HeLIX
The Scientific Explorer

Helix (2022) 12 (1): 22-26

https://doi.org/10.29042/2022-12-1-22-26

In silico Annotation and Mutational Hotspot Prediction in NL63 Gene of SARS CoV2

*[1]Indra Riyanto,[2]G, Jyothsna, [3]Sucharita Nayak

Faculty of Engineering, Universitas Budi Luthur, Jakarta, Indonesia
[2]Bioaxis DNA Research Centre, Hyderabad, India
[1]Khallikote Autonomous College, Brahmapur, Odisha, India
ndra.riyanto@budiluhur.ac.id

Received: 24th December 2021, Accepted: 15th February 2022, Published: 28th February 2022

Abstract

The first case of Corona virus was reported in China which lightened the research study related to the pathogen and the sequencing of the viral genome. The SARS gene sequences were submitted to the Public data base in December 2019 which paved the way for enormous study related to the pathogen. From the time of its discovery and till date the virus is known to exhibit several genetic mutations and strain variation which makes its elimination more challenging. The use of viral genomic sequences and the application of various bioinformatics techniques helps in understanding the genetics of the organism and its mutational annotation. The current work involves the mutational annotation of NL63 one of the pathogenic proteins of virus. Identification of probable mutational hotspots in the protein sequence of the virus helps to develop suitable therapeutics targeting these regions and can overcome the problem of drug resistance by organism.

Key words: *SARS CoV, Mutational Hotspot, Viral genome, NL63*

Introduction

The emergence of the new virus name Corona causing COVID 19 disease was first identified in Bats of China which later spread to the humans in Wuhan city during early March 2019 followed by its complete spread globally. The outbreak can be assigned to eh negligence in taking the preventive measures and early actions against its spread.

At the later end of 2019 a focus was laid on this new viral infection with pneumonia like condition exhibiting unfamiliar etiology. A detailed description about the outbreak was officially reported in China during early days of 2020. Initially the causative organism was named as Novel Corona Virus 2019 by WHO based on its year of origin and completely new behavior. However it was renamed later as Severe Acute Respiratory Syndrome Corona virus 2: SARS CoV2 by the international committee of Corona Virus Study Group (CSG). The disease caused by this virus was termed as COVID-19 Corona Virus Disease 2019. Soon the disease was identified as a pandemic and thrown its claws universally infecting a huge number of people exponentially. Several studies were focused to identify he measures to prevent spread of the virus which resulted in a great failure. Isolation of infected individuals and the suspects was one of the measures to prevent contact infection but could not be satisfactory. Several approaches including lock down were practiced which tough could decrease the intensity of spread could not completely eliminate the infection.

Drug Repurposing and Vaccination were identified to be the better treatment and prevention protocols which were later focused. Other alternative treatments emphasized were Plasma therapy and steroid drug administration.

Materials and Methods

Protein Sequence retrieval: Annotation and detailed study of the protein essentially requires its amino acid sequence to be collected and analyzed. Thus the protein sequence of NL63 was retrieved from NCBI database.

NCBI is a premier database which is publicly available and free for the users. Being a redundant data base the results obtained for 1 single search hit would be multiple. The user has to screen and logically select the best result matching the requirement. NCBI is a huge archive of protein, gene, SNP etc with its direct connections to the several other related databases and some basic annotation tools. Each data available in NCBI is sorted, structured, cross checked and stored thus would not have false information. Further it accepts even updates for the submitted data by the same author to facilitate on time information and accuracy.

Physicochemical characterization of protein: The query NL63 protein sequence was subjected for characterization of sequence based on the amino acids present. The tool used for the purpose was protparam. The characteristics studied include sequence length, molecular weight, positive amino acid residues, negative residues, half life of the protein, its stability, polarity in terms of hydropathicity etc. This information is used for analytical studies on the protein. All the properties provided by protparam are the calculated values based on the protein sequence submitted to the tool.

Annotation of structural confirmations in the protein: Based on the sequence of amino acids present in protein it undergoes conformational changes by rotations and coiling to produce its secondary structure. Various secondary structural confirmations in a protein include helix, coil, loop turn, extended strand etc. All the polar amino acids are packed in the superficial areas of the structure where as he non polar hydrophobic amino acids are pushed towards the interior of the protein yielding a secondary structure of a linear peptide. Identifying these confirmations helps in several docking studies and targeting of the protein. Tool used to identify the secondary structural confirmations in the protein is SOPMA.

Tertiary structure prediction of the protein: Apart from the secondary structural confirmations, it is essential that the final 3D structural package of the protein must be known for any further structure based studies. The exact packaging of the protein in three dimensions would enable the user to identify the characters like binding sites, pockets, loops etc in the protein. RCSB Protein data bank is an archive that stores the 3D structural details of all the known proteins along with their basic level annotations. Each entry in PDB is provided a unique code called PDB ID which is universal and can be used as a structure identifier for the protein. The data base helps to identify the structural and functional along with some basic evolutionary details of the protein.

Domain annotation: Each protein is composed of one or more independently functioning units called domains. Each domain is specific for a unique function and can work independent of the other protein regions. By studying the domains of any protein it would be easy to trace back its evolution and function easily. Domain annotation also helps to identify the vital regions in the sequence. SMART is a domain annotation tool which provides an insight into the various structural and function regions of importance in the user entered protein sequence. It also provides additional links and sources for further study of these regions and their evolution.

Disorder prediction: Each protein posses some sensitive regions based on their amino acid composition and structural confirmation which make them prone for mutations. Such reins are called mutational hotspots. Apart from being hotspots these regions may also prone for disorders. Thus it becomes a pivotal step to identify these regions of sensitivity. GLOBPLOT is an online tool that identifies the globular regions within the protein that have high probability of disorder. Globplot is also helpful for domain hunting and identification of unstructured motifs within the protein.

Identification of antigenic sites: For an organism to be pathogenic it poses some specialized regions called antigenic regions within its protein sequence. These regions would elicit an immune response upon interaction with the host body. Within these immunogenic regions there are certain sites called antigenic sites which possess highest antigenic propensity. EMBOSS antigenic is a tool that predicts these antigenic sites, regions and their score. The data can be used to select the best regions of antigenicity for vaccine development. The tool is available in Protein Variability server at http://imed.med.ucm.es/Tools/antigenic.html.

Comparison of GLOBPLOT regions with antigenic site results: The results of both Globplot and PVS are compared to select the regions of antigenicity falling in the disorder prone regions of the sequence. Based on the common sequences available in all tools the best antigenic sites can be filtered.

Effect of substitution mutation in the proposed mutational regions: After finalizing the regions of high anigenecity falling in the domain regions, these sites were substituted with all possible mutations and checked for their effects on the stability and functionality of the protein. Polyphen is an online tool that shows the effect of substitution mutations at the user selected sites and provides an insight to the stability changes of the protein.

Results and Discussion

Protein sequence was retrieved from NCBI and the following details were obtained. Length of the sequence was found to be 1356amino acids with the corresponding molecular weight to be 149864.45. The isoelectric point of the protein was calculated to be 6.83. It was found to contain early same no of acidic and basic amino acids making it a neutral protein.

Secondary structural annotation of the protein revealed the protein to be globular with high no of coil and helical confirmations and only little no of beta turns. The protein was found to be hydrophilic and polar.

After performing the structural, domain and antigenic site determinations some of the important regions in domains and most antigenic with high mutational chances are selected and tabulated below.

Table 1: Selected Peptides From PVS (Antigenic) And Globplot (Disorderd)

S.No	GLOBPLOT	PVS	COMMON REGION SELECTED
1	33-40	36-68	36-40
2	206-216	198-209	206-209
3	221-237	216-222	221-222
4	262-287	268-275	268-275
5	331-351	340-353	340-351
6	494-498	493-507	994-498
7	531-545	540-554	540-545
8	550-555	NO	NO
9	574-578	562-583	574-578
10	604-613	606-615	606-613
11	650-665	648-655	650-655
12	715-727	724-730	724-727
13	738-753	733-746	738-746
14	794-799	771-796	796-799
16	1103-1117	1114-1121	1114-1117
17	1233-1237	1236-1254	1236-1237
18	1322-1343	1297-1331	1322-1331

