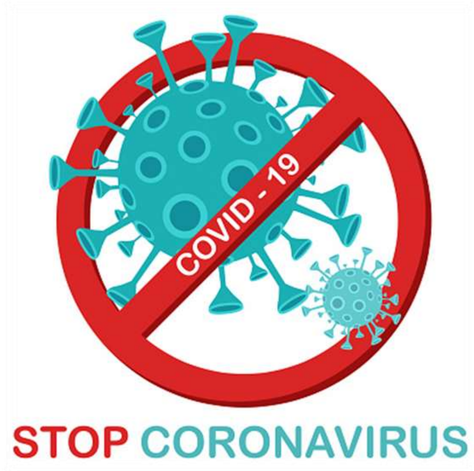
Structure Of Gene Mutation *On Covid Coronavirus*

Andrew Medhat, Nada Yassen, Omar Ahmed, Sara Mostafa and Lina Bassel

Supervisor: Dr. Amin Allam



Abstract

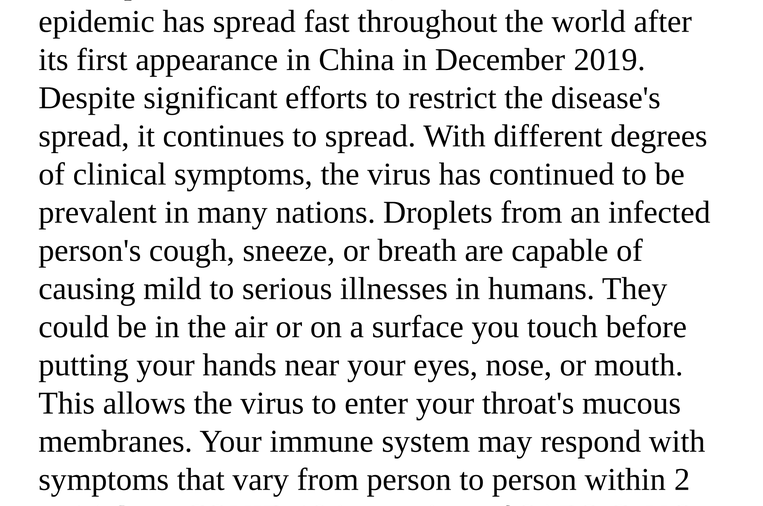
Although the majority of mutations in the SARS-CoV-2 genome are predicted to be detrimental and quickly purged or generally neutral, it is normal It takes a virus to mutate, causing variants So, we want to discover several new variants using machine learning instead of discovering them in genomics laboratories .

Finding COVID-19 vaccines is regarded as a major accomplishment .In the biological community, COVID-19 variants pose a significant this goal is an impediment to this goal. When researchers discover a new vaccine for a COVID-19 variant, they should consider developing vaccines for other COVID-19 mutations as well. It will be easier for them to do so if they already know the COVID-19 variants' sequence and structure.

To be prepared for any variants and develop an effective vaccine for it. We can collect and analyses all possible mutations using this method. Whether or not they attach to human receptors using our machine learning model to determine if it is tied to a human receptor or not, researchers can use it. Use this model to recommend potential medications we will use protein docking tool to ensure the binding between ace2 human receptor and mutated sequences through 3D structure , docking score and RMSD. Covid-19 mutations interact with human receptors and cause harm to humans.

Introduction

Our main research area is SARS-CoV-2 and its mutations. When an amino acid in the spike protein changes, the virus develops a new structure and feature known as a variant. This happened with SARS-CoV-2, and we now have numerous variants such as alpha, beta, gamma, and delta, as well as many spike protein sequences. Scientists map the genetic content of viruses (a process called "sequencing") and then compare them to discover if they've changed. Since the SARS-CoV-2 virus, which produces COVID-19, has been around. The epidemic has spread fast throughout the world after its first appearance in China in December 2019. Despite significant efforts to restrict the disease's spread, it continues to spread. With different degrees of clinical symptoms, the virus has continued to be prevalent in many nations. Droplets from an infected person's cough, sneeze, or breath are capable of causing mild to serious illnesses in humans. They could be in the air or on a surface you touch before putting your hands near your eyes, nose, or mouth. This allows the virus to enter your throat's mucous membranes. Your immune system may respond with symptoms that vary from person to person within 2 to 14 days. COVID-19 is a variant of SARS-CoV-2 and therefore has numerous variants, such as omicron. Our role is to use machine learning and a recurrent neural network to uncover various novel variants of COVID-19 (RNN model).



Methods

researchers proved that “spike” is responsible for human infection with virus by binding to the ACE2 receptor of the human cell, so we are interested in the spike (S) protein, once the virus interacts with the host cell, extensive structural rearrangement of the S protein occurs, allowing the virus to fuse with the host cell membrane. The total length of Spike protein is 1273 amino acids, it is composed of 2 subunits (s1 and s2).The S1 subunit (14–685 residues), and the S2 subunit (686–1273 residues), they are responsible for receptor binding and membrane fusion.s1 subunit there is an N-terminal domain (14–305 residues) and a receptor-binding domain (RBD, 319–541 residues) that recognizes and binds to human cell through a receptor called ACE2 receptor which is a protein on the surface of many human cells, s2 subunits promotes the fusion between human cell and RBD in s1 subunit.

