ACTA SCIENTIFIC MICROBIOLOGY (ISSN: 2581-3226)

Volume 4 Issue 1 January 2021

Research Article

On SARS-CoV-2, Tropical Medicine and Bioinformatics: Analysis of the SARS-Cov-2 Molecular Features and Epitope Prediction for Antibody or Vaccine Development

Joshua Angelo Hermida Mandanas*

Tropical Medicine Scientist and Immunobiologist, Philippine Medical Technology Professional/National Lecturer, University of the Philippines Manila, Philippines

***Corresponding Author:** Joshua Angelo Hermida Mandanas, Tropical Medicine Scientist and Immunobiologist, Philippine Medical Technology Professional/National Lecturer, University of the Philippines Manila, Philippines.

Received: November 24, 2020

Published: December 08, 2020

Abstract

Introduction: SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) is the cause of COVID-19 which is the pandemic as of the current time. It is a global public health emergency and is still on the loose of spreading more infections and deaths. Bioinformatics offer extensive visualization and analysis of combined molecular, cellular, biochemical and immunobiologic aspects of SARS-CoV-2 which is indeed vital for antibody and vaccine development.

Objectives: This paper presents the SARS-CoV-2 molecular virologic, biochemical, cellular and immunobiologic features and gives a basic B cell linear epitope prediction of the SARS-CoV-2 spike (S) glycoprotein using bioinformatics which both can serve as a guide for development of vaccines or antibody-based treatments.

Methods: Several bioinformatic methods were used, from sequence analysis (Uniprot), structural correlation (PDB), molecular modelling (UCSF Chimera) and illustration (Biorender), B cell linear epitope prediction tools, conservancy analysis and search of related epitopes (IEDB) together with determination of disordered regions (GlobPlot and DisEMBL).

Results: Host enzymes such as furin, cathepsin L and TMPRSS2 enable pre-processing of SARS-CoV-2 prior to infecting cells such as monocytes, macrophages and alveolar cells. Sialylated cells can potentially be infected with SARS-CoV-2. A 29 and an 11-residues long B cell linear epitope candidates were predicted from the SARS-CoV-2 spike (S) glycoprotein to be the potential target of antibodies or vaccines.

Conclusion: SARS-CoV-2 binds to many host receptors and is enzymatically processed leading to infection of many human cells. Prediction of B cell linear epitopes using bioinformatics can be vital for antibody or vaccine development. Integration of Tropical Medicine and Bioinformatics can offer new horizons and can turn the battle against SARS-CoV-2.

Keywords: SARS-Cov-2; COVID-19; Tropical Medicine; Bioinformatics; Epitope; Vaccine

Introduction

Background

SARS-CoV-2 infected patients were initially reported in Wuhan, Hubei province, People's Republic of China last year (2019) and an increase in outbreaks have been observed which escalated at a worldwide scale [1,2]. The World Health Organization (WHO) officially named the disease COVID-19 and the International Committee on Taxonomy of Viruses named the virus SARS-CoV-2 [3].

Chan., *et al.* [4] noted that coronaviruses are scattered globally and are under the family Orthocoronaviridae and order Nidovirales. The coronaviruses are composed of four types- alpha (α), beta (β), gamma (γ) and delta (δ). Previous coronaviruses which

caused outbreaks include SARS-CoV and MERS-CoV, altogether with SARS-CoV-2 they belong to beta coronaviruses. SARS-CoV were acquired from bat to civet and civet to human transmission with a case fatality rate of 9.5% [5]. MERS-CoV (Middle East respiratory syndrome coronavirus) had a case fatality rate of 30% and were acquired from human to human transmission from contact with camels [6].

Epidemiology

Bats are considered to be the natural hosts of SARS-CoV-2 while pangolins and snakes are the potential intermediate hosts [7]. Homologous recombination possibly explains the transmission of animal coronaviruses to humans [8]. Other domesticated animals such as dogs and cats are more affected by alpha coronaviruses [9]. The most common route of transmission include respiratory droplets [5-10um], close contact [3,10,11] and possibly aerosol droplets (<5 um) [12] especially in closed, confined spaces [13]. Vertical transmission of the SARS-Cov-2 can be possible in the last weeks of pregnancy [14].

The acute viral infection of COVID-19 generally presents with fever, cough, fatigue, diarrhea and vomiting (Wang., *et al*. 2020) with the acute respiratory distress occurs about 9 days after. Wang., *et al*. (2020) also noted that in general susceptible individuals such as elderly adults (especially males) with underlying diseases such as heart disease, diabetes, etc. can be more affected. The RO (R naught) is the measure of the average number of secondary infections that patients may cause in a completely susceptible population without intervention (Wang., *et al*. 2020) [3], the Philippines had decreased the RO from 1.19 in July 2020 to less than 1 this September 2020 [15-18]- possibly through joint intervention, increased capacity training and mobilization. It is imperative that sustained, consistent approaches and interventions together with preventive measures must be applied.

In the Philippines, about 56% of affected cases were males aged 20-29 years old and among the total confirmed deaths- about 61% are males aged 60-69 years old [15-18]. The average incubation range [3,10,11] of COVID-19 is about 2-14 days (and can extend up to 24 days in some cases) with a median incubation of 3 days, with symptoms can first appear as early or more than 3 days

Bioinformatic methods (analysis softwares)

Bioinformatics is defined as the emerging field that utilizes application of analytical or computational tools for the interpretation of a biologic data [19]. It also encompasses different fields of biology, physics, mathematics, computer science, molecular biology, biochemistry and genetics. With the use of different bioinformatic tools, one can identify genetic profiles, polymorphisms, structure determination as well as ligand binding analysis which can aid scientific exploration for development of drugs or vaccine and in terms of the SARS-CoV-2 to understand the basic molecular features and to generate the epitope target candidates for antibody and vaccine development. This paper focuses in the analysis of the molecular virologic, biochemical, cellular and immunobiologic features of SARS-CoV-2 together with B cell linear epitope prediction of the SARS-CoV-2 spike (S) glycoprotein and determination of disordered regions.

This research paper utilized the SARS-CoV-2 spike (S) glycoprotein sequence obtained in Uniprot and associated structures from Protein Data Bank (PDB) and includes visualization of an illustrative molecular modelling (Figure 1A) of the actual spike and other potential mechanisms (Figure 1B) using Biorender and UCSF Chimera. The research paper also performed B cell linear epitope predictions of the SARS-CoV-2 spike (S) glycoprotein utilizing the multiple softwares from Immune Epitope Database (IEDB) coupled with the determination of the disordered regions using DisEMBL and GlobPlot2.3 (Figure 1C) and finally, locating related epitopes using IEDB as well as epitope conservancy analysis (Figure 1D). Figure 1 below delineates the research flow. The findings in this research can serve as guide for scientists and learners with proclivity in basic bioinformatics and also to serve as initial finding or a guide towards antibody or vaccine development and exploration of other molecular features.

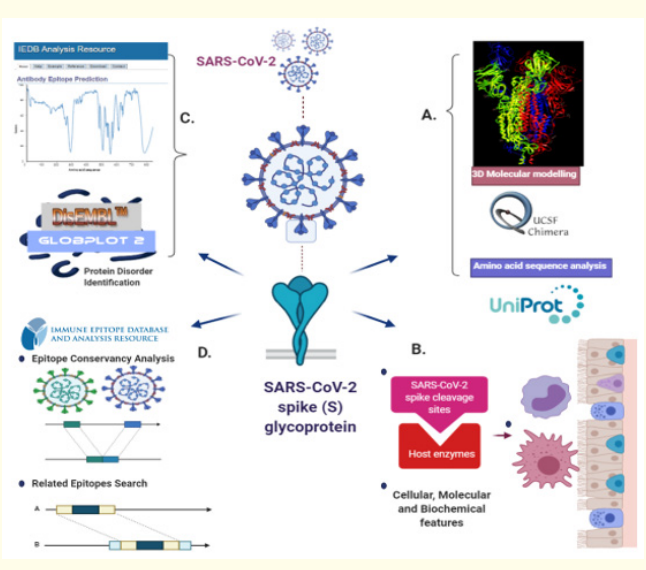


Figure 1: General scheme of the study.

Results and Discussion

In this section, we will be identifying the basic molecular virology of SARS-CoV-2 using bioinformatics tool or analysis softwares. A brief description of the basic features of each bioinformatic methods or softwares will be discussed along with correlation with SARS-CoV-2 virology.

Structure of SARS-CoV-2

SARS-CoV-2 is a type of single-stranded RNA virus (ssRNA) with about 30 kilobases in its genome [20]. It has a positive polarity (positive sense strand) meaning that the base sequence is in 5'-3' orientation [21]. It has a size range of about 70-90 mm [22]. Budding of SARS-CoV-2 from host cells develops the host-derived envelope (E) from which the SARS-CoV-2 spike (S) and membrane glycoproteins (M) are embedded. A helical nucleocapsid (N) then surrounds the viral RNA [20]. Figure 2 below shows the constructed image from Biorender which illustrates the basic anatomy of the SARS-CoV-2. An actual 3D model is being incorporated in Figure 2 courtesy of Eckert A. and Higgins D. of Center for Disease Control and Prevention (CDC). Figure 2 also shows some major types of cells infected by the SARS-CoV-2 as per binding on specific receptors that will be discussed in the succeeding contexts.

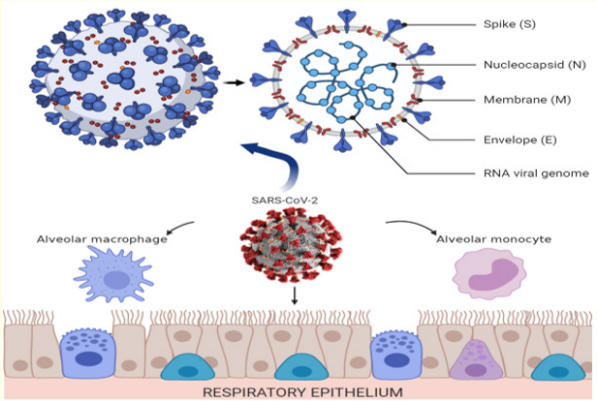


Figure 2: The SARS-CoV-2 and the major cells affected.

