Review Paper:

A novel Bioinformatic analysis for SARS CoV-2 isolated by Public Health Laboratory, USA

Hemalatha N.* and Melcy Philip
St Aloysius Institute of Management and Technology, Mangalore, INDIA
*hemasree71@gmail.com

Abstract

The outbreak of novel SARS CoV-2 has outraged people all over the world due to its astonishing capabilities to be pathogenic. Considering the severity of this disease, WHO declared COVID-19 as one of the pandemic diseases. Since this is an emerging, rapidly evolving situation, many nations have called for technology driven solutions, along with the acceleration of clinical research and trials.

In this scenario of dire to find the solution, we have considered the whole genome of SARS CoV-2 strain, with GenBank id MT188339 and subjected it to the bioinformatic analysis like sequence similarity, phylogenetic analysis which shows evolutionary relationships and structure prediction of its protein.

Keywords: COVID-19, BLAST, *SARS*, Clustal Omega.

Introduction

Coronavirus disease is a highly infectious disease identified in Wuhan part of China first in Dec 2019. The World Health Organization Emergency Committee on January 20, 2019 declared a global health emergency depending on the growing case of this disease. Corona Virus Disease (COVID-19) is the new name announced by WHO for the epidemic disease 2019-nCov on February 11, 2020. They have characterized COVID-19 as a pandemic on March 11, 2020 after assessing the spread of this disease around the world. Ever since it has been spreading globally and according to the latest report, the COVID-19 has affected more than 2,19,345 people in 176 countries with 8969 deaths.

Coronaviruses are single-stranded RNA viruses which are enveloped and belong to the subfamily *Coronavirinae*, family *Coronavirdiae*, order *Nidovirales*. Four genera of CoVs exist namely, Alphacoronavirus (αCoV), Betacoronavirus (βCoV), Deltacoronavirus (δCoV) andGammacoronavirus (γCoV)[1]. Evolutionary studies have shown that the gene sources for first two generas are bats and rodents whereas for the last two are avian species[2].

CoVs have repeatedly crossed the species barriers and presently have emerged as the most important human pathogens. Severe acute respiratory syndrome CoV (SARS-CoV) is the best known example which evolved in China in 2002-2003 causing large scale epidemic killing many people and Middle east respiratory syndrome CoV (MERS-CoV), causing a persistent epidemic in the Arabian peninsula since 2012[3,4].

In both these epidemics, this virus has likely originated from bats and then jumped to mammalian host the Himalayan palm civet (Paguma larvata) for SARS-CoV and the dromedary camel (Camelus dromedarius) for MERS-CoV before crossing species barriers to infect humans[2].

Coronaviruses (CoV) are a large family of viruses causing illness that range from common cold to severe diseases such as MERS-CoV and SARS-CoV. The strain of COVID-19 was not previously found in human beings and was discovered in 2019[5].

These strains are zoonotic which mean transmission of this strain happens between animals and humans. Earlier studies have shown that SARS-CoV was transmitted from civet cats to humans whereas MERS-CoV transmitted from dromedary camels to humans. In animals there are several known Corona virus which have not yet affected humans.

In this study, we have made an attempt to study the COVID-19 virus by doing some bioinformatics analysis and also trying to predict the 2D and 3D structure of the ten genes isolated from the whole genome sequence.

Dong et al, [6] summarized on various ongoing clinical trials for the deadly COVID-19. Different drugs such as chloroquine, arbidol, remdesivir and favipiravir have undergone many clinical studies to validate their efficacy and safety in the treatment of COVID-19. Some of the drugs showed a promising result, which can be improvised on further research.

Woo Chan et al,[7] worked on genomic characterization of the novel SARS-CoV-2, isolated from a patient with a typical pneumonia after visiting Wuhan[7]. Their undeniable studies showed that the genome of 2019-nCoV possessed 89% nucleotide identity with the sequence of bat; SARS-like-CoVZXC21 and 82% with that of human SARS-CoV. The phylogenetic trees of different proteins also showed close relation with those of the bat, civet and human SARS coronaviruses.

The secondary structure prediction and comparison in the 5′-untranslated region (UTR) and 3′-UTR using the RNAfold Web-Server, are demonstrated in the figure 1.

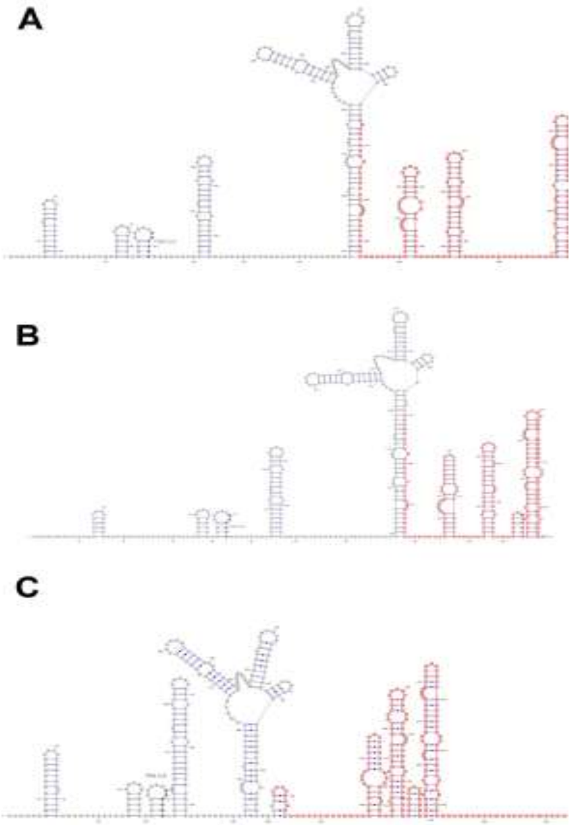


Figure 1: Secondary structure prediction and comparison in the 5′-untranslated region (UTR) and 3′-UTR using the RNAfold Web-Server (with minimum free energy and partition function in Fold algorithms and basic options. The SARS 5′- and 3′- UTR was used as a reference to adjust the prediction results. (A) SARS-CoV 5'-UTR; (B) 2019-nCoV (HKU-SZ-005b) 5'-UTR; (C) ZC45 5'-UTR

Sah et al,[8] studied the novel SARS-CoV-2, isolated from a Nepalese and obtained complete genome sequence from an oropharyngeal swab specimen. The sequencing was done using the Illumina MiSeq system with the Burrows-Wheeler Aligner MEM algorithm (BWAMEM) 0.7.5a-r405 assembly method. The whole genome of our concern was amplified directly from the RNA extract of the original specimen, using gene-specific primers. They have been successfully deposited the sequence in GenBank of NCBI with the accession number MT072688 and at the GISAID EpiCoV newly emerging coronavirus SARSCoV- 2 platform, with the id EPI_ISL_410301.

