

Identification of hostile genes in COVID variants

Kanisettyalli Karthik

B Srivathsan

Samirit Saha

Department of Computer
Science and Engineering

Department of Computer
Science and Engineering

Department of Computer
Science and Engineering

Amrita School of Engineering,
Bengaluru

Amrita School of Engineering,
Bengaluru

Amrita School of Engineering,
Bengaluru

Amrita Vishwa Vidyapeetham
Amrita University, India

Amrita Vishwa Vidyapeetham
Amrita University, India

Amrita Vishwa Vidyapeetham
Amrita University, India

[bl.en.u4aie20025@bl.students
.amrita.edu](mailto:bl.en.u4aie20025@bl.students.amrita.edu)

[bl.en.u4aie20006@bl.students
.amrita.edu](mailto:bl.en.u4aie20006@bl.students.amrita.edu)

[bl.en.u4aie20058@bl.students
.amrita.edu](mailto:bl.en.u4aie20058@bl.students.amrita.edu)

Abstract:- Following the transmission of the coronavirus on an epidemic level to various regions of the globe, owing to the process of random evolution, the original genome sequence of this virus has undergone quite a lot of mutations to give rise to different variants, that have altered several prominent features of this virus, such as its' infectivity, or the symptoms which are caused by the variants that arise from the original strain. Despite the randomness of the nature of mutation, owing to the biological process of transcription and translation which occur in the conversion of DNA to RNA and then to proteins, one can find certain noticeable patterns in mutations that might occur in the sequence while it is going through these processes. In this paper, we shall explore a few ways in which those regions in the genome and the subsequent proteins that are synthesized by them, wherein the chance of mutation is great. To perform this task, we make use of ORF detection to identify regions of the COVID-19 virus' genome sequence wherein lethal mutations may have occurred, giving rise to the proteins that may have improved the lethal properties of the coronavirus, such as the symptoms it creates or its' infectivity rate. This will also enable us to find the possible locations of the spike protein component of the COVID-19 virus, which is responsible for the effective infectivity of the COVID-19 virus.

Keywords— COVID-19, ORF, Genome Sequence, Mutation

I. INTRODUCTION

In terms of its' genetic nature, the COVID-19 virus contains a gene sequence which is made up of 4 principal nucleotides, which are: Adenosine (A), Thiamine (T), Guanine (G), Cytosine (C). The idea on which this paper is based is to run gene sequencing algorithms and subsequently protein sequencing algorithms upon the sequence of nucleotides in the coronavirus to easily sort out those sections of the COVID-19 genome which are shown to have a greater number of similarities with the other sections of the COVID gene sequence, as well as with other variants of the COVID-19 virus that have appeared in the world in recent times. A similar strategy can also be implemented for the proteins synthesized by this virus, following which Once this is done, mutation can be detected in these specific sections once we detect the nucleotides which have undergone changes in the target genome. The task of detecting mutations can be improved with the introduction of detecting open reading frames (ORFs). An ORF is the strand of DNA that exists between a start and stop codon of a genome, wherein the start codon serves as the checkpoint

wherein DNA translation can commence, and the stop codon serves as the region wherein DNA translation can finish. However, it should be noted that the process of transcription (conversion of DNA to RNA) stops at a site located beyond the ORF, to synthesize a complete functioning protein. Due to the importance of the presence of an ORF to synthesize a functional protein successfully, ORFs have been found to play a crucial role in the occurrence of mutations and evolution within the genetic structure of an organism, since any change in the region of the ORF can cause a change in the nature of the protein that is being synthesized, leading to a mutation in the genetic structure of the organism.