The SARS-CoV-2 spike (S) glycoprotein is important for binding (viral attachment) to host receptors and is deemed crucial for infection [23] determined that the spike (S) glycoprotein is a homotrimer (Figure 3B) composed of three (3) similar chains (3) also noted that the "spike" (S) glycoprotein of the SARS-Cov-2 has the characteristic receptor-binding domain (RBD) subunit or (S1), a fusion domain or (S2) which are thus separated [24] by a cleavage site (S1/S2) and another cleavage site [25] within (S2) that is called (S2'). Figure 3 below illustrates the 3D model of the SARS-CoV-2 spike (S) glycoprotein. The details of the sequence will be tackled in the SARS-CoV-2 spike (S) glycoprotein sequence analysis.

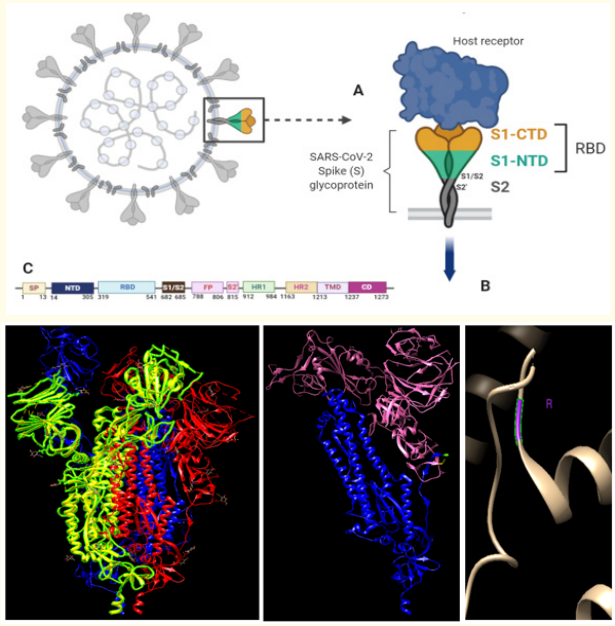
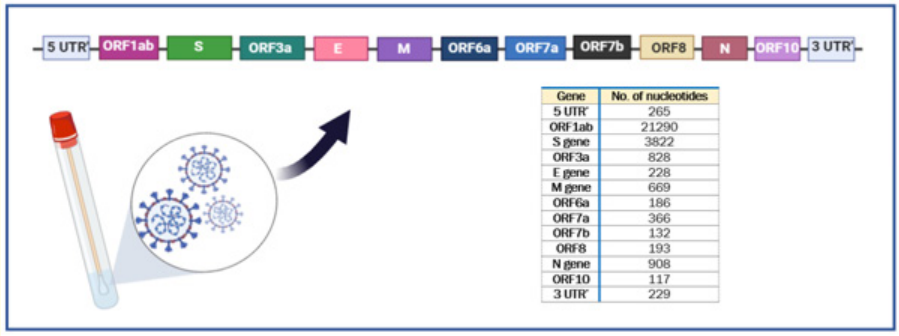


Figure 3: 3D Model of the SARS-CoV-2 spike (S) glycoprotein.

Figure 3A shows the schematic binding of the SARS-CoV-2 spike (S) glycoprotein on a general host receptor- there are different types of receptors in humans by which the SARS-CoV-2 spike (S) glycoprotein can bind that will be discussed in the succeeding contexts. The image on Figure 3A were created from Biorender. Molecular modelling for Figure 3B (B.1, B.2 and B.3) were generated from UCSF Chimera; Figure 3B.1 shows the 3D image of the SARS-CoV-2 spike (S) glycoprotein in open state from Protein Data Bank (PDB 6VYB) [108] to indicate exposure to binding on host receptors- highlighting the homotrimeric chain composed of chains A (shaded in red), B (shaded in blue, in open state) and chain C (shaded yellow); Figure 3B.2 highlights the important parts- S1 (pink), S1/S2(yellow), and S2(blue) using chain A of PDB 6ZP0 [109] (closed state, since some models in open state have residue S2' cleaved by enzymes or in processing), residue 685 (serine is highlighted) for S1/S2 since this is the available model as of the current time; Figure 3B.3 uses PDB 6VYB to indicate S2' (shaded in violet). Figure 3C shows the general representation of the SARS-CoV-2 spike (S) glycoprotein amino acid sequence organization which will be discussed in detail as we go along. The arrangement of the sequence were modified from [1,25-27] and modelled using Biorender.

Wu., *et al*. [28] noted that the SARS-Cov-2 genome has 14 ORFs (open reading frames) which thus encode 27 proteins. Kumar., *et al*. [20] noted that the first ORFs (ORF1ab) of SARS-CoV-2 encodes for about 67% of non-structural proteins such as cysteine proteases, papain-like protease (nsp3), chymotrypsin-like, 3C-like/ main protease (nsp5), RNA-dependent RNA polymerase (nsp12), helicase (nsp13) with other transcription and replication factors. About one-thirds of the genome mainly encodes the structural proteins such as the nucleocapsid (N), spike (S) envelope (E), and membrane (M) glycoproteins. Overall, the SARS-CoV-2 has 29,903 nucleotides [29]. Just like other coronaviruses, it has 5' and 3' UTRs (untranslated regions). Figure 4 below shows the genomic organization of the SARS-CoV-2. Biorender was used to model the genomic organization, details were modified from [20,28,29].



Gene	No. of nucleotides
5 UTR	265
ORF1ab	21290
S gene	3822
ORF3a	828
E gene	228
M gene	669
ORF6a	186
ORF7a	366
ORF7b	132
ORF8	193
N gene	908
ORF10	117
3 UTR	229

Figure 4: Genomic organization of the SARS-CoV-2.

Xu., *et al*. [30] noted that genome of SARS-CoV-2 is more than 85% similar to the genome of the bats SARS-like virus ZC45, they also determined that there are about regions of difference (RD) in the genome between SARS-CoV and SARS-CoV-2- RD1, RD2, and RD3 (448 nucleotides, 55 nucleotides, and 278 nucleotides, respectively) which are partial coding sequences of the ORF 1ab gene; RD4 and RD5 (315 nucleotides and 80 nucleotides, respectively) which are partial coding sequences of the S gene and RD6 which is 214 nucleotides and part of the coding sequence of the ORF7b and ORF8 genes. Lu., *et al*. [31] also determined that at the genomic level, the SARS-CoV-2 shares an 87.99% sequence identity with the bat-SL-CoVZC45 and 87.23% sequence identity with the bat-SL-CoVZXC2. SARS-CoV-2 is less genetically similar to the SARS-CoV (about 79%) and MERS-CoV (about 50%) [31,32]. Kindly see Figure 4 for the SARS-CoV-2 genomic reference.

Biorender is a bioinformatic tool that allows users to create high-quality figures, posters, presentations, structures, pathways and biologic modelling. It was developed by Shiz Aoki from the John Hopkins University [33]. Biorender was utilized in this paper to visualize molecular modelling and to generate the integrative illustration of the molecular features of the SARS-CoV-2.

UCSF Chimera is also a bioinformatic tool which is widely used for intensive visualization of 3D molecular structures. It also allows users to utilize its vast applications which includes identification and analysis of density maps, sequence alignments, molecular docking or binding as well as conformational ensembles [34].

It was developed by the Resource for Biocomputing, Visualization, and Informatics (RBVI) at the University of California, San Francisco. Molecular images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco. UCSF Chimera was utilized in this paper to visualize 3D molecular structures as well as visualization of molecular binding.

SARS-CoV-2 host processing and binding

Furin specifically cleaves at S1/S2 site of the SARS-Cov-2 spike (S) glycoprotein [35], the cleavage site is determined at arginine-arginine-alanine-arginine (polybasic) site, while both TMPRSS2 and cathepsin (L) can cleave either S1/S2 or at S2' site. Note that furins released extracellularly by host can preactvate the virus through initial cleavage [36] and SARS-CoV-2 has more cleavage efficiency by furins compared to SARS-CoV [37].

Experimental data by Jaimes., *et al*. [25] using a peptide containing amino acids (threonine-asparagine-serine-proline-arginine-arginine-alanine-arginine-serine-valine-alanine) is the one cleaved by cathepsin L at S1/S2 site, although at low pH (<6) caspases generally can further cleave at S2' [38]. TMPRSS2 has preference for arginine or lysine at residues 682, 683 and 685 in the S1/S2 and at residue 815 of the S2' [39].

SARS-Cov-2 can bind to TMPRSS2 (transmembrane protease, serine 2) which are highly expressed in nasal and goblet ciliated cells as well as in alveolar type II epithelial cells, cornea, esophagus, ileum, colon, liver, gallbladder and common bile duct [41]. TMPRSS2 is also an extracellular surface receptor enzyme which can activate the SARS-Cov-2 by proteolytic cleavage and promote viral fusion, this receptor is not usually expressed in monocytes but most commonly found in alveolar macrophages [42] which are noted to produce furins and cathepsin L. Figure 5 below schematically illustrates the enzymatic cleavage of furins, cathepsins and TMPRSS2 on the spike (S) glycoprotein of SARS-CoV-2.

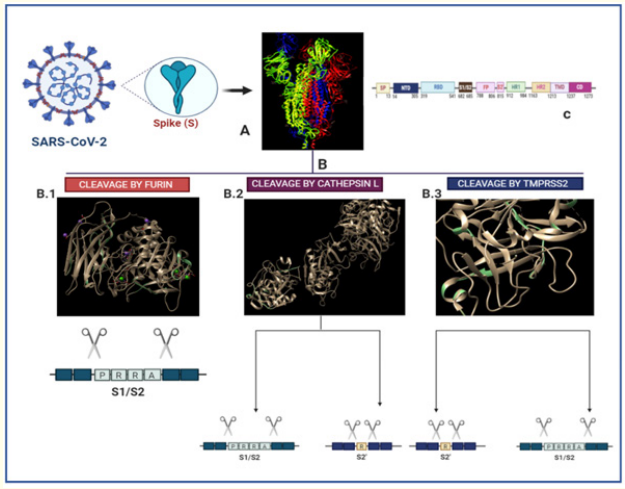


Figure 5: SARS-CoV-2 spike (2) glycoprotein processing by furin, cathepsin L and TMPRSS2.