Chang et al,[9] used molecular docking to regenerate HIV protease inhibitors and nucleoside analogues for COVID-19 and tested it based on docking scores calculated by AutoDock Vina and Rosetta Commons specimen. In a nutshell, according to their docking studies Indinavir and Remdesivir scored the best and comparison of the docking sites of the two drugs showed a near perfect.

Further studies unveiled that Indinavir does not dock on any active sites of the protease, which questions the reliability of the drug Indinavir. Apart from this, Remdesivir is not compatible with any known functional regions, including template binding motifs, polymerization motifs etc. However, testing the active form of Remdesivir, the docking site showed a perfect dock in the overlapping region of the NTP binding motif. Hence, they could conclude that the Remdesivir could be a potent therapeutic drug for COVID-19.

GenBank: The whole genome used for this particular analysis was from GenBank of NCBI which possessed the accession id MT188339.1. The genome under consideration is novel, hence the availability of data regarding this, was constrained into ten genes and its products (structural protein, membrane protein etc.). The whole genome taken from the GenBank was subjected to various bioinformatic analysis such as BLAST, Clustal Omega, SWISS-MODEL and GOR-Ⅳ.

BLAST (Basic Local Alignment Search Tool): BLAST is a search tool which finds similarity between sequences[10]. BLAST tool is popular since it's simple and finds similar regions for nucleotide sequences and protein sequences with its statistical evidences. Blast uses different strategies for its multiple types and consists of advanced search with many significant parameters. This work comprises of conventional nucleotide blast and customized nucleotide blast.

Clustal Omega: Clustal Omega is a multiple sequence alignment tool which can align more than 4000 sequences and generate alignment using HMM profile-profile techniques[11]. It can be used in finding evolutionary relationship between various sequences, by the means of homology. In this work, Clustal omega was used to align the first hundred sequences that got hit in the conventional blast results of SARS-CoV-2. Clustal Omega also gives output such as summary, phylogram, phylogenetic tree (cladogram) and many more which make the analysis much fundamental and rigid.

SWISS-MODEL: It is a free homology-modelling server, particularly designed for proteins and is fully automated[12]. SWISS-MODEL helps in generating various homology models for novel sequences with no wet lab confirmed structures. In this server, the submitted sequence is taken for a hit against various databases and finds the best matches. Once the matches are listed out, it also gives rankings based on various aspects such as GMQE. The user can select the reference sequence or the template and based on the selected reference sequence the model is built. Thus, the products (proteins) of the eight genes of this particular genome SARS-CoV-2 were modelled using SWISS-MODEL server.

GOR-IV: GOR is a free online server which predicts the secondary structure of proteins[13]. GOR-IV is an improvised version of the conventional GOR. The program gives a clear output as the sequence and its predicted secondary structure, where H stands for helix, E stands for extended/beta strand and C for coils. The output is highly compatible, since the best out of all predictions is submitted back to the user. Though we could not generate models for two proteins (due to absence of a reference sequence and the length of the submitted sequence), coded by two genes of our concern, the secondary structure was predicted using GOR-IV. Figure 2 explains the methodology followed in our work.

Results and Discussion

The whole genome sequence with the accession id MT188339.1 was analyzed with the conventional nucleotide blast, (high similarity ones only) and the top hundred sequences were listed out as shown in the figure 3. All these sequences show a query coverage of 99%, 0.0 as the E-value and a percentage identity of 99.8%.

A customized hit was performed with the organism HIV to study the similarity of HIV with COVID-19. The result showed least similarity and hence the search was widened to the aspect of somewhat similar search. Top nine sequences were listed out with a poor E-Value (3.8), least query coverage (0%) and percentage identity of 86.84%. Figure 4 shows the details of this result.

The results obtained from the nucleotide blast were further analysed with the help of Clustal omega. The Clustal omega resulted in a phylogenetic tree as shown in figure 5, which depicts the evolutionary relationship between these top hundred sequences of SARS-CoV-2 from various sources.

