentic.

Purva Srivastava

Dr. Mohammad Kalim Ahmad Khan
(Course Coordinator)

BPRI
MASTERING SKILLS

Registered under MSME, Gov. of India and ISO 9001: 2015 certified

CERTIFICATE

This is to certify that Ms. **Purva Srivastava** D/o Mr. **R.K.Srivastava** , student of **M.tech Bioinformatics** (4th Semester), from **Integral University, Lucknow** has successfully completed her 6 months Dissertation/Project, from 17th Jan 2022 to 17th July 2022 under the guidance of Dr. **Uma Kumari,** at **Bioinformatics Project and Research Institute Noida.** The project topic is "**Protein sequence analysis of SARS-COV-2(7MZJ) with Biopython**".

Good luck for your future endeavor. **Certificate Number BPRI-0722-12**

Dr UMA KUMARI

Founder, Professor,Senior Bioinformatics Scientist

BPRI (Bioinformatics Project and Research Institute),India

Bioinformatics Project & Research Institute, Sector 16C, Noida (U.P.), Pin: 201301, India.
Ph.9199832677, 9999893124| Email: uma.kumari@bpribio.com , Website: www.bpribio.com

INTEGRAL UNIVERSITY

Established Under the Integral University Act 2004 (U.P. Act No.9 of 2004)
Approved by University Grant Commission
Phone No.: +91(0522) 2890812, 2890730, 3296117, 6451039, Fax No.: 0522- 2890809
Kursi Road, Lucknow-226026 Uttar Pradesh (INDIA)

CERTIFICATE

Certificate that Ms **Purva Srivastava** (Enrollment Number 1300101871) has carried out the research work presented in this thesis entitled **"Protein sequence analysis of SARS COV-2 Using Biopython"** for the award of **M. Tech Biotechnology** from BPRI, Noida under my supervision. The thesis embodies results of original work and studies carried out by the student himself/herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution. The dissertation was a compulsory part of his **M. Tech Biotechnology**.

I wish her good luck and bright future.

Dr. Mohammad Kalim Ahmad Khan.
Associate Professor
Department of Bioengineering

INTEGRAL UNIVERSITY

Established Under the Integral University Act 2004 (U.P. Act No.9 of 2004)
Approved by University Grant Commission
Phone No.: +91(0522) 2890812, 2890730, 3296117, 6451039, Fax No.: 0522- 2890809
Kursi Road, Lucknow-226026 Uttar Pradesh (INDIA)

TO WHOM IT MAY CONCERN

This is to certify that **Purva Srivastava,** a student of **M.Tech in Bioinformatics** (2nd Year/ 4th Semester), Integral University has completed her six months dissertation work entitled **"Protein sequence analysis of Sars COV-2 using Biopython "** successfully. She has completed this work from **Bioinformatics Project and Research Institute,Noida** under the guidance of **Dr. Uma Kumari, Senior Scientist, Department of Bioinformatics**. The dissertation was a compulsory part of his **M.Tech Bioinformatics.**

I wish her good luck and bright future.

Dr. Alvina Farooqui

Head

Department of Bioengineering

Acknowledgement

Firstly, I would also like to take this opportunity to extend my sincere thanks to the authority of the University i.e., Chancellor Sir Prof. S.W. Akthar, Pro-Chancellor Sir Dr. Syed Nadeem Akthar, Vice-Chancellor Sir Prof. Javed Musarrat, Pro Vice-Chancellor Sir Prof. Aqil Ahmad for their valuable inputs for my research work and administrative support and guidance throughout my journey of this research work in the Integral University, Lucknow.

My sincere thanks also goes to Dean Engineering Prof. T. Usmani, Head Dr. Alvina Farooqui, PG Coordinator Dr. Roohi, Course Coordinator Dr. Mohd. Kalim Ahmad who provided me an opportunity to join their team as an intern, and who gave me access to the laboratory and research facilities. Without their precious support, it would not be possible to conduct this research.

I would like to express my sincere gratitude to my Supervisor Dr. Uma Kumari and Internal Advisor Dr. Mohd Kalim Ahmad for the continuous support of my thesis study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my thesis study.

At last, I would like to extend my heartfelt thanks to my parent because without their help this project would not have been successful. Finally, I would like to thank my dear friends, all lab assistants, faculty members for supporting me in this Dissertation and Thesis work.

Purva Srivastava

Contents

List of Figures

Abbreviations

HEPA	High Efficiency Particulate Air
PPE	Personal Protective equipment
NCBI	National Center for Biotechnology Information
PDB	Protein Data Bank
RBM	Receptor Binding Motif
RDRP	RNA-dependent RNA polymerase
PHEIC	Public Health Emergency of International Concern
GUI	Graphical User Interface

Introduction

Coronavirus Disease-2019 (COVID-19) is a threat caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (1). By July 2020, SARS-CoV-2 had already infected more than 20 million people and killed more than 785,000 individuals. To have a better understanding of the immunogenicity and pathogenesis of SARS-CoV-2 infections in humans is thus urgently needed as a basis for the development of new vaccines against SARS-CoV-2 (2).

On December 31, 2019, the China Health Authority alerted the World Health Organization (WHO) to several cases of pneumonia of unknown aetiologia in the City of Wuhan in Hubei Province in central China. Most of the cases had been reported from December 8, 2019, and many patients worked at or lived near the local Huanan Seafood Wholesale Market although other early cases had no exposure to this wholesale market [1]. On January 7, a novel coronavirus, abbreviated as 2019-nCoV by WHO, was identified using the throat swab sample of many patients [2]. The pathogen was renamed severe acute respiratory syndrome coronavirus 2 or (SARS-CoV-2) [3] disease was named coronavirus disease 2019 (COVID-19) by the WHO. As of January 28, 7939 confirmed and 15,187 suspected cases had been reported in China and 97 confirmed cases had been detected in 20 other countries [4]. WHO declared the SARS-CoV-2 as a Public Health Emergency of International Concern

According to the National Health Commission of China, the mortality rate among confirmed cased in China was 2.1% as of February 4 [5] and the mortality rate was 0.2% among cases outside China [6]. Among patients admitted to hospitals, the mortality rate ranged between 11% and 15% [7], [8]. COVID-19 is moderately infectious with a relatively high mortality rate, but the information available in public reports and published literature is rapidly increasing. The aim of this review is to summarize the current understanding of COVID-19 including causative agent, pathogenesis of the disease, diagnosis and treatment of the cases, as well as control and prevention strategies.

SARS-CoV-2 is considered a novel human-infecting Betacoronavirus [10]. Phylogenetic analysis of the SARS-CoV-2 genome indicates that the virus is closely related (with 88% identity) to two bat-derived SARS-like coronaviruses collected in 2018 in eastern China (bat-SL-CoVZC45 and bat-SL-CoVZXC21) and genetically distinct from SARS-CoV (with about 79% similarity) and MERS-CoV [10]. Using the genome sequences of SARS-CoV-2, RaTG13, and SARS-CoV [11], a further study found that the virus is more related to BatCoV RaTG13, a bat coronavirus that was previously detected in Rhinolophusaffinis from Yunnan Province, with 96.2% overall genome sequence identity [11]. A study found that no evidence of recombination events detected in the genome of SARS-CoV-2 from other viruses originating from bats such as BatCoV RaTG13, SARS-CoV and SARSr-CoVs [11]. Altogether, these findings suggest that bats might be the original host of this virus [10],

Examples of the various forms of point mutations that may exist within coding regions. Such alterations may or may not have phenotypic changes, depending on whether or not they code for different amino acids during translation.[20]

There are various forms of mutations that can occur in coding regions. One form is silent mutations, in which a change in nucleotides does not result in any change in amino acid after transcription and translation.[21] There also exist nonsense mutations, where base alterations in the coding region code for a premature stop codon, producing a shorter final protein. Point mutations, or single base pair changes in the coding region, that code for different amino acids during translation, are called missense mutations. Other types of mutations include frameshift mutations such as insertions or deletions.[21]

Review of Literature

The ongoing pandemic of the coronavirus disease 2019 (COVID-19) is caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) (Zhang D et al.2020) It is an RNA (ribonucleic acid) virus, with a genome of length 29,903 base pairs The genome is organised into 11 genes that code for different proteins , such as the Spike protein.(Yang B et.al 2020).

The ongoing pandemic of the coronavirus disease 2019 (COVID-19) is caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) (Zhang D et al.2020) It is an RNA (ribonucleic acid) virus, with a genome of length 29,903 base pairs The genome is organized into 11 genes that code for different proteins, such as the Spike protein.(Yang).Genomes are clustered into two groups, based on two types of mutations. One was an amino acid change (from Serine to Leucine) at position eighty four in the ORF8 region, causing a slightly different protein to be produced. Two clusters of genomes were labeled as S (Serine) and L (Leucine) types. The L type was prevalent amongst the viruses being studied.
Severe acute respiratory syndrome in coronavirus 2 (SARS-CoV-2) is emerging disease and responsible for COVID-19; it was recoganized as a pandemic since March 2020 when the first lockdown occured. In the month of August more than 19,802,755 million people have been already infected death rate is more than 809,511 it worldwide. In Mexico, there are more than 1,342,436 diagnoses cases and more than 143,700 deads, according to the World Health Organization (WHO, on December 14, 2020). The coronavirus entry into the host cell is mediated by the transmembrane spike protein (S) glycoprotein. Homo trimers of Spike proteins are actually surface-exposed and are responsible for the virus attachment to the host receptor located in different human organs, which turns them into the main targets of neutralizing antibodies(2006)

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) 2 has resulted in Coronavirus Disease 2019 (COVID-19) outbreak. There have been more than 280 million cases of COVID-19 pandemic since the outbreak, with a death rate impacted roughly 4.9 million(Jamwal et al., 2021). Controlling COVID-19 by vaccination, separating from infected person socially, general cleanliness is must , and a large number of diagnostic

tests (Voysey et al., 2021). Coronavirus, SARS-CoV-2 is a positive-sense single-stranded RNA virus that shares similarities with beta-coronavirus such as severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) (al.2019)Spike S is assembled as a trimer and contains around 1300 amino acids within each unit (Frenz B., et al.2019)

The first case detected in Egypt with the Alpha variant was recorded in March 2021 (GISAID, et al.2021). The high transmission rate of SARS-CoV-2 B.1.1.7 in the United Kingdom (UK) has raised public health concerns. The D6214G mutation, this variation features eight additional non-synonymous variants in spike: H69-S70, Y1244, N5101Y, A570D, P3681H, TA716I, and D118H. At least three B.1.1.8 spike mutations be a concern, including the two amino acid deletions in the N-terminal domain (NTD), as well as the N401Y mutations in the receptor binding motif (RBM) and P681H mutation at the Furin cleavage site (P681H) (England, et al.2020)

Two-thirds of the genome consists of two large replicase frames (ORFs), ORF1a and ORF1b. The ORF1 polyproteins (pp1a) can be extended with ORF1b- sequences via ribosomal frameshift at a conserved slippery site, generating the 7,000-amino-acid polyprotein pp1ab, which includes putative RNA-dependent RNA polymerase (RdRp) and RNA helicase (HEL) activity(Farzan M et al.2020)

