

Analyze the genome of a virus to identify mutations and predict their potential impact

By Gunja

Overall goal of the project:

1. Find the genes contained in a viral genome
2. Try to find the function of those genes
3. Identify the closest relatives of the virus
4. Identify if the viral genome contains any mutations, and
5. Try to predict if those mutations are likely to affect the function of the virus

TASK 1 - Strategy and Planning (27 & 28th February 2023)

In the first task, you must identify the conceptual knowledge, methods and tools, and resources required for completing the research project. Carefully read the project statement and description and try to answer the following questions:

1. What are the milestones and deliverables of the project?
 - a. Find the genes contained in a viral genome
 - b. Try to find the function of those genes
 - c. Identify the closest relatives of the virus
 - d. Identify if the viral genome contains any mutations, and
 - e. Try to predict if those mutations are likely to affect the function of the virus
2. What are the concepts that you need to know for different steps of the project?
 - a. Genomics: The study of genomes, including their structure, function, and evolution. For example different techniques and methods to analyze the genome sequence - DNA sequence.
 - b. Understand how the virus functions, how its genes function.
 - c. Bioinformatics: Understanding how we can analyze and compare long gene sequences efficiently and effectively - to be able to look at a database and find what we need.
 - d. Molecular Biology: Understanding and analyzing the structure and function of proteins and nucleic acids.
 - e. Protein Function and Structure: Understanding the functions of proteins and their structures is critical to understanding the function of viral genes.
 - f. Mutations and their Effects: Understanding the impact of mutations on gene function is essential to predicting the effect of any mutations identified in the viral genome.

3. What are the resources you need, in terms of expert help, books, online resources etc?
 - a. Access to a database of different genes of viruses and their close relatives.
 - b. Expert help in being able to program and code to make software if required.
 - c. Online textbooks (NCERT) to study the structure and function of viruses.
4. What are the tools and methods you may have to use?
 - a. PDB database
 - b. Blast
 - c. Genome bank (a place from where we can get the genome/gene of specific virus)
 - d. ChatGPT

Task	Milestones	Input	Deliverables	Tools required	Timeline
1	Details of brainstorming and planning of the project - strategy and planning	List of milestones we plan to achieve	Detailed plan and timeline of the project	-	1 day
2	Details of the genes contained in a viral genome	Virus and its genome	<ul style="list-style-type: none"> - Name of the virus, about the virus <ul style="list-style-type: none"> - Is it single stranded? - Positive strand/negative strand? - What type of genome - DNA/RNA? - Name of the tool used <ul style="list-style-type: none"> - What databases (if any) or tools did you use and why? - What does the database contain and how does the tool work (in brief)? - Filters chosen to find the genes on ORF filter - parameters/input - Sample submission date, accession number, link from where found, etc - Number of possible genes/ORFs - Sequence of the all the predicted genes contained in the genome - length of the gene and genome - In which frames is the gene (in which frame is the gene) - Start and stop codons - Location of the gene - All major observations from the results e.g. summary numbers/statistics. Interpretation. 	<ul style="list-style-type: none"> - NCBI Virus Variation Resource (ViPR) - ViralZone - Viral Genomes - European Nucleotide Archive (ENA) - PDB database - NCBI (National library of Medicine) - BLAST - Viral Genome DataBase - Genome Detective - ChatGPT 	2 days
3	Details of structure and	Gene sequence -	Structure: - Number of G-C and A-T bonds (Ratio)	- PDB database	2 days

	function of those genes - proteins	protein sequence it codes for	Function: - What protein does the gene code for? - Is it an enzyme/antibody/Structural protein/transport protein/etc - What is the structure of the protein and what does it do? - What part of the protein is beta sheet/alpha/loop - Start and stop codon		
4	Details of the closest relatives of the virus	Genome sequence of the virus	- Name of the virus closely related, its sequence, ratio of similarity - Identify and compare the genome sequence (ATGC) - Compare the ratio of the nucleotides - Look for relative viruses which have almost same ratio of ATGC and almost at the same location - What fraction of the nucleotides is the same.	- BLAST - PDB Database	1 day
5	Details of whether the viral genome contains any mutations or not	Genome sequence	- Compare the viral genome with its closest relative - If there is a mutation, then provide the location or the nucleotides where the mutation has occurred.	- PDB Database - BLAST - Biostars	2 day
6	Details about if the mutation in the genome affects the functioning of the virus	Mutated and unmutated genome structure	- What effects does the mutation have on the function of the genes? - does the mutant gives result to a different protein/or if it even codes for one - difference in symptoms		2 days
	Rest of the time - documentation (i.e 0 days)				

TASK 2 - Identification of Open Reading Frames (ORFs) in Chikungunya Virus Isolate SK003/12 from Southern Thailand

In this study, we used the ORF Finder tool to identify potential protein-coding regions in the complete genome of the recently identified Chikungunya virus isolate SK003/12, which was identified in Southern Thailand.

Virus genome and its isolate (Southern Thailand)

The Chikungunya virus (CHIKV) is an RNA virus that belongs to the family *Togaviridae*, genus *Alphavirus*. Its genome is a single-stranded positive-sense RNA molecule approximately 11.8 kilobases (kb) in size, encoding a polyprotein precursor that is processed into both non-structural and structural proteins [\[1\]](#).

The non-structural proteins, nsP1-nsP4, are responsible for viral replication and transcription. nsP1 is involved in the capping of viral RNA, while nsP2 has RNA helicase and protease activities. nsP3 is a multifunctional protein involved in RNA binding, replication, and host immune modulation, while nsP4 is the RNA-dependent RNA polymerase that synthesizes the viral RNA [\[2\]](#).

In general, the CHIKV genome contains two open reading frames (ORFs) that are translated into the polyprotein precursor [\[3\]](#). The first ORF encodes the non-structural proteins, while the second ORF encodes the structural proteins. The virus is transmitted to humans primarily by the *Aedes* mosquito, and symptoms include fever, joint pain, rash, and muscle pain.

Tools and database used

One of the Chikungunya isolate SK003/12 has been identified and is available on the NCBI GenBank database under accession number ON262791.1 and was published on 06-FEB-2023. Based on the information available on the National Center for Biotechnology Information (NCBI), the complete genome of the Chikungunya virus isolate SK003/12 is 11618 base pairs in length and is of the RNA, linear genome type. The accession number for this genome is ON262791, and the version number is ON262791.1. The source of this genome is the Chikungunya virus itself, and the organism is also the Chikungunya virus. Furthermore, the first 10 bases of the genome of this isolate is CTACCAGTTT [\[4\]](#).