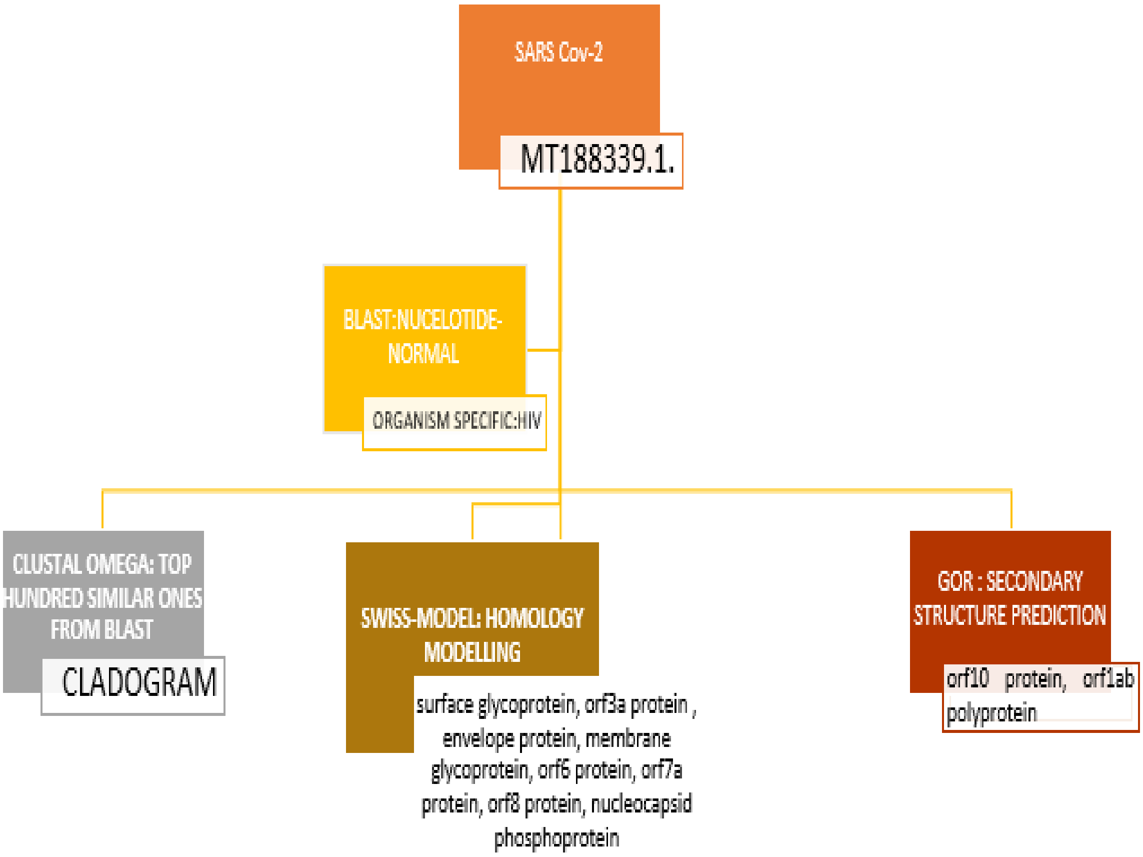
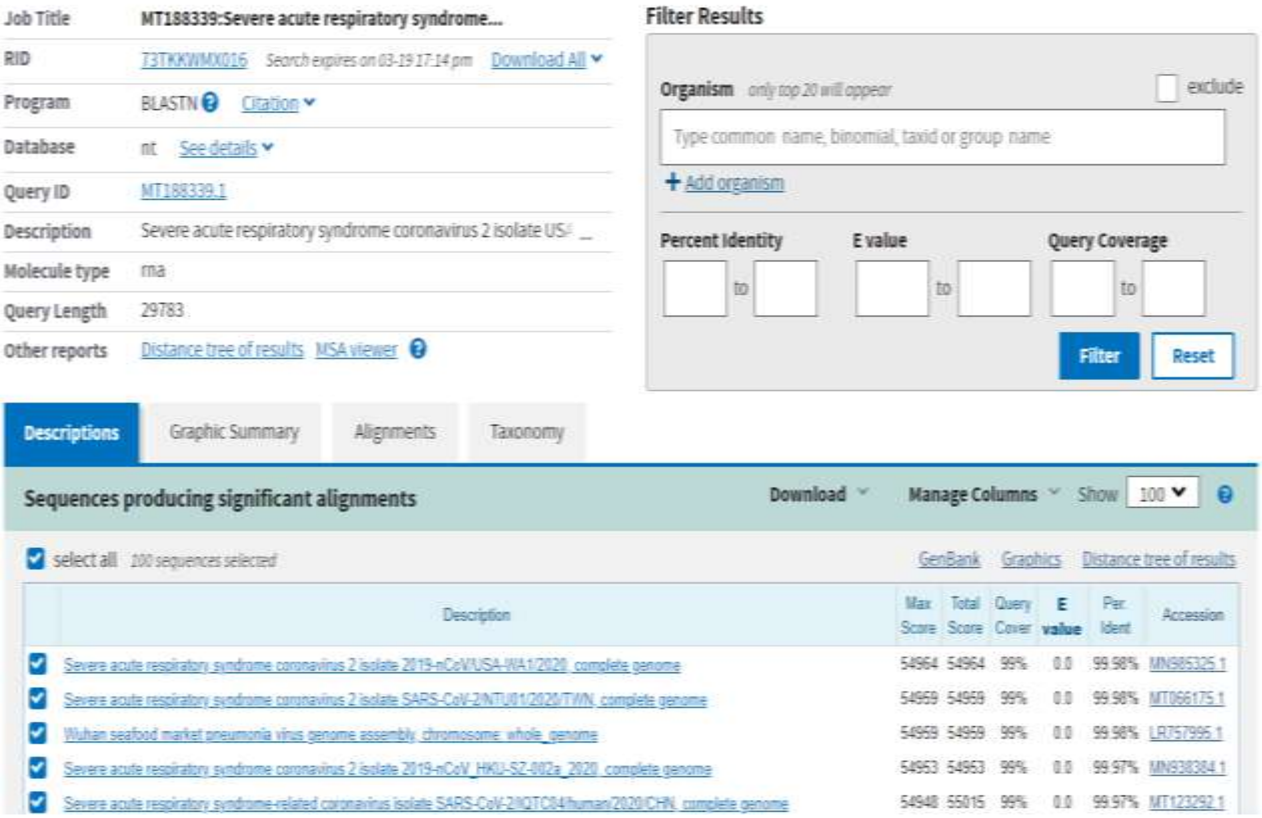

SARS Cov-2
MT188339.1.
BLAST:NUCELOTIDE-NORMAL
ORGANISM SPECIFIC:HIV
CLUSTAL OMEGA: TOP HUNDRED SIMILAR ONES FROM BLAST
CLADOGRAM
SWISS-MODEL: HOMOLOGY MODELLING
surface glycoprotein, orf3a protein , envelope protein, membrane glycoprotein, orf6 protein, orf7a protein, orf8 protein, nucleocapsid phosphoprotein
GOR : SECONDARY STRUCTURE PREDICTION
orf10 protein, orf1ab polyprotein