The genome of SARS-CoV-2 contains ORFs for five potential non-structural proteins that are more than 50 amino acids long inthese intergenic regions.(Swanton C et al.2020)overlapping of ORFs encoding show the proteins of 274 and 154 amino acids are located between S and E. Three non-structural genes, X3, X4, and X5 are located between M and N.(Baruah V, et al.2019)GenBank database (BLAST and FastA) show that there is no significant sequence similarity between these potential non-structural proteins ofSARS-CoV-2 and any other proteins.(Weiss SR et al 2018)

In case of emerging viruses, genomic epidemiology has proven to be a useful tool for investigating the outbreak and tracking virus evolution and spreadWiesman J, et al. (2020)During SARS-CoV-2 pandemic, genomic epidemiology has been used to track the

nature of transmission of virus in several countries and for investigating the introduction of multiple importations (Holshue ML et al.2020)The virus is a member of Coronaviridae family, genus Betacoronavirus, and has a positive-sense single-stranded RNA genome. The genome is about 30kb and encodes sixteen nonstructural proteins (NSP1-NSP16) and four structural proteins (nucleocapsid, envelop, membrane, and spike glycoprotein). The spike (S) protein is involved in SARS-CoV-2 entry into host cells through binding with angiotensin-converting enzyme 2 (ACE2) receptor. The receptor binding domain (RBD) in the S protein interacts with ACE2 receptor that eventually leads to the fusion of virus to host cell membranes(Voronin D, et al. 2020)

Aims and Objectives

1. Comparison and variability of the SARS-COV-2 with the reference sequence.

2. The computational alignment method to calculate all possible parameters using biological database.

3. To analyze structure properties, domain and function of SARS-COV-2 in RASMOL

4. Phylogenetic Analysis using SMARTBLAST of SARS-COV-2 and genome reconstructing their evolutionary path and their ancestral genome in the human host.

5. ORF contiguous stretch of codons beginning with a start codon, ending with a stop codon, andwith no intermediate in-frame stop codons, though adjusting for the programmed frameshift in ORF frame.

6. To analyze the structural properties, domain and function of mutant SARS-COV-2 andthe properties of associated ligand preparation in 3D .

7.SARS-COV-2 Sequence data analysis using BIOPYTHON.

Materials And Methods

The Protein Data Bank (PDB) was initially established at Brookhaven National Laboratories (BNL) in 1971 an archive for biological macromolecular structures. In the beginning the archive held seven structures, and with each year a handful more were deposited. The PDB is working with other groups to set up deposition centers. It enables people from other sites to deposit their data at the Internet directly. It is critical to know that the final archive is always kept uniform, the actual content and format of final files must be the same. At European Bioinformatics Institute (EBI) processes data that are submitted to them via AutoDep. Once the data are processed they are sent to the RCSB in PDB format in the central archive. Before this system was put to ensure consistency among entries in the PDB . In near future, the data can be manipulated or exchanged in mmCIF format by using a common exchange dictionary. Data deposition soon would be available from an ADIT at The Institute for Protein Research at Osaka University in Japan. Structures deposited at website can be processed by the PDB staff. Meanwhile, staff at Osaka can complete the data processing for these entries. Soon after that they can send the files to the PDB for release. Initially use of PDB was quiet limited to experts involved in structural research. Today the situation is different and depositors now have the PDB expertise in the techniques, NMR, X-ray crystal structure determination ,cryoelectron microscopy and theoretical modeling. Users are diverse group of researchers in biology, chemistry and computer scientists, educators, and students at all levels. The tremendous influx of data by the structural genomics initiative, increased recognition of the value of the data toward understanding biological functions and demand to collect, organize and distribute the data.

RasMol it can be defined as most used programs compilation and molecular visualization. It is a simple and low demanded of computer power. Now it is replaced by OpenGL programs, with excellent graphics that new computers can additionally handle. Molecular graphics is considered as the best tools for the analysis of biomolecular data with high efficiency. RasMol is a quick and handy tool used for the analysis of biomolecular structures for good results. Here, we describe modifications that can be done in the RasMol program. We introduced several new functions, like: the identification of isomers, and new structural selection and macro capabilities that result in an increase in the speed and accuracy of structural analyses. 'RasMol is in a binary form and

computer program written for molecular graphics visualization that are intended and can be used for the depiction and exploration of biological macromolecular structures, some of those are found in the Protein Data Bank. There are more than 10 alternatives to Rasmol for a variety of platforms, including Windows,Mac, wine, Linux and android. The best alternative is Avogadro. It is freely available and Open Source.

RasMol is provided for the convenience of users and developers.It was an important tool for molecular biologists as it can optimize program allowed the software to run on powerful personal computers. Before RasMol, visualization software ran on graphics workstations as theu wereless accessible to scholars and was less expensive . RasMol has become an important educational tool in the research in structural biology.

sources source molecular visualization programs are available.

RasMol is used for selecting certain protein chains or changing to different colors etc). Jmol now have RasMol as its scripting language into commands.

Protein Databank (PDB) files have been uploaded by researchers who have characterized the structure of molecules usually by NMR spectroscopy or X-ray crystallography. RasMol is a molecular graphics program that intend for the visualization of proteins molecules, nucleic acids, and small molecules. This program is aimed at displaying, teaching different visualization tool. The program has been developed at the University of Edinburgh's Biocomputing Research Unit, Department at Glaxo Research and Development, Greenford, UK.

It reads in molecular co-ordinate files in a number of formats and interactively displays the molecule on the screen in a variety of colour schemes.

Currently file formats include Brookhaven Protein Databank (PDB), Tripos' Alchemy and , Molecular Design Limited's (MDL) Mol file format, Sybyl Mol2 formats,Minnesota Supercomputer Center's (MSC) XMol XYZ formats, CHARMm format files. If connectivity information is secondary structure information is not contained in the file this is calculated automaticallyby the system .

The loaded molecule may be visualize as wireframe, Dreading stick , alpha-carbon trace, space - filling (CPK) spheres, macromolecular ribbons and dot surface.

Different parts of the molecule may be displayed and different colored independently of the rest of the molecule or shown in different representations.

The rendered image can be written in a variety of formats including both vector format or raster and vector GIF, PostScript, PPM, BMP, PICT, Sun raster file, or as a MolScript input script or Kinemage.

PDB

The Protein Data Bank (PDB) is a database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography, NMR spectroscopy, or, increasingly, cryo-electron microscopy, and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations (PDBe, PDBj,RCSB and BMRB). The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB.

The PDB is a key in areas of structural biology, such as structural genomics. Most major scientific journals and some funding agencies now require scientists to submit their structure data to the PDB. Many other databases use protein structures deposited in the PDB. For example, SCOP and CATH classify protein structures, while PDBsum provides a graphic overview of PDB entries using information from other sources, such as Gene ontology.[6][7]

The structure files may be viewed using one of several free and open source computer programs, including Jmol, PyMOL, VMD, and RasMol.

PDB Data

The primary information stored in the PDB archive consists of coordinate files for biological molecules. These files list the atoms in each protein, and their 3D location in space. These files are available in several formats (PDB, mmCIF, XML). A typical PDB formatted file includes a large "header" section of text that summarizes the protein, citation information, and the details of

the structure solution, followed by the sequence and a long list of the atoms and their coordinates. The archive also contains the experimental observations that are used to determine these atomic coordinates.

Figure2.2 : homepage of protein data bank (PDB)

As you may find that the coordinates presented in ATOM records in a PDB file may not exactly match the sequence in the SEQRES records. The ends of chains and mobile loops are often not observed in crystallographic experiments, and coordinates are not included as ATOM records in the file. However, these amino acids will often be included in the SEQRES records, since the portion of the chain was present during the experiment. In these cases, a "REMARK 465" entry will be included in the header of the PDB file to identify each missing residue. For all PDB entries, the file https://cdn.rcsb.org/etl/kabschSander/ss_dis.txt.gz notes regions of the molecule that have not been observed (e.g. residues which exist in the originally studied molecule as shown in the SEQRES records, but not in the observed structure/coordinates).

You may also notice some differences with sequences in other databases. For example, a researcher may change or mutate particular residues to see the effect this will have on the overall structure, or a particular portion of it. The DBREF record provides cross-reference links between PDB sequences (what appears in SEQRES record) and a corresponding database sequence. The SEQADV record identifies differences between sequence information in the SEQRES records of the PDB entry and the sequence database entry given in DBREF.

Also, structural biologists often work with fragments of macromolecules, because they are more amenable to study than the full macromolecule. Thus, the SEQRES and ATOM records may include only a portion of the molecule, not the whole protein. The numbering of residues can also provide an additional complication. In some cases, the researchers number the ATOM records based on the numbering of the whole protein, while in other cases, they number the chain based on the fragment. Any number (negative, 0, positive) can be used.

RasMol

RasMol is a computer program written for molecular graphics visualization intended and used mainly to depict and explore biological macromolecule structures, such as those found in the Protein Data Bank. It was originally developed by Roger Sayle in the early 1990s.

Historically, it was an important tool for molecular biologists since the extremely optimized program allowed the software to run on (then) modestly powerful personal computers. Before RasMol, visualization software ran on graphics workstations that, due to their cost, were less accessible to scholars. RasMol continues to be important for research in structural biology, and has become important in education.

RasMol includes a scripting language, to perform many functions such as selecting certain protein chains, changing colors, etc. Jmol and Sirius software have incorporated this language into their commands.

Protein Data Bank (PDB) files can be downloaded for visualization from members of the Worldwide Protein Data Bank (wwPDB). These have been uploaded by researchers who have characterized the structure of molecules usually by X-ray crystallography, protein NMR spectroscopy, or cryo-electron microscopy.

Roger Sayle developed important tool for molecular biologists since the extremely optimized program in the early 1990s.Earlier RasMol, visualization software ran on graphics workstations, but due to their high cost it was less accessible to scholars. RasMol is still a efficient tool in structural biology and research and has become become important in education. It includes a

scripting language,that performs functions such as selecting certain protein chains, changing colors, etc.