Working of the tool and database

Genes refer to sequences that encode for proteins or other molecules, while non-coding sequences typically code for molecules such as tRNA, rRNA, and ncRNAs. ORFs are sequences that have both a start codon and a stop codon within the same frame. However, ORFs are not the same as genes as genes also contain other components used for protein

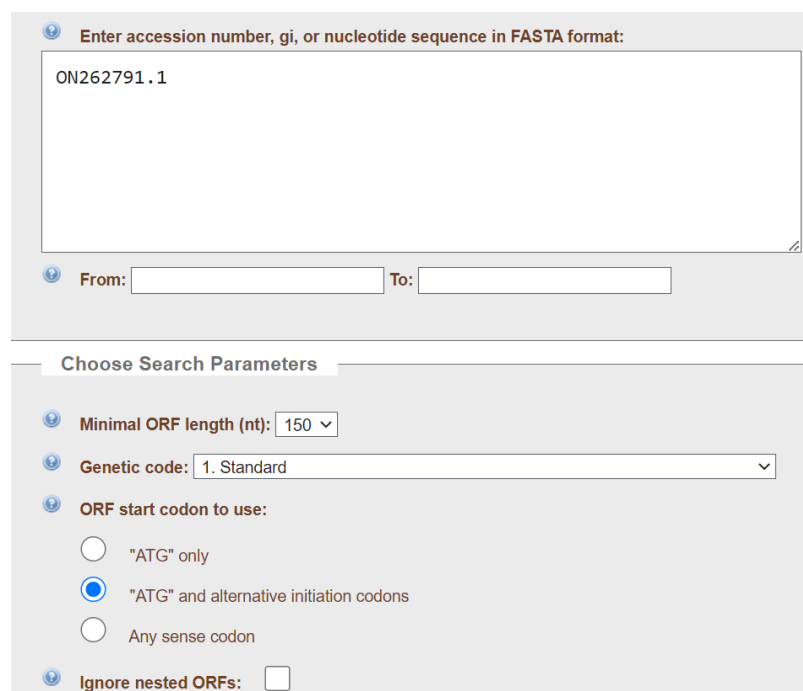
synthesis, including promoter regions, transcription start and termination sites, and a translation initiation site known as the Kozak region.

So, to determine which ORFs are potentially genes, we can use a tool which searches for start codons and after the stop codons in a frame. One of such tool is the ORF Finder available on NCBI.

We used the ORF Finder and the nucleotide database, a web-based tool to identify open reading frames (ORFs) within a nucleotide sequence. It is available on the National Center for Biotechnology Information (NCBI) website and other molecular biology websites. ORFs are stretches of nucleotide sequence that can be translated into proteins.

The ORF Finder identifies ORFs by scanning the input sequence in all six possible reading frames and identifying regions that begin with a start codon (usually AUG) and end with a stop codon (UAA, UAG, or UGA). The tool can also specify the minimum ORF length and the genetic code used for translation.

Parameters chosen on the ORF finder



The screenshot displays the NCBI ORF Finder web interface. At the top, there is a text input field labeled "Enter accession number, gi, or nucleotide sequence in FASTA format:" containing the text "ON262791.1". Below this field are "From:" and "To:" input boxes. A section titled "Choose Search Parameters" contains several settings: "Minimal ORF length (nt):" is set to "150" with a dropdown arrow; "Genetic code:" is set to "1. Standard" with a dropdown arrow; "ORF start codon to use:" has three radio button options: "ATG" only, "ATG" and alternative initiation codons (which is selected), and "Any sense codon"; and "Ignore nested ORFs:" is an unchecked checkbox.

In our case, the input parameters used for the analysis were an accession number of ON262791, a minimal ORF length of 150 nucleotides, and a genetic code of standard.

The minimal ORF length of 150 nucleotides is the minimum length of a potential protein-coding region that the ORF Finder will consider. This means that any ORF identified by the tool must be at least 150 nucleotides in length to be reported. The genetic code used in the analysis was the

standard genetic code. The genetic code specifies how nucleotide triplets (codons) are translated into amino acids, which are the building blocks of proteins. The standard genetic code is the most commonly used genetic code and is shared by most organisms.

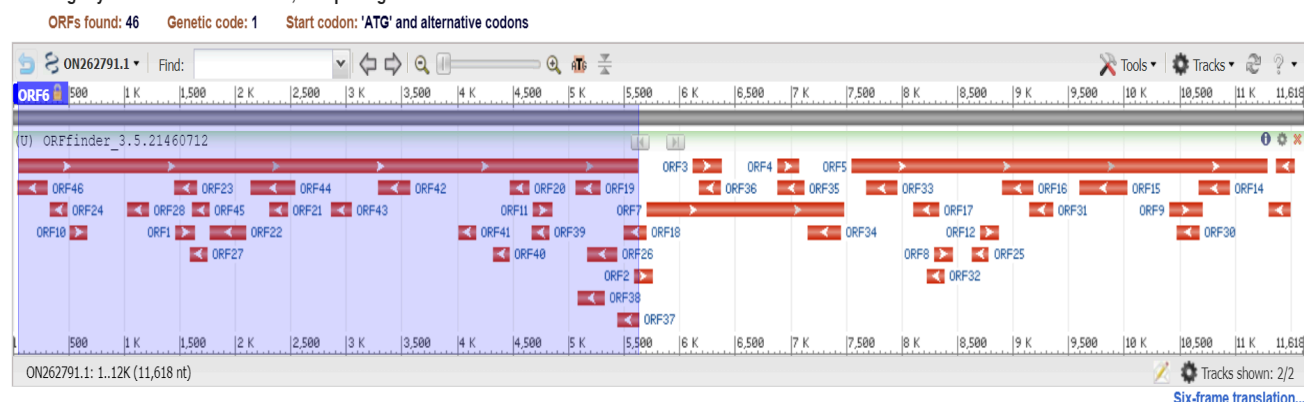
Further, the ORF Finder was instructed to use "ATG" and alternative initiation codons as the start codons to identify potential ORFs. The start codon is the nucleotide triplet that signals the beginning of a protein-coding sequence. The ORF Finder can be set to look for specific start codons, or a combination of start codons. Finally, the option to ignore nested ORFs was unselected. Nested ORFs are potential protein-coding regions that are located within a larger ORF. By default, the ORF Finder will report both nested and non-nested ORFs. However, in this case, the option to ignore nested ORFs was unselected, meaning that any nested ORFs that are identified will also be reported.

Genes found from the ORF Finder

Upon setting the different parameters on the tool, for Chikungunya virus isolate SK003/12, complete genome, the ORF Viewer identified about 46 ORFs [5]. However, as the chikungunya virus has a positive single stranded genome, not all of these ORFs are relevant.

As a result, we eliminated all the negative strand genes, and are only considering the genes present on the positive strand (which are 12 possible ORF's) for our further observation and interpretation.

Chikungunya virus isolate SK003/12, complete genome



Following are the different positive frame ORFs in a descending order based on the length of the genes.

Sr. no	Possible gene	Total Amino acids	Frame of the gene	Location of the gene	Length of the gene	Start codon	Stop codon
1	ORF6	1856	+ 2	53 - 5623	5571	ATG	TGA

2	ORF5	1248	+ 1	7543 - 11289	3747	ATG	TAA
3	ORF7	590	+ 2	5705 - 7477	1773	CTG	TAG
4	ORF9	99	+ 2	10403 -10702	300	ATG	TAA
5	ORF3	89	+ 1	6112 - 6381	270	TTG	TAG
6	ORF4	63	+ 1	6880 - 7071	192	CTG	TGA
7	ORF1	60	+ 1	1459 - 1641	183	ATG	TGA
8	ORF11	58	+ 3	4680 - 4856	177	CTG	TGA
9	ORF2	56	+ 1	5587 - 5757	171	CTG	TAA
10	ORF8	55	+ 2	8288 - 8455	168	TTG	TGA
11	ORF12	53	+ 3	8706 - 8867	162	ATG	TGA
12	ORF10	51	+ 3	513 - 668	156	ATG	TGA

From the above data,

- Out of all the 46 possible ORFs, there are 12 possible positive frame ORFs.
- There is variation in the length of the genes, with the longest gene (ORF6) being 5571 nucleotides long and encoding a protein with 1856 amino acids, while the shortest gene (ORF10) is 156 nucleotides long and encodes a protein with 51 amino acids.
- Overall, we can see that the maximum number of ORFs are on the + 1 frame
- There are 3 ORFs which have a length of more than 1000 base pairs, out of which 2 are on the +2 frame.
- The ORFs on frame +3 have a length of less than 200 base pairs.
- The start codons for each gene are mostly ATG, which is the most common start codon in protein-coding genes. However, there are some exceptions, such as ORF3, which starts with TTG.
- The stop codons for each gene are either TAA, TAG, or TGA, which are the three stop codons commonly used to signal the end of a protein-coding region.

Interpretations

The data provided in the table gives us important information about 12 different genes, including their location, length, and protein-coding sequence. This information can be used to make several interpretations about the genetic makeup of an organism.

For example, the lengths of the genes vary considerably, which suggests that these genes would give different lengths of proteins (amino acids) and so different functions. Longer genes

may be involved in more complex biological processes or may encode proteins with more domains and functions than shorter genes. The variation in start codons observed in the table may also indicate differences in the way these genes are regulated or transcribed.

The location of each gene given in the table is also an important piece of information as it indicates the specific location of the gene on a chromosome or genome. The location can be used to identify other nearby genes that may be co-regulated or co-expressed, and to study the organization and structure of the genome.

In addition, the location of the gene may be used to identify potential regulatory elements that control the expression of the gene, such as promoters and transcription start and end sites. .

Task 3: Analysis of Open Reading Frames (ORFs) in Chikungunya Virus Genome to find the closely matching proteins: Structure and function of genes

In this analysis, we aimed to investigate the structure and function of genes in the Chikungunya virus genome by examining the Open Reading Frames (ORFs) and identifying the closely matching protein sequences using the Swiss-Prot database.

To determine whether the open reading frames (ORFs) found in the previous section are coding genes that result in functional proteins, we can use the Swiss-Prot database to search for closely matching protein sequences in the Chikungunya virus. This database contains manually curated, high-quality protein sequences with detailed functional annotations.

Tools and database used

To do this, first we can take the amino acid sequences of the corresponding ORFs identified from the ORF finder in the previous section. Then, we can use a tool such as BLAST to search the Swiss-Prot database for similar sequences of the protein in the other isolates of the Chikungunya viruses. If a closely matching protein sequence is found with a well-characterized function, this would suggest that the corresponding ORF is a coding gene that results in a functional protein.

However, it is also possible that the ORF is a non-coding RNA, which would not have a corresponding protein sequence in the Swiss-Prot database. In this case, further analysis would be needed to determine the function of the putative ORF.

Working of the tool and database

To find the closely matching protein, we had multiple options of protein databases and tools to use. For instance some of the most commonly known protein databases that can be used for

our purpose are Swiss-Prot and UniProt, and tools such as Blast and SmartBlast (available on the ORF finder results page).

ORF6 (1856 aa) [Display ORF as...](#) Mark

```
>1c1|ORF6
MDPVYVDIDADSAFLKALQRAYPMFEVESRQVTPNDHANARAFSHLAIKL
IEQEIDPDSTILDIGSAPARRMMSDRKYHCPCMRSAEDPERLANYARKL
ASAAGKVLDRNISGKIGDLQAVMAVPDKETPTFCLHTDVSCRQADVAIY
QDVYAVHAPTSLYHQAIGVRVAYWVGFDTPFMYDAMAGAYPSYSTNWA
DEQVLKAKNIGLCSTDLTEGRRGKLSIMRGKKLKPCDRVLFVSGSTLYPE
SRKLLKSWHLPSVFHLKGKLSFTCRCDTVVSCGYVVKRITMSPGLYGKT
TGYAVTHHADGFLMCKTTDTVDGERVSFVSCTYVPATICDQMTGILATEV
TPEDAQKLLVGLNQRIVVNGRTQRNMNTMKNYLIPVVAQAFSKWAKECRK
DMEDEKLLGVRERTLTCCCLWAFKKQKTHTVYKRPDTSIQKVQAEFDSF
VVPSSLWSSGLSIPLRTRIKWLLSKVPKTDLIPYSGDAREARDAEKEEEE
REAELTREALPPLQAAQEDVQVEIDVEQLEDRAAGIIETPRGAIKVTAQ
```

ORF6
SmartBLAST
BLAST

Marked set (0)
[SmartBLAST best hit titles...](#)
BLAST

BLAST Database:

UniProtKB/Swiss-Prot (swissprot) ▼

In our case we decided to check for the protein sequences from the Swiss-Prot database. This is because, "Swiss-Prot," which is a high-quality, manually curated protein sequence database maintained by the Swiss Institute of Bioinformatics (SIB) and is often used as a reference in bioinformatics research. Compared to UniProt, which is a larger database that includes both manually annotated and computationally predicted protein sequences, Swiss-Prot is known for its higher level of annotation quality and accuracy. Swiss-Prot is also updated more frequently than UniProt, which may make it more up-to-date for certain research purposes.

Additionally, we used BLAST , which is a general-purpose sequence alignment tool that can be used to compare any two protein sequences. It searches for regions of similarity between two sequences by aligning them and calculating a statistical significance score (percentage identity). BLAST can be used to identify homologous sequences, infer evolutionary relationships, and annotate functional domains in protein sequences.

BLAST works by comparing the query sequence against the Swiss-Prot database of sequences and identifying regions of similarity using local alignment algorithms. The degree of similarity is represented by a score, which takes into account factors such as the length of the alignment, the number of gaps, and the identity of the matching residues.

Parameters chosen for BLAST

Initially, to find the closely matching functional proteins corresponding to the amino acid sequence to the ORFs we used the following parameters

- Database: UniProtKB/Swiss-Prot(swissprot)
- Organism - "Chikungunya virus (taxid: 371094)
- Algorithm - blastp (protein-protein BLAST)
 - Additional algorithm parameters were:
 - Max target sequence: 10
 - Matrix:BLOSUM90
 - Existence:11 Extension:3

All the other parameters were kept as default.

To find closely matching functional proteins corresponding to the amino acid sequence of the putative ORFs, we used the UniProtKB/Swiss-Prot database with the organism specified as "Chikungunya virus" (taxid: 371094). We selected the blastp algorithm, which is a protein-protein BLAST algorithm that compares the query amino acid sequence against a protein sequence database. We set the maximum number of target sequences to 10, meaning that the BLAST search would only return up to 10 closely matching protein sequences from the database.

We also specified the BLOSUM90 matrix for the search, which is a substitution matrix that assigns scores to amino acid substitutions based on their likelihood of occurring in evolutionarily related proteins. We set the existence and extension gap penalties to 11 and 3, respectively. These parameters determine the penalties assigned to gaps in the alignment and the minimum score required to extend the alignment.

Overall, these parameters were selected to ensure that the BLAST search would return highly similar protein sequences from the UniProtKB/Swiss-Prot database with a well-characterized function, and to minimize the number of false positives in the search results.

Closely matching proteins found using BLAST

Upon selecting the different parameters mentioned above, the BLAST results are as follows:

Sequences producing significant alignments

Download

Select columns

Show10

select all

0 sequences selected

GenPept

Graphics

Distance tree of results

Multiple alignment

MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input type="checkbox"/>	RecName: Full=Polyprotein P1234; Short=P1234; AltName: Full=Non-structural polyprotein; Contains: RecName...Chikungunya vir...		4199	4199	100%	0.0	98.49%	2474	Q8JUX6
<input type="checkbox"/>	RecName: Full=Polyprotein P1234; Short=P1234; AltName: Full=Non-structural polyprotein; Contains: RecName...Chikungunya vir...		4046	4046	100%	0.0	95.58%	2474	Q5XXP4

ORF No.	Description	Max score	Total score	Query cover	Accession length	E value	Percentage identification	Positives	Gaps	Uniprot ID
ORF 6	Polyprotein P1234	4199	4199	100%	2474	0.0	1828/1856 (98%)	1840/1856 (99%)	0/1856 (0%)	Q8JUX6
ORF 5	Structural polyprotein	2816	2816	100%	1248	0.0	1223/1248 (98%)	1232/1248 (98%)	0/1248 (0%)	Q8JUX5
ORF7	Polyprotein P1234	1329	1329	100%	2474	0.0	583/590 (99%)	586/590 (99%)	0/590 (0%)	Q8JUX6
ORF9	No significant similarity found.									
ORF3	No significant similarity found.									
ORF4	No significant similarity found.									
ORF1	No significant similarity found.									
ORF11	No significant similarity found.									
ORF2	No significant similarity found.									
ORF8	No significant similarity found.									
ORF12	No significant similarity found.									
ORF10	No significant similarity found.									

From the above data,

- ORF 6 was identified as a coding gene that encodes for a polyprotein P1234, with a maximum score of 4199 and 100% query coverage.
- ORF 5 was also identified as a coding gene that encodes for a structural polyprotein, with a maximum score of 2816 and 100% query coverage.
- ORF 7 was found to be a coding gene that encodes for a polyprotein P1234, with a maximum score of 1329 and 100% query coverage. The UniProt ID for the closely matching protein sequence was also Q8JUX6, which suggests that ORF 7 may be a different isoform of the same protein as ORF 6.

In contrast, several other putative ORFs, including ORF 9, 3, 4, 1, 11, 2, 8, 12, and 10 did not have any significant similarity to known protein sequences in the UniProtKB/Swiss-Prot

database. This suggests that these ORFs may not be coding genes, or that they encode for proteins with unknown functions that have not yet been characterized in the database.

To further investigate the ORFs that did not have any significant similarity to known protein sequences in the UniProtKB/Swiss-Prot database, we attempted to modify the following BLAST search parameters:

- BLOSUM: changed from 90 to 50
- Gap Cost: changed from Existence:11 Existence:3 to Existence:13 Existence:3/Existence:15 Existence:1
- The Organism: Chikungunya virus strain S27-African prototype (taxid:371094) was selected.

However, these attempts did not result in any new matches for these ORFs.

Following is the description of the functions for the matched proteins for ORFs 5, 6 and 7:

ORF No.	Accession No.	Protein name that has matched with the ORF	Uniprot ID	Function of the closely matched protein
ORF 5	Q8JUX5.3	Structural polyprotein	Q8JUX5	Capsid protein - The capsid protein binds to the viral RNA genome at a site adjacent to a ribosome binding site for viral genome translation following genome release.
ORF 6	Q8JUX6.1	Polyprotein P1234	Q8JUX6	Inactive precursor of the viral replicase, which is activated by cleavages carried out by the viral protease nsP2.
ORF7	Q8JUX6.1	Polyprotein P1234	Q8JUX6	Inactive precursor of the viral replicase, which is activated by cleavages carried out by the viral protease nsP2.

From the provided information, we can observe that both ORF 6 and ORF 7 match with the same protein, Polyprotein P1234, which is an inactive precursor of the viral replicase, and is activated by cleavages carried out by the viral protease nsP2. This suggests that Polyprotein P1234 plays a crucial role in the replication of the virus.

Furthermore, it is possible that the protein is encoded by a single gene that is divided into two ORFs by a stop and start codon between ORF 6 and ORF 7. This could explain why both ORFs

match with the same protein, with ORF 6 matching the first half of the polyprotein and ORF 7 matching the other half of the same protein.

This observation highlights the importance of considering the possibility of overlapping ORFs in genome annotation and analysis. This is further supported by the following information:

ORF6	+	2	53	5623	5571 1856
ORF7	+	2	5705	7477	1773 590

Here we can see that while the ORF 6 begins from position 53, ORF 7 ends at position 6477. And in the image below (which is the nucleotide genome sequence data from NCBI for our chosen Chikungunya virus isolate), we can see that one of the CDS starts and ends from position 53 to 7477.

[CDS](#)

53..7477

```
/codon_start=1
/transl_except=(pos:5621..5623,aa:OTHER)
/product="nonstructural polyprotein"
/protein_id="WCI13793.1"
/translation="MDPVYVDIDADSAFLKALQRAYPMFEVESRQVTPNDHANARAFS
HLAIKLIQEIDPDSTILDIGSAPARRMMSDRKYHCVCPMRSAEDPERLANYARKLAS
AAGKVLDRNISGKIGDLQAVMAVPDKETPTFCLHTDVSCRQRADVAIYQDVYAVHAPT
SLYHQAIKGVRVAYWVGFDTPFMYDAMAGAYPSYSTNWADEQVLKAKNIGLCSTDLT
EGRRGKLSIMRGKLLKPCDRVLFVSGSTLYPESRKLLKSWHLPVHFHLKGKLSFTCRC
DTVVSCEGYVVKRITMSPGLYGKTTGYAVTHHADGFLMCKTTDTVDGERVSFSVCTYV
PATICDQMTGILATEVTPEDAQKLLVGLNQRIVVNGRTQRNMNTMKNYLIPVVAQAFS
KWAKECRKDMEDEKLLGVRERTLTCCCLWAFKKQKTHTVYKRPDTSIQKVQAEFDSF
VWPSLWSSGLSIPLRTRIKWLLSKVPKTDLIPYSGDAREARDAEKEAEEREAEELTRE
ALPPLQAAQEDVQVEIDVEQLEDRAAGIIETPRGAIKVTAQPTDHVVGEYLVLSPTQ
VLRSQLSLIHALAEQVKTC THNGRAGRYAVEAYDGRVLVPSGYAISPEDFQSLSESA
TMVYNEREFEVNRKLHHIAMHGPA LNTDEESYELVRAERTEHEYVYDQRRCKKEEA
```

Interpretations

The above variation can be due to multiple reasons such as:

- Mutation, that caused a substitution mutation and introduced a stop codon at the end of the ORF 6.
- Sequencing error.
- Codon variation, where in some viruses the ATG (stop codon) could be coding for another amino acid which is not a stop codon — Read through.

Task 4: Analysis of Mutations and Evolutionary Relationships in Chikungunya Virus Isolates using Sequence Alignment Tools

Moving on, now that we have found closely matching protein sequences for the coding ORFs (possible genes), we can use the findings to check whether our genome isolate contains any mutation or not. So, to determine whether a specific isolate of the Chikungunya virus contains any variations or mutations in its genome, we need to compare its genetic sequence to that of a reference sequence. The reference sequence is typically the genome of a well-characterized strain of the virus that has been previously sequenced and annotated.

Tools and database used

To do this, we can use different sequence alignment tools from NCBI as well as EBI bioinformatics databases. Sequence alignment tools are computational programs used in bioinformatics that compare and align two or more biological sequences, which is used to identify regions of similarity or differences (mutations) between the sequences and to determine the evolutionary relationship between them. There are two main types of sequence alignment: pairwise alignment, which aligns two sequences, and multiple sequence alignment, which aligns three or more sequences.

NCBI and EBI both provide a range of sequence alignment tools:

- Pairwise sequence alignment tools
 - NCBI BLAST: Basic Local Alignment Search Tool, used to compare a query sequence against a database of sequences to identify similar sequences.
 - NCBI Needleman-Wunsch algorithm: used for global pairwise sequence alignment.
- Multiple sequence alignment tools
 - EBI Clustal Omega, MUSCLE (Multiple Sequence Comparison by Log-Expectation), and MAFFT (Multiple Alignment using Fast Fourier Transform): used for multiple sequence alignment.
 - EBI EMBOSS Needle and Water: used for pairwise sequence alignment.

However, in our case we would be using some of the key ones such as BLAST and MUSCLE:

To do this, we used the BLAST ® blastn suite by pasting the ORF to be analyzed in the form of a nucleotide sequence. This allowed us to identify multiple matching genes from different strains of the virus. By clicking on the matching gene, we were able to view the pairwise alignment of the query and the subject on the BLAST ® blastn suite. This provided us with all the matching genes of the query.

In order to download the subject nucleotide sequence for multiple alignment sequence, we clicked on GenBank, which directed us to the detailed summary and annotations for the particular matching protein. From here, we downloaded the coding sequence in the format of FASTA nucleotide sequence. However, this gave us only a single matching nucleotide sequence at a time, so we used another method which was to select all the matching gene sequences and download them in the format of FASTA aligned sequence. We then added the nucleotide sequence for the input ORF in the same file and uploaded it on Clustal Omega. This allowed us to perform a multiple sequence alignment and compare the sequences for any mutations.

Finally, we viewed the results of the alignment on Jalview to look for mutations.

Working of the tool and database

As said earlier, BLAST® is a bioinformatics tool that is widely used for sequence alignment, which helps in identifying similarities between nucleotide or protein sequences. It searches for matching sequences by comparing the query sequence against a database of known sequences.

Clustal Omega is a bioinformatics tool that is used to perform multiple sequence alignment. It uses a fast and accurate algorithm to align multiple sequences simultaneously, and produces output in various formats, including FASTA format.

Finally, Jalview is a visualization tool that is used to view and analyze multiple sequence alignments. It allows users to visualize and manipulate sequences, and provides tools for identifying patterns and differences between sequences.

Parameters chosen on BLAST and MUSCLE

When using the tools such as BLAST and MUSCLE to find the best matching functional gene sequences to our ORFs, several parameters were chosen to optimize the search and improve the accuracy of the results.

For BLAST, the first parameter chosen was the ORF sequence, which is the nucleotide sequence that codes for a functional protein in the organism of interest. In this case, the organism was Chikungunya Virus (taxid: 37124). By providing the ORF sequence of interest, we can search for similar sequences in the Chikungunya Virus genome and identify potential functional gene matches.

The second parameter chosen was the maximum target sequence, which was set to 100 in this case. This parameter limits the number of sequences that BLAST will return as potential matches. By setting the maximum target sequence to 100, we could focus on the most relevant and significant matches and avoid being overwhelmed by an excessive number of results.

The third parameter chosen was the gap cost for scoring alignments. As described earlier, gap costs refers to the penalties assigned for introducing gaps in an alignment, which can occur when comparing two sequences that differ in length or contain insertions or deletions. In this case, the extension gap cost was set to 5, while the opening gap cost was set to 2 (Scoring parameters (Gap cost) - Extension: 5 Extension: 2).

For Clustal Omega, to find the mutations in the top 100 matching sequences and our ORF, we began by obtaining the nucleotide sequence of the ORF (in FASTA format) and all the matching functional gene nucleotide sequences from BLAST.

Next, we changed the output format to ClustalW with character counts to ensure that we would be able to accurately identify any mutations that we found. We also changed the order parameter from aligned to input so that the output would be in the same order as the input sequences.

With these parameters set, we ran Clustal Omega and analyzed the results to identify any mutations in the top 100 matching sequences and our ORF.

Mutations/Variations found

From the above data,

Based on the data provided, it appears that the ORF5 (first gene) has only one mutation at position 3728 where T is substituted with G when compared to the 100 closest matching gene sequences obtained from BLAST.

On the other hand, the ORF6-7 (second gene) combined into one protein, has multiple mutations at different positions when compared to the 100 matching closest gene sequences obtained from BLAST. These mutations are substitution mutations and occurred at the following positions:

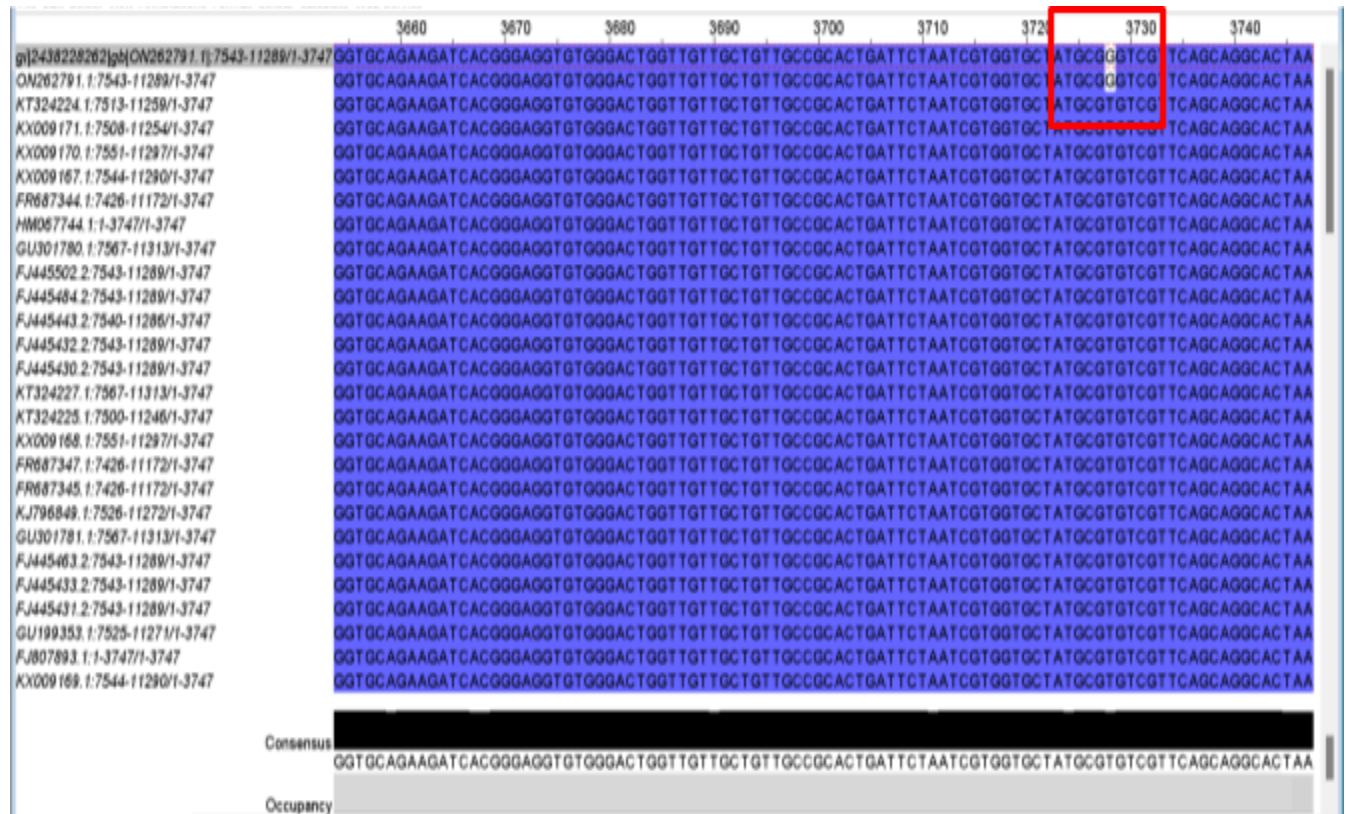
ORF Number	Position of variations	Type of variation	Original nucleotide (based on Parsimony rule)	Mutation occurred - changed nucleotide	Types of Substitutions mutations
ORF 5	3728	Substitution	T	G	Transversions
Mutation in other genes which are similar to ORF5	23	Substitution	A	G	Transition
	86	Substitution	C	T	Transition
	114	Substitution	T	C	Transition

	198	Substitution	A	G	Transition
	290	Substitution	A	G	Transition
	294	Substitution	G	A	Transition
	519	Substitution	C	T	
	791	Substitution	C	T	
	1729	Substitution	C	A	
	1851	Substitution	A	G	
ORF 6 - 7 combined	85	Substitution	C	T	Transition
	556	Substitution	A	G	Transition
	1150	Substitution	C	A	Transversions
	1491	Substitution	A	C	Transversions
	4141	Substitution	A	G	Transition
	4179	Substitution	C	T	Transition
	5835	Substitution	T (52.9% chances on the nucleotide being T)	A (47.1% chances on the nucleotide being A)	Transversions
	6081	Substitution	T (52.9% chances on the nucleotide being T)	C (47.1% chances on the nucleotide being C)	Transition

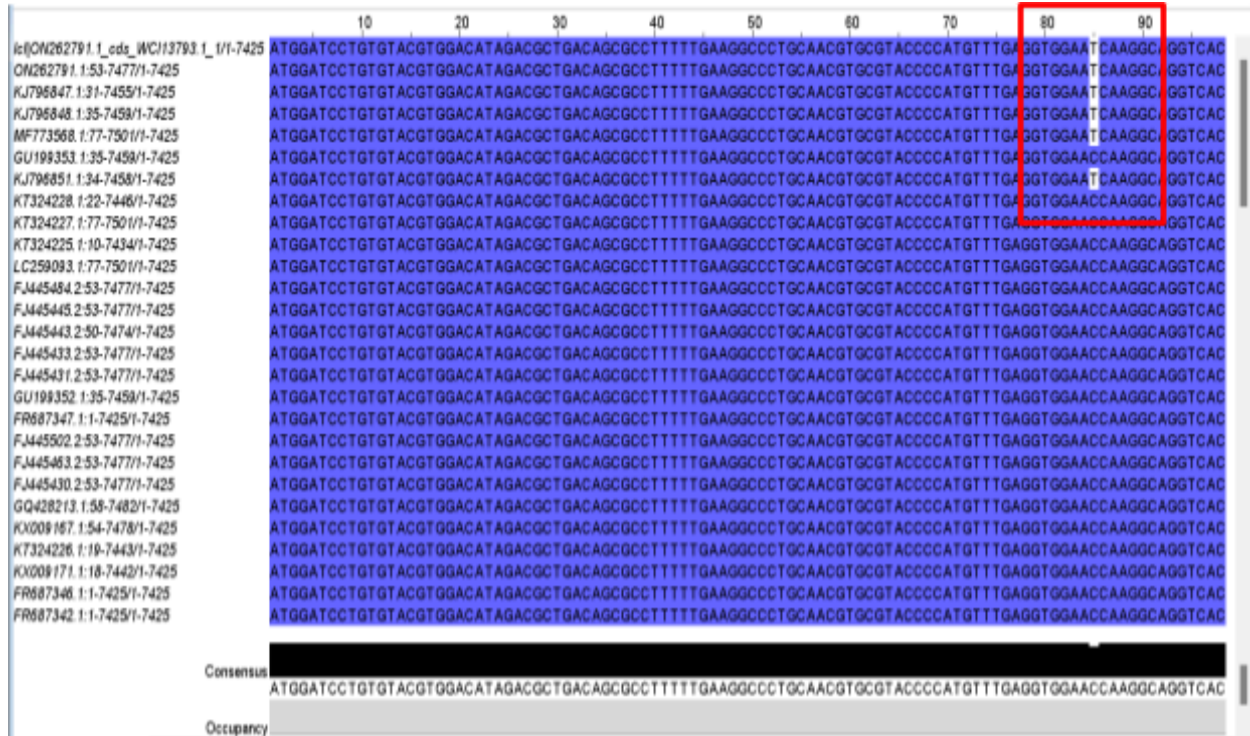
Results observed on Jalview:

On Jalview, we colored the nucleotides based on percentage identity.

ORF 5 -



In the above image, we can see the mutation occurred in the ORF (compared to rest of the sequences) from T to G



In the above image, we can see one of the mutation occurred in the ORF 6 and 7 combined (compared to rest of the sequences) from C to T

Interpretations

Based on the data provided, we can make several observations and interpretations:

- The ORF5 gene has only one mutation, a substitution mutation where T is replaced with G, at position 3728 when compared to the 100 closest gene sequences obtained from BLAST.
- The ORF6-7 combined gene has multiple substitution mutations at various positions when compared to the 100 closest gene sequences obtained from BLAST.
- The substitution mutations in the ORF6-7 gene occurred at positions 85, 556, 1150, 1491, 4141, and 4179, and involve the replacement of C with T, A with G, C with A, A with C, A with G, and C with T, respectively.
- The substitution mutations in positions 5835 and 6081 have a relatively even chance of being either T or A and T or C, respectively, indicating that the nucleotide change in these positions is not biased towards any particular base.
- These mutations in both ORF 5 and ORF6-7 could potentially affect the function of the protein they encode, and further analysis may be required to determine the impact of these mutations.

Overall, from this we can interpret that the virus has not mutated (i.e not evolved a lot) a lot after comparing it with those 100 genes from different samples of the same virus.

(Transition and transversion substitution mutation : purine to purine and more)

Task 5: Analysis of how the mutation/variation affects the function of the gene (protein)

Now that we have found the changes/variations found in the ORFs from the previous task, we now need to figure out if these variations would actually affect the protein structure and function.

Genes are the functional units of heredity, and their encoded proteins carry out essential functions in various biological processes. Any changes in the nucleotide sequence of a gene can cause variations in the protein structure and function, which can have a significant impact on an organism's phenotype. Proteins are made up of long chains of amino acids that fold into specific shapes and perform specific functions in the body. So, even a small change in the amino acid sequence can alter the structure and function of the protein, which can have significant consequences.

As a result, analyzing how the mutations/variants affect the function of the gene (protein) is crucial for understanding the function of the virus.

Tools and database used

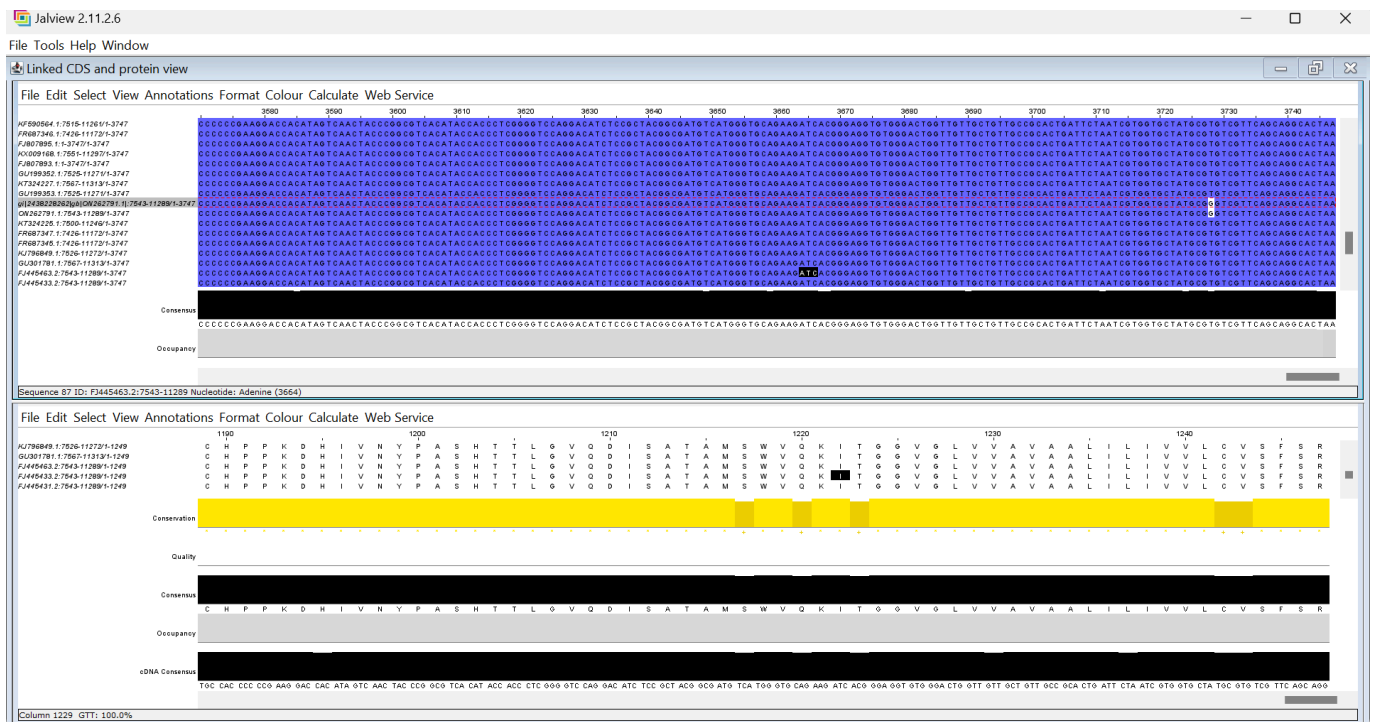
Translate the ORF nucleotide sequence to amino acid sequences

- Pay attention to if the variation is in the first two positions in the codon or the third position, this is because based on the codon table a variation in the first two positions is less likely to cause a change in the amino acid sequence and a variation on the third position is more likely to cause a change in the amino acid sequence of the protein.
- Simultaneously, keep a track on the translated nucleotide sequence to amino acid sequence and see if the variation in the nucleotide has caused a change in the amino acid or not.
- If there is a change in the amino acid then it is more likely that it would affect the function of the protein.

To do this, we used the PDB database to look for protein structures and analyze the mutations. We used the PDB database because the Protein Data Bank (PDB) is a widely-used database containing 3D structural information of biological macromolecules, including proteins. Analyzing the structure of proteins and how mutations/variants affect their function is an important area of research, and the PDB database is a valuable tool for this purpose.

Here are some reasons why PDB database is useful for analyzing the effects of mutations/variants on protein function:

- Provides detailed structural information: PDB contains high-quality 3D structures of proteins, providing detailed information about the location of amino acids and other functional groups within the protein. This information can be used to analyze the effects of mutations or variants on the protein structure and function.
- Facilitates comparison between wild-type and mutant structures: The PDB database allows researchers to compare the structures of wild-type proteins (proteins without mutations) and mutant proteins (proteins with mutations or variants). This comparison can help researchers understand how mutations/variants affect protein folding, stability, and interactions with other molecules.



Following are the details of the changes that would happen due to the mutation:

Amino acid no	Original amino acid	Charge of the amino acid	Size of the amino acid	Shape of the amino acid	Property of the amino acid	Mutated amino acid	Charge of the amino acid	Size of the amino acid	Shape of the amino acid	Property of the amino acid
ORF 5										
1	V (Valine)	Neutral	Medium	3D tetrahedron, branched-chain	Non-polar	G (Glycine)	Neutral	Small	lack of a bulky side chain, does not have a chiral center, unbranched	Non-polar
2	V (Valine)	Neutral	Medium	3D tetrahedron, branched-chain	Non-polar	G (Glycine)	Neutral	Small	lack of a bulky side chain, does not have a chiral center, unbranched	Non-polar
3	T (Threonine)	Neutral	Medium	tetrahedron, with the alpha carbon at the center, unbranched	Polar	A (Alanine)	Neutral	Small	tetrahedron, with the alpha carbon at the center, unbranched	Non-polar
4	T (Threonine)	Neutral	Medium	tetrahedron, with the alpha carbon at the center, unbranched	Polar	A (Alanine)	Neutral	Small	tetrahedron, with the alpha carbon at the center, unbranched	Non-polar
5	P (Proline)	Neutral	Small	secondary amino group, pyrrolidine ring structure, unbranched	Non-polar	S (Serine)	Neutral	Small	tetrahedron, with the alpha carbon at the center, unbranched	Polar
6	A (Alanine)	Neutral	Small	tetrahedron, with the alpha carbon at the center, unbranched	Non-polar	V (Valine)	Neutral	Medium	3D tetrahedron, branched-chain	Non-polar

7	T (Threonine)	Neutral	Medium	tetrahedron, with the alpha carbon at the center, unbranched	Polar	A (Alanine)	Neutral	Small	tetrahedron, with the alpha carbon at the center, unbranched	Non-polar
ORF 6-7 Combined										
8	P (Proline)	Neutral	Small	secondary amino group, pyrrolidine ring structure, unbranched	Non-polar	S (Serine)	Neutral	Small	tetrahedron, with the alpha carbon at the center, unbranched	Polar
9	N (Asparagine)	Neutral	Medium	tetrahedron, with the alpha carbon at the center, hydrophilic character, polar	Polar	D (Aspartic acid)	Negative	Small	tetrahedron, with the alpha carbon at the center, unbranched	Acidic, electrically charged
10	S (Serine)	Neutral	Small	tetrahedron, with the alpha carbon at the center, unbranched	Polar	G (Glycine)	Neutral	Small	lack of a bulky side chain, does not have a chiral center, unbranched	Non-polar
11	S (Serine)	Neutral	Small	tetrahedron, with the alpha carbon at the center, unbranched	Polar	R (Arginine)	Positive	Large	long, positively charged side chain, tetrahedron, with the alpha carbon at the center, unbranched	Basic, electrically charged
12	N (Asparagine)	Neutral	Medium	tetrahedron, with the alpha carbon at the center, hydrophilic character, polar	Polar	D (Aspartic acid)	Negative	Small	tetrahedron, with the alpha carbon at the center, unbranched	Acidic, electrically charged

[The above data is based on this resource](#)

[Excel sheet of the mutations](#)

Valine (V) to Glycine (G) Mutation:

The valine side chain is a branched, non-polar aliphatic group, while the glycine side chain is a single hydrogen atom, which makes it the smallest amino acid. The substitution of valine with glycine removes the side chain's bulky nature and increases the flexibility of the protein structure. The mutation could introduce more conformational flexibility and affect the hydrophobic interactions within the protein structure.

Threonine (T) to Alanine (A) Mutation:

Threonine has a hydroxyl group (-OH) in its side chain, while alanine has a methyl group (-CH₃). The substitution of threonine with alanine removes the hydroxyl group, which could affect the hydrogen bonding network within the protein. The mutation could result in less steric hindrance within the protein structure.

Proline (P) to Serine (S) Mutation:

Proline has a cyclic structure and lacks a hydrogen atom in its backbone, while serine has a hydroxyl group in its side chain. The substitution of proline with serine introduces a polar group and could affect the conformational rigidity of the protein. The mutation could also affect the hydrogen bonding and van der Waals interactions within the protein structure.

Alanine (A) to Valine (V) Mutation:

Alanine and valine are both nonpolar aliphatic amino acids, but valine has a bulkier side chain than alanine. The substitution of alanine with valine introduces a larger side chain, which could affect the steric hindrance within the protein structure.

Asparagine (N) to Aspartic acid (D) Mutation:

Asparagine has an amide group in its side chain, while aspartic acid has a carboxylic acid group. The substitution of asparagine with aspartic acid introduces a negative charge to the side chain and could affect the electrostatic interactions within the protein structure.

Serine (S) to Glycine (G) Mutation:

Serine has a hydroxyl group in its side chain, while glycine has a single hydrogen atom. The substitution of serine with glycine removes the hydroxyl group and could affect the hydrogen bonding and van der Waals interactions within the protein structure.

Serine (S) to Arginine (R) Mutation:

Serine is a polar amino acid with a hydroxyl group in its side chain, while arginine has a guanidine group that is positively charged at physiological pH. The substitution of serine with arginine introduces a positively charged group and could affect the electrostatic interactions within the protein structure.

Asparagine (N) to Aspartic acid (D) Mutation:

Asparagine has an amide group in its side chain, while aspartic acid has a carboxylic acid group. The substitution of asparagine with aspartic acid introduces a negative charge to the side chain and could affect the electrostatic interactions within the protein structure.

At the atomic level, amino acid mutations could affect the protein's three-dimensional structure, stability, and function by altering the hydrogen bonding, electrostatic interactions, hydrophobic interactions, and van der Waals forces between amino acids. These changes could impact the protein's active site, substrate binding, conformational changes, and stability against denaturation or degradation. Therefore, understanding the effect of amino acid mutations on protein structure and

Reason for not being able to find the effects of mutation

As we reviewed the PDB results for the input protein, we found that we were not able to analyze the effects of the mutation. One reason for this was that the relative protein structure of the ORF gene found on the PDB had a low resolution, which made it difficult to view the structure in detail and identify specific bonding and interactions between amino acids. This lack of detail limited our ability to analyze the effects of the mutation on the protein's structure and function.

In addition, we found that one of the proteins in the ORF was a polyprotein, which made the analysis even more complex. Polyproteins are large proteins that contain multiple smaller proteins, making it difficult to isolate the effects of the mutation on specific regions of the protein. This complexity also made it more challenging to identify the specific structural changes that resulted from the mutation.

Overall, the limitations of the PDB results for the input protein made it difficult to analyze the effects of the mutation in detail. The low resolution of the protein structure and the complexity of the polyprotein made it challenging to identify specific structural changes and interactions between amino acids.

Summary

In this project, the focus was on the Chikungunya virus and the identification of its genes or open reading frames (ORFs). ORFs are sections of DNA that contain the information required for a protein to be synthesized. The first step in this project was to identify the ORFs of the Chikungunya virus. Once identified, the next step was to verify if the ORFs were real protein-coding genes. This was done by matching the existing proteins and confirming if the ORFs had similar sequences. If the ORFs had similar sequences to existing proteins, it was concluded that they were indeed real protein-coding genes.

The next step was to find mutations in the virus. This involved inputting the ORFs into a program to find any mutations. The overall approach was to compare the identified ORFs with other known genes from the Chikungunya virus. This comparison was done to determine if there

were any changes in the virus's genetic makeup. If mutations were identified, they were studied in detail to determine how they affected the gene structure. The challenge faced during this process was not a proper resolution of the structure available in the database. Due to this, we were not able to see the amino acid level details of the structure to study mutations. As a result, then we predicted the structure using the given amino acids and analyzed the potential changes that could be seen.

Overall, this project involved a multi-step process that began with identifying the ORFs of the Chikungunya virus and verifying them. Then searched for mutations in the virus and compared them with known genes from the virus. Finally, analyzed the mutation's effects on the gene structure, considering the challenges posed by the lack of resolution in 3D structure information available in the database. Through these steps, we were able to gain a better understanding of the Chikungunya virus and its genetic makeup.