Figure 2: Workflow of this bioinformatics work



Job Title	MT188339:Severe acute respiratory syndrome...
RID	73TKKWMX016 Search expires on 03-19 17:14 pm Download All
Program	BLASTN Citation
Database	nt See details
Query ID	MT188339.1
Description	Severe acute respiratory syndrome coronavirus 2 isolate US/ ...
Molecule type	rna
Query Length	29783
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to E value to Query Coverage to

Filter Reset

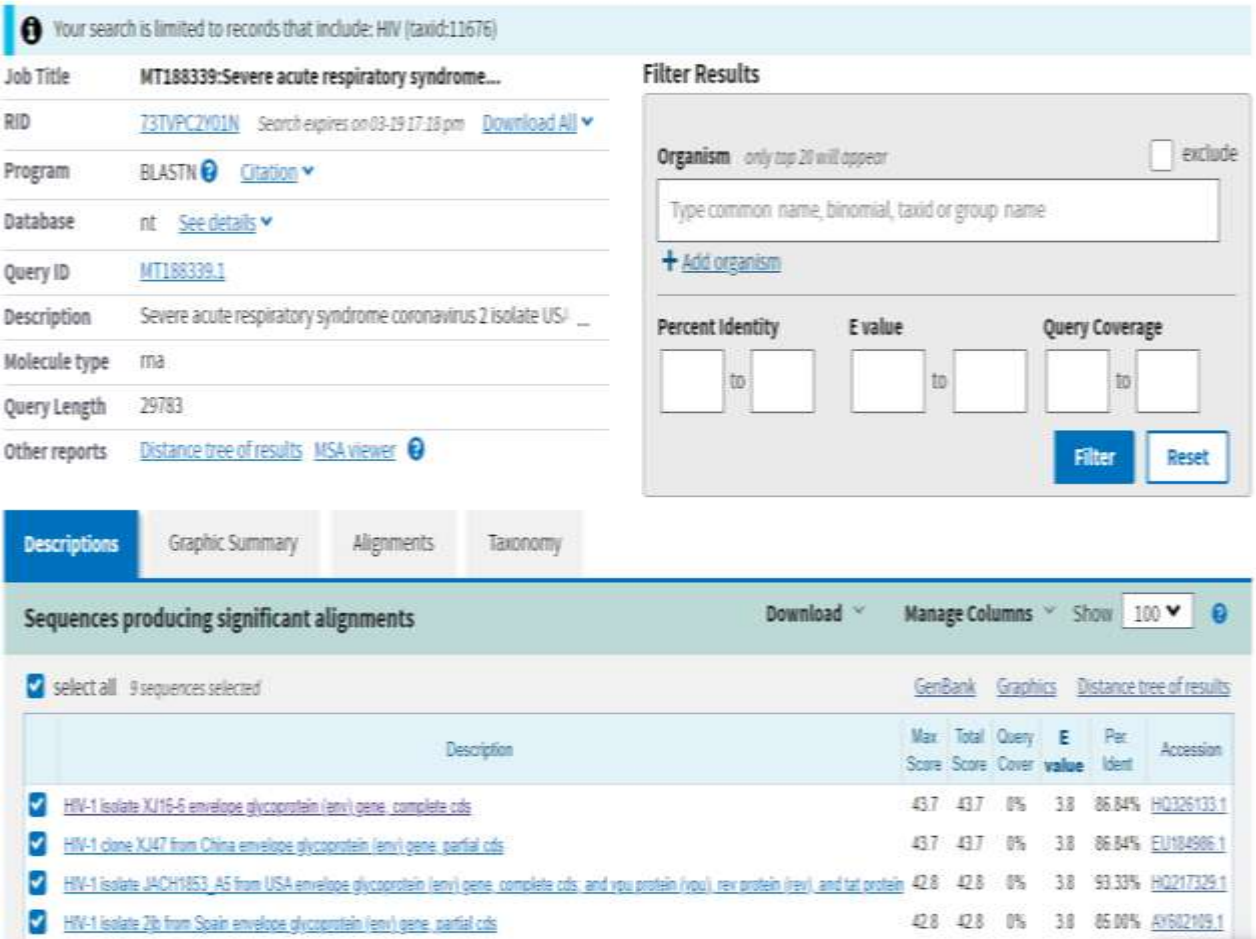
Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

☑ select all *100 sequences selected* GenBank Graphics Distance tree of results

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-WA1/2020, complete genome	54964	54964	99%	0.0	99.98%	MN985325.1
☑	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/NTU01/2020/TWN, complete genome	54959	54959	99%	0.0	99.98%	MT066175.1
☑	Wuhan seafood market pneumonia virus genome assembly, chromosome: whole_genome	54959	54959	99%	0.0	99.98%	LR757995.1
☑	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV_HKU-SZ-002a_2020, complete genome	54953	54953	99%	0.0	99.97%	MN938384.1
☑	Severe acute respiratory syndrome-related coronavirus isolate SARS-CoV-2/NQTC04/human/2020/CHN, complete genome	54948	55015	99%	0.0	99.97%	MT123292.1

Figure 3: Nucleotide BLAST result for MT188339.1 (SARS-CoV-2)



Your search is limited to records that include: HIV (taxid:11676)

Job Title	MT188339:Severe acute respiratory syndrome...
RID	73TVPC2Y01N Search expires on 03-19 17:18 pm Download All
Program	BLASTN Citation
Database	nt See details
Query ID	MT188339.1
Description	Severe acute respiratory syndrome coronavirus 2 isolate US/ ...
Molecule type	rna
Query Length	29783
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to E value to Query Coverage to

Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

☑ select all *9 sequences selected* GenBank Graphics Distance tree of results

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	HIV-1 isolate XJ16-6 envelope glycoprotein (env) gene, complete cds	43.7	43.7	0%	3.8	86.84%	HQ326133.1
☑	HIV-1 clone XJ47 from China envelope glycoprotein (env) gene, partial cds	43.7	43.7	0%	3.8	86.84%	EU184986.1
☑	HIV-1 isolate JACH1853_A5 from USA envelope glycoprotein (env) gene, complete cds; and vpu protein (vpu), rev protein (rev), and tat protein	42.8	42.8	0%	3.8	93.33%	HQ217329.1
☑	HIV-1 isolate 2ib from Spain envelope glycoprotein (env) gene, partial cds	42.8	42.8	0%	3.8	85.00%	AY682109.1