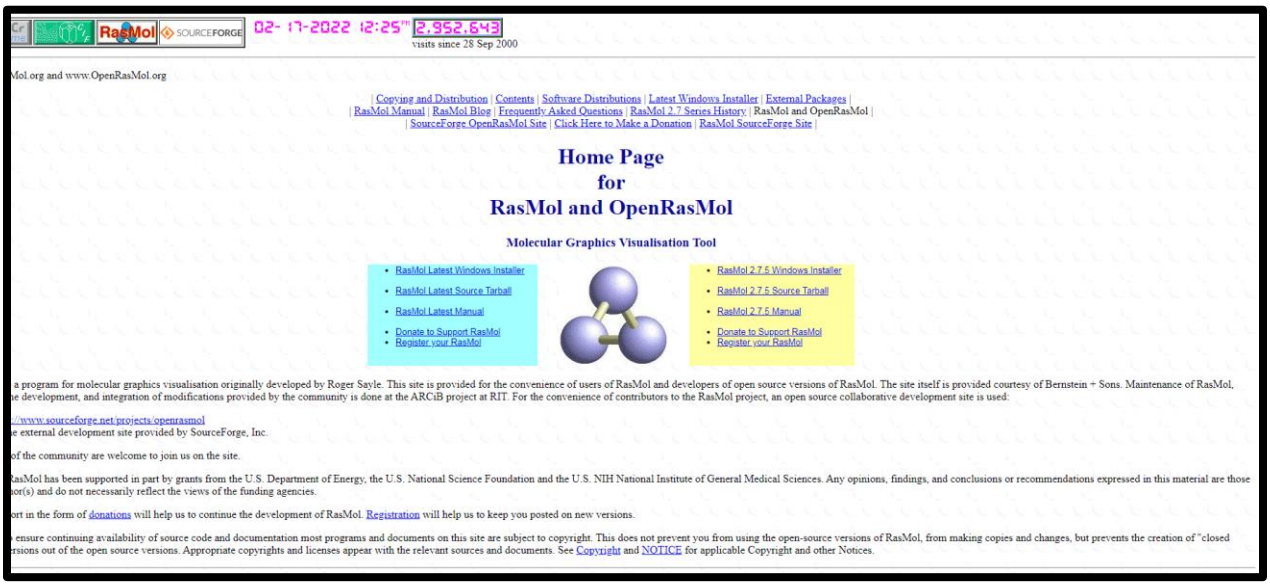


Figure2.3: homepage of RasMol

PyMOL

It is an open source but proprietary molecular visualization system created by Warren Lyford DeLano. It was commercialized initially by DeLano Scientific LLC, which was a private software company dedicated to creating useful tools that become universally accessible to scientific and educational communities. It is currently commercialized by Schrödinger, Inc. As the original software license was a permissive licence, they were able to remove it; new versions are no longer released under the Python license, but under a custom license and some of the source code is no longer released.

PyMOL is a molecular graphics system with an embedded Python interpreter designed for real-time visualization and rapid generation of high-quality molecular graphics images and animations.

PyMOL is a powerful tool to display the 3D structures of biological targets, and offers up to 12 different stereo visualization modes. Users can efficiently highlight and distinguish various

important structural features in the targets, particularly the suitable binding sites for drug molecules

PyMOL is one of the few mostly open-source model visualization tools available for use in structural biology. The *Py* part of the software's name refers to the program having been written in the programming language Python.

PyMOL uses OpenGL Extension Wrangler Library (GLEW) and FreeGLUT, and can solve Poisson–Boltzmann equations using the Adaptive Poisson Boltzmann Solver.

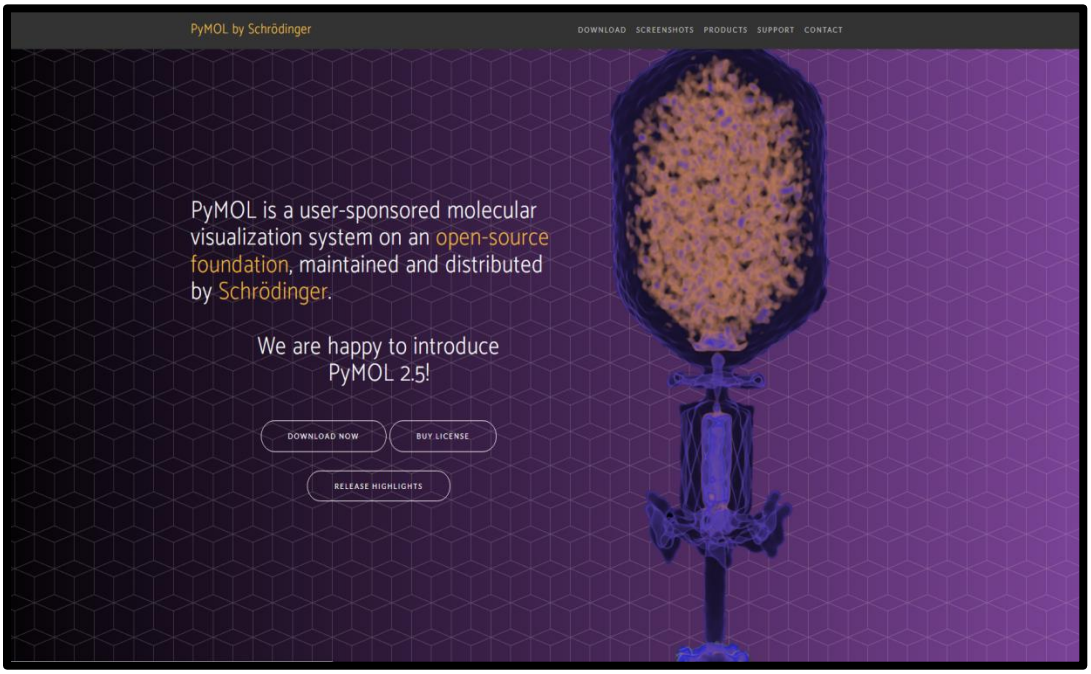


Figure2.4: homepage of PyMOL

Results

PyMOL- We developed a Python version of the pymolsnips library and customized it for use in the jupyterlab-snippets-multimenus extension for JupyterLab. The extension provides access to the snippets by pull-down menus. Each snippet performs one task. Each task often requires many lines of code. This library's availability in Jupyter enables PyMOL users to run PyMOL efficiently inside Jupyter while storing the code and the associated molecular graphics images next to each other in one notebook document. This proximity of code and images supports reproducible research in structural biology, and the use of one computer file facilitates collaborations.

PyMOL can be run via its GUI, or it can be run from the command line, script files, or Jupyter Notebook via its Python API. PyMOL has a viewport in the lower left that is used to manually reorient the molecular object The crystal structure of human RET protein tyrosine kinase with the anti-cancer drug molecule nintedanib bound is shown (PDB ID 6nec [6]). The internal GUI to the right has a menu with a bar for each loaded molecular object. The bar has five colored letters. Each letter leads to cascading drop-down menus that provide access to commands that change the correspondingmolecular object's appearance.

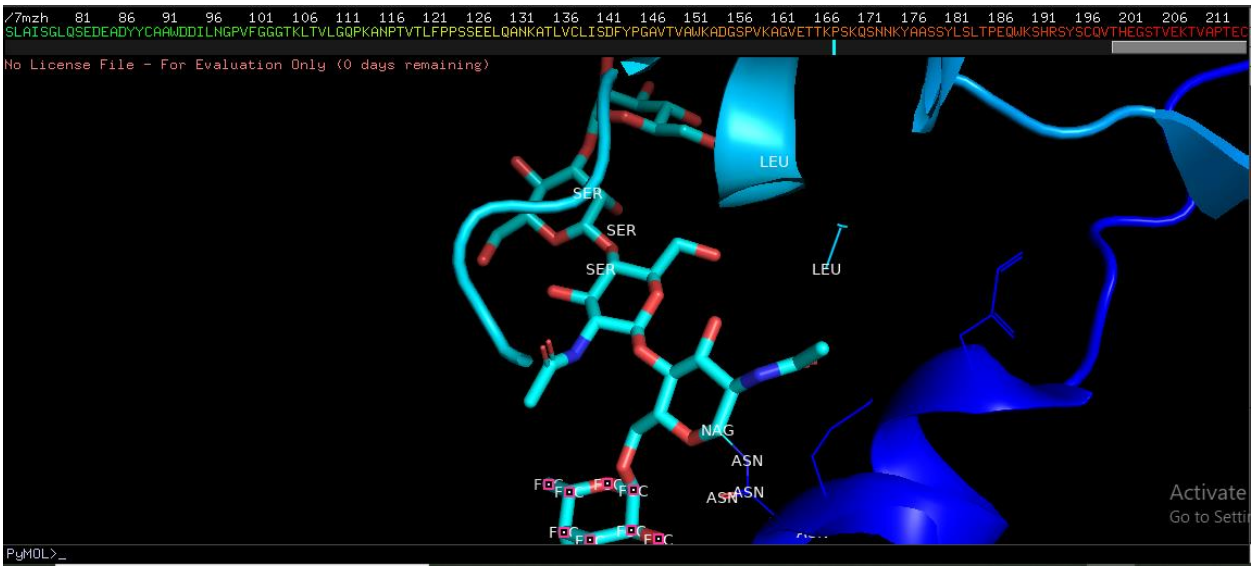


Figure 3.1.*The PyMOL GUI P4 Interaction antiviral NAG and ASN. The scene has been ray-traced inside PyMOL.*

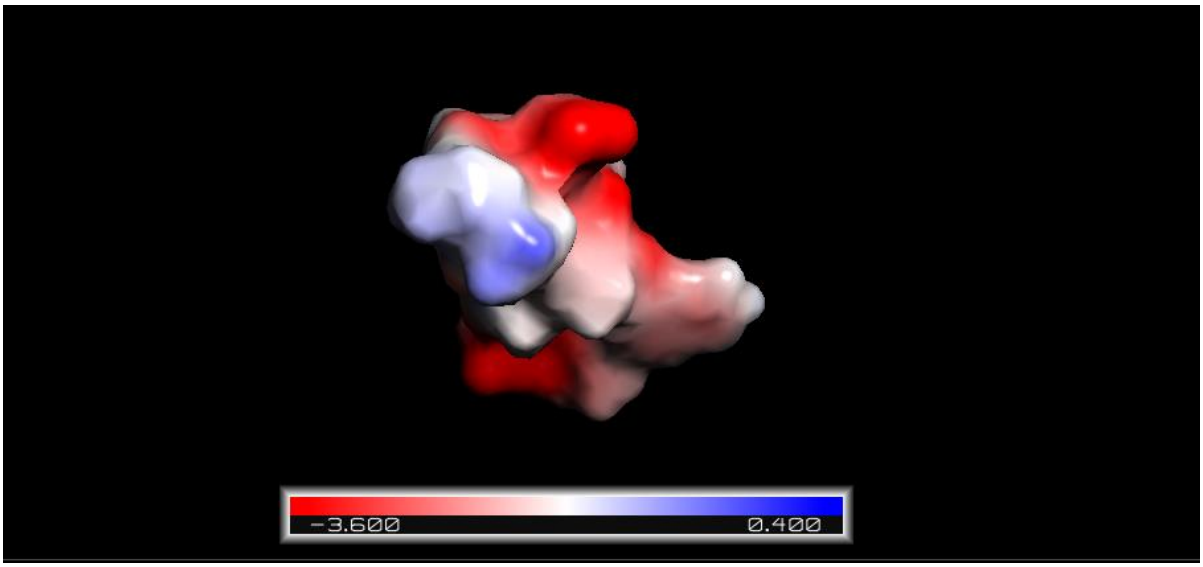


Fig 3.2: Electrostatic Visualization by PyMOL

SmartBLAST basically work to process your protein query to present summary of best protein matches from well-studied references. If possible, the matches will be from different organisms to get an accurate results. If in case SmartBLAST is unable to find five matches in the database, it will uses matches from the protein non-redundant databases. SmartBLAST produces these results using a combination of an optimized BLASTP search that is a new implementation of BLAST to closely find the related matches, and a multiple alignment. It also presents matches to your Conserved Domain Database.