When change happens in amino acid in spike protein know as a mutation and the virus will have a new structure and characteristic known as a variant, that happened in SARS-COV- 2

We will work on the effective part in spike protein called RBD to predict if we have another sequence if it causes infection or not in other word if it will bind with human cell or not.So, we got the sequence of each variant from online database called NCBI through searching by the variant’s scientific name . The genbank file consists of regions of the genome of the variant with it’s name and range of each region in the genome , we used code written in python which takes a fasta file containing sequence of genome and splits it according to ranges written in the genbank file.

Then we arranged these regions in an excel sheet which contains each region and a positive or negative sign which specify whether this region bind to human cell or not and RBD in the spike protein is the only region that binds to human cell. We need to increase our positive dataset so that our model won’t be biased because of the difference between number of positive and negative regions.

Next steps:

1-splitting the RBD of each variant into segments.

2-mutating randomly by code in the RBD of the 4 variants then entering them to swiss model tool to get PDB files then entering these files to docking tool with ACE2 receptor to see whether they will bind or not.

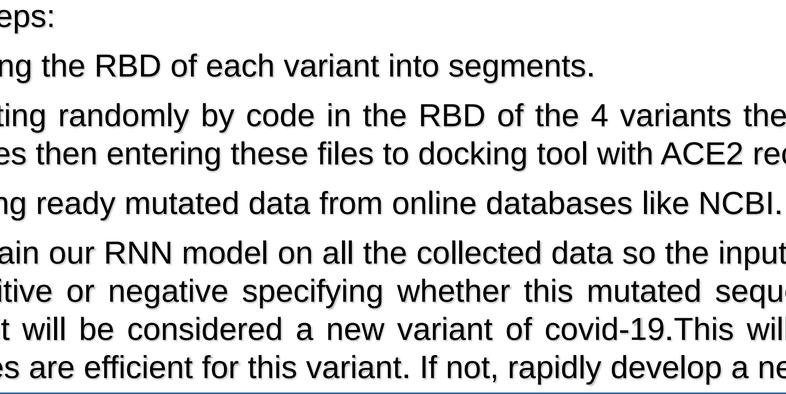
3-Getting ready mutated data from online databases like NCBI.

Then train our RNN model on all the collected data so the input will be a mutated sequence and output will be positive or negative specifying whether this mutated sequence will bind to a human cell or not. If it binds, it will be considered a new variant of covid-19.This will help researchers determine if the current vaccines are efficient for this variant. If not, rapidly develop a new vaccine for it.

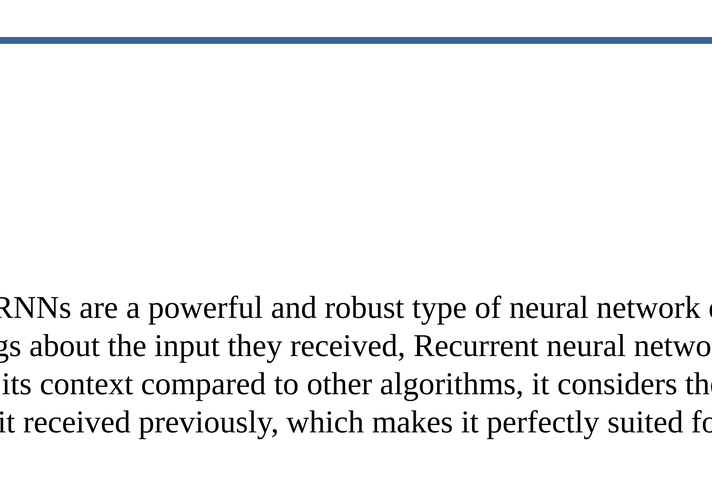
Primarily Design

Why to use RNN:

Our input is a sequence of DNA like ACTGACGTG and RNNs are a powerful and robust type of neural network due to an internal memory, RNN’s can remember important things about the input they received, Recurrent neural networks can form a much deeper understanding of a sequence and its context compared to other algorithms, it considers the current input and also what it has learned from the inputs it received previously, which makes it perfectly suited for machine learning problems that involve sequential data.



Architecture For Our Project



Conclusion

Our project should help researches in :

1-Knowing the sequence of variants that may appear in the future.

2-Determining whether the current vaccines are efficient for these variants.

3-If not, Rapidly develop new vaccine for these variants

Team Members:-

1-Andrew Medhat Email :-andrewmedhat@gmail.com
2-Nada Yassen Email:-Nadayassen0@gmail.com
3-Lina Bassel Email:-Lina_bio2018@outlook.com
4-Omar Ahmed Email:-omaraboyoussef94@gmail.com
5-Sara Mostafa Email:-sara.elrais@yahoo.com