Inference: From the above table it can be explained that all the above 18 peptide regions are those vital regions in the protein that are involved in functioning of the protein, present in the hot spot regions of mutation and also the disorder prone regions of the sequence. Thus, they can be used for targeting and vaccine development. Among all the above 18 peptide regions the one with highest antigenic score was identified to be 1322 to 1343. Within this peptide the site 1340 was identified to be most antigenic in nature. It was subjected for Polyphen analysis to predict the effect of substitutions a site.

```
(33) Score 1.128 length 19 at residues 1335->1353
                     *
 Sequence: RGCCDCGSTKLPYYEFEKV
           |                   |
         1335                1353
Max score pos: 1340

(34) Score 1.127 length 14 at residues 705->718
                       *
 Sequence: LQNLLQLPNFYYVS
           |             |
         705           718
Max score pos: 714

(35) Score 1.126 length 33 at residues 37->69
                     *
 Sequence: STIVTGLLPTHWFCANQSTSVYSANGFFYIDVG
           |                                 |
          37                                69
```

Fig 1: Results of EMBOSS ANTGENIC showing antigenic peptide and score

Inference: The above results of EMBOSS ANTIGENIC shows the antigenic score of the selected peptide 1335-1353 as 1.128 and the antigenic site was identified to be C at the position 1340. This was further subjected for Polyphen analysis to detect the effect of mutation at this site.

PolyPhen-2 report for AFV53148.1 C1340N

Query

Protein Acc	Position	AA_1	AA_2	Description
AFV53148.1	1340	C	N	protein S [Human coronavirus NL63]

Results

Prediction/Confidence	*PolyPhen-2 v2.2.2r406*

HumDiv

This mutation is predicted to be **PROBABLY DAMAGING** with a score of **0.998** (sensitivity: **0.27**; specificity: **0.99**)

HumVar

Details

Multiple sequence alignment	*UniProtKB/UniRef100 Release 2011_12 (14-Dec-2011)*
3D Visualization	*PDB/DSSP Snapshot 25-May-2021 (178229 Structures)*

Fig 2: Polyphen analysis at 1340 site

Inference: The above pictures show that the change n the amino acid C to N at position 1340 was shown to be highly damaging by polyphen tool. Thus this is the most sensitive region in the protein.

Conclusion

In view of the intensity of spread and damage caused by the pandemic the current work was undertaken to annotate the NL63 protein of SARS COv2.

This protein is one of the most important structural proteins in SARS CoV2 hence it has been subjected for complete molecular analysis to identify the mutation prone hot spot regions which may be the cause for the drug resistance and high stability of the organism. The *in silico* work included the collection of protein sequence, annotation of its physicochemical properties, Functional region recognition and analysis, Structural study, identification of mutation hot spots and disordered regions within the sequence etc. The above work was successful in identifying the hot spot regions which are more sensitive and mutation prone. Further using polyphen tool, effect of the mutations on this site 1340 C was identified to be potentially hazardous.

References

1. Chen L, Liu W, Zhang Q, Xu K, Ye G, Wu W, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg Microbes Infect. 2020;9:313–9.
2. Kashyap, A., Gunjan, V. K., Kumar, A., Shaik, F., & Rao, A. A. (2016). Computational and clinical approach in lung cancer detection and analysis. Procedia Computer Science, 89, 528-533.
3. Holshue M L D, Lindquist S L, Wiesman J B, Spitters C E, Wilkerson S T. First Case of 2019 Novel Corona virus in the United States. NEnglJ Med 2020 101056/NEJMoa2001191. 2020
4. Huang et al. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance. 2020;25(3)
5. Shaik, F., Sharma, A. K., Ahmed, S. M., Gunjan, V. K., & Naik, C. (2016). An improved model for analysis of Diabetic Retinopathy related imagery. Indian J Sci Technol, 9, 44.
6. Korber et al., 2020, Cell 182, 812–827 August 20, 2020 Published by Elsevier Inc. https://doi.org/10.1016/j.cell.2020.06.043
7. Wise J. Covid-19: New corona virus variant is identified in UK. BMJ. 2020 Dec 16; 371:m4857. doi: 10.1136/bmj.m4857. PMID: 33328153.