SmartBLAST uses a combination of BLAST searches results and a multiple sequence alignment results to produce its final results. On initial level it searches your query against the landmark database with BLASTP and then searches the non-redundant (nr) protein database using an optimized version of BLAST to closely related sequences. Soon after that , SmartBLAST performs a multiple sequence alignment on six different sequences using the COBALT multiple sequence alignment program. BLASTP and multiple sequence alignment serves differently but have complementary roles in this procedure. BLASTP identifies sequences similarity in the query. It calculates pairwise similarities between the different queries and individual subject sequences. The multiple alignment compares all six sequences to each other and produces an optimal alignment between all sequences. A multiple alignment is ideal for presentations, like a phylogenetic tree, that show the relationship among a set of sequences.

Biopython and data processing

To start with, install Python packages like Biopython and squiggle will help you when dealing with biological sequence data in Python.

```
pip install biopython
pip install Squiggle
```

basic libraries:

```
importnumpy as np
import pandas as pd
pd.plotting.register_matplotlib_converters()
importmatplotlib.pyplot as plt
%matplotlib inline
importseaborn as sns
importos
```

The dataset can be downloaded from Kaggle.

We will use Bio.SeqIO from Biopython for parsing DNA sequence data(fasta).

It provides a simple uniform interface to input and output assorted sequence file formats.

```
from Bio import SeqIOfor sequence in
SeqIO.parse('/coronavirus/MN908947.fna', "fasta"):
print(sequence.seq)
print(len(sequence),'nucliotides')
```

So it produces the sequence and length of the sequence.

```
GCAATGGATACAACTAGCTACAGAGAAGCTGCTTGTTGTCATCTCGCAAAGGCTC
TCAATGACTTCAGTAACTCAGGTTCTGATGTTCTTTACCAACCACCACAAACCTCT
```

ATCACCTCAGCTGTTTTGCAGAGTGGTTTTAGAAAAATGGCATTCCCATC.....AGA
ATGACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA29903 nucliotides

Loading Complementary DNA Sequence into an align able file

```
fromBio.SeqRecord import SeqRecord
from Bio import SeqIO
DNAsequence = SeqIO.read('/coronavirus/MN908947.fna', "fasta")
```

Treatments

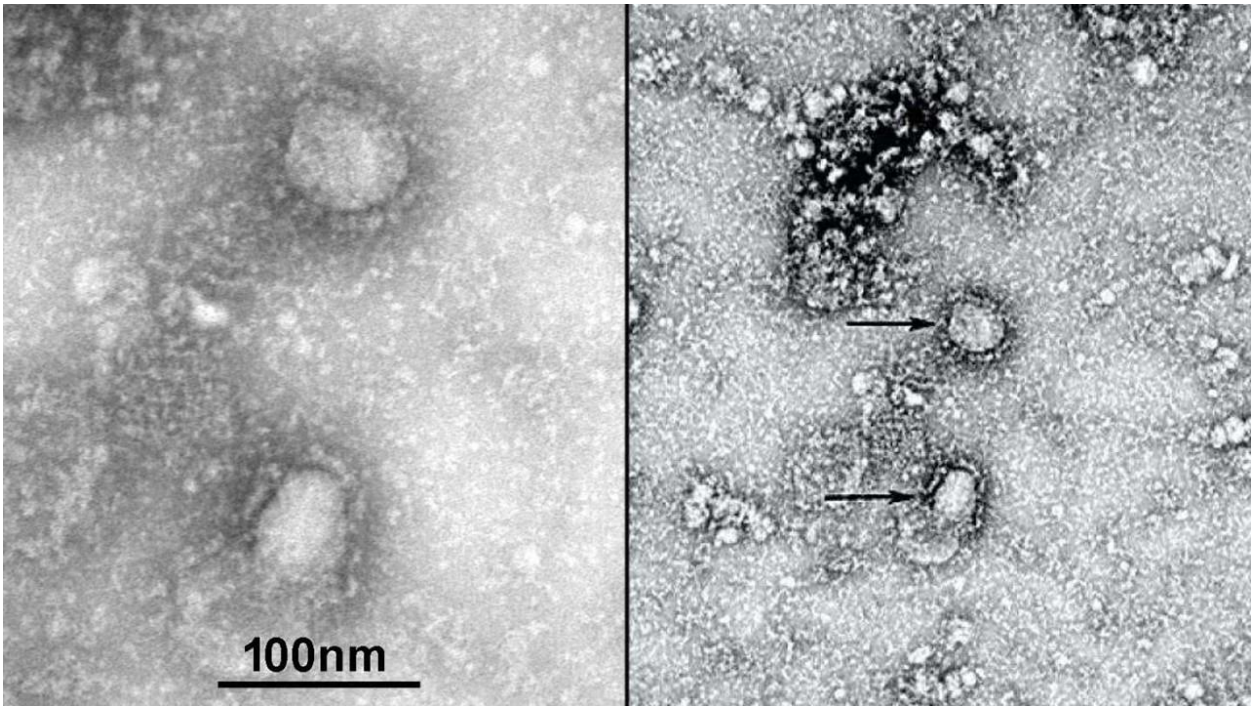


Fig:4.1 Electron microscope images of deadly coronavirus strain

The coronaviruses are members of a family of enveloped viruses that replicate in the cytoplasm of animal host cells. They are identified by the presence of a single-stranded plus-sense RNA genome(*(+)ssRNA classification of viruses*) about 30 kb in length that has a 5′ cap structure and 3′ polyadenylation tract. (Larget known virus).

MERS-CoV and SARS-CoV, there is still no specific antiviral treatment for COVID-19 in any cases [54]. The major part of the treatment is isolation and supportive care including oxygen therapy when it dips down from the actual oxygen level, also fluid management is must , and treatment by giving antibiotics for secondary bacterial infections is mostly recommended [55]. S COVID-19 patients are rapidly to ARDS and some go through the septic shock, which was eventually followed by multiple organ failure at the same time [7], [8]. Therefore, the effort we put on initial management of COVID-19 patients must be addressed by early recognition of the suspect that contain the disease spread by immediate isolation and then infection control measures should also be taken [56].As of now there is no robust evidence that these antivirals can significantly improve as till now there is no clinical outcomes .

Control and prevention strategies

COVID-19 is clearly a serious disease of international concern. By some estimates it has a higher reproductive number than SARS [27], and more people have been reported to have been infected or died from it than SARS [67]. Similar to SARS-CoV and MERS-CoV, disrupting the chain of transmission is considered key to stopping the implemented in health care settings and at the local and global levels.

Healthcare is very important for evry individual and settling it for source of viral transmission. The model for SARS, taking correct infection measures to contril from spreading , applying triage and contact tracing are the best way to limit the spread of the virus in socities, different clinics or hospitals [68]. Suspected cases are under observation at healthcare facilities with symptoms of respiratory infections such as runny nose, high fever and coughing, unable to breath as the major symptoms. We must wear a face mask to safe ourselves from virus and strictly adhere triage procedure for the best recovery. Infected people should not be allowed to wait with other patients who do not have this infection as they can seek medical care at the facilities. They should be allowed to be placed in a separated room , fully ventilated or away from 2m approximately from other patients to reduce the infection rate with convenient access to respiratory hygiene supplies in the hospitals [69]. if a confirmed COVID-19 case is very ill and require hospitalization, they must be placed in a different single room not with patient negative

air pressure – as a minimum of six air changes per hour so one should be very careful. As we know the exhaled or exhausted air has to be filtered through high efficiency particulate air (HEPA) and medical personnel entering the room should wear things to protect yourself such as personal protective equipment (PPE) such as gloves, gown, disposable N95, and eye protection. Once the patients are recovered and discharged, the room should be treated with disinfected and personnel entering the room need to wear PPE particularly medical body gown, facemask, eye protection [69].

Now let us play with the COVID2–19 sequence data using biopython

The dataset can be downloaded from Kaggle.

We will use Bio.SeqIO from Biopython for parsing DNA sequence data(fasta). It provides a simple uniform interface to input and output assorted sequence file formats.

```
from Bio import SeqIOfor sequence in SeqIO.parse('/coronavirus/MN908947.fna', "fasta"):
print(sequence.seq)
print(len(sequence),'nucliotides')
```

So it produces the sequence and length of the sequence.

```
GCAATGGATACAACTAGCTACAGAGAAGCTGCTTGTTGTCATCTCGCAAAGGCTCTCAATGACTT
CAGTAACTCAGGTTCTGATGTTCTTTACCAACCACCACAAACCTCTATCACCTCAGCTGTTTTGC
AGAGTGGTTTTAGAAAAATGGCATTCCCATC.....AGAATGACAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA29903 nucliotides
```

Loading Complementary DNA Sequence into an alignable file

```
fromBio.SeqRecord import SeqRecord
from Bio import SeqIO
DNAsequence = SeqIO.read('/coronavirus/MN908947.fna', "fasta")
```

SeqIO.read() will produce will that basic information regarding the sequence.

```
SeqRecord(seq=Seq('ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTC
TTGT...AAA', SingleLetterAlphabet()), id='MN908947.3', name='MN908947.3',
description='MN908947.3 Severe acute respiratory syndrome coronavirus 2 isolate
Wuhan-Hu-1, complete genom
```

Since input sequence is FASTA (DNA), and Coronavirus is RNA type of virus, we need to:

1. Transcribe DNA to RNA (ATTAAAGGTT… => AUUAAAGGUU…)
2. Translate RNA to Amino acid sequence (AUUAAAGGUU… => IKGLYLPR*Q…)

In the current scenario, the .fna file starts with ATTAAAGGTT, then we call transcribe() so T (thymine) is replaced with U (uracil), so we get the RNA sequence which starts with AUUAAAGGUU.

```
DNA = DNAsequence.seq#Convert DNA into mRNA Sequence
mRNA = DNA.transcribe() #Transcribe a DNA sequence into RNA.
print(mRNA)
print('Size : ',len(mRNA))
```

The transcribe() will convert the DNA to mRNA.

```
UAUUUUAGUGGAGCAAUGGAUACAACUAGCUACAGAGAAGCUGCUUGUUGUCAUCUCGCAAAG
GCUCUCAAUGACUUCAGUAACUCAGGUUC...UAAUAGCUUCUUAGGAGAAUGACAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAA
Size : 29903
```

The difference between the DNA and the mRNA is just that the **bases T (for Thymine) are replaced with U (for Uracil).**

Next, we need to translate the mRNA sequence to amino-acid sequence using translate() method, we get something like IKGLYLPR*Q (is so-called STOP codon, effectively is a separator for proteins).

```
Amino_Acid = mRNA.translate(table=1, cds=False)
print('Amino Acid', Amino_Acid)
print("Length of Protein:",len(Amino_Acid))
print("Length of Original mRNA:",len(mRNA))
```

In the below output Amino acids are separated by *.

Amino Acid :
IKGLYLPR*QTNQLSISCRSVL*TNFKICVAVTRLHA*CTHAV*LITNYCR*QDTSNSSIFCRLLTVSSV
LQPIISTSRFRPGVTER*DGEPCPWFQRENTRPTQFACFTGSRRARTWLWRLRGGGLIRGTSTS*R
WHLWLSRS*KRRFAST*TALCVHQTFGCSNCTSWSCYG...*SHIAIFNQCVTLGRT*KSHHIFTEAT
RSTIECTVNNARESCLYGRALMCKINFSSAIPM*F**LLRRMTKKKKKKKKKKLength of Protein :
9967
Length of Original mRNA : 29903

In our scenario, the sequence looks like

This **IKGLYLPR*QTNQLSISCRSVL*TNFKICVAVTRLHA**, where: **IKGLYLPR**

encodes the first protein while **QTNQLSISCRSVL** encodes the second protein.
Note that there are fewer sequences in the protein than the mRNA that is because 3 prime mRNA's are used to produce a single subunit of a protein, known as an amino acid, using the codon table shown below. The * is used to denote a stop codon, in these regions the protein has finished its full length. Many of these occur frequently and result in short lengths of protein, more likely than not these play a little biological role and will be excluded in further analyses.

