Taibah University

Journal of Taibah University Medical Sciences

Student Article

In-silico analysis of recombinant protein vaccines based on the spike protein of Indonesian SARS-CoV-2 through a reverse vaccinology approach

Riska A. Febrianti, S.Pd. [a] and Erlia Narulita, Ph.D [a, b, *]

[a] *Department of Biotechnology, Postgraduate Program, University of Jember, Indonesia*
[b] *Department of Biology Education, University of Jember, Indonesia*

الملخص

أهداف البحث: تهدف هذه الدراسة إلى إنتاج لقاح بروتين تجميعي مرشح مع حاتمة بروتين الشوكة من فيروس كورونا سارس -2 الإندونيسي الذي يمكن أن يكون مناعيا ووقائيا من العدوى المستقبلية.

طرق البحث: تم استخدام نهج لقاح عكسي في هذا البحث لتحديد اللقاحات المحتملة المرشحة عن طريق فحص جينوم العامل الممرض من خلال التحليلات الحسابية.

النتائج: بناء على نتائج هذه الدراسة، تم اختيار لقاح الحاتمة المرشحة مع تسلسل الأحماض الأمينية المحبة للماء، وليس لها القدرة على إحداث تفاعلات المناعة الذاتية والحساسية، مستضدات، مصنفة كبروتينات مستقرة، ومن المتوقع تقديمها في غشاء الخلية. تم تصنيع الحواتم المحددة في ناقل البلازميد (الثرومبين). إضافة ميزات لتركيبات اللقاح مثل إنزيم ترميز الجين ناقلات الغلوتاثيون، والذي يمكن أن يزيد من التعبير عن المستضد وقابلية ذوبانه، وبروتين ترميز الجين للبروتين غير الهيكلي 1-4، الذي يشفر البروتين الأساسي في التكاثر، تصبح قيمة مضافة للبروتين التجميعي الذي سينتج.

الاستنتاجات: لقاح البروتين المؤلف مع الحواتم التي تمت تجربتها لديهم معلمات كافية ليتم إنتاجها على نطاق معملي ولإجراء مزيد من الاختبارات.

الكلمات المفتاحية: بروتين تجميعي؛ بروتين الشوكة؛ فيروس كورونا سارس -2؛ إندونيسيا؛ اللقاح العكسي.

Abstract

Objectives: This study aimed to produce a recombinant protein vaccine candidate based on an epitope of spike protein from Indonesian SARS-CoV-2 to provide immunogenicity and protection against future infection.

Methods: A reverse vaccinology approach was used to identify potential vaccine candidates by screening the pathogen's genome through computational analyses.

Results: Epitope vaccine candidates with the amino acid sequence of FKNHTSPDV were selected. This peptide is hydrophilic, does not induce autoimmune and allergic reactions, is antigenic, is classified as a stable protein, and is predicted to be present in the cell membrane. The selected epitope sequences were inserted into the plasmid vector pcDNA3.1(+) N-GST (thrombin). Inclusion of additional features of the gene encoding glutathione-S transferase, which can increase antigen expression and solubility, and the genes encoding NSP 1−4 proteins, which are essential in replication, added value to the produced recombinant protein.

Conclusion: Recombinant protein vaccine candidates with the FKNHTSPDV epitope have parameters sufficient for production on a laboratory scale for further testing.

Keywords: Indonesia; Recombinant protein; Reverse vaccinology; SARS-CoV-2; Spike protein

* Corresponding address: Center for Development of Advance Science and Technology, University of Jember, JL. Kalimantan 37, Kampus Tegalboto, Jember 68121, Indonesia.
E-mail: erlia.fkip@unej.ac.id (E. Narulita)
Peer review under responsibility of Taibah University.