Figure 5 above generally shows the enzymatic processing of the SARS-CoV-2 spike (2) glycoprotein upon infection in the body. Figure 5A illustrates the spike (S) glycoprotein in both schematic representation (created using Biorender) and in 3D form (generated from UCSF Chimera) with PDB 6VYB [108], the homotrimeric chains of SARS-CoV-2 spike (S) glycoprotein were composed of chains A (shaded in red), B (shaded in blue, open) and C (shaded yellow). Figure 5B (B.1, B.2, and B.3) shows the 3D molecular models of the enzymes, furin (PDB 5MIM) [110], cathepsin L (PDB 3HHA) [111] and hepsin (PDB 5CE1) [112] which were also generated using UCSF Chimera. Rensi., *et al*. [43] noted that TMPRSS2 crystal structure isn't available at this time, therefore another similar protease hepsin was used. The mechanisms of cleavage were discussed in the previous texts- with furin cleaving at the polybasic site or S1/S2 junction [35] while both cathepsin L and TMPRSS2 can cleave at S1/S2 or at the S'2 residue. The highlighted (green) portions in furin, cathepsin L and TMPRSS2 indicate the potential active sites and docking regions as determined by UCSF chimera. The molecule chosen for binding in all enzymes is PDB 6ZP0 [109], which has been processed in UCSF Chimera to indicate the corresponding amino acids for cleavage. Figure 5C (also shown in Figure 3C) shows the SARS-CoV-2 spike (S) glycoprotein sequence organization for reference in Figure 5B.

SARS-CoV-2 can also potentially bind to sialic acid expressing tissues and cells [44,45]. Figure 6 below generally shows the schematic diagram. Some receptors and proteins in the human body are also sialylated or with sialic-acid binding sites such as vWF, CD33, CD34, CD43 and CD169 [46,49] and cells expressing these proteins or receptors can potentially be affected, with this- future exploration of the ligand binding and cellular effects of SARS-CoV-2 is imperative. Monocytes and macrophages which are key cell players during COVID-19 both express some of the receptors like CD43 [50,51], the latter also are noted to express CD169 [52].

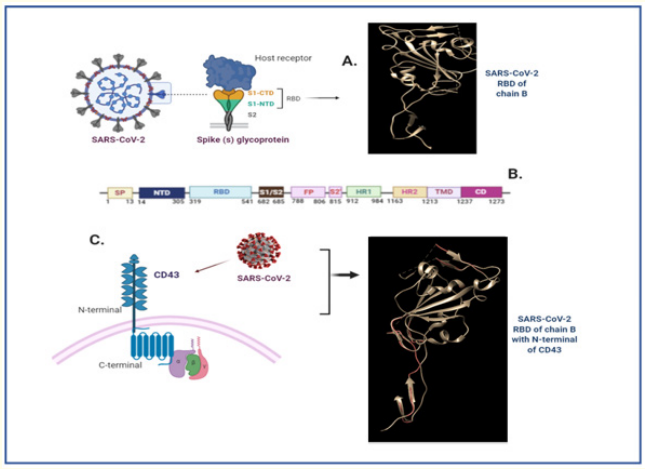


Figure 6: Molecular binding of a sialylated protein (CD43) to spike (S) glycoprotein of SARS-CoV-2.

Figure 6 above shows the molecular binding of a sialylated protein CD43 to spike (S) glycoprotein of SARS-CoV-2. CD43 is a sialylated transmembrane protein [53,54] with G-protein signalling complexes. UCSF Chimera was used to generate the 3D molecular structures. Figure 6A depicts the schematic illustration of the SARS-CoV-2 spike (S) glycoprotein receptor binding domain of S1 subunit to its 3D image, figures were created using Biorender; Figure 6B shows the SARS-CoV-2 spike (S) glycoprotein sequence organization for reference (also shown in Figures 3C and 5C); Figure 6C shows the schematic binding of SARS-Cov-2 in a CD43 expressing cell and using UCSF Chimera simulation, the highlighted areas in pink depicts the possible binding sites of CD43 to the receptor binding domains of the SARS-Cov-2 S1 subunit. Host receptors can possibly utilize the whole RBD residues or to bind to specific residues only. Note that the RBD is used by the SARS-CoV-2 in binding to host receptors [55], which is composed of residues 319 to 541 which forms the 3D image seen in Figure 6A and 6C. In the model, the open chain B (open state) of PDB 6VYB relating to SARS-CoV-2 spike (S) glycoprotein was used and bound to the available sequence of CD43 obtained from PDB 2EMS [113]. As of the current time, there is no available amino terminal of CD43 available so the N-terminal sequence used in this paper is derived from the PDB 2EMS using UCSF Chimera.

Note that SARS-Cov-2 have higher binding affinity by 2-fold to its receptor (ACE2) as compared to the binding affinity of SARS-CoV [56] and this high affinity binding can be mediated by higher electrostatic interactions due to SARS-CoV-2 amino acid residues. The higher binding can be evenly promoted by proteases such as TMPRSS2, furins and cathepsin L- note that these enzymes do enhance binding by specific cleavage of the viral "spike" proteins regions which promotes fusion and activation. Bioinformatic analysis can also show that TMPRSS2 can also cleave ACE2 at amino acids 697-716 which can support viral attachment and infectivity [40]. However, Sungnak., *et al*. [41] noted that SARS-Cov-2 have been shown to infect even those cells without expression of TMPRSS2, making an assertion that the higher affinity of the virus even without being cleaved can further support its entry leading to fusion.

SARS-Cov-2 principally gain entry through the angiotensin-converting enzyme 2 (ACE2) receptor present primarily in the surface of monocytes, alveolar macrophages and type II alveolar cells. ACE2 serves as a regulatory receptor as it promotes cleavage of angiotensin and promotes formation of a vasodilator peptide. Sungnak., *et al*. [41] noted that ACE2 can also be expressed in adipose tissue, kidneys, heart, small intestine, testis and thyroid. It can also be found in adrenal gland, fetal liver and thymus, liver, gallbladder, colon, placenta/decidua and most especially in the lungs, nasal and bronchial epithelium. The exploration of the cellular effects of SARS-CoV-2 to other ACE2 expressing cells is also imperative. Figure 7 below generally shows the molecular binding of the SARS-CoV-2 spike (S) glycoprotein to the ACE2 receptor. Figure 7A shows schematic representation of the SARS-CoV-2 spike (S) glycoprotein bound to the ACE2 receptor created in Biorender with the 3D structure generated using UCSF Chimera of the retrieved PDB 7A97 [114] representing the ACE2 (shaded in pink) and homotrimeric chain of the spike (S) glycoprotein composed of chains A (shaded in red, open), B (shaded in blue, open) and chain C (shaded yellow, unbound/closed); Figure 7B shows the RBDs of chains A and B shaded in white and the bound ACE2 (shaded in pink), chain C is omitted.

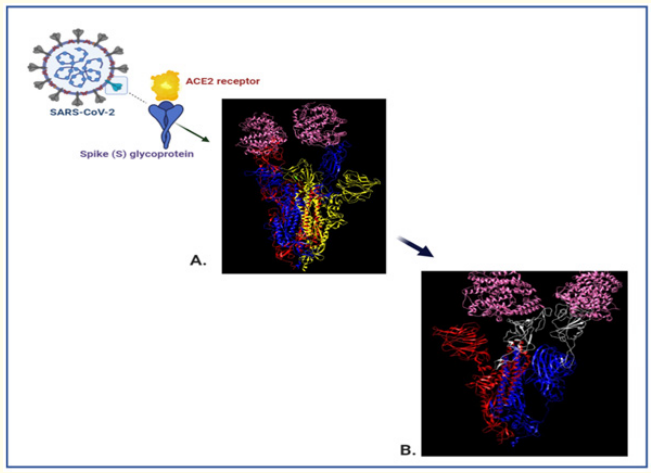


Figure 7: Molecular binding of SARS-CoV-2 spike (S) glycoprotein to the ACE2 receptor.

In addition, Koyama., *et al*. [57] also showed that SARS-Cov-2 can have 1.12×10^{-3} mutations per site-year with several clades (groups observed based on evolutionary aspect) identified- such as the basal clade of L84S in China, European clades of D614G/Q57H and the more common clade, D614G. The clades mentioned were amino acid substitutions which can provide viral fitness for example the D614G mutation (aspartic acid/D replaces glycine/G) in the spike protein have the characteristic mutation in relation for proteins in SARS-CoV-2 replication [57], which can enhance replication making it more infectious and transmissible.

Most RNA viruses have RNA dependent polymerase with no existing proof-reading mechanisms, however this is different in terms of coronavirus- an exonuclease with a proofreading function (ExoN) has been found [28], [58] in both SARS-CoV and SARS-CoV-2 (similar/homologue) which increases the nucleotide processivity and transcription fidelity of the RNA dependent RNA polymerase machinery. In a study by Pachetti., *et al*. [59], mutations in position 14408 of the RNA polymerase can alter binding of cofactors such as ExoN-affecting its proofreading function and might have an impact in the mutation rate.

SARS-CoV-2 spike (S) glycoprotein amino acid sequence analysis:

Huang., *et al*. [60] noted that the SARS-CoV-2 spike (S) glycoprotein is 1273 amino acids long (See Figure 3C) that has a signal peptide (SP) of about 1-13 amino acids from the N-terminus, followed by the S1 subunit (residues 14-685) and the S2 subunit (residues 686 to 1273). The S1 subunit has an N-terminal domain residues 14-305 which is then followed by the RBD residues 319-541. The S2 subunit is then composed of fusion peptide (residues 788-806), heptapeptide repeat sequence 1/HR1 (residues 912-984), HR2 (residues 1163-1213), transmembrane domain (residues 1213-1237) followed by the cytoplasmic domain (residues 1237-1273). The S1/S2 junction is made up of polybasic site made up of residues PRRA at the residues 682-685 [25,26]. S2' cleavage region is found at residue 815 [27]. For reference, kindly see Figure 3C.