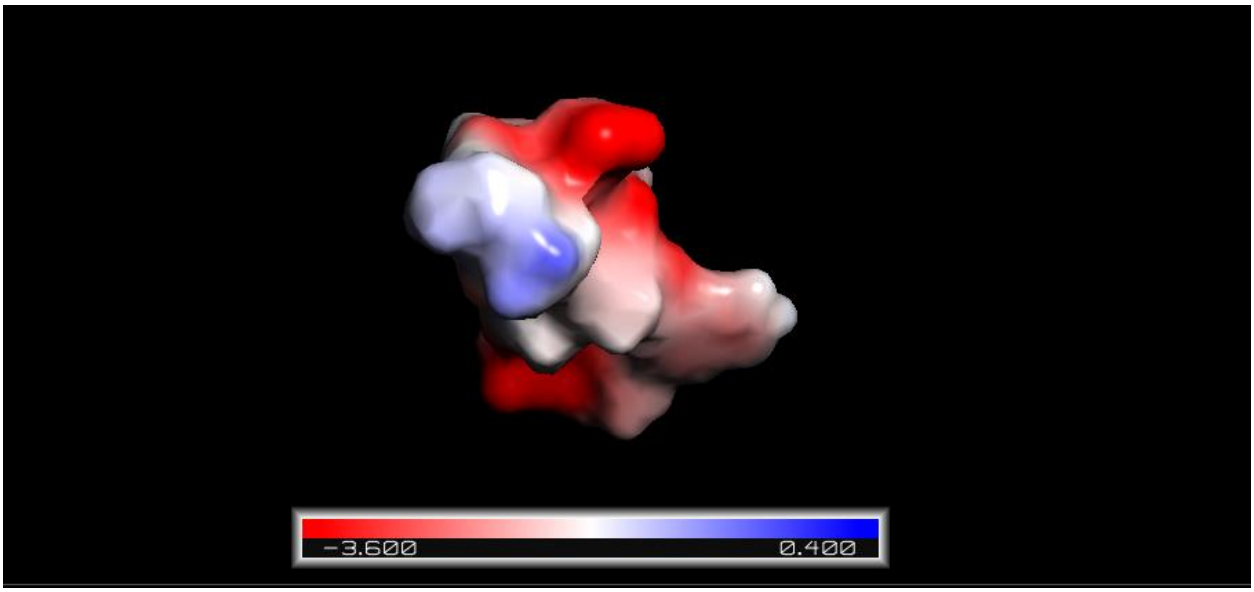


Fig :5.1 The PyMOL molecular graphics software package can both run APBS and visualize resulting electrostatic potentials.

Before proceeding, you must load the electrostatic potential data into PyMOL. Under the Visualization tab of the PyMOL APBS Tools window, press the Update button.

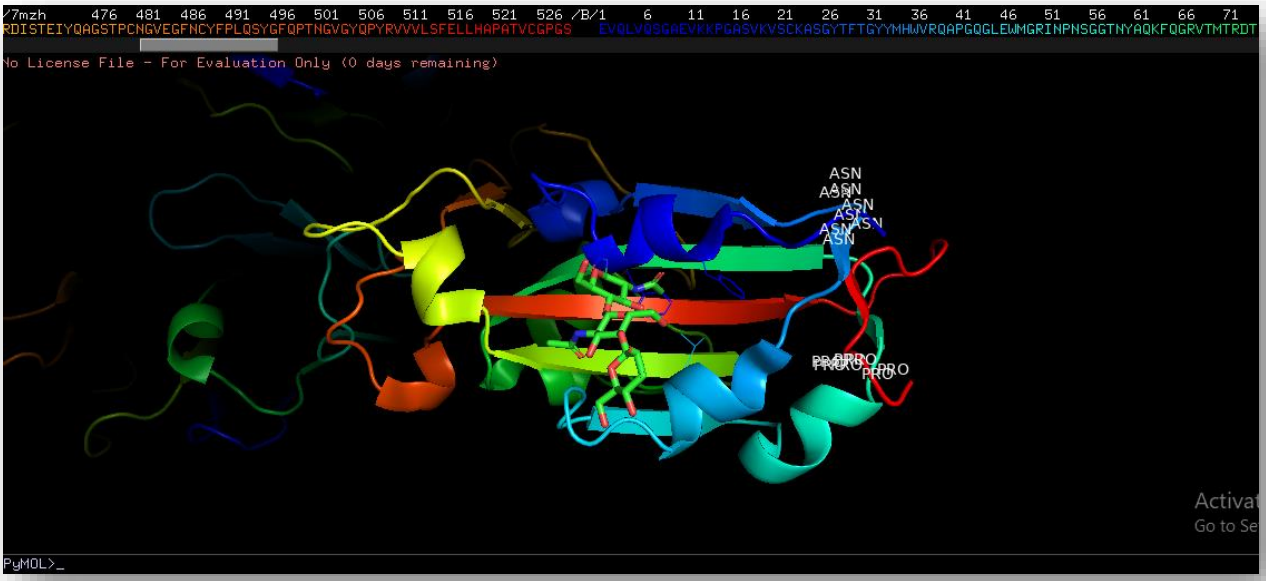


Fig 5.2: Pymol N-terminal and C-terminal Visualization

PyMOL can simultaneously provide geometric information (from the molecular surface) and useful electrostatic potential information (from the solvent-accessible surface). To visualize the molecule in this way, simply uncheck the “Solvent accessible surface”box and check the “Color by potential on sol. acc. surf.” box on the Visualization tab.

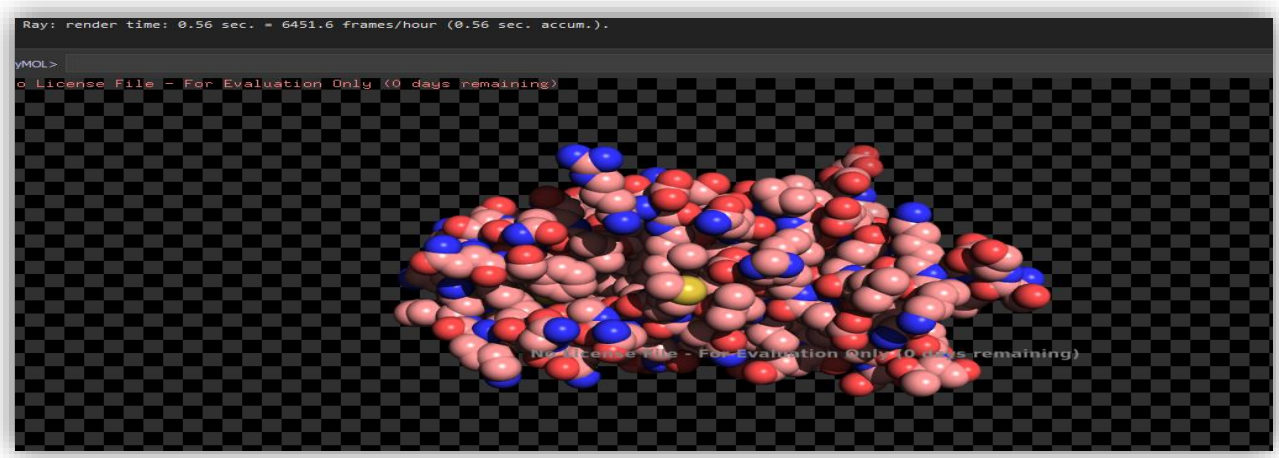


Fig 5.3: Ray Tracing molecular graphics snap in PYMOL

Volume Rendering Visualization in PyMOL

To bring your structural and volumetric (e.g., density) data into the workspace, select *File* > *Open*, and select your files. Then, for the newly loaded volumetric data, select select *A* > *Volume*> *Default.* Your data will then be shown as a volume. To control the colors and opacity, just click on the new *Volume* button in the Upper Control Window.

Volume data is represented by color and transparency that is graded according to the value of the volume. By default, volumes with higher scalar values (e.g., high electron density) will appear more opaque red and volumes with lower scalar values (e.g. low more transparent with lower electron density will appear in a less opaque blue. However, coloring and transparency are fully customizable using a new built-in interface. To control the color ramp for your data, select the *Volume* button from the Upper Control Panel.

To set transparency levels for any number of iso-surfaces, CTRL-click on the desired iso-level in the upper control panel and an iso-mesh will be created. To remove a color point, just middle click on it. To update density or opacity values for a color, just drag the color point with your mouse. PyMOL updates the colors on the fly.

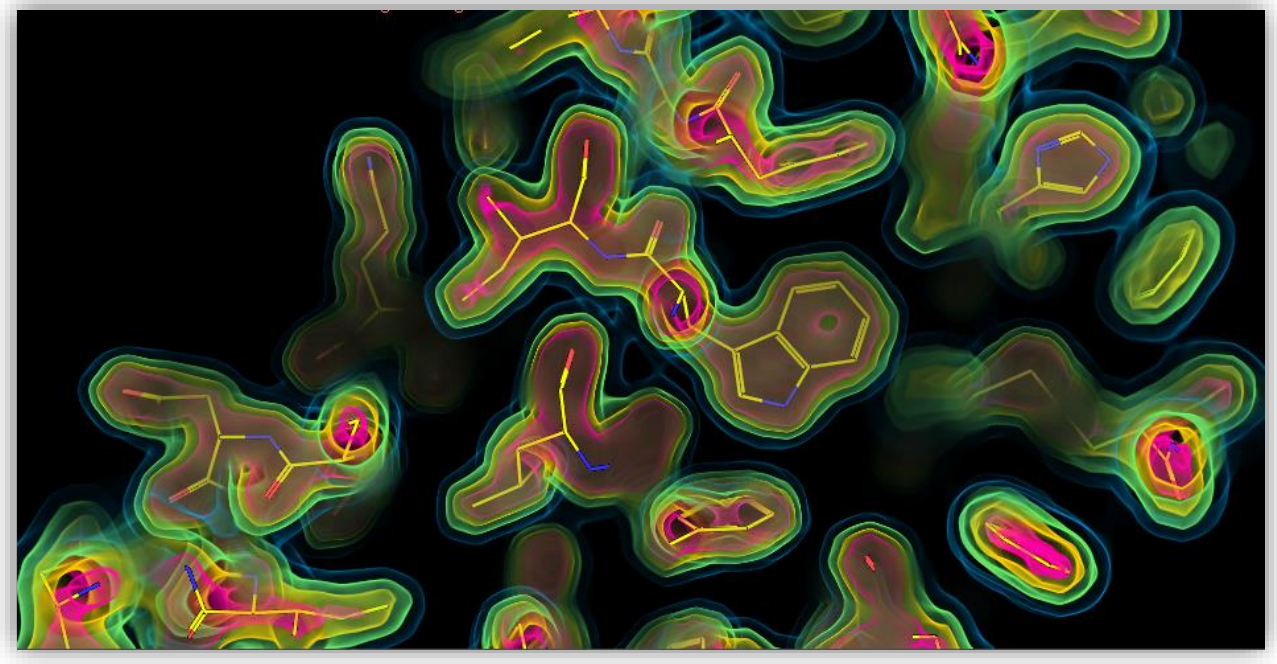


Fig 5.4: Volume Rendering Visualization in PyMOL

Rendering settings can be modulated to improve performance or quality. The volume representation to improve performance. There are two main settings that can be controlled to modulate the quality-speed tradeoff - volume_layers (default = 256) and volume_bit_depth (default = 16). Change the volume_layers variable by typing set volume_layers, X, where X is some number like 256 or 512--or even 4096 if your machine is fast enough. Try setting X to 32 if your machine isn't that powerful. Change the volume_bit_depth variable to 4, 8, 16, or 32 by typing set volume_bit_depth, X, where X is 4, 8, 16 or 32, at the command line. The default values for these variables should enable reasonable quality and performance on a fairly new computer