Figure 4: Customized nucleotide BLAST result for MT188339.1 (SARS-CoV-2) – organism specific (HIV)

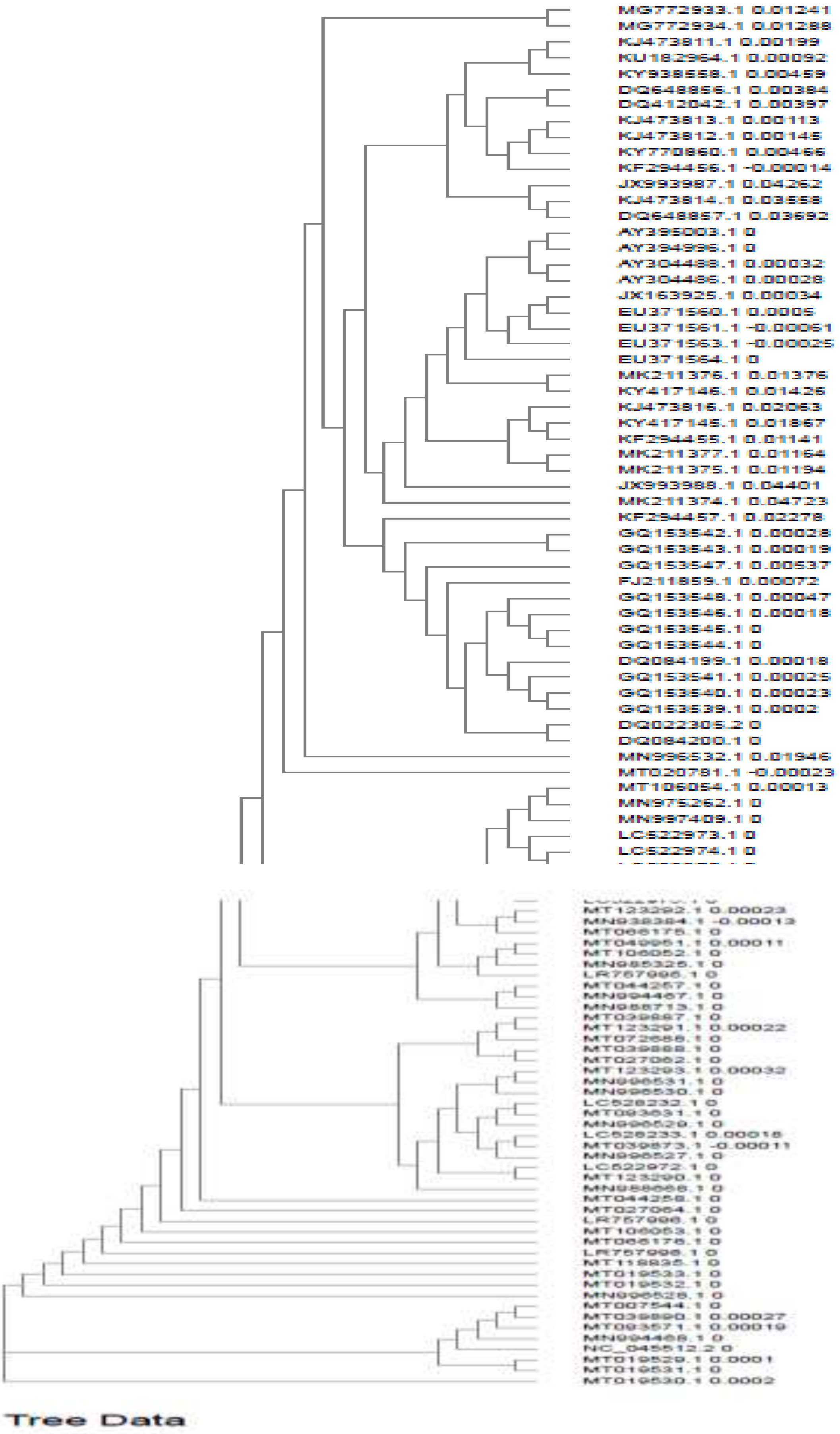


Figure 5: The Clustal Omega cladogram for top hundred sequences for MT188339.1 (SARS-CoV-2)

Using SWISS-MODEL, eight protein structures were predicted and two could not be predicted due to the reasons such as length of the sequence and absence of a reference sequence. Figure 6 – figure 13 show the predicted models of the proteins under consideration.

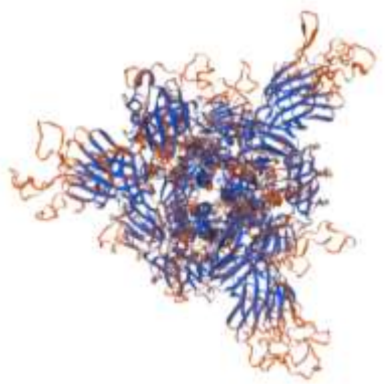


Figure 6: gene="S", product="surface glycoprotein", protein_id="QIK02944.1"

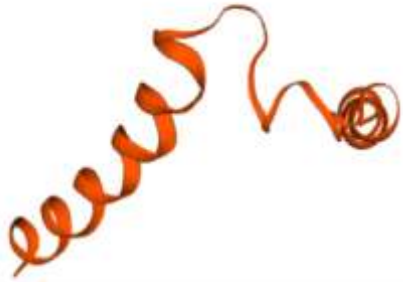


Figure 7: gene="orf3a", product="orf3a protein", protein_id="QIK02945.1"

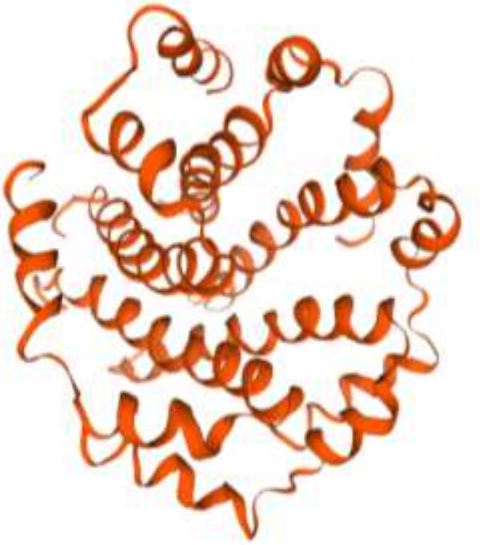


Figure 8: gene="E", product="envelope protein", protein_id="QIK02946.1"

Remaining two proteins, that is orf1ab polyprotein (gene: orf1ab) and orf10 protein (gene: orf10) were predicted using GOR tool and its secondary structure is shown in figure 14 and figure 15. Orf10 protein shows that the coils and extended strands are more often in its structure as shown in figure 15, whereas in orf1ab polyprotein the occurrence of extended strand, alpha helix and coil is more (Figure 14).

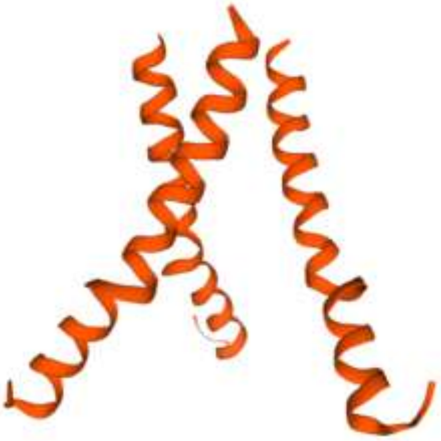


Figure 9: gene="M", product="membrane glycoprotein", protein_id="QIK02947.1"

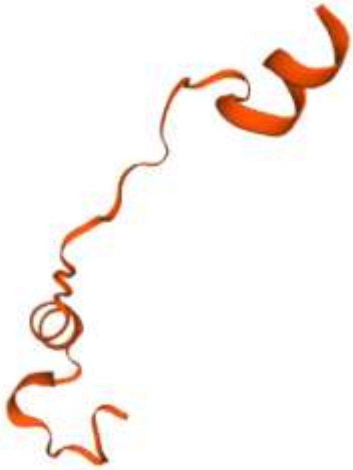


Figure 10: gene="orf6", product="orf6 protein", protein_id="QIK02948.1"

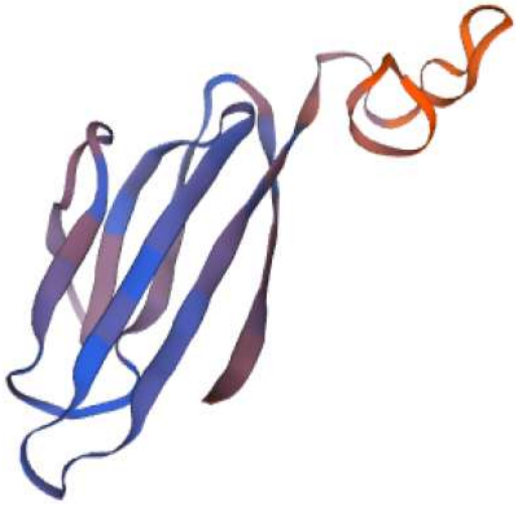


Figure 11: gene="orf7a", product="orf7a protein", protein_id="QIK02949.1"

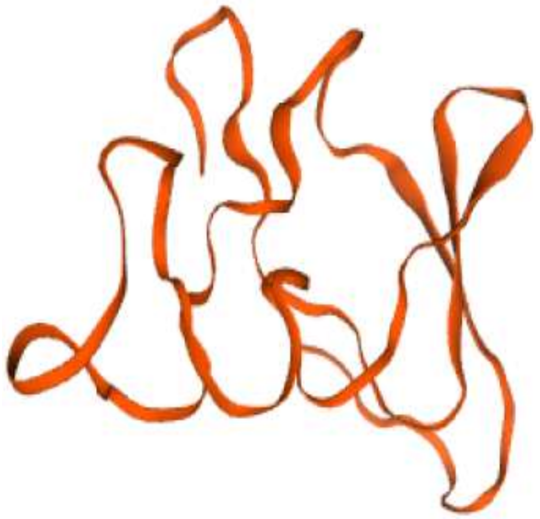


Figure 12: gene="orf8", product="orf8 protein", protein_id="QIK02950.1"

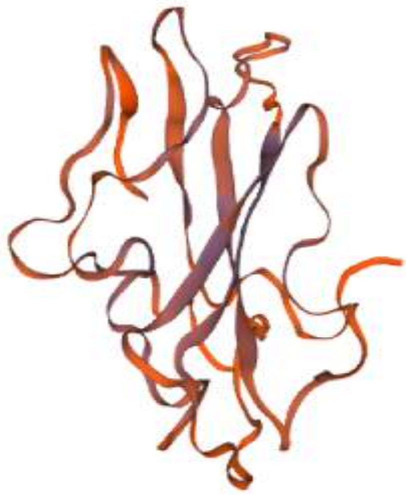
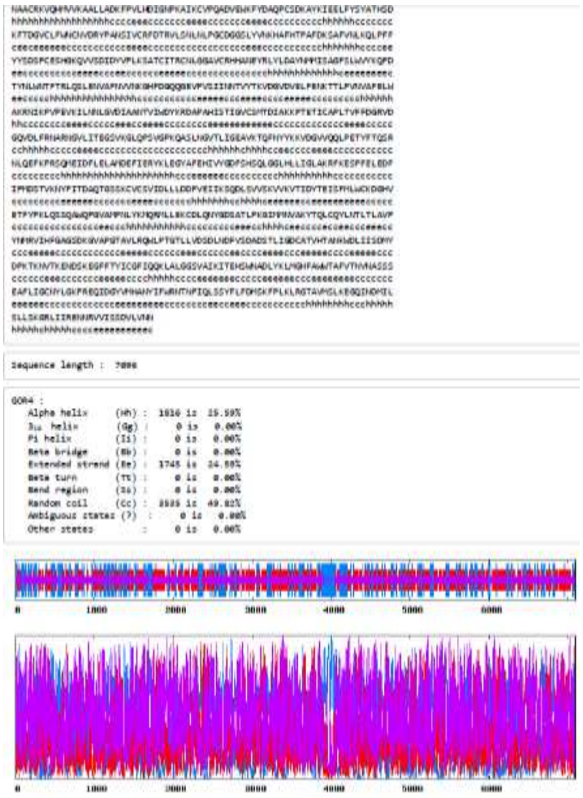


Figure 13: gene="orf8", product="orf8 protein", protein_id="QIK02950.1"

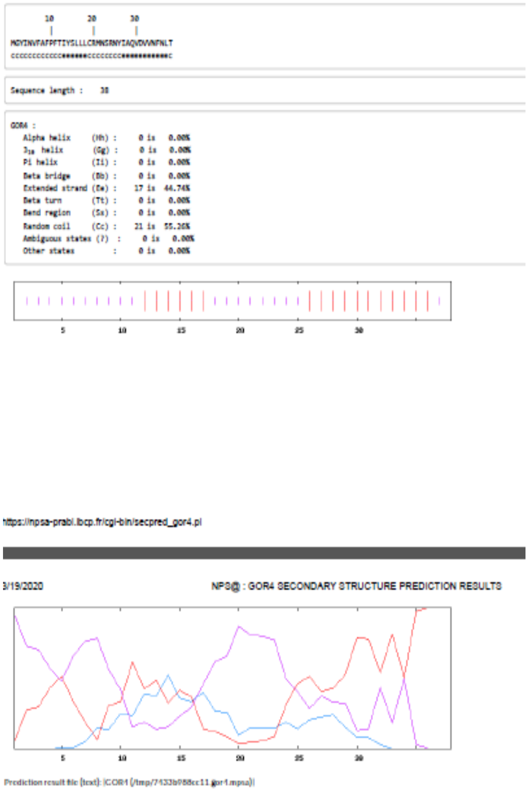


```
NAACRKVQHMVVKAALLADKFPVLHDIGNPKAIKCVPQADVEWKFYDAQPCSDKAYKIEELFYSYATHSD
hhhhhhhhhhhhhhhhhhccceeeccccccceeeeccccccceeeeccccccccccchhhhhhcccccccc
KFTDGVCLFWNCNVDRYPANSIVCRFDTRVLSNLNLPGCDGGSLYVNKHAFHTPAFDKSAFVNLKQLPFF
ceeceeeeeecccccccccccceeeecccccccccccccccccceeeeccccccccccchhhhhhhccccee
YYSDSPCESHGKQVVSDIDYVPLKSATCITRCNLGGAVCRHHANEYRLYLDAYNMMISAGFSLWVYKQFD
eecccccccccceeeeeccceeccccccceeeeecccccccccccccchhhhhhhhhhhhhhceeeeeeeec
TYNLWNTFTRLQSLENVAFNVVNKGHFDGQQGEVPVSIINNTVYTKVDGVDVELFENKTTLPVNVAFELW
eecccchhhhhhhhhhhhhhhhcccccccccccceeeeeccceeeeeccceeeeecccccccccchhhhhhh
AKRNIKPVPEVKILNNLGVDIAANTVIWDYKRDAPAHISTIGVCSMTDIAKKPTETICAPLTVFFDGRVD
hhcccccccceeeecccccceeecceeeecccccccceeeeeeeeeeeecccccccccccccceeeeccccc
GQVDLFRNARNGVLITEGSVKGLQPSVGPKQASLNGVTLIGEAVKTQFNYYKKVDGVVQQLPETYFTQSR
cchhhhhcccceeeeecccccccccccccccccccchhhhhhchhhhcceeccccceeeecccccccccccc
NLQEFKPRSQMEIDFLELAMDEFIERYKLEGYAFEHIVYGDFSHSQLGGLHLLIGLAKRFKESPFELEDF
ccccccccchhhhhhhhhhhhhhhhhhhhhccceeeeeecccccccccchhhhhhhhhhhcccccccccccc
IPMDSTVKNYFITDAQTGSSKCVCSVIDLLLDDFVEIIKSQDLSVVSKVVKVTIDYTEISFMLWCKDGHV
ccccccccceeeeeecccccccceeeecccccchhhhhhhhchhhhheeeeeeecceeeeeeeeeeccccc
ETFYPKLQSIQAWQPGVAMPNLYKMQRMLLEKCDLQNYGDSATLPKGIMMNVAKYTQLCQYLNTLTLAVP
ccccccccccccccceeecchhhhhhhhhhhhhccccccccccceeecchhhhceecccceceecccceeeec
YNMRVIHFGAGSDKGVAPGTAVLRQWLPTGTLLVDSDLNDFVSDADSTLIGDCATVHTANKWDLIISDMY
ccceeeeeccccccccccceeeeecccccceeeccccceecccceeeeccceeeeeecccceeeeeeccc
DPKTKNVTKENDSKEGFFTYICGFIQQKLALGGSVAIKITEHSWNADLYKLMGHFAWWTAFVTNVNASSS
cccccceeeccccccceeeeeccchhhhhcccceeeeeeeccccceeeeeeccceeeeeeeeeccccccc
EAFLIGCNYLGKPREQIDGYVMHANYIFWRNTNPIQLSSYSLFDMSKFPLKLRGTAVMSLKEGQINDMIL
eeeeeecccccccccccceeeeeeeeeecccccccceecceeeccccccccccchhhhhhhhcchhhhhhh
SLLSKGRLIIRENNRVVISSDVLVNN
hhhhhchhhhhcccceeeeeeeeeec
```

```
Sequence length :  7096
```

```
GOR4 :
   Alpha helix     (Hh) :  1816 is  25.59%
   3₁₀ helix       (Gg) :     0 is   0.00%
   Pi helix        (Ii) :     0 is   0.00%
   Beta bridge     (Bb) :     0 is   0.00%
   Extended strand (Ee) :  1745 is  24.59%
   Beta turn       (Tt) :     0 is   0.00%
   Bend region     (Ss) :     0 is   0.00%
   Random coil     (Cc) :  3535 is  49.82%
   Ambiguous states (?) :     0 is   0.00%
   Other states         :     0 is   0.00%
```