Figure 8 below shows the amino acid sequence of the SARS-CoV-2 spike (S) glycoprotein obtained from Uniprot. It has the ID: UniProtKB - P0DTC2 (SPIKE_SARS2).

Figure 8: Sequence of the SARS-CoV-2 spike (S) glycoprotein. (Figure generated from Uniprot in FASTA format)

Uniprot is a database of proteins [61] and sequences as well as their biologic or functional information. It is divided into four (4) databases [66]- UniPROTKB which is subdivided into Swiss-Prot and TrEMBL, UniParc and UniRef.

Swiss-Prot is a manually annotated protein sequence database that combines information extracted from scientific literature and biocurator-evaluated computational analysis. It specifically aims to provide all known relevant information about a particular protein [61]. TrEMBL provides high-quality computationally analyzed records that are enriched with automatic annotation [63]. UniParc contains only protein sequences, with no annotation, it allows further information about the protein to be retrieved from the source databases [64]. UniRef are composed of three (3) databases from UniProtKB and UniParc records which then combines identical sequences and fragments into a single UniRef entry [65]. UniProt is funded by the National Human Genome Research Institute, the National Institutes of Health (NIH), the European Commission, the Swiss Federal Government through the Federal Office of Education and Science, National Cancer Institute and the US Department of Defense [66].

The UniProtKB/Swiss-Prot entry name, ID or accession is composed of up to 11 uppercase alphanumeric characters with a naming convention that can be symbolized as X_Y, where "X" is a mnemonic protein identification code of at most 5 alphanumeric character; "_" is the separator and "Y" is a mnemonic species identification code of at most 5 alphanumeric characters [61].

PDB (Protein Data Bank) is a database by which the 3D structure/molecular data of biomolecules such as proteins and nucleic acids can be found and analyzed, it is also overseen by the Worldwide Protein Data Bank [67]. Imaging methods range from X-ray diffraction, NMR, electron microscopy or a combination of these methods. The protein then can be viewed by open source computer programs, in this study UCSF Chimera was used. In addition, the PDB entry 6VYB [108] which can be accessed in the protein data bank database can offer us the protein summary (Figure 9), a 3D view, annotations and sequence identifier. RCSB stands for Research Collaboratory for Structural Bioinformatics [68]. PDB numbers are automatically assigned and do not have meaning [68]. Figure 9 below shows the sequence identifier using the chain A of PDB 6VYB [108], through RCSB- a user can identify the Uniprot alignment, outliers, sequences missing, covalent interactions of amino

acids such as disulphide bridges, types of chains, glycosylation, and other domain analysis.

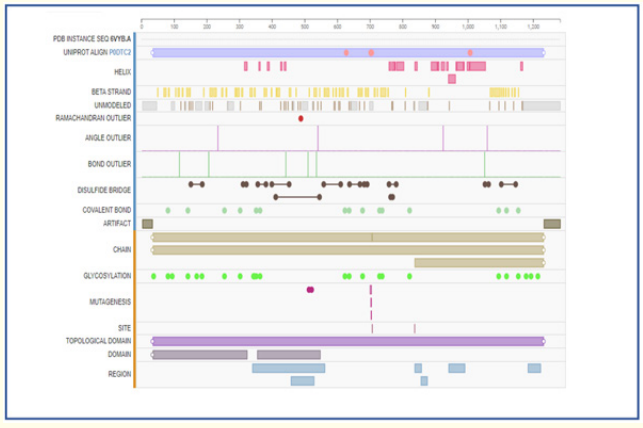


Figure 9: Amino acid sequence of the SARS-CoV-2 spike (S) glycoprotein.

SmartBLAST is a bioinformatic tool that processes the protein accession inquiry (by which users can also upload a FASTA sequence) and the software provides a summary of about five best protein matches from well-studied reference species in the database [69,70]. SARS-CoV-2 spike (S) glycoprotein has a total mass of 141,178 daltons as provided from Uniprot P0DTC2. By using the SmartBLAST, Figure 10 below shows the output. The sequence obtained from Uniprot P0DTC2 (shaded in yellow) was used to generate the sequence homologies from SmartBLAST and as of the current time the five (5) matches were also surface glycoprotein obtained from isolates in the Wuhan seafood market.

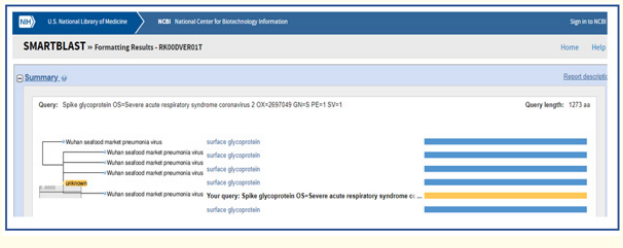


Figure 10: SmartBLAST output.

To determine the degree of conservancy of between SARS-CoV-2 and SARS-CoV receptor binding domains, the IEDB (Immune Epitope Database) Epitope Conservancy Analysis was used. It is a bioinformatic tool which computes the degree of conservancy of an epitope within a given protein sequence set at a given identity level.

In this study, the receptor binding domain (RBD) of SARS-CoV obtained from Uniprot accession/ID: P59594 with the residues 424-494 of the SARS-CoV spike (S) glycoprotein [72] with specific sequence (NTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCTPPALNCYWPLNDYGFYTTTGIGYQPY) was aligned to the SARS-CoV-2 spike (S) glycoprotein receptor binding domain containing residues 319-541 [60] containing sequences (RVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNF) acquired from the Uniprot accession ID (P0DTC2) using IEDB Epitope Conservancy Analysis tool.

Since the study is focused on determination of linear epitopes, the Epitope Linear Conservancy option was selected. In the tool, conservancy is defined as the fraction of protein sequences that contain the epitope, and identity is the degree of correspondence (similarity) between two sequences. The summary view shows for each epitope, the calculated degree of conservancy (percent of protein sequence matches a specified identity level) and the matching minimum/maximum identity levels within the protein sequence set. IEDB is supported by a contract from the National Institute of Allergy and Infectious Diseases, a component of the National Institutes of Health in the Department of Health and Human Services [73].

Figure 11 below generally shows the IEDB Epitope Conservancy Analysis. Figure 11A shows the schematic representation of the SARS-CoV-2 and SARS-CoV-2 spike (S) glycoprotein receptor binding domains [60,72] created using Biorender, UCSF was used to model the 3D image of the homotrimeric chains [23,74] of SARS-CoV-2 [PDB 6VYB [108], chain B in open state, shaded in red] and SARS-CoV [PDB 5XLR [115], chain B in prefusion conformation shaded in blue]. Figure 11B shows the output from the IEDB Conservancy Analysis tool generated from tools.iedb.org. An epitope length composed of 71 amino acid residues with sequence (NTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCTPPALNCYWPLNDYGFYTTTGIGYQPY) has a maximum identity of about 14% possibly conserved within the SARS-CoV-2 and SARS-CoV-2 spike (S) glycoprotein receptor binding domains (RBD), IEDB Epitope Conservancy Analysis tool was used to predict the conserved regions within the RBDs of the two viruses which infected humans.

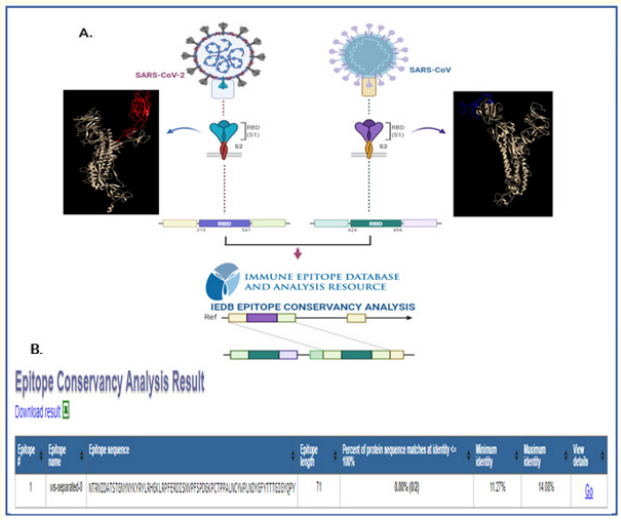


Figure 11: IEDB Epitope Conservancy Analysis tool output.

SARS-CoV-2 spike (S) glycoprotein epitope prediction:

Epitope as defined by Sette., *et al*. [75] are the determinants or parts of the molecule (e.g., amino acid sequences) recognized by antibodies which are glycoproteins produced by plasma cells (a type of B cell) that can neutralize or render a target antigen [76] inactive by recognition of the binding sites and through antibody-dependent cell cytotoxic mechanisms. B cell can also express antibodies (IgG, IgD) on their surface which can interact with antigens [77] by recognition of epitopes and exhibit their immune functions. Paratopes are the sequences or determinants found in the antibodies [78] in recognizing its counterpart in antigens, epitopes.

Epitopes can either be linear or conformational/discontinuous [79]. The former is defined by the sets of amino acids in close position to one another while the latter are residues brought about by protein folding. Linear epitopes are the choice for antibody development relating to the protein target which can be wholly or partially denatured during the sample preparation prior to the immunoassays, such as Western blot (WB), immunohistochemistry (IHC) or in polyclonal antibody development while conformational epitopes is preferred for applications involving protein targets in their native state such as for flow cytometry [79]. Linear epitopes for vaccine development may also offer an advantage as it targets specific regions of antigens, provide accomodation for occurrence of antigenic mutations and most importantly, as the binding depends on conformation affinities- antibody to linear epitopes might be less damaged by denaturation and loss of activities [80]. Linear epitope prediction leading to antibody or vaccine development against SARS-CoV-2 is imperative.

The study is focused on the prediction of the B cell linear epitopes for the SARS-CoV-2 spike (S) glycoprotein. As of the current time, many pharmaceutical companies either had developed vaccines or still on the process of constructing an effective vaccine against SARS-CoV-2. The efficacy of SARS-CoV-2 vaccines still needs to be validated and assessed. Poh., *et al*. [81] had determined a pool of B cell linear peptides utilizing Mimotopes, ELISA and cell culture from IgG immunodominant regions on SARS-CoV-2 spike (S) glycoprotein. Lin *et al*. [82] combined epitope predictions of both T cells and B cells as well as characterization of interactions. Trolle., *et al*. [83] also noted that cytotoxic cells (CD8+) also recognize linear epitopes presented by virally infected cells expressing MHC I [83] containing processed peptides.

IEDB Epitope Prediction and Analysis tools was utilized in this study for the prediction of linear epitopes. The IEDB Epitope Database Analysis Resource provides a collection of tools for the prediction and analysis of immune epitopes, several bioinformatics tools can be found here, T cell, B cell and an extensive analysis for a group of sequences. For the B cell analysis, it is divided into two parts- a collection of methods using different principles and the ElliPro prediction method.

In the study, ElliPro prediction was utilized to generate the B cell linear epitope predictions in combination to the following: Bepipred Linear-1.0 Epitope Prediction; Kolaskar and Tongaonkar Antigenicity and, Emini Surface Accessibility Prediction. The selection of methods is based on identifying potential fixed B cell linear epitopes.

From the IEDB home [85], the general methods can be identified- parameters such as accessibility, hydrophilicity, flexibility, turns, exposed surface, polarity and antigenic propensities/properties of polypeptides chains will be correlated with the location of linear or continuous epitopes. A set of empirical rules would allow the position of epitopes to be predicted from certain features of the protein sequence. All prediction calculations are based on propensity scales for each of the 20 amino acids. Each scale consists of 20 values assigned to each of the amino acid residues on the basis of their relative propensity to possess the property described by the scale which differs in the methods mentioned (refer to Figure

12). Note that when computing the score for a given residue "i", the amino acids in an interval of the chosen length, centered around residue "i", are considered. In other words, for a window size n, the "i" - (n-1)/2 neighboring residues on each side of residue "i" were used to compute the score for residue "i". Unless specified, the score for residue "i" is the average of the scale values for these amino acids. In general, a window size of 5 to 7 is appropriate for finding regions that may potentially be antigenic. Table 1 in the proceeding selection summarizes the three (3) selected linear predicted epitope peptides per methods.

Ellipro can predict linear and discontinuous antibody epitopes based on a protein antigen's 3D structure [86], generally it accepts proteins in PDB format but can also cater protein sequences that needs processing in the modelling and docking section. Ponomarenko., *et al*. [86] also mentioned that ElliPro associates each predicted epitope with a score which is defined as a PI (protrusion index) value averaged over epitope residues. In the method, the protein's 3D shape is approximated by a number of ellipsoids, thus that the ellipsoid with PI = 0.9 would include within 90% of the protein residues with 10% of the protein residues being outside of the ellipsoid; while the ellipsoid with PI = 0.8 would include 80% of residues with 20% being outside the ellipsoid. For each residue, a PI value is defined based on the residue's center of mass lying outside the largest possible ellipsoid; for example, all residues that are outside the 90% ellipsoid will have score of 0.9. Residues with larger scores are associated with greater solvent accessibility. Discontinuous epitopes are defined based on PI values and are clustered based on the distance R (in Å between residue's center of mass). The larger R is associated with prediction of larger discontinuous epitopes. Fleri [87] noted that ElliPro computes a score for each residue of the input sequence (PI) which is defined as the percentage of the protein atoms enclosed in the ellipsoid at which the residue first become lying outside the ellipsoid.

For the ElliPro prediction, the chain B of PDB 6VYB [108] was used since it denotes the open state and accessible to host receptor binding. Chain B from PDB 6VYB [108] has 949 residues. ElliPro can predict both linear and discontinuous epitopes however this paper focuses on the B cell linear epitopes. The output is arranged in from highest to lowest scores. A 3D option is also available to view the residues predicted in ElliPro.

Figure 12 below shows the different scales and thresholds for the methods Emini Surface Accessibility Prediction, Kolaskar and Tongaonkar Antigenicity Scale and Bepipred-1.0 Linear Epitope Prediction from the IEDB database [85].

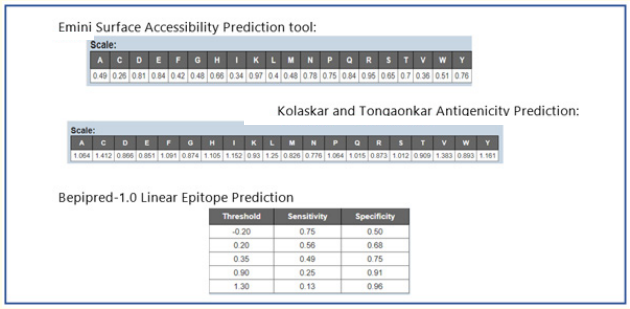


Figure 12: IEDB B cell epitope prediction scales and thresholds.

In the Emini Surface Accessibility Prediction tool, the calculation was based on surface accessibility scale on a product instead of an addition within the window. The accessibility profile was obtained using the formula Sn = (n+4+"i") (0.37)-6 where Sn is the surface probability, dn is the fractional surface probability value, and "i" vary from 1 to 6. A hexapeptide sequence with Sn greater than 1.0 indicates an increased probability for being found on the surface. (Emini., *et al*. 1985). The sequence of the SARS-CoV-2 spike (S) glycoprotein obtained from Uniprot (ID: P0DTC2) was utilized in predicting the possible linear epitopes under the method Emini Surface Accessibility Prediction tool. Figure 16 below shows the output of the prediction. The peptides are arranged by prediction from the start to end of sequence. Figure in the next page shows the score charts (P0DTC2).

Note that for the methods of Bepipred Linear-1.0 Epitope Prediction, Kolaskar and Tongaonkar Antigenicity and Emini Surface Accessibility Prediction, the SARS-CoV-2 spike (S) glycoprotein sequence obtained from Uniprot (ID: P0DTC2) was used for generating the prediction of linear epitopes since the methods mentioned require a FASTA sequence in addition that some models have missing residues which can affect the epitope prediction. The outputs generated indicate peptides arranged by prediction from the start to end of sequence. Kindly refer to table 1 for the summary of the results generated from these B cell epitope prediction tools together with the results from determination of disordered regions of the SARS-CoV-2 spike (S) glycoprotein.

In the Emini Surface Accessibility Prediction tool, the calculation was based on surface accessibility scale on a product instead

of an addition within the window. The accessibility profile was obtained using the formula Sn = (n+4+"i") (0.37)-6 where Sn is the surface probability, dn is the fractional surface probability value, and "i" vary from 1 to 6. A hexapeptide sequence with Sn greater than 1.0 indicates an increased probability for being found on the surface [88].

In the method of Kolaskar and Tongaonkar Antigenicity Prediction, it utilizes the physicochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes to predict antigenic determinants on proteins). It has about 75% more accuracy compared to other methods [90].

Bepipred-1.0 Linear Epitope Prediction [89], predicts the location of linear B- cell epitopes using a combination of a hidden Markov model and a propensity scale method. The residues with scores above the threshold (default value is 0.35) are predicted to be part of an epitope and is marked on the graph.

The linear epitopes of interest were analyzed and determined from combinations of structural analysis, especially the RBD of the S1 that is used in viral interaction to host receptors together with the B cell linear epitopes generated by IEDB prediction tools- ElliPro Prediction, Bepipred Linear-1.0 Epitope Prediction**,** Kolaskar and Tongaonkar Antigenicity together with Emini Surface Accessibility Prediction. The identification of the disordered regions of the SARS-CoV-2 spike (S) glycoprotein which will be discussed below was performed to compliment the process and serve as a guide for the determination of the linear epitopes, from all of these techniques the two (2) best linear epitope candidates are selected. The chosen linear epitopes are then assessed in the IEDB resource database to locate similar sequences (see figure 15 for the output). In the succeeding contexts, Table 1 gives the summary of the linear epitope predictions incorporating all the bioinformatics analysis and figure 16 illustrates the bioinformatics process for the SARS-CoV-2 spike (S) glycoprotein epitope prediction for antibody or vaccine development.

SARS-CoV-2 spike glycoprotein predicted linear epitope and degree of disorder:

Deng., *et al*. [91] noted that disordered regions in a protein are segments or residues which do not form a stable structure and is important for protein-DNA binding. MacRaild., *et al*. [92] also noted that disordered regions in proteins are smaller than the ordered counterparts but are more efficient in terms of antibody binding. Guy., *et al*. [93] also noted that disordered regions in proteins tend to cluster with tandem repeats which can be immunodominant and effectively induce an immune response, which has implications for vaccine design. In this study, both DisEMBL and GlobPlot2.3 was used to determine the disordered regions in the SARS-CoV-2 spike (S) glycoprotein using Uniprot obtained sequence (ID: P0DTC2).

DisEMBL is composed of three network algorithms that predict the presence of disorder: COILS, HOTLOOPS, and REM465, referred to by Disorder Atlas as DisEMBL-C, DisEMBL-H, and DisEMBL-R, respectively [94]. Vincent., *et al*. [95] noted that DisEMBL-C utilizes secondary structure prediction to assign disorder/order classifications, and classifies residues as disordered if they are present in loops. Residues are predicted to belong to loops if they do not belong to either alpha-helices, 3_{10}-helices, or beta-strands.

Vincent., *et al*. [95] also discussed that DisEMBL- H classifies residues (contained within loops) as disordered only if its alpha-carbon has a high temperature factor (B factor). DisEMBL-R network correlated non-assigned electron densities from X-ray crystallography data contained within the PDB and assumes residues with missing XRC coordinates (defined by REMARK 465) as disordered (Vincent., *et al*. 2019). Linding., *et al*. [94] also mentioned that GlobPlot is a simple approach of identifying disordered regions based on a running sum of the propensity for amino acids to be in an ordered or disordered state, it can also be used in the design of constructs corresponding to globular proteins as well as identification of new functional sites. Pillay., *et al*. [96], and Chan., *et al*. [97] also noted that the RBD of the S1 has a globular shape. DisEMBL and GlobPlot can complement one another in terms of disordered region determination [94]. Figures 13 and 14 shows the output from DisEBML and GlobPlot (version 2.3 was used in the study), respectively. SARS-Cov-2 spike (S) glycoprotein sequence obtained from Uniprot (ID: P0DTC2) was used for analysis. Both outputs generated disordered regions shaded in blue and are capitalized.

Table 1 summarizes the selected B cell linear epitopes (amino acid sequences) as determined by the ElliPro Prediction, Bepipred Linear-1.0 Epitope Prediction**,** Kolaskar and Tongaonkar Antigenicity Emini Surface Accessibility Prediction together with DisEMBL and GlobPlot2.3. For each bioinformatics method, the possible B cell linear epitopes (shaded in black) is considered for the final selection. Structural and sequence analysis of the SARS-CoV-2 spike (S) glycoprotein, contribution of the receptor binding domains (RBD) to human immune cell receptors and alignment of the disordered regions are carefully weighed to arrive at the two (2) selected B cell linear epitopes (highlighted in yellow, amino acid sequences shaded in blue).

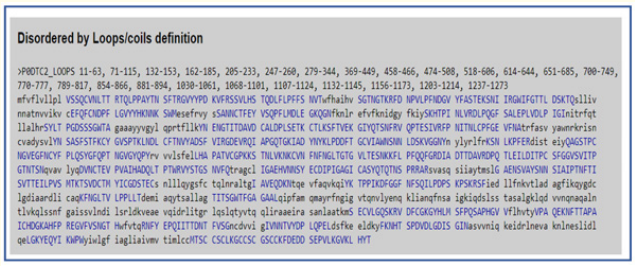


Figure 13: Disordered regions of SARS-CoV-2 spike (S) glycoprotein as determined by DisEMBL.

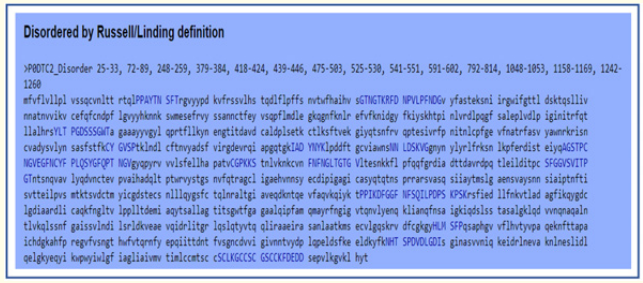


Figure 14: Disordered regions of SARS-CoV-2 spike (S) glycoprotein as determined by GlobPlot2.3.

IEDB Prediction tool	No.	Predicted B cell Linear epitopes	Position	Residue length
ElliPro	1	LCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDD FTGCVIAWNSNNLDSKVGGNYNYLYRKPFERPLQSYG FQPTNGVGYQPYRVVVLSFATVCG	390-526	100
	2	FPNITNLCPFGEVFNATRFASVYAWNRKRISNC VADYSVLYNSASFSTFKCYG	329-381	53
Emini Surface Accessibility	1	NSNNLD	437-442	6
	2	YGFQPT	495-500	6
Kolaskar and Tongaonkar Antigenicity	1	CYFPLQSY	488-495	8
	2	YQPYRVVVLSFELLHAPATVCGP	505-527	23
Bepipred Linear-1.0	1	YQAGSTPCNGV	473-483	11
	2	YGFQPTNGVGYQ	495-506	12
DisEMBL	1	YNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQI APGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNY	369-449	81
	2	QAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY	474-508	35
GlobPlot2.3	1	NNLDSKVG	439-446	8
	2	AGSTPCNGVEGFNCYFPLQSYGFQPTNGV	475-503	29
Selected B cell linear epitopes	1.	CNGVEGFNCYFPLQSYGFQPTNGVGYQPY	480-508	29
	2.	NNLDSKVGGNY	439-449	11

Table 1: Summary of the selected linear epitopes for the SARS-CoV-2 spike (S) glycoprotein.

Polyclonal antibodies recognize multiple epitopes in an antigen [89] as compared to monoclonal antibodies which recognize single epitope. Polyclonal antibodies are hypothesized to be developed if the objective is to bind both linear epitopes. Lipman., *et al.* [98] also noted that polyclonal antibodies can do cross linking and possibly complement activation as multiple epitopes are recognized and due to Fc receptors proximity. Bioinformatics can also serve as a guide for monoclonal antibody development as it offers great success for target and binding analysis [99]. Epitope predictions using bioinformatics are also utilized in terms of vaccine design and development as well as identification of immune interactions with high specificity [100] through analysis of vast amount of data, modelling, binding interactions, sequence investigation and determination of functional areas and immunobiologic correlations with the use of bioinformatic methods. Cornick., *et al.* [101] developed vaccine candidates against *Streptococcus pneumoniae* with bioin-

formatic analysis using epitope prediction and structural modelling. In a previous paper published [102], binding of a monoclonal antibody to *Streptococcus pyogenes* serotype M18 superantigen SpeC was analyzed using bioinformatics-guided epitope prediction tools. Alam., *et al*. [103] also designed vaccine candidates against Zika using predicted MHC I epitopes. Different bioinformatic researches for SARS-CoV-2 are also available- Chen., *et al*. [104] also used IEDB softwares for B cell (ABCpred, Bepipred) and T cell linear and discontinuous (Discotope 2.0) epitope identification. Poran., *et al*. [105] identified HLA alleles using Virus Pathogen Database and Analysis Resource (ViPR) database.

SARS-CoV-2 spike (S) glycoprotein predicted linear epitope database search

The IEDB database was utilized for the two (2) selected B cell linear epitopes to identify similar sequences. Filters used to identify if there are similar epitopes include "organism" (SARS-CoV-2); which infects "host" (humans); including searches for all "assays" (Positive assay, B cell assay, T cell assay and Major Histocompatibility Ligand/MHC assay); "MHC restriction" (any restriction); and "disease" (any disease). Figures shows the result of the IEDB search for both (2) selected B cell linear epitopes- CNGVEGFNCYFPLQSYGFQPTNGVGYQPY and NNLDSKVGGNY. No similar sequences are found in humans, regardless of any assays for immune cells such as B cells, T cells and MHC. Note that this bioinformatics process was done in order to check cross-specificity between the selected linear epitopes of interest to the immune cells and proteins of the human body, this can possibly indicate that by using the epitopes of interest no cross-reaction may occur. Different *in vivo* and *in vitro* tests must be performed to compliment and to validate the findings.

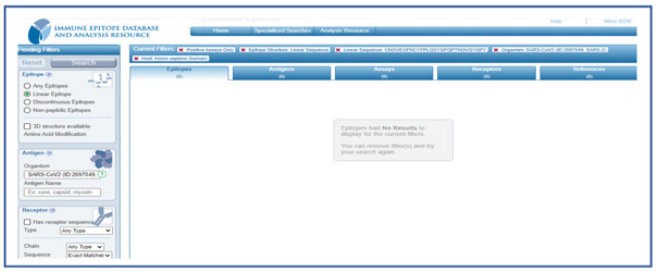


Figure 15: IEDB Related Epitope Search Database for the predicted B cell linear epitopes.

Figure 16 below illustrates the overall process of predicting the B cell linear epitope of the SARS-CoV-2 spike (S) glycoprotein. Biorender was used in modelling the figures while UCSF Chimera was utilized in generating the 3D molecular structure of the SARS-CoV-2 spike (S) glycoprotein highlighting the selected B cell linear epitopes CNGVEGFNCYFPLQSYGFQPTNGVGYQPY (highlighted in pink) and NNLDSKVGGNY (highlighted in cyan). The prediction of the B cell linear epitopes were generated from different bioinformatics tools or software analysis- ElliPro Prediction, Bepipred Linear-1.0 Epitope Prediction, Kolaskar and Tongaonkar Antigenicity Emini Surface Accessibility Prediction together with DisEMBL and GlobPlot2.3. Figure 16A schematically shows the B cell-derived antibody and SARS-CoV-2 spike (S) glycoprotein interaction. Figure 16B depicts the bioinformatics tools utilized and Figure 16C illustrates the 3D model of the selected peptides highlighted form the chain, for comparison the PDB 7A97 [114] chain B was used and the ACE2 (shaded in orange) was still bound to serve as reference, indicating that the chosen B cell linear epitopes are part of the receptor binding domains (RBD) which can potentially be neutralized by antibodies or vaccine-induced antibody responses. Figure 16D indicates the linear epitope search from IEDB which is important to identify possible cross-reactivities when utilizing the epitopes of interest. More importantly, the predicted linear epitopes adopt a helix conformation and Kim., *et al*. [106] noted that antibodies directed against helix strands binds with high affinity and can be advantageous in relation to vaccine development.

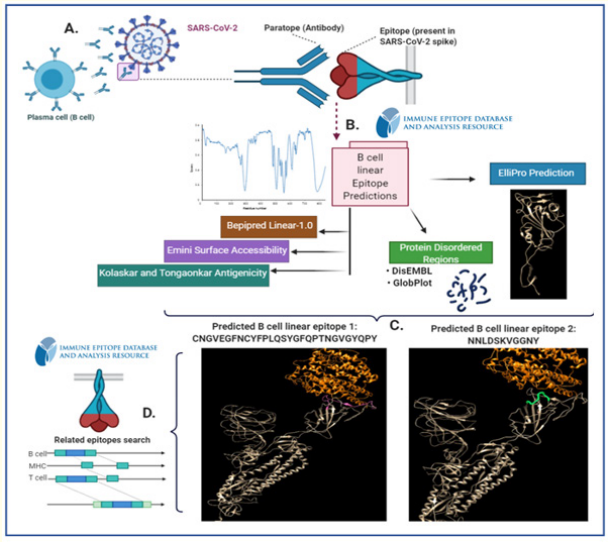


Figure 16: SARS-CoV-2 spike (S) glycoprotein B cell linear epitope prediction.