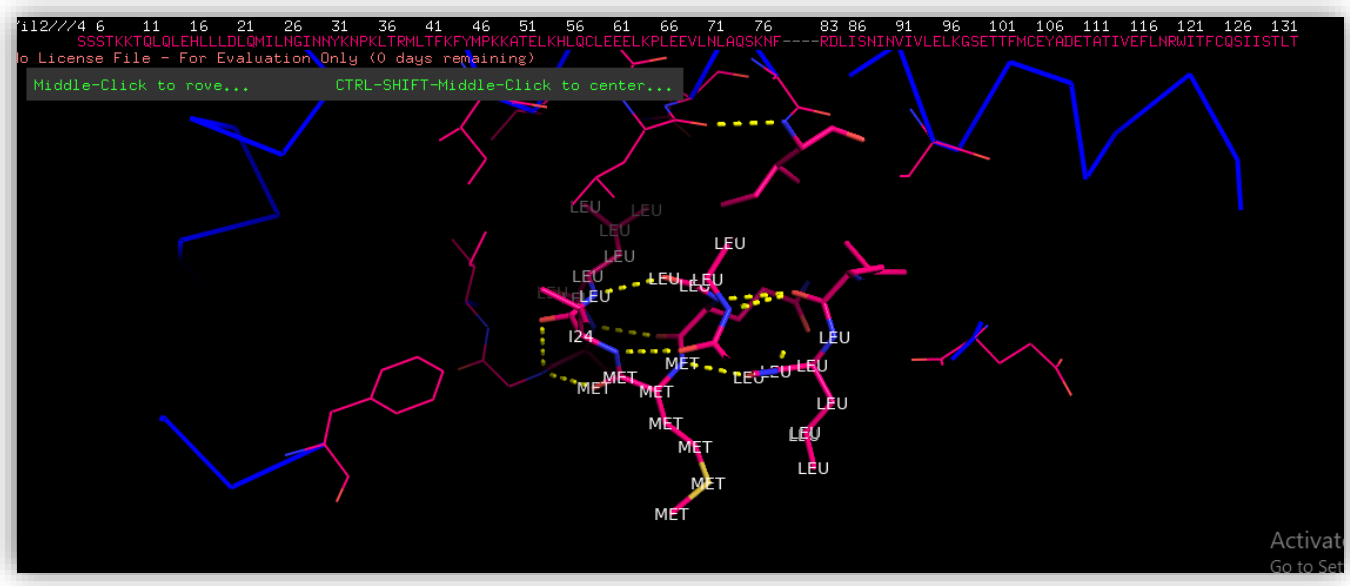


Fig 5.5: Roving Density in PyMOL

DISCUSSION

The genomic analysis of the sequence by applying bioinformatics tools NCBI,,PDB,RASMOL, PyMOL and BioPython. The blast represents the basic local alignment search tool is an alignment for comparing primary biological sequences information such as beta coronavirus protein sequence . value is increased frm default value, larger lists with more low scoring his can be reported based on the quality of alignment RNA whose reference sequence showing the BLAST output result to study its protein-protein interaction. The FASTA format sequence is used for representing nucleotide sequence of its genome with NCBI reference sequence . where the discription line is distinguished from the sequence data by a greater symbol at the beginning.

PDB is database for three dimentional structural data of large biological molecules such as nucleic acid. Many other databases use protein structures deposited in the PDB. RasMol is the computer programed written for molecular graphics visualization.therasmol amino color scheme colors amino acids according to traditional amino acid properties. The purpose of colouring is to identify amino acids in an unusual or surprising environment. ORF is the part of reading frames that has the potential to be translated. PyMOL is a molecular graphics system with an embedded Python interpreter designed for real-time visualization and rapid generation of high-quality molecular graphics images and animations. PyMOL is a powerful tool to display the 3D structures of biological targets, and offers up to 12 different stereo visualization modes Users can efficiently highlight and distinguish various important structural features in the targets, particularly the suitable binding sites for drug molecules.

MINOR PROJECTS GRAPHICAL CODING

```
import turtle


sc = turtle.Screen()
pen = turtle.Turtle()
def semi_circle(col, rad, val):


    pen.color(col)
    pen.circle(rad, -180)
    pen.up()
    pen.setpos(val, 0)
    pen.down()
    pen.right(180)
col = ['violet', 'indigo', 'blue',' green', 'yellow', 'orange', 'red']


sc.setup(600, 600)
sc.bgcolor('black')
pen.right(90)
pen.width(10)
pen.speed(7)
for i in range(7):
    semi_circle(col[i], 10*(i + 8), -10*(i + 1))
pen.hideturtle()
```

```
from turtle import *
color('red', 'yellow')
begin_fill()
while True:
    forward(200)
    left(170)
    if abs(pos()) < 1:
        break
end_fill()
done()
```