ELSEVIER **Production and hosting by Elsevier**

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes Coronavirus disease 2019 (COVID-19), has led to a global pandemic with rapid spread. Limited preventive measures in the form of vaccination are available.[1] Until August 2021, according to Worldmeters data (https://www.worldometers.info/), the total number of COVID-19 cases was 219 million, and the death toll was nearly 4 million worldwide. In Indonesia, the percentage of deaths due to infection with the SARS-CoV-2 virus is higher than that in many other countries, at approximately 9.36%.

Measures for treatment and prevention of COVID-19 to date have included the antiviral drug remdesivir, vaccines, and convalescent plasma transfusion, thus providing potential options for treating patients with severe COVID-19. However, the clinical benefits of remdesivir in patients with severe disease are limited. Recent studies have shown that the severity of COVID-19 is associated with an increase rather than a decrease in IgG response, and that convalescent plasma transfusions can be beneficial for patients only when given 14 days after the onset of the disease rather than later.[2]

The key to fighting this pandemic is understanding the viral receptor recognition mechanisms that regulate infectivity, pathogenesis, and host range. SARS-CoV-2 and SARS-CoV recognize the same receptor, ACE2, in humans. The SARS-CoV-2 spike glycoprotein determines viral binding and invasion of target cells via the ACE2 receptor. Many studies have used this glycoprotein to develop a vaccine for COVID-19, because of its crucial role and high surface exposure. Therefore, the S protein interaction region was selected for inclusion in the structure of the designed vaccine. This region has a specific and highly conserved glycosylation pattern, which makes it a suitable antibody target. Neutralizing antibodies binding this site on spike protein block viral binding and entry into host cells.[3]

In contrast, the roles of T cells in the immune response to coronavirus infection are much more important than those of B cells. More than 70% of the T cell immune response targets the structural proteins of the coronavirus. T cells are stimulated by antigen-presenting cells (APCs). When APCs phagocytose viruses, only structural proteins from the coronavirus undergo processing via the major histocompatibility complex (MHC), because of the lack of non-structural proteins (NSPs) in viral particles.[4]

In the past, vaccine development relied heavily on immunological trials, in an expensive and time-consuming process. However, recent advances in immunological bioinformatics have provided a viable tool to significantly decrease the time and costs involved, and the risk of errors, when used in laboratory settings for vaccine development. This tool can decrease the costs of discovery and simulation in testing the antigenicity of vaccine candidates in early research stages.[4]

One development associated with this framework is the production of recombinant protein vaccines through a reverse vaccinology approach to identify recombinant vaccine epitope candidates. Recombinant protein vaccines stimulate the production of antibodies that interact with antigen proteins or virus particles.[5] Recombinant protein vaccines do not replicate and do not have an infectious component of viral particles. These vaccines are thus considered to be safer than live virus vaccines. This technology has been tested extensively and found to elicit only very mild adverse effects.[6]

Materials and Methods

Identification and download of SARS-CoV-2 Indonesian isolate sample sequences

The complete genome sequences of SARS-CoV-2 Indonesian isolates, amounting to 29 sequences representing the mutations in various provinces in Indonesia, were retrieved on the date on which this project started, March 5, 2021, from the GISAID website (https://www.gisaid.org/) without filtering coverage and with download of coding nucleotide sequences. SARS-CoV-2 wild-type spike protein originating from Wuhan (reference No. NC_045512.2) was used as a reference sequence for mapping the region of the wild-type spike protein-encoding gene in the complete genome sequence of Indonesian SARS-CoV-2, which had not been annotated at that time.[4]

Alignment of Indonesian SARS-CoV-2 nucleotide sequences with the Wuhan reference sequence (wild type)

A total of 29 nucleotide sequences of the complete genome of Indonesian SARS-CoV-2 with spike protein were aligned with the wild type Wuhan sequence via the online tool Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/). The alignment was aimed at determining the positions of Indonesian SARS-CoV-2 spike protein nucleotides with reference to the Wuhan spike protein nucleotide sequence taken from the annotated complete genome.[4]