Figure 14: gene="orf1ab", product="orf1ab polyprotein", protein_id="QIK02943.1"



```
        10        20        30
         |         |         |
MGYINVFAFPFTIYSLLLCRMNSRNYIAQVDVVNFNLT
ccccccccccceeeeeecccccccceeeeeeeeeeecc
```

```
Sequence length :    38
```

```
GOR4 :
   Alpha helix     (Hh) :    0 is   0.00%
   3₁₀ helix       (Gg) :    0 is   0.00%
   Pi helix        (Ii) :    0 is   0.00%
   Beta bridge     (Bb) :    0 is   0.00%
   Extended strand (Ee) :   17 is  44.74%
   Beta turn       (Tt) :    0 is   0.00%
   Bend region     (Ss) :    0 is   0.00%
   Random coil     (Cc) :   21 is  55.26%
   Ambiguous states (?) :    0 is   0.00%
   Other states         :    0 is   0.00%
```

https://npsa-prabi.ibcp.fr/cgi-bin/secpred_gor4.pl

3/19/2020 NPS@ : GOR4 SECONDARY STRUCTURE PREDICTION RESULTS

Prediction result file (text): [GOR4 (/tmp/7433b988cc11.gor4.mpsa)]

Figure 15: gene="orf10", product="orf10 protein", protein_id="QIK02952.1

Conclusion

This work upgrades understanding of novel SARS-COV-2 (MT188339.1) to another level and touches the unexplored opportunities of bioinformatics. Numerous recent studies on this vibrant topic, SARS-COV-2 were studied, analysed and concluded with perspective of contributing more to the research community, in this rapidly evolving emergency situation.

COVID-19 is a deadliest pandemic disease and hence the need for the solution is crucial. Bioinformatic analysis would be one of the promising approaches for the advancements in this area of research. Considering these aspects, the whole genome of SARS-COV-2 (MT188339.1) was subjected to bioinformatic analysis which resulted in eight novel protein structures and two secondary structures.

Further through the cladogram, we could also conclude the evolutionary relationship between 100 strains of SARS COV2 virus.

References

1. Chan J.F. et al, Interspecies transmission and emergence of novel viruses: lessons from bats and birds, *Trends Microbiol*, **21(10)**, 544–555 **(2013)**

2. Jasper Fuk-Woo Chan et al, Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan, *Emerging Microbes and Infections*, **9**, 221-236 **(2020)**

3. Cheng V.C. et al, Severe acute respiratory syndrome coronavirus as an agent of emerging and remerging infection, *Clin Microbiol Rev.*, **20(4)**, 660–694 **(2007)**

4. Chan J.F. et al, Middle East respiratory syndrome coronavirus: another zoonotic betacoronavirus causing SARS-like disease, *Clin Microbiol Rev*, **28(2)**, 465–522 **(2015)**

5. https://www.who.int/health-topics/coronavirus

6. Dong L., Hu S. and Gao J., Discovering drugs to treat coronavirus disease 2019 (COVID-19), *Drug Discov Ther*, **14(1)**, 58–60 **(2020)**

7. Jasper Fuk-Woo Chan, Kin-Hang Kok, Zheng Zhu, Hin Chu, Kelvin Kai-Wang to, Shuofeng Yuan and Kwok-Yung Yuen, Genomic characterization of the 2019 novel humanpathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan, *Emerging Microbes and Infections*, **9(1)**, 221-236 **(2020)**

8. Sah Ranjit, Rodriguez-Morales Alfonso, Jha Runa, Chu Daniel, Gu Haogao, Peiris Joseph S., Bastola Anup, Lal Bibek, Ojha Hemant, Rabaan Ali, Zambrano Lysien, Costello Anthony, Morita Kouichi, Pandey Basu, Poon Leo and Hopkins Johns, Healthcare, Aramco & Dhahran, Saudi & Arabia, Complete Genome Sequence of a 2019 Novel Coronavirus (SARS-CoV-2) Strain Isolated in Nepal, *ASM Science Journal*, 10.1128/MRA.00169-20, **9 (2020)**

9. Chang Y., Tung Y., Lee K., Chen T., Hsiao Y., Chang H., Hsieh T., Su C., Wang S., Yu J., Shih S., Lin Y., Lin Y., Tu Y.E., Tung C. and Chen C., Potential Therapeutic Agents for COVID-19 Based on the Analysis of Protease and RNA Polymerase Docking, *Preprints*, doi: 10.20944/preprints202002.0242.v1 **(2020)**

10. Boratyn G.M., Thierry-Mieg J., Thierry-Mieg D., Busby B. and Madden T.L., Magic-BLAST, an accurate RNA-seq aligner for long and short reads, *BMC Bioinformatics*, **20(1),** 405 **(2019)**

11. Madeira Fábio et al, The EMBL-EBI search and sequence analysis tools APIs in 2019, *Nucleic Acids Research*, **47(W1)**, W636–W641 **(2019)**

12. Waterhouse A., Bertoni M., Bienert S., Studer G., Tauriello G., Gumienny R., Heer F.T., De Beer T.A.P., Rempfer C., Bordoli L., Lepore R. and Schwede T., SWISS-MODEL: homology modelling of protein structures and complexes, *Nucleic Acids Res*., **46(W1)**, W296-W303 **(2018)**

13. GOR secondary structure prediction method version IV, Garnier J., Gibrat J.F. and Robson B., *Methods in Enzymology*, Doolittle R.F. Eds., **266**, 540-553 **(1996)**.

(Received 20th March 2020, accepted 31st May 2020)