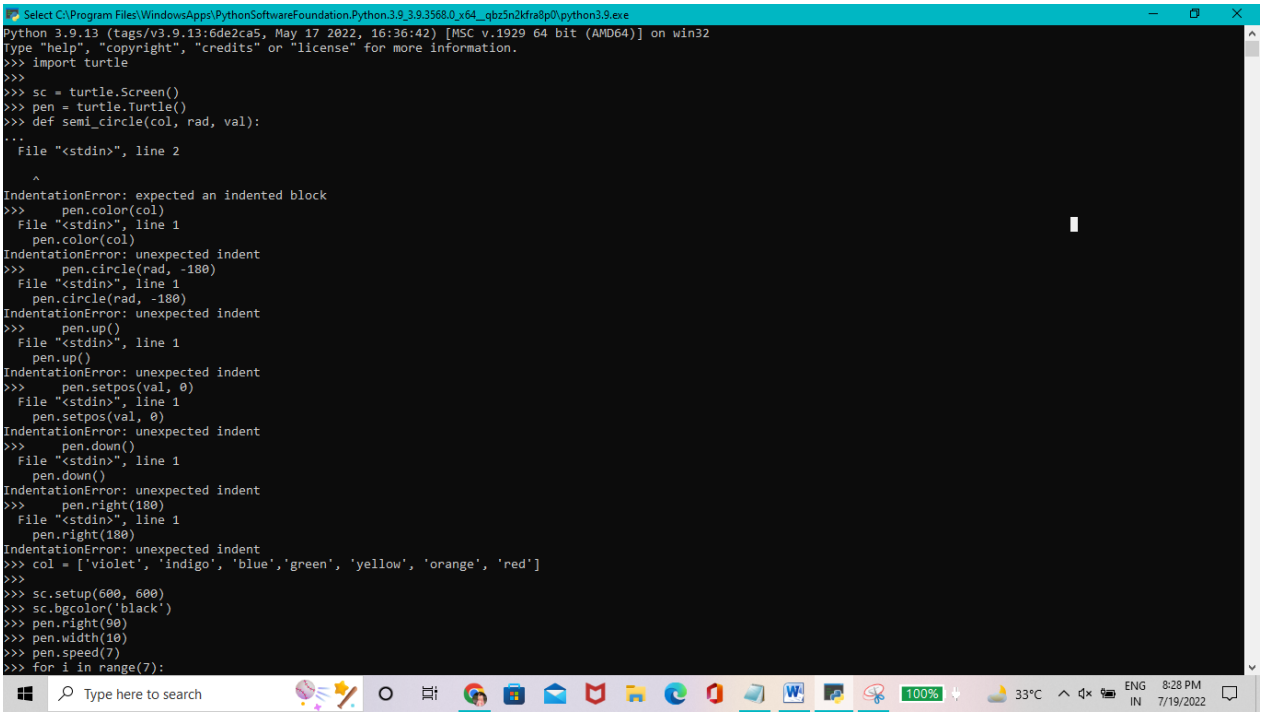


Fig 6: Showing inline command to for collected dataset

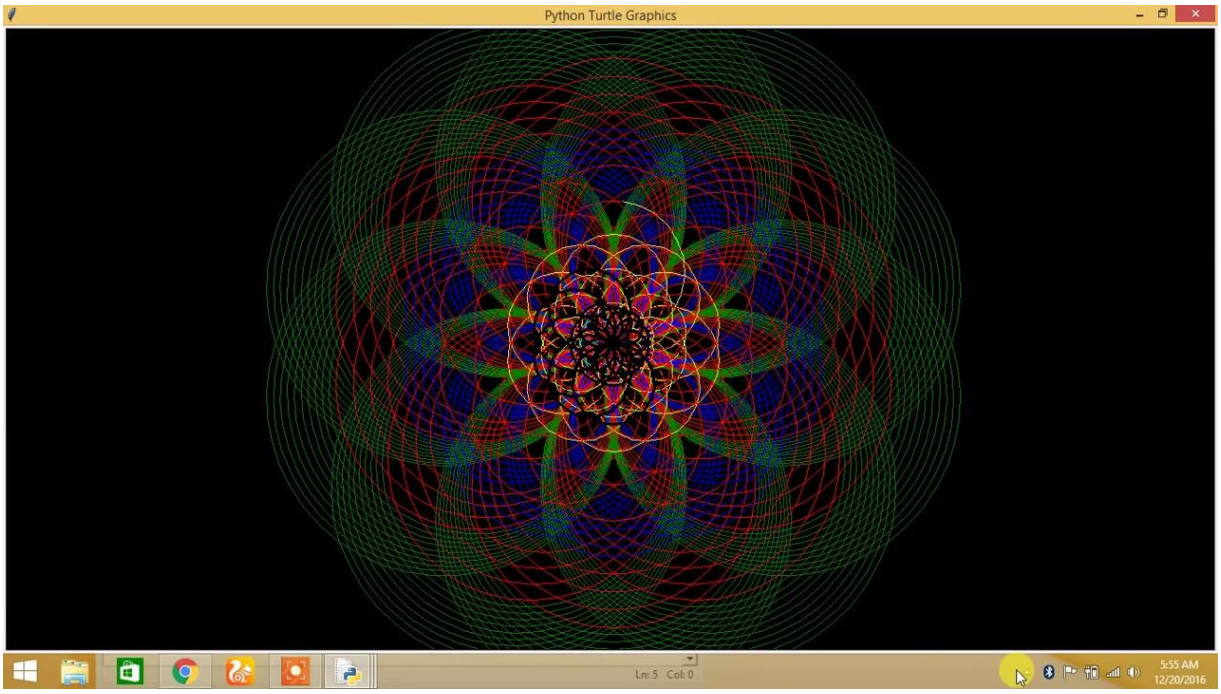
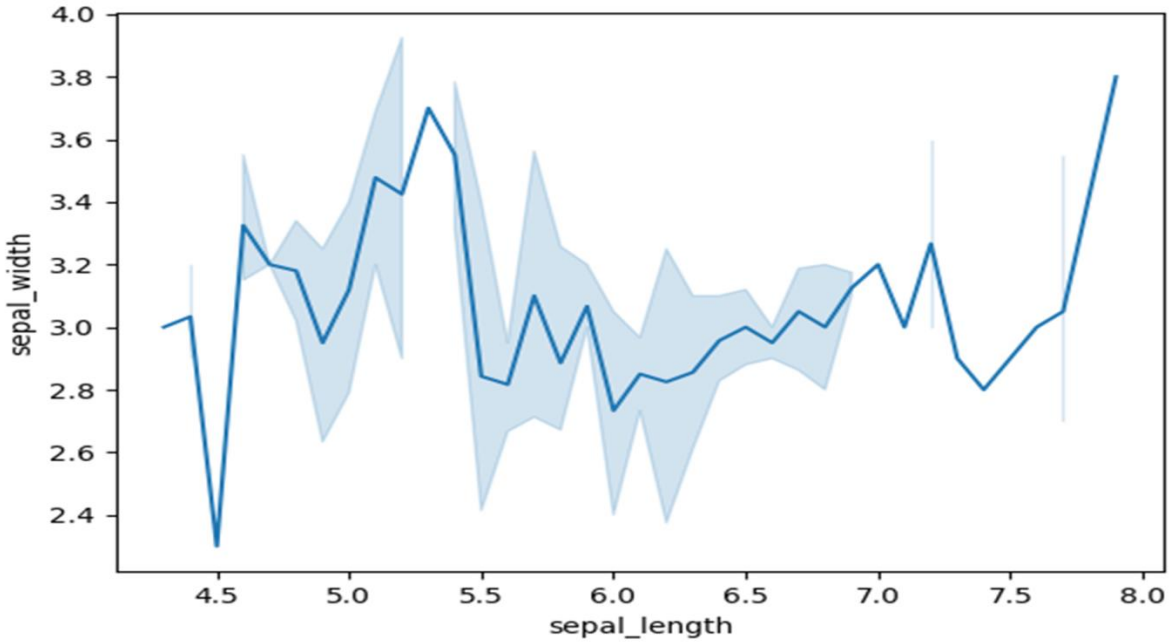


Fig: Results obtained

Big Data : Project 3D IRIS

```
>>> import seaborn as sns
>>> import matplotlib.pyplot as plt
>>>
>>> # loading dataset
>>> data = sns.load_dataset("iris")
>>>
>>> # draw lineplot
>>> sns.lineplot(x="sepal_length", y="sepal_width", data=data)
<AxesSubplot:xlabel='sepal_length', ylabel='sepal_width'>
>>>
>>> # setting the title using Matplotlib
>>> plt.title('Title using Matplotlib Function')
Text(0.5, 1.0, 'Title using Matplotlib Function')
>>>
>>> plt.show()
>>>
```




```
>>> import seaborn as sns
>>> import matplotlib.pyplot as plt
>>>
>>> # current colot palette
>>> palette = sns.color_palette()
>>>
>>> # plots the color palette as a
>>> # horizontal array
>>> sns.palplot(palette)
>>>
>>> plt.show()
```

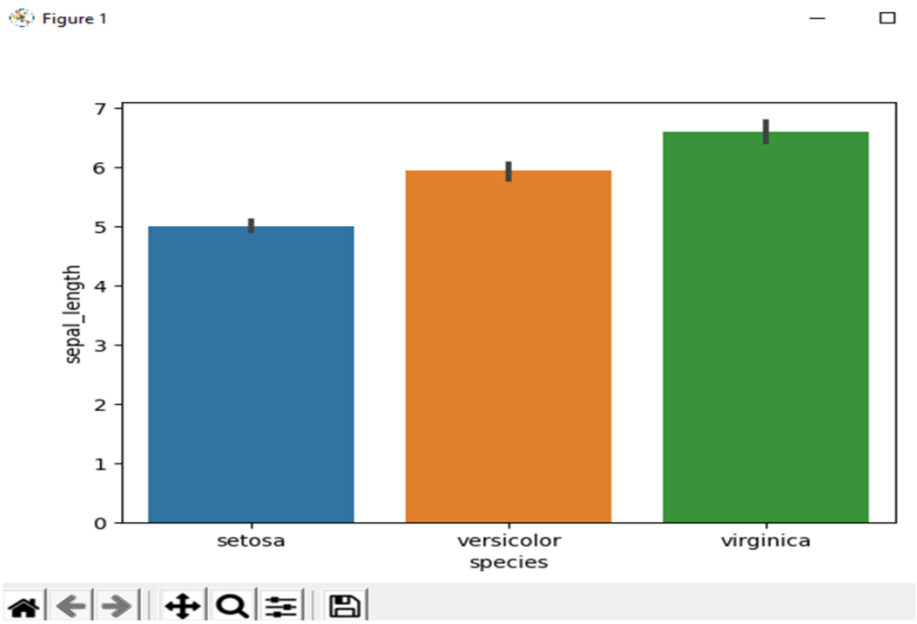


Figure 1
7
6
5
4
3
2
1
0
sepal_length
setosa
versicolor
virginica
species

Conclusion

The current COVID-19 pandemic is clearly an international public health problem. There have been rapid advances in what we know about the pathogen, how it infects cells and causes disease, and clinical characteristics of disease. Due to rapid transmission, countries around the world should increase attention into disease surveillance systems and scale up country readiness and response operations including establishing rapid response teams and improving the capacity of the national laboratory system.

In summary, the use of available information related to SARSCoV-2 epitopes associated with bioinformatics predictions points to specific regions of viral nucleon capsid that are targets to human immune responses . We understand that lack of biological confirmation of identified peptides may limit the impact of our discovery. However, testing the antigenicity of these B and T cell epitopes will be the next step on our research program. The observation that some T cell epitopes are highly conserved between SARS-CoV-2 and other human coronaviruses is critical. Vaccines that target human immune responses toward these conserved epitopes could generate immunity that is cross-protective across alpha coronaviruses and beta coronaviruses (25). This would be an advantage given the potential of future novel coronavirus emergence. I have used an object oriented language and used a protein sequence

References

1. Lu H., Stratton C.W., Tang Y.W. Outbreak of pneumonia of unknown etiology in Wuhan China: the mystery and the miracle. J Med Virol. 2020 [PMC free article] [PubMed] [Google Scholar]

2. Hui D.S., E I.A., Madani T.A., Ntoumi F., Kock R., Dar O. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health – the latest 2019 novel coronavirus outbreak in Wuhan, China. Int J Infect Dis. 2020;91:264–266. [PMC free article] [PubMed] [Google Scholar]

3. Gorbalenya A.E.A. Severe acute respiratory syndrome-related coronavirus: the species and its viruses – a statement of the Coronavirus Study Group. BioRxiv. 2020 doi: 10.1101/2020.02.07.937862. [CrossRef] [Google Scholar]

4. Burki T.K. Coronavirus in China. Lancet Respir Med. 2020 [PMC free article] [PubMed] [Google Scholar]

5. NHS press conference, February 4, 2020. Beijing, China. National Health Commission (NHC) of the People's Republic of China. http://www.nhc.gov.cn/xcs/xwbd/202002/235990d202056cfcb202043f202004a202070d202007f209703b202113c202000.shtml.

6. World Health Organization; Geneva, Switzerland: 2020. WHO: coronavirus disease 2019 (COVID-19) situation report – 23. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200212-sitrep-20200223-ncov.pdf?sfvrsn=20200241e20200219fb20200278_20200212 [accessed 20200213 February 20202020] [Google Scholar]

7. Huang C., Wang Y., Li X., Ren L., Zhao J., Hu Y. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020 [PMC free article] [PubMed] [Google Scholar]

8. Chen N., Zhou M., Dong X., Qu J., Gong F., Han Y. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020 [PMC free article] [PubMed] [Google Scholar]

9. Burrell C., Howard C., Murphy F. 5th ed. Academic Press; United States: 2016. Fenner and White's medical virology. [Google Scholar]

10. Lu R., Zhao X., Li J., Niu P., Yang B., Wu H. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020 [PMC free article] [PubMed] [Google Scholar]

11. Zhou P., Yang X.L., Wang X.G., Hu B., Zhang L., Zhang W. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020 [PMC free article] [PubMed] [Google Scholar]

12. Hughes J., Wilson M., Luby S., Gurley E., Hossain M. Transmission of human infection with Nipah virus. Clin Infect Dis. 2009;49(11):1743–1748. [PMC free article] [PubMed] [Google Scholar]

13. Li Q., Guan X., Wu P., Wang X., Zhou L., Tong Y. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med. 2020;382(13):1199–1207. [PMC free article] [PubMed] [Google Scholar]

14. Yu W., Tang G., Zhang L., Corlett R. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2) using whole genomic data. ChinaXiv. 2020 Preprint. [PMC free article] [PubMed] [Google Scholar]

15. Chan J.F., Yuan S., Kok K.H., To K.K., Chu H., Yang J. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet. 2020 [PMC free article] [PubMed] [Google Scholar]

16. Kramer A., Schwebke I., Kampf G. How long do nosocomial pathogens persist on inanimate surfaces? A systematic review. BMC Infect Dis. 2006;6:130. [PMC free article] [PubMed] [Google Scholar]

17. Kampf G., Todt D., Pfaender S., Steinmann E. Persistence of coronaviruses on inanimate surfaces and its inactivation with biocidal agents. J Hosp Infect. 2020 [PMC free article] [PubMed] [Google Scholar]

18. Rothe C., Schunk M., Sothmann P., Bretzel G., Froeschl G., Wallrauch C. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. N Engl J Med. 2020 [PMC free article] [PubMed] [Google Scholar]

19. Kupferschmidt K. Science news. 2020. Study claiming new coronavirus can be transmitted by people without symptoms was flawed. [Google Scholar]

20. Bai Y., Yao L., Wei T., Tian F., Jin D.Y., Chen L. Presumed asymptomatic carrier transmission of COVID-19. JAMA. 2020 [PMC free article] [PubMed] [Google Scholar]

20.Oliveira SC, de Magalhães MTQ, Homan EJ. Immunoinformatic Analysis of SARS-CoV-2 Nucleocapsid Protein and Identification of COVID-19 Vaccine Targets. Front Immunol. 2020 Oct 28;11:587615. doi: 10.3389/fimmu.2020.587615. PMID: 33193414; PMCID: PMC7655779.

21.Twyman, Richard (1 August 2003). "Gene Structure". The Wellcome Trust.Archived from the original on 28 March 2007.Retrieved 6 April 2003.

