Machine Learning Approach to Identify Receptor Binding Domain of Spike Glycoprotein as A Potential Vaccine Candidate for COVID-19

Aryan Prajapati[*], Kamal Rawal and Preeti P

Amity Institute of Biotechnology, Amity University, Uttar Pradesh, India

[*]Corresponding Author: Prajapati A, Amity Institute of Biotechnology, Amity University, Uttar Pradesh, India.

Abstract

Recent SARS-CoV-2 outbreaks have spurred continuing efforts to exploit different viral protein targets for therapy, but preventing viral proteins, including in therapeutic and vaccine research, has largely failed. In the lack of clear clinical proof for COVID-19 pathogenesis, a comparison of previous pandemic HCoVs-related immune system reactions could provide insight into COVID-19 pathogenesis. Authors summarize the possible genesis and method of spread of COVID-19, in addition to our present understanding of the viral genome integrity of known outbreak viruses against SARS-CoV-2 in this study. The COVID-19 pandemic continues to be a major concern for health-care systems globally. Accurate and timely identification of SARS Coronavirus 2 (SARS-CoV-2) infection is crucial for limiting dissemination and commencing therapy. The gold standard among test procedures is real-time reverse-transcriptase polymerase-chain reaction (rRT-PCR). Despite the fact that this test has a high specificity and sensitivity, the incidence of erroneously negative findings in patients with symptoms and/or having a positive CT scan remains a difficulty. In this article authors analyze the receptor binding domain of spike glycoprotein to be potential vaccine candidates.

Keywords: SARS-CoV-2; Respiratory diseases; Omicron; RNA-dependent RNA-polymerase; Covishield.

Introduction

COVID-19 is a viral infection induced by the Sars cov-2 virus. Maximum individuals affected with the viral infection will suffer from mild moderate to severe respiratory illness; however, some people will be severely impacted while others will heal quickly. People above the age of 65, as well as those with other health issues such as chronic respiratory disease, diabetes, cardiovascular disease, or cancer, are at a greater risk of developing severe illness. People of any age can become severely ill or die as a result of COVID-19.

The best method to protect yourself is to learn everything you can regarding the virus and

the way it spreads. To be safe, keep 1 meter away from people, make use of a mask and sanitizer, and lastly, get vaccinated. The virus can spread through droplets released from the nostrils and throat while speaking, sneezing, singing, and so on. These particles can be tiny or large respiratory particles.

COVID-19 has different effects on distinct people. The majority of infected individuals will experience mild moderate to severe symptoms. Fever, cough, exhaustion, loss of scent and flavor, sore throat, headache Pain and Aches Diarrhea. A skin rash or discoloration of the epidermis and toes, eyes that are red or inflamed Chest pain caused by breathing difficulties or shortness of breath, trouble with speech or mobility, or confusion.

Human coronavirus structure, pathogenicity. Human coronavirus is a collection of viruses known as coronaviruses that cause a variety of respiratory disorders such as pneumonia, the common cold, as well as bronchitis. Because of their high genome substitution and recombinant rate, CoVs are renowned for their rapid evolution. The infectious agent and its host held a common connection that included multiple viral and host factors for viral infection and, potentially, pathogenesis. Because there is no possible therapeutic drug or vaccine for HCoVs, an immediate implementation of high vigilance, such as for SARS-CoV-2, has been suggested for prevention and infection control. According to the findings of the whole genome research, the SARS COV-2 is an associate of BETA CoV linkage B [1].

Similarly, MERS-COV has been discovered to be a new Beta CoV Linkage C virus with high pathogenicity. Phylogenetic study based on RNA-dependent RNA polymerase (RdRP) along with spike (S) gene sequences identified SARSCoV as a member of the BetaCoV subgroup. The sequence study of SARS-CoV-2 revealed 88% identity to the sequences of two bat-derived SARS-like CoVs: bat-SL-CoVZC45 and bat-SLCoVZXC21, as well as 79 and 50% similarity to SARS-CoV and MERS-CoV, respectively [2].

This includes phosphorylated nucleocapsid protein as the core, surrounded by phospholipid bilayers to form globular or pleomorphic particles with an exterior surface projected spike (S) protein. The polybasic breakdown site (RRAR/S) of SARS-CoV and SARSCoV-2 is located at the junction of two S protein components. (S_1 and S_2). Adding the furin cleavage site (FCS) at the S_1-S_2 interface improved cell-cell fusion without affecting viral entrance, and efficient digestion enabled MERS-CoV and CoVs from bats to enter human cells, according to study with SARS-CoV and MERS-CoV. Furthermore, M protein is abundant in virion structure and is required for viral form, assembly, and the creation of mature viral coats [2].

During viral particle formation, the N protein is engaged in genomic RNA packaging. SARS CoV-2 follows a normal replication and translation approach before infecting the target cell. Their life cycle starts with association to the host cell and concludes with the discharge of new offspring from the infected cell, which is the host.

This is the first and most important stage in viral infection, determining the intensity of the infection and pathogenesis. An improved knowledge of viral binding to a receptor on a host cell's surface can aid in the prediction of human transmission. Infection begins with

the attachment of a densely glycosylated S protein to the host cell surface, which is a trimeric class 1 fusion protein with two primary subunits: a receptor binding domain (S_1) and another domain (S_2) that mediates virus fusion with the host cell membrane. When the S_1 domain attaches to a host-cell receptor, such as angiotensin-converting enzyme 2 (ACE_2) for SARS-CoV and SARS-CoV-2, the process of fusion of the virus membrane with the host cell membrane begins [3].

As a result, the cell's proteases break the spike glycoprotein, enabling the virus's S_2 domain to interact with it and infiltrate the cell. The ACE_2 receptor is located on trachea, bronchi, bronchial serous duct, and alveolar epithelial cells, as well as alveolar macrophages and monocytes.

After SARS-CoV invaded these cells, mature virions were released to enter new target cells. ACE_2 is also found on endothelial cells in arteries and veins, immune cells, neurons, kidney tube epithelium cells, mucosal cells in the gut, and kidney tubule epithelial cells, providing a varied collection of targets for SARS-CoV infection [3-4].

According to studies, SARS-CoV-2 utilizes ACE_2 as the primary receptor with greater affinity than SARS-CoV infection, implying that the same group of host-cells is being attacked and infected. The virus now penetrates the host cell after attachment. The virus enters the host cell via two distinct routes, both of which rely on the target cell protease activating receptor-attached spike protein. CoVS enters the recipient cell as an endosome via clathrin-dependent and clathrin-independent endocytosis in the first route. The gold standard technique to determine and distinguish SARS-CoV-2 from various beta-coronaviruses including SARS and MERS is presented using the rRT-PCR method and particular primers and probes [3].

PCR is a highly sensitive laboratory method that has been demonstrated to be beneficial for biological and medical sciences and can provide both quantitative and qualitative findings. rRT-PCR is a diagnostic PCR modification that is used to identify target RNAs in clinical samples, specifically for detecting pathogens in molecular diagnostic labs.

In addition, during the COVID-19 pandemic, the rRT-PCR method was used to identify the SARS-CoV-2 genome in biological materials. Despite its middling sensitivity and high precision, as well as its approval [5]. Omicron, a novel form of SARS-CoV-2, has been discovered globally [6]. Globally, the number of verified Omicron-related cases has grown considerably. The novel variant is rapidly expanding and has passed many boundaries all over the globe. This novel variation has been found to spread much faster than other SARS-CoV-2 variants. The new SARS-CoV-2 variation has unique epidemiological and biological features, rendering it more transmissible than other SARS-CoV-2 varieties. In just two weeks, it impacted 2152 individuals in 57 nations.

The SARSCoV-2 Omicron variant's mortality rate has not yet been documented. It infects more young and middle-aged individuals than earlier variants. To reduce the illness load on humankind, global health institutions should take urgent preventive steps to halt outbreaks of this new and returning pathogenic variant around the world [7].

Vaccines

The SII/Covishield, The Pfizer/BioNTech Comirnaty vaccine, AstraZeneca/AZD1222 vaccines, the Johnson & Johnson Janssen/Ad26. COV 2. S vaccines are all available for COVID-19. Moderna COVID-19 vaccine (mRNA 1273), Sinopharm COVID-19 vaccine, Bharat Biotech BBV152 COVAXIN vaccine, Sinovac-CoronaVac vaccine, Covovax vaccine (NVX-CoV2373). The immunization Nuvaxovid (NVX-CoV2373). Vaccination reduces the chance of contracting COVID-19. It also guarantees that the illness does not spread from person to person. Vaccines/vaccination protects one's health.

VaxELAN

Background

This system/tool identifies possible vaccine candidates. This system combines a reverse vaccinology tool and an immuno-informatics tool to analyze genomic and proteomics data sets from various diseases and to identify vaccine targets. The SARS-CoV-2 outbreak is an example of cellular defense failing.

Methodology

Authors retrieved the protein sequence of the receptor binding domain of spike glycoprotein [P0DTC$_2$] from NCBI. In the Violin database (https://www.violinet.org/vaxgen/gene_detail.php?c_gene_id=789), this protein was reported to have length of 1255 amino acids but the NCBI record provided only 1273 residue long sequences. The Pi was 5.6 and the molecular weight was 130077.89 Daltons, according to NCBI molecular weight is 141176.816 Kda. As per Vaxign prediction, the localization probability was found to be unknown. Authors found Partial protein sequence from National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov). The protein sequence is analyzed in Vax-Elan and Vax-DL respectively. (https://vac.kamalrawal.in/vaxelan/v2) (https://vac.kamalrawal.in/vaxidl/). The tool will evaluate the protein and will provide the results.

The features used by Vax-Elan are reported in Table 1.

S.No	Features	Tools	Cut-off	Reference
1	Adhesion Property	FungalRv	≥ -1.2	Monterru bio-López et al. (2015)
2	Cellular Localization (Target P)	Targetp	≥ 0.8	Goodswen et al. (2014)
3	Secretory/Non-Secretory Protein	Signal d value	≥ 0.5	Liebenberg et al. (2012)
4	Stability	ProtParam	<40	Solanki et al. (2018)
5	TM Helices	TmPred TMHMM HMMtop	≤ 1	Monterrubi o-López et al. (2015) Solanki et al. (2018) Naz et al. (2019)
6	MHC Class-1 Binding	NetMHC	≥ 4.9(for Higher binders) ≥ 5.05(weak binders)	Schroeder and Aebischer (2011)
7	Cleavage Sites	NetChop	≥ 110	Dhanda et al. (2017)
8	CTL Epitope Prediction	NetCTL	<7.5	Solanki et al. (2018)

9	Essential Genes	DEG Database	e-value:10_e -5, identity>30%, query coverage ≥ 70%	Solanki et al. (2018)
10	Molecular Weight	ProtParam	<110 kDa	Naz et al. (2019)
11	Non-Bacterial Pathogen	Blastp With Gutflo raDB	e-value:10_e -5, identity>30%, query coverage ≥ 70%	Naz et al. (2019)
12	Non-Homology	BLAST with human proteome	e-value:10_e -5, identity>30%, query Coverage ≥ 70%	Pearson et al. (2013)
13	Virulence Factor	Blast p with VFDB	e-value:10_e -5, identity> 30%	Solanki et al. (2018)
14	Non-Allergen	NetChop	e-value:10_e -5, identity>30%	Pearson et al. (2013)
15	IEDB MHC-1 Binding Predictions	NetMHC	>50 nM	Schroeder and Aebischer (2011)
16	Cellular Localization (PSORTb Gram+)	Psortb	Cell wall Extracellular	Naz et al. (2019) Solanki et al. (2018) Muruato et al. (2017)
17	Cellular Localization (PSORTb GraM-)	Psortb	Outer membrane, extracellular and periplasmic	Naz et al. (2019)
18	Cellular Localization (Wolf PSORT)	Wolf Psort	Extracellular or plasma membrane	Watanabe et al. (2021)

Table 1: List of tools used by Vax-Elan.

Results

The protein sequence of "receptor binding domain of spike glycoprotein" [ID-PoDTC$_2$] was extracted from the NCBI protein database. This protein is 1273 amino acids long and belongs to SARS Cov-2.

Next, researchers used Vax-Elan server to evaluate important properties to check the suitability of this protein as potential vaccine candidate. First, authors use three bioinformatics tools to examine cellular localization: TargetP [8], PSORTB, along with Wolf PSORT. The TargetP technique detects the existence of N-terminal sorting signals. The pattern is classified as a signal peptide, a mitochondrial transit peptide, a chloroplast transit peptide or a thylakoid lumen transit peptide. PSORTB analyses a protein sequence for amino acid composition, resemblance to known localization proteins, the existence of a signal peptide, transmembrane alpha-helices, and motifs matching to specific localizations.

Based on sorting cues, amino acid composition, and functional patterns such as DNA-binding motifs, Wolf PSORT transforms protein sequences of amino acids into numerical localization characteristics.

As per VaxElan algorithm label 1 for TargetP was defined when the score of TargetP tool crosses the threshold 0.8. For "the receptor binding domain of spike glycoprotein" the TargetP prediction is 0 suggesting that it is not a signal peptide. As per VaxElan, the

protein is predicted to be 0 for TargetP. Receptor binding domain of spike glycoprotein is not predicted to be a signal peptide as it scored 0. Being a signal peptide there is a criterion for potential vaccine candidates. Results obtained for TargetP are SP [signal peptide] 0.997, mTP [mitochondrial transit peptide] 0.0003, cTP [chloroplast transit peptide].

Adhesion prediction

Authors used FungalRV to forecast adherence [9]. FungalRV is an adhesin prediction program that analyzes expected adhesins and adhesin-like proteins in eight human pathogenic fungus species.

As per VaxElan, the protein is predicted to be adhesin protein if it scores ≥ -1.2 but our protein receptor binding domain of spike glycoprotein scores -0.70. Therefore, it is not an adhesin protein.

Stability and molecular weight

Next, researchers computed the molecular weight and instability index of the protein using ProtParam. ProtParam is used to compute the physico-chemical properties of protein such as instability index, molecular weight, extinction coefficient, isoelectric point, composition of amino acids, half-life etc. As per VaxElan results, the receptor binding domain of spike glycoprotein partial is stable as it has an instability index <40 and a molecular weight of 141176.816 kda.

Signal peptide prediction

Authors performed signal peptide prediction using SignalP. SignalP forecasts the existence and location of signal peptides in the sequences of amino acids from archaea, gram+ve, and gram -ve bacteria, as well as eukaryotes. As per VaxElan, the receptor binding domain of spike glycoprotein has a signal peptide as its Dvalue is 0.837 which is greater than the threshold Dvalue (0.450).

Transmembrane helices

The receptor binding region of spike glycoprotein is then tested over transmembrane proteins using TMpred [9].

The TMpred program forecasts membrane-spanning areas and their direction. The method is built on a statistical study of the database TMbase, which contains naturally found transmembrane proteins.

According to VaxElan findings, the receptor binding region of spike glycoprotein protein contains transmembrane proteins. Inside to outside helices: 12 discovered; outside to inside helices: 11 discovered.

Essential genes

Next, authors analyzed whether" the receptor binding domain of spike glycoprotein" is an essential gene or not using the DEG database.

The DEG database contains information about all the essential genes with which our query gene can be compared using BLAST to check whether the query gene is essential or not based on the homology with database genes. As per VaxElan, the target protein is not an essential gene.

Non-homology

Next, VaxElan used BLAST to find homologous sequences for the "receptor binding domain of spike glycoprotein". For our protein, "receptor binding domain of spike glycoprotein" BLAST did not find any homologous sequence. Therefore, our protein is a vaccine candidate.

Virulence factor

Next, authors analyzed the receptor binding domain of spike glycoprotein for virulence factor which is an important parameter while identifying a vaccine candidate. Virulence factor is determined by using BLAST which finds a similar virulent protein. For the receptor binding domain of spike glycoprotein, no virulent protein was found. Thus, it is a vaccine candidate.

Non-allergen

Using BLAST, authors determined whether the receptor binding region of spike glycoprotein was allergen-free. According to VaxElan findings, the receptor binding region of spike glycoprotein is not allergenic because BLAST found no subject sequence. As a result, it is a vaccine option.

Non-bacterial pathogen

This parameter runs BLAST against a bacterial pathogen database to find whether the query protein is bacterial pathogen or not. As Vax-Elan predicted 1 hence, my protein is not a bacterial pathogen. Thus, a vaccine candidate.

Results obtained for receptor binding domain of spike glycoprotein protein from Vax-Elan is given below in Table 2.

Sequence_id	So_QIG55955.1
Adhesion Property	0
Cellular localization (TargetP)	0
Secretory/Non-Secretory Protein	1
Stability	1
TM Helices	1
MHC Class-I Binding	1
Cleavage Sites	1
CTL Epitope Prediction	1
Essential Genes	0
Molecular Weight	0
Non-bacterial Pathogen	1
Non-Homology	1
Non-Allergen	1
Virulence Factor	0
IEDB MHC-I Binding Predictions	1
Cellular localization (PSORTb Gram+)	0
Cellular localization (PSORTb Gram-)	0.5
Cellular localization (Wolf PSORT)	1
Si Value	11.5
Pi Value	0.68

Table 2: Results obtained for receptor binding domain of spike glycoprotein protein from Vax-Elan.

Vaxi-DL section

Vaxi-DL was assessed on the receptor binding region of the spike glycoprotein gene and found to be a poor vaccine candidate (labeled as '1'). This indicates that the receptor binding region of spike glycoprotein possesses the necessary antigenic characteristics.

Spike Glycoprotein protein has a 61.06 percent chance of becoming a possible vaccine candidate.

Number of Sequences Classified as Vaccine Candidate	Number of Sequences Classified as Non-Vaccine Candidate
0	1

Sequence ID	Predicted Class	Probability (%)
So_sp|PoDTC2.1|SPIKE_SA RS2	Not a vaccine candidate	61.06

Table 3 a,b: Results obtained for Spike Glycoprotein from Vax-DL.

Discussion

The worldwide pandemic of the coronavirus disease 2019 unfolds over the world, with several varieties emerging, some of which are of concern.

COVID-19 is a fast and communicable disease which is transferred via touch, inhaling droplets released from infected persons, surface, etc.

Omicron (B.1.1.529), a recent variant of coronavirus with a high number of variations in the receptor-binding domain, or RBD, of the spike protein, has gotten a lot of scientific and popular interest. The experiment described here employs a thorough approach to find disease vaccination targets using bioinformatics and computer algorithms. Various approaches were used to narrow down possible vaccine candidates.

To find a potential vaccine candidate for COVID-19, computational analysis using VaxElan and VaxDL pipelines were performed on Spike Glycoprotein. As a result, VaxElan and VaxDL bioinformatics pipelines suggested that our protein could not be a vaccine candidate.

Conclusion

COVID-19 is caused by the SARS COV-2 virus. According to the Vax-dl the receptor binding domain of spike glycoprotein [ID-PoDTC2] is not a potential vaccine candidate.

The protein has molecular weight of 141176.816 Kda according to NCBI, it consists of 1273 amino acids with peptide length of 1265.

Its Adhesion Prediction came to be -0.7024651 which signifies it as not a vaccine candidate. TargetP location is M, Signalp D-Value is 0.837 which classified it as a vaccine candidate. Instability Index is 33.01 which is less than 40 hence it is a vaccine candidate as predicted by ProtParam. This protein contains transmembrane helices as predicted by Vax-Elan so it is not a vaccine candidate. The protein has 18 weak binders, 12 high binders and 391 cleavage sites. This protein

has no essential genes and PSORTb Gram+ location is Cytoplasmic, PSORTb Gram+ location is outer membrane and the Wolf PSORT location is plasma membrane.

Spike Glycoprotein was found to be not a promising vaccine candidate using the bioinformatics approach and the tools Vax-Elan and Vaxi-DL.

References

1. Kirtipal N, Bharadwaj S, Kang SG. From SARS to SARS-Cov-2, Insights on Structure, Pathogenicity and Immunity Aspects of Pandemic Human Coronaviruses. Infect Genet Evol. 2020;85:104502. PubMed | CrossRef

2. Zhang S, Zhou P, Wang P, Li Y, Jiang L, Jia W, et al. Structural Definition of a Unique Neutralization Epitope on The Receptor-Binding Domain of MERS-Cov Spike Glycoprotein. Cell Rep. 2018;24(2):441-52. PubMed | CrossRef

3. Lehrer S, Rheinstein PH. Ivermectin Docks to the SARS-Cov-2 Spike Receptor-Binding Domain Attached to ACE2. In Vivo. 2020;34(5):3023-6. PubMed | CrossRef

4. Han P, Li L, Liu S, Wang Q, Zhang D, Xu Z, et al. Receptor Binding and Complex Structures of Human ACE2 to Spike RBD from Omicron and Delta SARS-Cov-2. Cell. 2022;185(4):630-40. PubMed | CrossRef

5. Rahbari R, Moradi N, Abdi M. rRt-PCR for SARS-Cov-2: Analytical Considerations. Clin Chim Acta. 2021;516:1-7. PubMed | CrossRef

6. Karim SS, Karim QA. Omicron SARS-Cov-2 Variant: A New Chapter in the COVID-19 Pandemic. Lancet. 2021;398(10317):2126-8. PubMed | CrossRef

7. Ferré VM, Peiffer-Smadja N, Visseaux B, Descamps D, Ghosn J, Charpentier C. Omicron SARS-Cov-2 Variant: What We Know and What We Don't. Anaesth Crit Care Pain Med. 2022;41(1):100998. PubMed | CrossRef

8. Goodswen SJ, Kennedy PJ, Ellis JT. Vacceed: A High-Throughput in Silico Vaccine Candidate Discovery Pipeline for Eukaryotic Pathogens Based on Reverse Vaccinology. Bioinform. 2014;15;30(16):2381-3. PubMed | CrossRef

9. Monterrubio-López GP, Ribas-Aparicio RM. Identification of Novel Potential Vaccine Candidates Against Tuberculosis Based on Reverse Vaccinology. Biomed Res Int. 2015;2015:483150. PubMed | CrossRef