II. Related Works

I. LITERATURE SURVEY

"Effect of the HP0159 ORF mutation on the lipopolysaccharide structure and colonizing ability of *Helicobacter pylori*." [1] by Altman, Eleonora, et al provides us with an insight into how mutation within an ORF can create changes in the protein structures which are synthesized following the process of transcription. It also provides a case study of the *Helicobacter pylori* (a species of bacteria which causes stomach infection) whose ability to effectively propagate and colonize the stomach for its' survival is affected by a change in the sequence of the ORF. The paper "Isolation of SARS-CoV-2 strains carrying a nucleotide mutation, leading to a stop codon in the ORF 6 protein." by Delbue, Serena, et al., [2] presents the unique case of how small changes in the ORF region of the SARS-CoV-2 virus (another name for the COVID-19 virus) is able to account for inhibition of certain proteins in the strand of the virus, such as the ORF 6 protein, which creates a significant change in the processes of transcription and translation, leading to the rise of a completely new variant of COVID-19 itself, which is found to be more infectious than the original strain. The paper, which is titled 'A new coronavirus associated with human respiratory disease in China - PubMed (nih.gov) by Wu F et al [3] provides information on the initial research carried out on the original strain of COVID-19 and sheds light on the genetic nature of this virus. From this

paper we learn that the Metagenomic RNA sequencing of the COVID-19 genome sequence shows us that this virus is an RNA virus, which belongs to the family of Coronaviridae. This paper also further informs us that this virus, which consists of around 29,903 nucleotides, bears an 89.1% nucleotide similarity to the group of SARS-like coronaviruses (which belong to the genus of Beta coronavirus, and subgenus of Sarbecovirus), which so far could only be found in the bats of China. Thus, this paper highlights how the phenomenon of the cross-species jumps of this virus (from bat to human) opened the way to the possibility of mutation in COVID-19. The paper affirms that the best way to understand the important characteristics of the virus ‘ virulence, infectivity and so on.

II. CONCEPTS USED

A. A Brief Introduction to SARS COV-2

To have a better understanding of how mutation takes place within a coronavirus, first we should have a look at the structure of SARS COV-2, which causes COVID-19 in humans. Studies have shown that SARS COV-2 is an RNA virus. Research has shown that the SARS COV-2 virus has a great affinity for attaching onto the ACE2 receptors which exist on the outer surface of human cells. When it comes to the structure of the COVID-19 virus, we see that the SARS COV-2 virus is made up of a couple of structural proteins which are named as follows: the Spike protein, the Nucleocapsid protein, the Membrane protein, and the Envelope protein. Understanding this structure of the virus is especially important when it comes to further understanding how mutation takes place. Understanding the structure of the virus further goes on in helping us understand which protein of the virus our vaccine should target for it to have maximum efficiency. For example, if the mutation in the COVID-genome takes place in the gene, which is synthesizing the Spike Protein, then the vaccine which is made can be designed in such a way that it specifically targets the gene for the Spike Protein. Also, there are 11 genes within a general COVID genome where mutation can occur. The 11 genes, and the protein which each gene synthesizes is given below:

gene=**ORF1ab**. This gene encodes the following proteins: ORF1a and ORF1ab. Several nonstructural proteins are created by this gene as-well.

gene=**S**. This **Spike Protein** is encoded by this gene. The Spike protein is responsible for docking with the ACE2 receptors of the human cell before endocytosis. The protein synthesized is a trimer protein with 2 subunits: S1 and S2.

gene=**ORF3a**. This gene encodes the ORF3a protein.
gene=**E**. This gene encodes for the **Envelope** protein.
gene=**M**. This gene encodes for the **Membrane** protein.

gene=**ORF6**. This gene encodes the ORF6 protein.
gene=**ORF7a**. This gene encodes the ORF7a protein.

gene=**ORF7b**. This gene encodes the ORF7b protein.
gene=**ORF8**. This gene encodes the ORF8 protein.
gene=**N**. This gene encodes for the Nucleocapsid phosphoprotein.

gene=**ORF10**. This gene encodes ORF10 protein.