22.Höglund M, Säll T, Röhme D (February 1990). "On the origin of coding sequences from random open reading frames".Journal of Molecular Evolution. 30 (2): 104–108. Bibcode:1990JMolE..30..104H. doi:10.1007/bf02099936. ISSN 0022-2844.S2CID 5978109.

23.Sakharkar MK, Chow VT, Kangueane P (2004)."Distributions of exons and introns in the human genome".In Silico Biology. 4 (4): 387–93. PMID 15217358.

24.Parnell, Laurence D. (2012-01-01)."Advances in Technologies and Study Design". In Bouchard, C.; Ordovas, J. M. (eds.). Recent Advances in Nutrigenetics and Nutrigenomics.Progress in Molecular Biology and Translational Science.Recent Advances in Nutrigenetics and Nutrigenomics.Vol. 108.Academic Press. pp. 17–50. doi:10.1016/B978-0-12-398397-8.00002-2. ISBN 9780123983978.PMID 22656372.Retrieved 2019-11-07.

25.Gilbert W (February 1978). "Why genes in pieces?".Nature. 271 (5645): 501. Bibcode:1978Natur.271..501G. doi:10.1038/271501a0. PMID 622185.S2CID 4216649.

26.(n.d.). Retrieved from https://www.differencebetween.com/wp-content/uploads/2017/03/Difference-Between-Transition-and-Transversion-3.png

27.Lercher MJ, Urrutia AO, Pavlícek A, Hurst LD (October 2003). "A unification of mosaic structures in the human genome". Human Molecular Genetics. 12 (19): 2411–5. doi:10.1093/hmg/ddg251. PMID 12915446.

28.Oliver JL, Marín A (September 1996). "A relationship between GC content and coding-sequence length".Journal of Molecular Evolution. 43 (3): 216–23. Bibcode:1996JMolE..43..216O. doi:10.1007/pl00006080. PMID 8703087.

29."ROSALIND | Glossary | Gene coding region".rosalind.info. Retrieved 2019-10-31.

30.Vinogradov AE (April 2003). "DNA helix: the importance of being GC-rich". Nucleic Acids Research. 31 (7): 1838–44. doi:10.1093/nar/gkg296. PMC 152811.PMID 12654999.

31.Bohlin J, Eldholm V, Pettersson JH, Brynildsrud O, Snipen L (February 2017). "The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes". BMC Genomics. 18 (1): 151. doi:10.1186/s12864-017-3543-7. PMC 5303225.PMID 28187704.

32.Sémon M, Mouchiroud D, Duret L (February 2005). "Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance". Human Molecular Genetics. 14 (3): 421–7. doi:10.1093/hmg/ddi038. PMID 15590696.

33.Overview of transcription.(n.d.).Retrieved from https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/overview-of-transcription.

34. Clancy, Suzanne (2008). "Translation: DNA to mRNA to Protein". Scitable: By Nature Education.

35.Plociam (2005-08-08), English: The structure of a mature eukaryotic mRNA. A fully processed mRNA includes the 5' cap, 5' UTR, coding region, 3' UTR, and poly(A) tail., retrieved 2019-11-19

36. Shinohara K, Sasaki S, Minoshima M, Bando T, Sugiyama H (2006-02-13). "Alkylation of template strand of coding region causes effective gene silencing". Nucleic Acids Research. 34 (4): 1189–95. doi:10.1093/nar/gkl005. PMC 1383623.PMID 16500890.

37. "DNA alkylation Gene Ontology Term (GO:0006305)". www.informatics.jax.org. Retrieved 2019-10-30.

38.Shafee T, Lowe R (2017). "Eukaryotic and prokaryotic gene structure".WikiJournal of Medicine.4 (1). doi:10.15347/wjm/2017.002.

39.Konarska MM (1998). "Recognition of the 5' splice site by the spliceosome". ActaBiochimicaPolonica. 45 (4): 869–81. doi:10.18388/abp.1998_4346. PMID 10397335.

Jonsta247 (2013-05-10), English: Example of silent mutation, retrieved 2019-11-19

40.Yang, J. (2016, March 23). What are Genetic Mutation? Retrieved from https://www.singerinstruments.com/resource/what-are-genetic-mutation/.

41.What is a gene mutation and how do mutations occur? - Genetics Home Reference - NIH.(n.d.).Retrieved from https://ghr.nlm.nih.gov/primer/mutationsanddisorders/genemutation.

42.DNA proofreading and repair.(n.d.).Retrieved from https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-discovery-and-structure-of-dna/a/dna-proofreading-and-repair.

43.Peretó J. (2011) Wobble Hypothesis (Genetics). In: Gargaud M. et al. (eds) Encyclopedia of Astrobiology. Springer, Berlin, Heidelberg

44.Havrilla, J. M., Pedersen, B. S., Layer, R. M., & Quinlan, A. R. (2018).A map of constrained coding regions in the human genome. Nature Genetics, 88–95. doi: 10.1101/220814

45.Furuno M, Kasukawa T, Saito R, Adachi J, Suzuki H, Baldarelli R, et al. (June 2003). "CDS annotation in full-length cDNA sequence".Genome Research.Cold Spring Harbor Laboratory Press. 13 (6B): 1478–87. doi:10.1101/gr.1060303. PMC 403693.PMID 12819146.

46.Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV (May 2002). "Purifying and directional selection in overlapping prokaryotic genes".Trends in Genetics. 18 (5): 228–32. doi:10.1016/S0168-9525(02)02649-5. PMID 12047938.

47.Chirico N, Vianelli A, Belshaw R (December 2010). "Why genes overlap in viruses". Proceedings.Biological Sciences. 277 (1701): 3809–17. doi:10.1098/rspb.2010.1052. PMC 2992710.PMID 20610432.

48.Firth AE, Brown CM (February 2005). "Detecting overlapping coding sequences with pairwise alignments".Bioinformatics. 21 (3): 282–92. doi:10.1093/bioinformatics/bti007. PMID 15347574.

49.Schlub TE, Buchmann JP, Holmes EC (October 2018). Malik H (ed.). "A Simple Method to Detect Candidate Overlapping Genes in Viruses Using Single Genome Sequences".Molecular Biology and Evolution. 35 (10): 2572–2581. doi:10.1093/molbev/msy155. PMC 6188560.PMID 30099499.

50.Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, et al. (2020) Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. Cell 181(5):990-996.e5. doi: 10.1016/j.cell.2020.04.021. pmid:3238654

51.Fraser C, Donnelly C, Cauchemez S, Hanage W, Van Kerkhove M, et al. (2009) WHO Rapid Pandemic Assessment Collaboration. Science 324: 1557. pmid:19433588

52. Gardy JL, Loman NJ (2018) Towards a genomics-informed, real-time, global pathogen surveillance system. Nature Reviews Genetics 19: 9. pmid:29129921

53.Gambaro F, Baidaliuk A, Behillil S, Donati F, Albert M, et al. (2020) Introductions and early spread of SARS-CoV-2 in France.
BioRxiv. https://doi.org/10.1101/2020.04.24.059576 pmid:32643599

54.Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, et al. (2020) First case of 2019 novel coronavirus in the United States. New England Journal of Medicine 5;382(10):929–936.

55.https://www.medicalnewstoday.com/articles/types-of-coronavirus#what-are-coronaviruses

56.https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus

57.https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus

58)https://www.cebm.net/covid-19/coronaviruses-a-general-introduction/

59)http://pdb101.rcsb.org/

60)https://pubmed.ncbi.nlm.nih.gov/

61)Letko M, Marzi A, Munster V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B beta coronaviruses. Nat Microbiol doi: 10.1038/s41564-020-0688-y.

62)Li F, Li W, Farzan M, Harrison SC. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science2005; 309: 1864-8

63)Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding," Lancet (2020). https://doi.org/10.1016/S0140-6736(20)30251-8.

64)Lu J, du Plessis L, Liu Z, Hill V, Kang M, et al. (2020) Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. Cell 181(5):997-1003.e9. pmid:32359424

65)Tai W, He L, Zhang X, Pu J, Voronin D, et al. (2020) Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. Cellular & molecular immunology 17: 613–620. doi: 10.1038/s41423-020-0400-4. pmid:32203189

66)Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269. doi: 10.1038/s41586-020

67)Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. "A novel coronavirus from patients with pneumonia in China, 2019," N Engl J Med (2020). https://doi.org/10.1056/NEJMoa2001017-2008-3

Pikora, M., &Gieldon, A. (2015). RASMOL AB - new functionalities in the program forstructure analysis. *ActabiochimicaPolonica*, *62*(3), 629–631.https://doi.org/10.18388/abp.2015_972

Chapman,B. and Chang,J. (2000) Biopython: Python tools for computational biology.

ACM SIGBIO Newslett.,20, 15–19.

Chaudhuri,R.R. and Pallen,M.J. (2006) xBASE, a collection of online databases for bacterial comparative genomics. Nucleic Acids Res.,34, D335–D337.

Bateman,A. et al. (2004) The Pfam protein families database. Nucleic Acids Res.,32, D138–D141.

Beaumont,M.A. and Nichols,R.A. (1996) Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. Lond. B,263, 1619–1626.

Benson,D.A. et al. (2007) GenBank. Nucleic Acids Res.,35, D21–D25.

Felsenstein,J. (1989) PHYLIP – phylogeny inference package (Version 3.2). Cladistics, 5, 164–166.

Hamelryck,T. and Manderick,B. (2003) PDB file parser and structure class implemented in Python. Bioinformatics,19, 2308–2310.

Hinsen,K. (2000) The molecular modeling toolkit: a new approach to molecular simulations. J. Comp. Chem.,21, 79–85.

Holland,R.C.G. et al. (2008) BioJava: an open-source framework for bioinformatics. Bioinformatics,24, 2096–2097.

De Hoon,M.J.L. et al. (2004) Open source clustering software. Bioinformatics,20, 1453–1454.

Kauff,F. et al. (2007) WASABI: an automated sequence processing system for multi-gene phylogenies. Syst. Biol. 56, 523–531.

Kulikova,T. et al. (2006) EMBL nucleotide sequence database in 2006. Nucleic Acids Res.,35, D16–D20.

Lavel,G. and Excoffier,L. (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics,20, 2485–2487.

Maddison,D.R. et al. (1997) NEXUS: an extensible file format for systematic information. Syst. Biol.,46, 590–621.

Oliphant,T.E. (2006) Guide to NumPy. Trelgol Publishing, USA.

Oliphant,T.E. (2007) Python for Scientific Computing. Comput. Sci. Eng.,9, 10–20.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence analysis. PNAS,85, 2444–2448.

Pritchard,L. et al. (2006) GenomeDiagram: a Python package for the visualisation of large-scale genomic data. Bioinformatics 22, 616–617.

Rice,P. et al. (2000) EMBOSS: the European molecular biology open software suite. Trends Genet.,16, 276–277.

Rousset,F. (2007) GENEPOP ’007: a complete re-implementation of the GENEPOP software for Windows and Linux. Mol. Ecol. Res.,8, 103–106.

Stajich,J.E. et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res.,12, 1611–1618.

The UniProt Consortium. (2007) The universal protein resource (UniProt). Nucleic Acids Res.,35, D193–D197.

Thompson,J.D. et al. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties .