Final Project

Bioinformatics

Jan 12, 2021

Submitted by:

1. **Ahmed Mahmoud Fawzi** **ID:1170523**
2. **Doaa Sherif** **ID:1170122**
3. **Hendwan Abozide** **ID:1170463**
4. **Rokaia Mohamed Ahmed** **ID:1170193**

Contents

Contribution
We all participated **equally** in this project, as we held meetings daily to help and share ideas with each other.

1 Introduction

In this project we conducted a comparative study related to SARS-Cov-2 **(COVID-19)**. The study was made between 13 sequences downloaded from **NCBI** database; 5 sequences from **Egypt** and 2 sequences from each of the following regions: **USA**, **China**, **Europe** and **Gulf area**.

To compare these sequences we did the following:
1) Computed multiple sequences alignment
2) Analyzed the phylogenetic tree
3) Calculated the percentages of chemical constituents (C, G, T, A and CG)
4) Knew the functional products and the interpretations of conserved and un-conserved regions

The next section will go through the details of this study.

2 Methods

2.1 ClustalO

We used the **ClustalO software** to align the 13 sequences plus an extra one called the **reference sequence**. The phylogenetic tree is generated automatically in the alignment results and will be explained in details in the next section.

2.2 Python

Python was used to get the percentage of the chemical constituents and get the conserved and un-conserved regions. Three output csv files were produced; one for the constituents' percentages and two for the conserved and un-conserved regions.

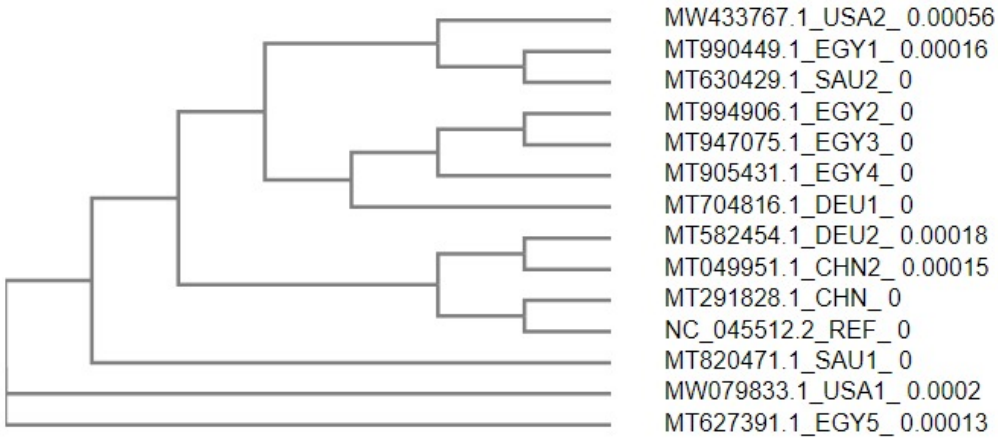
2.3 NCBI Database

After getting the regions (conserved and un-conserved) we compared them to the **GFF3** file of the reference sequence, to know the corresponding **GeneID** and **protein_id**. After that we searched for this gene or protein and extracted its function and some information about it.

3 Results and Discussion

3.1 Phylogenetic Tree

The following image is the phylogenetic tree of the **14** sequences:



The Phylogenetic Tree is divided into three groups (branches):

1) EGY5 sequence
2) USA1 sequence
3) The rest of the sequences + Reference sequence.

This means that the first and second branches are not similar to any of the groups, while the third branch divides into several branches which are:

1) SAU1 is similar to the branch which contains all of the following sequences: REF + CHN + CHN2 + DEU2 + DEU1 + EGY4 + EGY3 + EGY2 +SAU2 + EGY1 + USA2

2) DEU2 is similar to CHN2
3) CHN is similar to REF
4) 2 & 3 branches are similar to the branch which contains all of the following sequences: DEU1 + EGY4 + EGY3 + EGY2 + SAU2 +EGY1 + USA2
5) EGY2 is similar to EGY3
6) No.5 branch is similar to EGY4
7) No.6 branch is similar to DEU1
8) EGY1 is similar to SAU2
9) No.8 branch is similar to USA2
10) No.8 branch is similar to No.7 branch

3.2 Chemical Constituents Percentages

The following table shows the percentages of the constituents (C, G, T, A and the CG content)

Sequence Name	A Content	T Content	C Content	G Content	CG Content
MT990449.1(EGY1)	29.8452345	32.14489542	18.35364421	19.65622587	38.00987008
MT994906.1(EGY2)	29.86764952	32.1293349	18.37493718	19.62807841	38.00301558
MT947075.1(EGY3)	29.86764952	32.1293349	18.37493718	19.62807841	38.00301558
MT905431.1(EGY4)	29.86764952	32.1293349	18.37493718	19.62807841	38.00301558
MT627391.1(EGY5)	29.9250786	32.10248177	18.36243227	19.61000736	37.97243963
MT291828.1(CHN)	29.91158149	32.10194923	18.36023846	19.62623083	37.98646929
MT049951.1(CHN2)	29.95017222	32.08039327	18.36605023	19.60338428	37.9694345
MT704816.1(DEU1)	29.86273853	32.13927017	18.37294945	19.62504185	37.9979913
MT582454.1(DEU2)	29.87039151	32.12343026	18.3735142	19.62930629	38.0028205
MT820471.1(SAU1)	29.87248322	32.12416107	18.37919463	19.62416107	38.0033557
MT630429.1(SAU2)	29.94076107	32.10616152	18.34398742	19.60574316	37.94973058
MW079833.1(USA1)	29.85999464	32.13759378	18.3681672	19.63424437	38.00241158
MW433767.1(USA2)	29.79207821	32.02866039	18.27769779	19.54330867	37.82100646
$NC_045512.2(REF)$	29.94348393	32.08373742	18.36605023	19.60672842	37.97277865

3.3 Gene and Protein Outputs

We used python to get the regions of **conserved** and **nonconserverd** regions, and their corresponding genes, proteins and their functions. In this section we will show a sample of the results we had.

3.3.1 Conserved Regions

The following table shows a sample of the conserved regions after comparing them with the reference sequence:

Range	**R1**	**R2**
Range=1060-1396	ID=gene-GU280$_gp$01; $Dbxref$ = $GeneID$: 43740578; $Name$ = $ORF1ab$; $gbkey$ = $Gene$; $gene$ = $ORF1ab$; $gene_biotype$ = $protein_coding$; $locus_tag = GU280_gp01$	ID=cds-YP$_0$09724389.1; $Parent$ = $gene - GU280_gp01$; $Dbxref$ = $Genbank$: $YP_09724389.1$, $GeneID$: 43740578; $Name$ = $YP_09724389.1$; $Note$ = $pp1ab\%3Btranslatedby - 1ribosomalframeshift$; $exception$ = $ribosomalslippage$; $gbkey$ = CDS; $gene$ = $ORF1ab$; $locus_tag$ = $GU280_gp01$; $part$ = 1; $product$ = $ORF1abpolyprotein$; $protein_id$ = $YP_09724389.1$
Range=1398-1721	ID=gene-GU280$_gp$01; $Dbxref$ = $GeneID$: 43740578; $Name$ = $ORF1ab$; $gbkey$ = $Gene$; $gene$ = $ORF1ab$; $gene_biotype$ = $protein_coding$; $locus_tag = GU280_gp01$	ID=cds-YP$_0$09724389.1; $Parent$ = $gene - GU280_gp01$; $Dbxref$ = $Genbank$: $YP_09724389.1$, $GeneID$: 43740578; $Name$ = $YP_09724389.1$; $Note$ = $pp1ab\%3Btranslatedby - 1ribosomalframeshift$; $exception$ = $ribosomalslippage$; $gbkey$ = CDS; $gene$ = $ORF1ab$; $locus_tag$ = $GU280_gp01$; $part$ = 1; $product$ = $ORF1abpolyprotein$; $protein_id$ = $YP_09724389.1$
Range=1723-2142	ID=gene-GU280$_gp$01; $Dbxref$ = $GeneID$: 43740578; $Name$ = $ORF1ab$; $gbkey$ = $Gene$; $gene$ = $ORF1ab$; $gene_biotype$ = $protein_coding$; $locus_tag = GU280_gp01$	ID=cds-YP$_0$09724389.1; $Parent$ = $gene - GU280_gp01$; $Dbxref$ = $Genbank$: $YP_09724389.1$, $GeneID$: 43740578; $Name$ = $YP_09724389.1$; $Note$ = $pp1ab\%3Btranslatedby - 1ribosomalframeshift$; $exception$ = $ribosomalslippage$; $gbkey$ = CDS; $gene$ = $ORF1ab$; $locus_tag$ = $GU280_gp01$; $part$ = 1; $product$ = $ORF1abpolyprotein$; $protein_id$ = $YP_09724389.1$
Range=2144-2394	ID=gene-GU280$_gp$01; $Dbxref$ = $GeneID$: 43740578; $Name$ = $ORF1ab$; $gbkey$ = $Gene$; $gene$ = $ORF1ab$; $gene_biotype$ = $protein_coding$; $locus_tag = GU280_gp01$	ID=cds-YP$_0$09724389.1; $Parent$ = $gene - GU280_gp01$; $Dbxref$ = $Genbank$: $YP_09724389.1$, $GeneID$: 43740578; $Name$ = $YP_09724389.1$; $Note$ = $pp1ab\%3Btranslatedby - 1ribosomalframeshift$; $exception$ = $ribosomalslippage$; $gbkey$ = CDS; $gene$ = $ORF1ab$; $locus_tag$ = $GU280_gp01$; $part$ = 1; $product$ = $ORF1abpolyprotein$; $protein_id$ = $YP_09724389.1$
Range=16378-17090	ID=gene-GU280$_gp$01; $Dbxref$ = $GeneID$: 43740578; $Name$ = $ORF1ab$; $gbkey$ = $Gene$; $gene$ = $ORF1ab$; $gene_biotype$ = $protein_coding$; $locus_tag = GU280_gp01$	ID=cds-YP$_0$09724389.1; $Parent$ = $gene - GU280_gp01$; $Dbxref$ = $Genbank$: $YP_09724389.1$, $GeneID$: 43740578; $Name$ = $YP_09724389.1$; $Note$ = $pp1ab\%3Btranslatedby - 1ribosomalframeshift$; $exception$ = $ribosomalslippage$; $gbkey$ = CDS; $gene$ = $ORF1ab$; $locus_tag$ = $GU280_gp01$; $part$ = 2; $product$ = $ORF1abpolyprotein$; $protein_id$ = $YP_09724389.1$

The outputs of conserved regions were two as shown in the table. In Each range there was two different results, denoted by **R1** and **R2**. We searched for each GeneID in **NCBI gene database** and we got the corresponding proteins and their possible functions. The following table shows the gene names, corresponding proteins and their functions:

Gene_Name	**Protein_Name**	**Function**
ORF1ab	ORF1ab polyprotein,nsp2,ORF1a polyprotein	Severe acute respiratory syndrome coronavirus 2 causes coronavirus disease,Once inside the cell the infecting RNA is used to encode structural proteins that make up virus particles, nonstructural proteins that direct virus assembly, transcription, replication and host control and accessory proteins whose function has not been determined.
ORF1ab	helicase	Severe acute respiratory syndrome coronavirus 2,host is Homo sapiens,there are some regions with functions like region DEXXQc$_{UPf1-likenoteonitDEXXQ-boxhelicasedomainofUpf1-likehelicase}$

3.3.2 Nonconserved Regions

The following table shows a sample of the nonconserved regions after comparing them with the reference sequence:

Range	**R1**	**R2**
Range=22384-22481	ID=gene-GU280$_{gp02;Dbxref=GeneID:43740568;Name=S;gbkey=Gene;gene=S;gene_{b}iotype=protein_{c}oding;gene_{s}ynonym=spikeglycoprotein;locus_{t}ag=GU280_{gp02}}$	ID=cds-YP$_{0}$09724390.1;$Parent=gene-GU280_{gp02};Dbxref=Genbank:YP_{0}09724390.1,GeneID:43740568;Name=YP_{0}09724390.1;Note=structuralprotein\%3Bspikeprotein;gbkey=CDS;gene=S;locus_{t}ag=GU280_{gp02};product=surfaceglycoprotein;protein_{i}d=YP_{0}09724390.1$
Range=28881-28883	ID=gene-GU280$_{gp10;Dbxref=GeneID:43740575;Name=N;gbkey=Gene;gene=N;gene_{b}iotype=protein_{c}oding;locus_{t}ag=GU280_{gp10}}$	ID=cds-YP$_{0}$09724397.2;$Parent=gene-GU280_{gp10};Dbxref=Genbank:YP_{0}09724397.2,GeneID:43740575;Name=YP_{0}09724397.2;Note=ORF9\%3Bstructuralprotein;gbkey=CDS;gene=N;locus_{t}ag=GU280_{gp10};product=nucleocapsidphosphoprotein;protein_{i}d=YP_{0}09724397.2$

The outputs of nonconserved regions were two as shown in the table. The following table shows the gene names, corresponding proteins and their functions:

Gene_Name	**Protein_Name**	**Function**
S	surface glycoprotein	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an enveloped, positive-sense, single-stranded RNA virus that causes coronavirus disease 2019 (COVID-19).The structural proteins of SARS-CoV-2 include spike or surface glycoprotein (S),The spike glycoprotein is found on the outside of the virus particle and gives coronavirus viruses their crown-like appearance. This glycoprotein mediates attachment of the virus particle and entry into the host cell. S protein is an important target for vaccine development, antibody therapies and diagnostic antigen-based tests.
N	nucleocapsid phosphoprotein	The structural proteins of SARS-CoV-2 include the envelope protein (E), spike or surface glycoprotein (S), membrane protein (M) and the nucleocapsid protein (N). The nucleocapsid phosphoprotein is a structural protein that binds to, protects the viral RNA genome and is involved in packaging the RNA into virus particles. The N protein has been suggested as an antiviral drug target. Coronavirus nucleocapsid protein is a region of the protein sequence.

4 Conclusion

After this study, we conclude that:

1) All 13 sequences were very similar to each other and to the reference sequence, which resulted in several conserved regions and only two nonconserverd regions.

2) The percentages of the chemical constituents shows that all 14 sequences were very similar, as the results are almost identical.

3) After searching for gene and protein names, it was clear that **ORF1ab** gene was found several times in different regions and it is one of the main genes causing **COVID-19**.

4) Some of the output proteins were targeted as an antiviral drug target, and are used in vaccine development and antibody therapies, like the surface **glycoprotein (S)**, **membrane glycoprotein (M)** and **nucleocapsid phosphoprotein (N)**.

References

[1] National Center for Biotechnology Information https://www.ncbi.nlm.nih.gov/

[2] Biopython Website https://biopython.org/