A diagrammatic representation of the structure of the SARS COV-2 virus is given below:

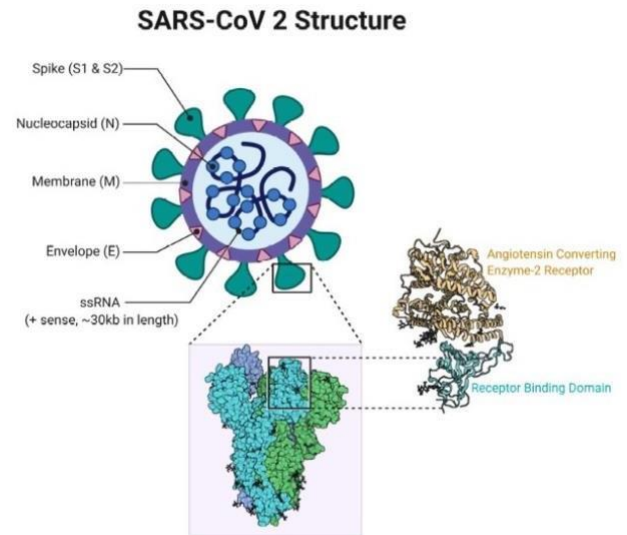


Photo by Rohan Bir Singh, MD from NCBI. Image Info: Features, Evaluation and Treatment Coronavirus (COVID-19). Authors — Marco Cascella; Michael Rajnik; Arturo Cuomo; Scott C. Dulebohn; Raffaella Di Napoli. Figure Contributed by Rohan Bir Singh, MD; Made with Biorender.com

Fig: Structure of COVID-19

B. Open Reading Frame

The open reading frame (ORF in short) is that region of the sequence of nucleotides which stretches from the start codon (which is the first three nucleotides of a nucleotide sequence from where the process of translation, or conversion of a DNA to protein starts, the nucleotide sequence for a start codon is usually ATG) to the stop codon (the region where the process of translation stops).

One usually searches for the open reading frame of an organism's gene sequence to find and pinpoint the location a particular gene in that organism. Usually, it is seen that the open reading frame of a DNA will start with the codon "ATG", though this may not always be the case, and will end with any of the three codons with the following sequences: TAA, TGA or TAG, which are accordingly referred to as 'stop codons'. Depending on the starting point, there are six possible ways of translating a sequence of nucleotides into its' respective amino acid sequence according to the genetic code.

The six ways are as follows: three on forward strand and three on complementary strand.

However, it must be noted that the start and ending of the ORF cannot be equated to the ends of the messenger RNA (mRNA). On the other hand, they are contained within the mRNA itself. In the case of a gene, ORFs are present, as mentioned earlier, between the start and stop codons. ORFs are found when traversing through the regions of a DNA in search of a specific gene. The ORFs are identified differently based on the variations of the start code sequence that can take place in an organism's genetic code due to mutation. A typical ORF finder will, based on the facts mentioned above, make use of algorithms which are based on existing genetic codes (including the altered ones) and all reading frames.

It has been found that the existence of an ORF is a good indicator of the location of a gene in the surrounding sequence around the desired gene. In this situation, the open reading frame constitutes a part of a longer sequence that is translated by the ribosomes, if it is long, which is noticed more in the case of prokaryotes (organisms without a well-defined nucleus), and in the, the ORF be present in sporadic gaps called introns. However, shorter ORFs can also be present outside the functional genes. However, these ORFs will be much shorter and will terminate faster.

After the sequencing of a gene, it is important to the correct ORF is determined so that the desired gene wherein mutation is believed to be taking place is located. The open reading frame is usually considered as the longest sequence without a gap or a stop codon. This scenario is noticed in eukaryotic messenger RNA which tends to have only one ORF. A problematic situation arises in the case of eukaryotic pre-mRNA wherein large parts of DNA are not properly translated. When the objective is to find ORFs for eukaryotic mRNA it is necessary to have a look at the spliced or split RNA first.

Fig: All possible open reading frames for a given genome sequence

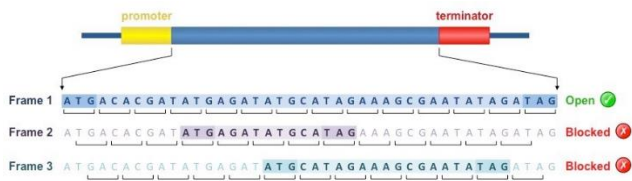


Fig: Usage of ORF in gene identification

C. Origin of Replication

The origin of replication is a specific sequence in the genome that initiates its' replication process. The replication of the genetic material and its' transfer to the next generation needs to be synchronized, and it must be semiconservative in nature so that the transfer of genes takes place appropriately, and each daughter cell receives the complementary chromosomes. The origin of replication helps in achieving these purposes. Also, since the origin of replication is the place where replication is initiated, therefore any mutation that occurs at or near the origin of replication has an extremely high chance of being transferred to the next generation. The given diagrams show the origin of replication and how it initiates replication, in the abacterial and eukaryotic chromosomes, respectively.

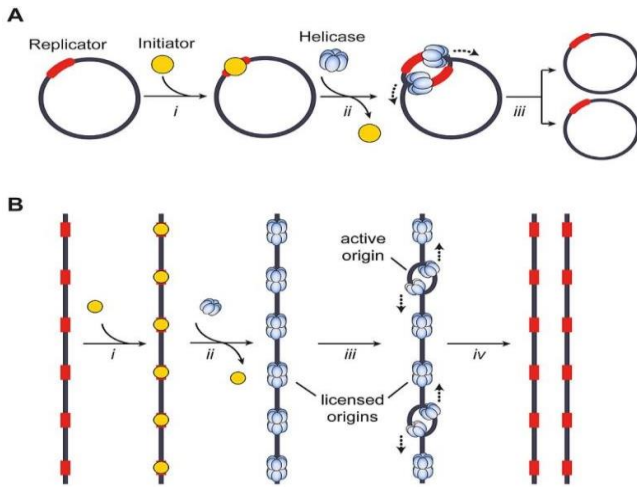


Fig : Origin of Replication

I. Type of Mutations

There are many types of mutations due to many reasons. Mutations can be harmful, have a mild effect or not affect anything.

The mutations we will be discussing are Point mutations which are caused by small shift of nucleotides.

They can give rise to 4 kinds of mutations.

1.Silent Mutations which are caused by substituting nucleotides in genome base which do not cause any affect or change in amino acid sequence as it forms a codon mapping to the same peptide.

Possible Amino Acid Sequences (Forward)	RSRAFYSPHSAAOSS*KAAPFTTHRASHRQPTAK
Nucleotide Sequence	ATGACACGATATGAGATATGCATAGAAAGCGAATATAGATAG
Possible Amino Acid Sequences (Reverse)	ATGACACGATATGAGATATGCATAGAAAGCGAATATAGATAG
Gene 1	ATGACACGATATGAGATATGCATAGAAAGCGAATATAGATAG
Gene 2	ATGACACGATATGAGATATGCATAGAAAGCGAATATAGATAG

2. Missense Mutations where a single amino acid is altered in sequence changing the function of a protein by a small margin but significant enough to be considered a variant.

3. Nonsense mutations are caused when insertion or inversion of a nucleotide causes a premature stop codon and shortening the Orf of the sequence and when in relation to structural proteins such as S protein in covid may affect their structural integrity.

4. **Frame Shift mutations** which is a significant mutation which causes a chain reaction in the reading frame due to a single codon change and changes many codons on the genome.

II. Methodology

1. Gathering of genomic sequence of COVID-19 and its' variants in notepad files: First we gather the genomic sequences of the original COVID-19 virus and its' variants and then save them in separate .txt files of notepad. This helps our Python program, which is written in Google Collab (an online Python notebook), to read the .txt file and run the code on it

2. Implementation of algorithm to identify origin of replication: Next, we run a series of algorithms present within our Python file. These include an algorithm to identify the origin of replication within the genetic sequence, which we perform since we know that those mutation that take place within the section of the genetic sequence which is close to the origin of replication has a higher chance to get transferred to the next generation of the COVID-19 virus, thus creating a variant of COVID-19.

3. Alignment of the genomes using global alignment to account for gaps and mismatches due to Frame Shift mutations using Needleman-Wunsch algorithm.

4. Transcription and Translation of DNA to obtain the protein sequence with multiple ORFs without filtering.

5. ORF's stored in an array by identifying start and stop codons and filtered to obtain the significant codons by length of the codons as the smaller proteins are not registered for a nonstructural function in NCBI Smart Blast search.

6. We compare the ORF of genomes obtained with each other to identify mutations


III. DATASET

1.creation of the dataset: To train the algorithm, we create a dataset, which contains the gene sequence of the original COVID-19

virus which started in Wuhan, China, and other variants of COVID-19

Which have sprung up in various parts of the world, such as USA, India, and so on. We find these gene sequences on the NCBI website. The given pictures show an example of how the dataset is created which contains the gene sequence for the original COVID-19 virus and its' variants.

1.The data hub in the NCBI site from where we find the gene sequence of COVID-19 and its' variants:



NCBI Virus

Impulses for Discovery

[About Us](#)
[Find Data](#)
[Help](#)
[How to Participate](#)
[Submit Sequences](#)

SARS-CoV-2 Data Hub

[Download](#)
[Quick Links](#)
[EpiCoV: SARS-CoV-2 Genomic Data](#)
[SARS-CoV-2 Articles in PubMed](#)
[NCBI SARS-CoV-2 Resources](#)

[Tabular View](#)
[Dashboard Visualizations](#)
[Mutations in SRA](#)
[Complete Tree](#)

[Selected Results: 0](#)
[Align](#)
[Build Phylog](#)

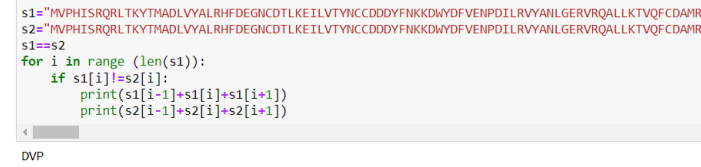
Nucleotide (3,533,763)		Protein (270,686,291)		RefSeq Genome (1)			
Accession	Submitters	Release Date	Pangolin	Isolate	Species		
NC_045512	Wu, F., et al.	2020-01-13	B	Wuhan-Hu-1	Severe acute respiratory s...	1	1
OM649125	Kurono, M., ...	2022-01-24	B.1.1.214	HirofH04	Severe acute respiratory s...	1	1
OM649126	Kurono, M., ...	2022-01-24	B.1.1.214	HirofH06	Severe acute respiratory s...	1	1
OM649127	Kurono, M., ...	2022-01-24	B.1.1.214	HirofH08	Severe acute respiratory s...	1	1
OM649128	Kurono, M., ...	2022-01-24	B.1.1.214	HirofH17	Severe acute respiratory s...	1	1
OM649129	Kurono, M., ...	2022-01-24	B.1.1.214	HirofH18	Severe acute respiratory s...	1	1
OM649130	Kurono, M., ...	2022-01-24	B.1.1.214	HirofH22	Severe acute respiratory s...	1	1

2. An example of the gene sequence of the original COVID-19 virus which is found on the NCBI data hub:

[illegible]

[illegible]

Fig: ORFs of Protein sequences of all variants



```
[0, 7, 21, 20, 7, 21]
[5, 0, 19, 18, 4, 19]
[65, 65, 0, 7, 65, 8]
[63, 63, 6, 0, 63, 9]
[6, 5, 20, 19, 0, 20]
[65, 65, 8, 10, 65, 0]
```

```
>>>
= RESTART: C:\Users\B SRIVATHSAN\AppData\Local\Programs\Python\Pyt
de.py
[56, 57, 58]
[236, 237, 238]
[226, 227, 228]
[214, 215, 216]
[208, 209, 210]
[198, 199, 200]
30130
30277
30271
30234
30262
30220
T A G
T A G
T A G
T A G
T A G
T A G
```

The Origin of Replication of the aligned sequences begin around the position of 210th nucleotide and are signified by a stop codon.

The GC Skew of all the sequences are of a similar amount despite the mutations and thus the hydrogen bonding structural integrity of the variants seem to have no significant effect.

The Biggest gene sequence of ORF1ab has undergone point mutations between the variant groups of Original, Alpha, Delta while maintaining most other genomes to be the same.

The gene of ORF1a, like ORF1ab has undergone point mutations between the variants of beta gamma and delta.

BOTH THE ORF1AB AND ORF1A PERFORM SYNTHESIS AND FUNCTIONS OF NONSTRUCTURAL PROTEINS IN THE REGION OF SUPPRESSING HOST IMMUNITY AND

response to growth.

Comparing the protein sequences, the mutations occur between Valine and Isoleucine which plays a part in immunity and strength of the host and thus a change in the peptide causes a difference response.

This mutation is due to Isoleucine and Valine being an important part of immune response and by adapting to it the virus tries to fool the immune system.

CONCLUSION

We have compared 5 Covid Genome variants by transcription and translating the given DNA sequence and identified ORF's.

We identified 5 genomes responsible for the hostility of the virus in their mutated form among different variants.

The Original, Alpha and Delta variants of the virus have undergone nonsense and missense mutations in their ORF1ab and ORF7 for suppression of host immunity by synthesis of nonstructural protein among themselves. Whereas they have undergone frame shift mutations with Beta Gamma Omicron and caused a lot of mutations in NS-p1 which is a protein derivative of ORF1ab and ORF1a which performs their functions. The other orfs such as ORF3a have influenced the life cycle incubation period and replication factor of the virus. We also compared the GC skew as they signify the oric and structural integrity of the virus due to the high no of hydrogen bonds shared between

them and noticed that the oric starts from the same place and the structural integrity of these strains are similar. We have arrived at the conclusion that the hostile mutations in the virus are of the ORF ORF1ab ORF3a ORF7 whereas the other mutations involve the envelope E protein and N protein which are studied for suppressing the virus by weakening its spike protein effect.

ACKNOWLEDGMENT

We would like to thank our course instructor Dr Amrita Thakur and Dr Nidhin Prabhakar T V for providing us with the opportunity to pursue this project.

REFERENCES

- Altman, Eleonora, et al. "Effect of the HP0159 ORF mutation on the lipopolysaccharide structure and colonizing ability of *Helicobacter pylori*." *FEMS Immunology & Medical Microbiology* 53.2 (2008): 204-213.
- [2] Delbue, Serena, et al. "Isolation of SARS-CoV-2 strains carrying a nucleotide mutation, leading to a stop codon in the ORF 6 protein." *Emerging microbes & infections* 10.1 (2021): 252-255.
- [3] Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, Ming-Li Yuan, Yu-Ling Zhang, Fa-Hui Dai, Yi Liu, Qi-Min Wang, Jiao-Jiao Zheng, Lin Xu, Edward C Holmes, Yong-Zhen Zhang. A new coronavirus associated with human respiratory disease in China, *Nature* 2020 Mar;579(7798):265-269. doi: 10.1038/s41586-020-2008-3. Epub 2020 Feb 3.
- [4] Marta Giovanetti , Francesca Benedetti , Giovanni Campisi , Alessandra Ciccozzi , Silvia Fabris , Giancarlo Ceccarelli , Vittoradolfo Tambone , Arnaldo Caruso , Silvia Angeletti , Davide Zella , Massimo Ciccozzi . Evolution patterns of SARS-CoV-2: Snapshot on its genome variants. *Biochem Biophys Res Commun*. 2021 Jan 29;538:88-91. . Doi: 10.1016/j.bbrc.2020.10.102. Epub 2020 Nov 6.
- [5] <https://www.frontiersin.org/articles/10.3389/fimmu.2021.708264/full>