Advances and Applications in Mathematical Sciences
Volume 20, Issue 12, 2021, Pages 3147-3156

STRUCTURE PREDICTION AND EXAMINATION OF SPIKE PROTEIN FROM SARS-COV2 USING COMPUTATIONAL AND SIMULATION TOOLS

T. SUDHA RANI and RAYUDU SRINIVAS

Computer Science and Engineering
Aditya Engineering College
India
E-mail: sudharani.tirukoti@aec.edu.in
srinivas.rayudu@aec.edu.in

Abstract

Recent emergence of COVID-19 corona virus has resulted in WHO-declared public health emergency of international concern. Research around the globe is working towards establishing a great understanding of this particular viruses and developing treatments and vaccines to prevent spread. Knowing structural properties of a protein provides an important resource for understanding how it functions. This paper provides structural details of Spike Protein(s) of novel corona virus. The Spike protein of nCOV2 is prominent for confining with a cell receptor which is act as a referee for the synthesis of virus and host membranes, and these activities being crucial for virus ingress in to host cell. This paper studies primary, secondary and tertiary structures of spike protein of SARS-COV2 predicted using Exapasy Program, PSIPRED, and Homology Modeling Methods respectively [1]. The predicted structure was validated using PROCHECK by Ramachandran plot and also validated through ProSAWeb tool. Finally, this predicted structure will helps to discovering efficient medicine against corona epidemic. In future, resultant structure of homology modeling would be energy minimized and can do MD (Molecular dynamics) simulations to examine how the expected model behave structurally, dynamically, using several computational and simulation tools.

I. Introduction

Corona Virus Disease (COVID-19) is a epidemic disease discovered recently known as corona virus (SARSCOV2). COVID-19 belongs to Coronaviride family, mainly comprise harmful bacterium with zoonotic

2010 Mathematics Subject Classification: 82M37.
Keywords: Spike Protein, Molecular Dynamics, Homology Modeling, PSIPRED, PROCHECK.
Received October 13, 2020; Accepted November 8, 2020

attribute, Severe Respiratory Syndrome (SARS-COV), Middle East Respiratory Syndrome (MERS-COV) of this group already started in 2003 and at present this COVID-19 has transpired in China. These are single stranded RNA germs which could be secluded in various zoological genus. This lengthy strand of ribonucleic acid (RNA), which serves as the viruses genetic material. When this virus infects a cell, it hijacks the molecular machinery to create long chains of proteins required by the virus to generate even more copies itself. The spike protein of SARS-COV2 can easily interact with host cell receptor ACE2.

The complete viral particle of a nucleocapsid (N) core encircled by an envelope contains 3 membrane proteins, spike (S), membrane (M), envelope (E) which are equivalent to whole members of genre. The Spike (S) Protein, it biologically appears like a projections on the exterior of the viral particle, intermediates confining to host cells and membrane fusion. This Spike glycoproteins comprises 2 sub units (S1 and S2). Newest study says that, a spike variation, possibly appeared in November 2019, activated and leaping to human beings.

The 3D structure of spike protein is essential to discovery and development of antiviral drugs. Plenty of methods are available to visualize 3D structure of a Protein like homology modeling, threading, and abinito methods. One of the most robust, strong and widely used methods for interpreting 3D Structure is homology modeling. It is Comparative modeling, provides atomic resolution model, which models a structure deploy on the sameness of query sequence with given target protein. The modelled Structures which are generated by above techniques are static but by nature all proteins are dynamic in nature.

II. Methods and Tools

A. Data Collection

The amino acid sequence of Spike protein of COVID-19 (Uniprot ID:P0DTC2) extracted from Uniprot Data Source, which is available in www.uniport.org.

B. Prediction of Primary Structure

Several physico-chemical attributes like composition of amino acid, atomic composition, Extinction Coefficient, Grand Average Hydrophobicity (GRAVY), etc relating to primary structure of P0DTC2 were anticipated by examining the amino acid sequence arrangement of the Spike Protein by a server called of Expasy's ProtParam server. This is available at web.expasy.org.

C. Prediction of Secondary Structure

The secondary structure of spike protein was anticipated using PSIPRED server. The secondary structure was analyzed with 3D model structure to anticipate the structural characteristics of amino acid residues in various structural regions of the model obtained by modeller 9.23.

D. Prediction of 3D Structure

The template for the modeling structure can be take out by applying BLAST-PROTEIN (BLASTp) sequence similarity/search tool, which is available in NCBI website. BLAST-PROTEIN, performing the sequence similarity/search by taking Uniprot: P0DTC2 as target sequence in the form of FASTA format. From the results of BLASTp, we extracted top three resultant proteins based on highest sequence similarity and their PDBID's are,6VSB,6VXX,6VYB(SARS-COV-2) and they shared identity with queried protein as 99.59%.99.50%,99.42% respectively. The three dimensional structure of the input protein sequence was obtained by modeller 9.23 Homology and Comparative tool.

E. Structure Analysis and Visualization [2]

The Predicted structures obtained by Homology modeling can be visualized by PyMol, python based molecular visualization tool. The RMSD analysis of this structure analyzed with the template and typically energy minimized structure helpful to determine the accuracy of the model. The resultant structure of steriochemical stability attribute can be analyzed with a tool called PROCHECK.

III. Result and Discussion

A. Primary Structure Prediction

To analyze primary structure of a protein, Prot Param tool from a Expasy Server was used. Prot Param computes different kinds of physico-chemical attributes which can be extrapolate from a given protein input sequence. The input protein sequence can be taken up as Swiss rot/TrEMBL accession number/ID or it may be in form of raw sequence. Here, we used raw sequence and uploaded as FASTA format. And these calculations performed by ProtParam, which are based on either *N*-terminal amino acid (or) structure data.

The result generated by Expasy's ProtParam for the spike protein, contains 1273 amino acid residues with approximate molecular weight of 141178.47.The theoretical pI(pH at which protein remains stable) was anticipated to be 6.24 which is < pI=7 which says that the protein is acidic by nature. Table 1 shows that several physico-chemical properties of spike protein from SARS-COV2 obtained from Expasy ProtParam Tool.

Table 1. Physico-chemical properties of spike protein from SARS-COV2 obtained from Expasy ProtParam Tool.

Parameters	Predicted Value
Molecular Weight	141178.47
Theoretical pI	6.24
Number of Positive Residues	103
Number of Negative Residues	110
Half-life mammalian reticulocytes(in vitro)	30hrs
Half-life yeast	>20hrs
Half-life E. Coil	>10hrs
Extinction coefficient	148960-146460
Instability Index	33.01
Aliphatic Index	84.67
GRAVY Index	-0.079

This approximation is helpful to develop the storage system of a buffer for the amino acid of regard. The total number of positive residues(Arg+Lys) are 103 and the total number negative residues(Asp+Glu) are 110.The half life cycle can be articulate as the time required to decay a protein to its half concentrated after the process of synthesis. The half life of P0DTC2 for 3 morphons (human, yeast, E-coil) was approximated to be 30 hours for mammalian reticluocytes in vitro and 20 years for yeast in vivo and more than 10 hours for *E*-coil in vivo. One of the prominent parameter given by Protparam is Extinction Co-efficient. There two types of Extinction Co-efficient, one that assumes that all cystine residues form cystine bonds and other assumes that all cystine residues as being reduced or unbounded. Cystine is a important amino acid that forms bonds with other amino acids mainly between cystine and it gives a stability to the protein. Formally, Extinction Co-efficient can be defined as the absorbance of a Protein sample at 280nm measured in water with a spectrophotometer. And Extinction Co-efficient of P0DTC2 was approximated to be scaling between 148960-146460 based on composition of cystine.

>sp|P0DTC2|SPIKE_SARS2 Spike glycoprotein OS=Severe acute respiratory syndrome coronavirus 2 OX=2697049 GN=S PE=1 SV=1

MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHST
QDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNII
RGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKS
WMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGY
FKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPG
DSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCT
LKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAW
NRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
DEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLY
RLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVG
YQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLT
ESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTS
NQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEH
VNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAY
SNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSF
CTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPS

KPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVL
PPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVT
QNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTL
VKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIR
AAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFL
HVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEP
QIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPD
VDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWP
WYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPV
LKGVKLHYT

FASTA Sequence of Spike Protein of SARS-COV2 (P0DTC2) in FASTA format take out from Uniprot Website.

Statistical review of 12 unstable and 32 stable proteins that exhibit significant difference in appearance of certain dipeptides (a peptide composed of 2amino acid residues) in unstable proteins as analyzed with stable proteins. An experiment is produce to correlate catabolic stability of proteins with attributes of their primary sequence here weight values of instability for a protein of template could accordingly be applied as an indicator for anticipating its stability characteristics.

Protparam also exhibits that the proteins with instability index of < 40 were stable and value > 40 were anticipated to be unstable. The instability index (II) of P0DTC2 was calculated as 33.01, indicating that protein will be stable in vacutainer. Another characteristic is aliphatic index, defined as relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine).Higher the value designated that higher stability of a protein. The aliphatic index of **P0DTC2** is anticipated to be 84.67. The GRAVY (Grand Average of Hydropathy) indicates that likelihood of the interaplay of protein with water. The lower value of GRAVY index yields higher likelihood that protein will interplay with water. The GRAVY index for **P0DTC2** is anticipated to be -0.079.This score exhibits that the protein will be interplaying with water, which says that the nature of the protein is immunogen.

B. Secondary Structure Prediction:

Generally, Secondary structure of a protein comprises helices, sheets, and

coils. To predict the secondary structure of (**P0DTC2**) PSIPRED was used. This server usually based on the analysis of result obtained from PSIBLAST (Position Specific Iterated-BLAST). PSIPRED returns the results of neural network based 3-state secondary structure predictor. The result of the PSIPRED shown in Figure 2 as Secondary Structure of SARS COV2 Spike Protein.

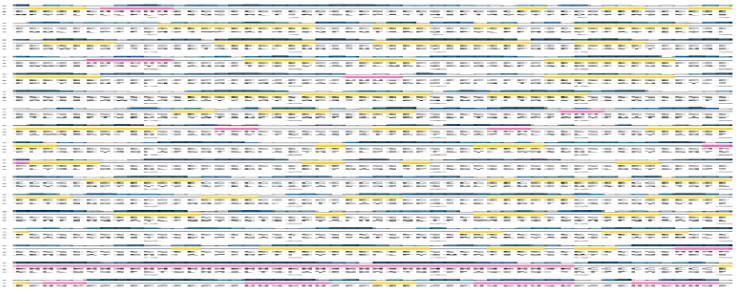

Figure 1. Secondary Structure of Spike Protein of SARS COV2.

C. Tertiary Structure Prediction

The 3D structure of P0DTC2 was obtained or modelled with Comparative and Homology Modeling tool, Modeller 9.23. The target Protein was given as query sequence to Blastp to determine the homologous sequences [3]. The top three results extracted and which are shared 99% identity, such as PDB ID's:6VSB, 6VXX, 6VYB with input sequence. Therefore, these are selected as a template structures for Homology Modeling. The obtained 3D structure of the target Spike glycoprotein from SARS-COV2 (Uniprot ID: P0DTC2) and template SARS-COV2 from PDB ID's:6VSB, 6VXX, 6VYB were superimposed and calculated RMSD (Root Mean Square Deviation) value, the template structure 6VXX got least RMSD value as 0.793,which shows predominantly good quality of the modeled structure [4]. The Figure 2 shows that superimposed structcture of target and 6VXX template generated by Pymol.

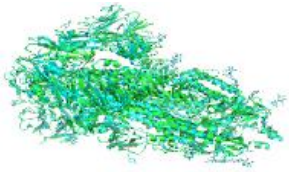

Figure 2. Super imposed structures of target PDBs 6VSB, 6VXX, 6VYB with template 6VXX by Pymol molecular viewer tool. Calculated RMS=0.793.

Figure 3. Tertiary Structure of target Protein extracted by PROCHECK.

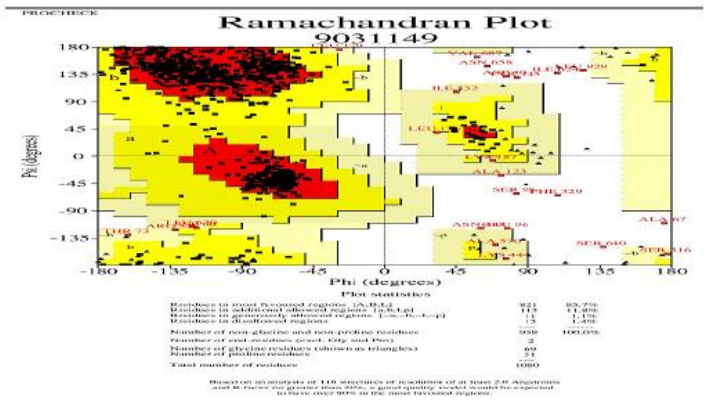

Figure 4. Ramchandran plot of P0DTC2 produced by PROCHECK rendering the occurrence of 85.7% amino acids in core part.

The plot of Ramachandran Figure 4 of the model obtained by tool called PROCHECK renders that number of residues in favoured region are 821 (85.7%), Number of residues in allowed region are 113 (11.8%), Number of residues in generously allowed region are 11 (1.1%), Number of residues in disallowed region are 13 (1.4%). These results tells that the model is stereo chemically stable. We can also verify 3D protein structure using PROCHECK. The Figure 3 shows that 3D Protein structure of predicted modelled structure generated by PROCHECK.

The validation is important step in Homology Modeling. To do this, ProSA-Web tool is used [5]. It is a Web-based Protein Structure Analysis tool, provides easily available online user interface for protein structure validation. Pro-SA computes an overall virtue score for a given input structure by means of Z-Score value. This Score also figured out in a graph as

plot, which contains Z-scores of all experimentally set protein chains in a given Protein Data Bank (PDB) input file. The Z-Score returned by this web server for modelled structure is -9.59. The Z-score plot shown in Figure 5, a class of structures from various sources (X-Ray, NMR) are differentiate by various colors. This will helps to verify whether Z-score modelled structure with in span of scores customarily retrieve for local proteins of similar size.

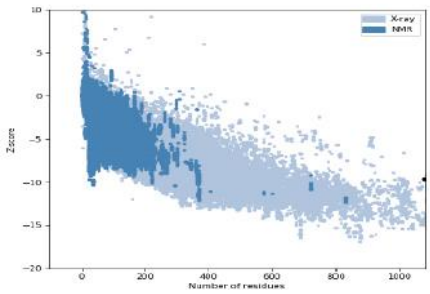

Figure 5. P0DTC2 Z-score plot generated by ProSA-Web tool.

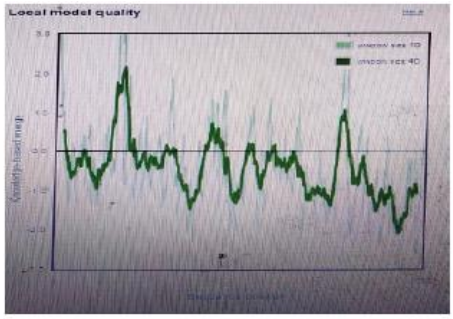

Figure 6. P0DTC2 residue scores plot generated by ProSA-Web.

Another graphical plot generated by ProSA web server is that position of amino acid sequence i. From this p of given input structure. And this plot is fragment of 40 residues, since single residue evaluation. Window size of 40-residue fragment denoted by the thick line, residue fragment is rendered in the plot as shown in the Figure 6.

Conclusion

The obtainability of robust 3D structure of molecular target is very crucial for drug discovery. This analysis, primary, secondary and 3D

structures of spike protein of SARS-COV2 predicted using Exapasy Program, PSIPRED, and Homology Modeling Methods correspondingly. The structure examination of predicted model was performed using ramachandran plot and also validated through ProSAWeb tool. The Ramachandran plot shows that 85.7% residues were fall in to core region says that the obtained resultant model structure is stereo-chemically stable. This structure examination of spike protein yields a valid platform for the designing specific antiviral medicine or drugs against SARS-COV2. In future, we can study molecular dynamics of above target protein to examine how the predicted model behave structurally, dynamically, thermodynamically by molecular dynamics and simulation tools.

References

[1] Rakesh Kr Pandit, Tapan K. Mukherjee, Ashok Kumar, Varinder Kumar and Palki Sahib Kaur, Structure Prediction and Assessment of Beta-Lactamase Tem-1 from S. Typhi Using Molecular Dynamics and Simulation studies, International Journal of Recent Scientific Research Research 7(3) (2016), 9509-9513.

[2] Emanuele Bramucci, Alessandro Paiardini, Francesco Bossa, Stefano Pascarella, Py Mod: sequence similarity searches, multiple sequence-structure alignments, and homology modeling within PyMOL, From Eighth Annual Meeting of the Italian Society of Bioinformatics (BITS) Pisa, Italy. 20-22 June 2011.

[3] Arun K. Shanker, Divya Bhanu and Anjani Alluri, Whole Genome Sequences Analysis and Homology Modeling of a 3C Like Peptidase and a Non-Structural Protein 3 of the SARS-CoV-2 Shows Protein Lig and Interaction with an Aza-Peptide and a Noncovalent Lead Inhibitor with Possible Antiviral, Pre prints February, 2020.

[4] Kuo-Yuan Hwa, Wan Man Lin, Yung-I Hou and Trai-Ming Yeh, Molecular Mimicry between SARS Coronavirus Spike Protein and Human Protein, Frontiers in the Convergence of Bioscience and Information Technologies 2007.

[5] Wiederstein and Sippl, ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. m Nucleic Acids Research 35, W407-W410 (2007).