Conclusion

The recent pandemic is caused by the SARS-CoV-2 which resulted in about 1 million deaths globally and increasing number of infections and deaths are still reported. The spike (S) glycoprotein of SARS-CoV-2 binds to many host receptors and is enzymatically processed leading to infection of many human cells. Prediction of B cell linear epitopes using bioinformatics can be vital for antibody or vaccine development. Bioinformatics allow an extensive visualization and analysis of combined molecular virologic, biochemical, cellular and immunobiologic aspects of SARS-CoV-2 for antibody or vaccine development. Further, Tropical Medicine and Bioinformatics allow us to explore new horizons and turn the tides of the battle against SARS-CoV-2.

Recommendations

Bioinformatics output in the paper offer an analysis of the molecular features of the SARS-CoV-2 spike (S) glycoprotein in terms of host receptor binding, amino acid sequences and host enzyme processing as well as basic prediction linear epitopes for antibody or vaccine development. Different *in vitro*, *in vivo* and other *in silico* methods can also be used to complement the results as well as visualize and measure the laboratory parameters. Different antibody purification and characterization methods can also be performed in the future. The analysis of other the molecular features such as viral transcription to translation and immunobiologic characteristics can also be further explored.

Conflict of Interest

None declared.

Funding Details

None declared

Ethical Approval

Not required.

Author Declarations

JAHM conceptualized the study as the sole author- from the selection of bioinformatic tools; sequence and structure analysis; correlation of molecular, biochemical, cellular and immunobiologic features; B cell linear epitope predictions; disordered region identification and related epitope search and analysis. The bioinformatics method utilized were discussed and cited properly. No competing interests are involved in the study.

Acknowledgements

The author would like to acknowledge University of the Philippines Manila. The author would also like to address his gratitude to the developers of the bioinformatic tools utilized- Biorender, Uniprot, PDB, UCSF Chimera, Smart BLAST, IEDB softwares, GlobPlot and DisEMBL.

Bibliography

1. Huang C., *et al*. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China". *Lancet* 395 (2020): 497-506.

2. Jin Y., *et al*. "Virology, Epidemiology, Pathogenesis, and Control of COVID-19". *Viruses* 12 (2020): 372.

3. Wang L., *et al*. "Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence". *International Journal of Antimicrobial Agents* 55.6 (2020): 105948.

4. Chan JF., *et al*. "Interspecies transmission and emergence of novel viruses: lessons from bats and birds". *Trends in Microbiology* 21 (2013): 544-555.

5. Corman VM., *et al*. "Hosts and Sources of Endemic Human Coronaviruses". *Advances in Virus Research* 100 (2018): 163-188.

6. Mackay IM and Arden KE. "MERS coronavirus: diagnostics, epidemiology and transmission". *Virology Journal* 12 (2015): 222.

7. Ji W., *et al*. "Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross-species transmission from snake to human". *Journal of Medical Virology* (2020): 92.

8. Ji W., *et al*. "Cross-species transmission of the newly identified coronavirus 2019-nCoV". *Medical Virology* 92.4 (2020): 433-440.

9. King AMQ., *et al*. "Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses". Amsterdam: Elsevier (2011): 806-828.

10. https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations#:~:text=Respiratory%20infections%20can%20be%20transmitted,droplets%20and%20contact%20routes

11. https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions

12. Baron. "Generation and behaviour of airborne particles" (2020).

13. Adam D., *et al*. "Clustering and superspreading potential of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections in Hong Kong (pre-print)". *Research Square* (2020).

14. Vivanti AJ., *et al*. Transplacental transmission of SARS-CoV-2 infection". *Nature Communication* 11 (2020): 3572.

15. https://reliefweb.int/sites/reliefweb.int/files/resources/WHO%20PHL%20SitRep%2052_COVID-19_8September2020.pdf

16. https://www.doh.gov.ph/covid-19/case-tracker

17. https://covid19.who.int/?gclid=Cj0KCQjwwuD7BRDBARIsAK_5YhUDSRs8OurDFY0qzMFOiqeeqdFWg2uzObwkHGlgp5qKydndFaw8pE0aAgwUEALw_wcB

18. https://covid19.who.int/

19. Bayat A. "Science, medicine, and the future: Bioinformatics". *BMJ* (Clinical research ed.), 324.7344 (2002): 1018-1022.

20. Kumar S., *et al*. "Morphology, Genome Organization, Replication, and Pathogenesis of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Coronavirus Disease 2019 (COVID-19): Epidemiology, Pathogenesis, Diagnosis, and Therapeutics" (2020): 23-31.

21. Woo PC., *et al*. "Coronavirus diversity, phylogeny and interspecies jumping". *Experimental Biology and Medicine (Maywood)* 234.10 (2009): 1117-1127.

22. Park WB., *et al*. "Virus Isolation from the First Patient with SARS-CoV-2 in Korea". *Journal of Korean Medical Science* 35.7 (2020): e84.

23. Kalathiya U., *et al*. "Highly Conserved Homotrimer Cavity Formed by the SARS-CoV-2 Spike Glycoprotein: A Novel Binding Site". *Journal of Clinical Medicine* 9.5 (2020): 1473.

24. Lau S Y., *et al*. "Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction". *Emerging Microbes and Infections* 9.1 (2020): 837-842.

25. Jaimes J A., *et al*. "Proteolytic Cleavage of the SARS-CoV-2 Spike Protein and the Role of the Novel S1/S2 Site". *iScience* 23.6 (2020): 101212.

26. Ou X., *et al*. "Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV". *Nature Communication* 11 (2020): 1620.

27. Trigueiro-Louro J., *et al*. "Unlocking COVID therapeutic targets: A structure-based rationale against SARS-CoV-2, SARS-CoV and MERS-CoV Spike" (2020).

28. Wu A., *et al*. "Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China". *Cell Host Microbe* 27.3 (2020): 325-328.

29. Khailany R A., *et al*. "Genomic characterization of a novel SARS-CoV-2". *Gene Reports* 19 (2020): 100682.

30. Xu J., *et al*. "Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV". *Viruses* 12.2 (2020): 244.

31. Lu R., *et al*. "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding". *Lancet (London, England)* 395.10224 (2020): 565-574.

32. Wang H., *et al*. "The genetic sequence, origin, and diagnosis of SARS-CoV-2". European journal of clinical Microbiology and Infectious Diseases: official publication of the European Society of Clinical Microbiology 39.9 (2020): 1629-1635.

33. Perkel. "The software that powers scientific illustration". *Nature* 582 (2020): 137-138.

34. Pettersen., *et al*. "UCSF Chimera--a visualization system for exploratory research and analysis". *Journal of Computational Chemistry* 25.13 (2019): 1605-1612.

35. Hoffmanm M., *et al*. "A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells". *Molecular Cell* 78.4 (2020): 779-784.e5.

36. Shang J., *et al*. "Cell entry mechanisms of SARS-CoV-2". *Proceedings of the National Academy of Sciences of the United States of America* 117.21 (2020): 11727-11734.

37. Johnson B A., *et al*. "Furin Cleavage Site Is Key to SARS-CoV-2 Pathogenesis". *bioRxiv* (2020).

38. Tang T., *et al*. "Coronavirus membrane fusion mechanism offers a potential target for antiviral development". *Antiviral Research* 178 (2020): 104792.

39. Tong Meng., *et al*. "The insert sequence in SARS-CoV-2 enhances spike protein cleavage by TMPRSS". *bioRxiv* (2020).

40. Heurich A., *et al*. "TMPRSS2 and ADAM17 cleave ACE2 differentially and only proteolysis by TMPRSS2 augments entry driven by the severe acute respiratory syndrome coronavirus spike protein". *Journal of Virology* 88 (2014): 1293-1307.

41. Sungnak W., *et al*. "SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes". *Nature Medicine* 26 (2020): 681-687.

42. Abassi Z., *et al*. "The Lung Macrophage in SARS-CoV-2 Infection: A Friend or a Foe?". *Frontiers in Immunology* 11 (2020): 1312.

43. Rensi S., *et al*. "Homology Modeling of TMPRSS2 Yields Candidate Drugs That May Inhibit Entry of SARS-CoV-2 into Human Cells". *ChemRxiv* (2020).

44. Morniroli D., *et al*. "Human Sialome and Coronavirus Disease-2019 (COVID-19) Pandemic: An Understated Correlation?". *Frontiers in Immunology* (2020).

45. Wei Hao., *et al*. "Binding of the SARS-CoV-2 Spike Protein to Glycans". *bioRxiv* (2020).

46. Ward S., *et al*. "von Willebrand factor sialylation-A critical regulator of biological function". *Journal of Thrombosis and Haemostasis* 17.7 (2019): 1018-1029.

47. Estus S., *et al*. "Evaluation of CD33 as a genetic risk factor for Alzheimer's disease". *Acta Neuropathology* 1380.2 (2019): 187-199.

48. Schwartz-Albiez R., *et al*. "Cell surface sialylation and ecto-sialyltransferase activity of human CD34 progenitors from peripheral blood and bone marrow". *Glycoconj Journal* 21 (2004): 451-459.

49. Kawasaki N., *et al*. "Targeted delivery of lipid antigen to macrophages via the CD169/sialoadhesin endocytic pathway induces robust invariant natural killer T cell activation". *Proceedings of the National Academy of Sciences of the United States of America* 110.19 (2020): 7826-7831.

50. Merah-Mourah F., *et al*. "Identification of Novel Human Monocyte Subsets and Evidence for Phenotypic Groups Defined by Interindividual Variations of Expression of Adhesion Molecules". *Scientific Report* 10 (2020): 4397.

51. Venosa A., *et al*. "Characterization of Distinct Macrophage Subpopulations during Nitrogen Mustard-Induced Lung Injury and Fibrosis". *American Journal of Respiratory Cell and Molecular Biology* 54.3 (2016): 436-446.

52. Grabowska J., *et al*. "CD169+ Macrophages Capture and Dendritic Cells Instruct: The Interplay of the Gatekeeper and the General of the Immune System". *Frontiers in Immunology* 9 (2018): 2472.

53. Kessel K U., *et al*. "Emergence of CD43-Expressing Hematopoietic Progenitors from Human Induced Pluripotent Stem Cells". *Transfusion medicine and hemotherapy: offizielles Organ der Deutschen Gesellschaft fur Transfusionsmedizin und Immunhamatologie* 44.3 (2017): 143-150.

54. Pedraza-Alva G., *et al*. "T cell activation through the CD43 molecule leads to Vav tyrosine phosphorylation and mitogen-activated protein kinase pathway activation". *The Journal of Biological Chemistry* 273.23 (1998): 14218-14224.

55. Wang Q., *et al*. "Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2". *Cell* 181.4 (2020): 894-904.e9.

56. Nguyen H L., *et al*. "Does SARS-CoV-2 Bind to Human ACE2 More Strongly Than Does SARS-CoV?". *The Journal of Physical Chemistry. B* 124.34 (2020): 7336-7347.

57. Koyama T., *et al*. "Variant analysis of SARS-CoV-2 genomes". *Bulletin of the World Health Organization* 98 (2020): 495-504.

58. Ma Y., *et al*. "Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex". *Proceedings of the National Academy of Sciences of the United States of America* 112.30 (2015): 9436-9441.

59. Pachetti M., *et al*. "Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant". *Journal of Translational Medicine* 18.1 (2020): 179.

60. Huang Y., *et al*. "Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19". *Acta Pharmaceutica Sinica* 41 (2020): 1141-1149.

61. The UniProt Consortium. "UniProt: a worldwide hub of protein knowledge". *Nucleic Acids Research* 47 (2019): D506-515.

62. https://www.uniprot.org/uniprot/P0DTC2

63. Apweiler R., *et al*. "Protein sequence databases". *Current Opinion in Chemical Biology* 8.1 (2004): 76-80.

64. Leinonen R Diez., *et al*. "UniProt archive". *Bioinformatics* 20.17 (2004): 3236-3237.

65. Suzek B., *et al*. "UniRef: Comprehensive and non-redundant UniProt reference clusters". *Bioinformatics* 23.10 (2007): 1282-1288.

66. Uniprot C. "The Universal Protein Resource (UniProt) in 2010". *Nucleic Acids Research* 38 (2009): D142-D148.

67. wwPDB Consortium. "Protein Data Bank: the single global archive for 3D macromolecular structure data". *Nucleic Acids Research* 47 (2019): 520-528.

68. https://www.rcsb.org/

69. NCBI Resource Coordinators. "Database resources of the National Center for Biotechnology Information". *Nucleic Acids Research* 46 (2018): D8-D13.

70. Introduction. SMARTBLAST.

71. Bui HH., *et al*. "Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines". *BMC Bioinformatics* 8 (2007): 361.

72. Li F., *et al*. "Structure of SARS coronavirus spike receptor-binding domain complexed with receptor". *Science* (New York, N.Y.), 309.574 (2005): 1864-1868.

73. http://tools.iedb.org/conservancy/help/#:~:text=This%20tool%20computes%20the%20degree, (similarity)%20between%20two%20sequences

74. Song H C., *et al*. "Synthesis and characterization of a native, oligomeric form of recombinant severe acute respiratory syndrome coronavirus spike glycoprotein". *Journal of Virology* 78.19 (2004): 10328-10335.

75. Sette A., *et al*. "Definition of epitopes and antigens recognized by vaccinia specific immune responses: their conservation in variola virus sequences, and use as a model system to study complex pathogens". *Vaccine* 27 (2009): G21-G26.

76. Forthal D N. "Functions of Antibodies". *Microbiology Spectrum* 2.4 (2014): 1-17.

77. Noviski M., *et al*. "IgM and IgD B cell receptors differentially respond to endogenous antigens and control B cell fate". *eLife* 7 (2018): e35074.

78. Frank SA." Immunology and Evolution of Infectious Disease". Princeton (NJ): Princeton University Press. Chapter 4, Specificity and Cross-Reactivity (2002).

79. Forsström B., *et al*. "Dissecting antibodies with regards to linear and conformational epitopes". *PloS One* 10.3 (2015): e0121673.

80. Oscherwitz J. "The promise and challenge of epitope-focused vaccines". *Human Vaccines and Immunotherapeutics* 12.8 (2016): 2113-2116.

81. Poh CM., *et al*. "Two linear epitopes on the SARS-CoV-2 spike protein that elicit neutralising antibodies in COVID-19 patients". *Nature Communication* 11 (2020): 2806.

82. Lin L., *et al*. "Epitope-based peptide vaccines predicted against novel coronavirus disease caused by SARS-CoV-2". *Virus Research* 288 (2020): 198082.

83. Trolle T., *et al*. "The Length Distribution of Class I-Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele-Specific Binding Preference". *Journal of Immunology* (Baltimore, Md.: 1950), 196.4 (2016): 1480-1487.

84. Hewitt E W. "The MHC class I antigen presentation pathway: strategies for viral immune evasion". *Immunology* 110.2 (2003): 163-169.

85. http://tools.iedb.org/bcell/help

86. Ponomarenko J., *et al*. "ElliPro: a new structure-based tool for the prediction of antibody epitopes". *BMC Bioinformatics* 9 (2008): 514.

87. https://help.iedb.org/hc/en-us/articles/114094151671-ElliPro-What-do-the-output-scores-mean

88. Emini EA. "Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide". *Journal of Virology* 55.1985 (1990): 836-839.

89. Larsen JE., *et al*. "Improved method for predicting linear B-cell epitopes". *Immunome Research* 2 (2006): 2.

90. Kolaskar AS and Tongaonkar PC. "A semi-empirical method for prediction of antigenic determinants on protein antigens". *FEBS Letter* 276 (1990): 172-174.

91. Deng X., *et al*. "A comprehensive overview of computational protein disorder prediction methods". *Molecular bioSystems* 8.1 (20121): 114-121.

92. MacRaild C A., *et al*. "Antibody Recognition of Disordered Antigens". *Structure* (London, England: 1993), 24.1 (2016): 148-157.

93. Guy A J., *et al*. "Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions". *PloS One* 10.10 (2015): e0141729.

94. Linding R., *et al*. "GlobPlot: Exploring protein sequences for globularity and disorder". *Nucleic Acids Research* 31.13 (2003): 3701-3708.

95. Vincent M and Schnell S. "Disorder Atlas: Web-based software for the proteome-based interpretation of intrinsic disorder predictions". *Computational Biology and Chemistry* 83 (2019): 107090.

96. Pillay TS. "Gene of the month: the 2019-nCoV/SARS-CoV-2 novel coronavirus spike protein". *Journal of Clinical Pathology* 73.7 (2020): 366-369.

97. Chan J F., *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* 9.1 (2020): 221-236.

98. Neil S., *et al*. "Monoclonal Versus Polyclonal Antibodies: Distinguishing Characteristics, Applications, and Information Resources". *ILAR Journal* 46.3 (2005): 258-268.

99. Denisova G F., *et al*. "Applying bioinformatics for antibody epitope prediction using affinity-selected mimotopes - relevance for vaccine design". *Immunome Research* 6 (2010): S6.

100. Oli A N., *et al*. "Immunoinformatics and Vaccine Development: An Overview". *ImmunoTargets and Therapy* 9 (2020): 13-30.

101. Cornick JE., *et al*. "The global distribution and diversity of protein vaccine candidate antigens in the highly virulent Streptococcus pneumoniae serotype 1". *Vaccine* 35.6 (2017): 972-980.

102. Mandanas, J. "Molecular Characterization of Tefimazumab Binding to the Streptococcus pyogenes Serotype M18 Superantigen SpeC as Basis for Vaccine Development". *Acta Scientific Microbiology* 2.5 (2019): 114-118.

103. Alam A., *et al*. "From ZikV genome to vaccine: In silico approach for the epitope-based peptide vaccine against Zika virus envelope glycoprotein". *Immunology* 149.4 (2016): 386-399.

104. Chen H., *et al*. "Bioinformatics analysis of epitope-based vaccine design against the novel SARS-CoV-2". *Infectious Diseases of Poverty* 9 (2020): 88.

105. Poran A., *et al*. "Sequence-based prediction of SARS-CoV-2 vaccine targets using a mass spectrometry-based bioinformatics predictor identifies immunogenic T cell epitopes". *Genome Medicine* 12 (2020): 70.

106. Kim J W., *et al*. "Application of antihelix antibodies in protein structure determination". *Proceedings of the National Academy of Sciences of the United States of America* 116.36 (2019): 17786-17791.

107. Vita R., *et al*. "The Immune Epitope Database (IEDB): 2018 update". *Nucleic Acids Research* 24 (2018).

108. Walls A C., *et al*. "Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein". *Cell* 181.2 (2020): 281-292. e6.

109. Xiong X., *et al*. "A thermostable, closed SARS-CoV-2 spike protein trimer". *Nature Structural and Molecular Biology* 27 (2020): 934-941.

110. Dahms S O., *et al*. "Structural Studies Revealed Active Site Distortions of Human Furin by a Small Molecule Inhibitor". *ACS Chemical Biology* 12.5 (2017): 1211-1216.

111. Asaad N., *et al*. "Dipeptidyl nitrile inhibitors of Cathepsin L". *Bioorganic and Medicinal Chemistry Letters* 19.15 (2009): 4280-4283.

112. Rao K., *et al*. "Crystal Structure of Serine protease Hepsin in complex with Inhibitor".

113. Takai Y., *et al*. "Structural basis of the cytoplasmic tail of adhesion molecule CD43 and its binding to ERM proteins". *Journal of Molecular Biology* 381.3 (2008): 634-644.

114. Benton D J., *et al*. "Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion". *Nature* (2020).

115. Gui M., *et al*. "Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding". *Cell Research* 27.1 (2017): 119-129.