Translation of Indonesian SARS-CoV-2 spike protein nucleotide sequences

The nucleotide sequences encoding the spike protein from each of the Indonesian SARS-CoV-2 sequences obtained in the previous stage were translated. The translation was aimed at obtaining the amino acid spike sequences with respect to the annotated Wuhan wild-type spike protein amino acid reference sequence, by using the online tool ExPASy (https://web.expasy.org/translate/).[4]

Determination of the sustainability of Indonesian SARS-CoV-2 spike protein sequences

Amino acid spike sequences of the entire sample were aligned for sustainability analysis with the Antigenic Variability Analyzer (AVANA) tool. The threshold was set to 95%. This step was aimed at determining conserved and non-conserved regions. In the determination of B cell and T cell epitopes, only sequences of peptides in conserved sites were considered as vaccine candidates.

Consensus sequence construction for Indonesian SARS-CoV-2 spike protein

Consensus sequence construction was performed from amino acid sequence alignment of the Indonesian SARS-CoV-2 spike protein with the online tool EMBOSS (https://www.ebi.ac.uk/Tools/msa/embosscons/).[4]

Determination of B cell epitopes

The consensus sequence of the Indonesian SARS-CoV-2 spike protein was mapped against B cells through prediction of the linear epitopes of B cells by using the Bepipred 2.0 parameter in IEDB (http://tools.iedb.org/bcell/). Bepipred 2.0 is an immunoinformatics tool for predicting B-cell epitopes from antigen sequences.[7] This step was aimed at determining regions that B cells can potentially recognize.[4]

Determination of the epitopes of CD4 T cells and CD8 T cells

CD8 T cell epitopes was determined from the mapping of B cell epitope sequences longer than 9-mers against human leukocyte antigen (HLA) class I, by using the online tool netCTLpan 1.1 (http://www.cbs.dtu.dk/services/NetCTLpan/). Epitope sequences recognized by CD8 T cells were then mapped against HLA class II to predict which epitopes are also identified by CD4 T cells, by using the online tool netMHCII 2.3 (http://www.cbs.dtu.dk/services/NetMHCII/) with a peptide length setting of 9-mer for each web tool used. Other settings followed the default for each web tool.[4] This study mapped peptide sequences against 56 class I HLA alleles and 22 class II HLA alleles found in the Indonesian population.[8]

Similarity analysis to human peptides

The detected epitope candidates with the potential to be recognized by B cells, CD8 T cells, and CD4 T cells were then tested for suitability through comparison against human non-redundant protein sequences [taxid: 9606] by using NCBI Blastp (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The NCBI Blastp parameters were as follows: 1) 30,000 expectation value; 2) PAM30 matrix; 3) disabled low complexity filter; 4) composition-based statistics category set to "no adjustment"; 5) cutoff from 10e-4 to 10e-3. Nonameric sequences with homologous identities equal or more than seven-ninths without gaps or mismatches to human self-peptides were eliminated from the vaccine epitope candidates.[9]

Analysis of hydrophobicity properties

Physicochemical analysis was performed by determining the solubility of the epitope according to the hydrophobicity value by using the ExPASy online tool (https://web.expasy.org/protparam/). This program predicts the hydrophobicity of short amino acid sequences (above six amino acids), isoelectric point, and molecular protein weight. It also uses a grand average of hydropathicity (GRAVY) score distribution, showing the value for each type of amino

acid, with a range of -4.5 for arginine and $+4.5$ for isoleucine.[4]

Novelty analysis of the candidate vaccine epitope

The epitope recency test was performed by determining the presence of the epitopes in the epitope database in the IEDB Analysis Resources (https://www.iedb.org/).[9]

Analysis of antigenicity and allergenicity, and determination of membrane topology for the vaccine epitope candidates

Antigenicity testing of epitope vaccine candidates was conducted with the online tool VaxiJen Server 2.0 (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html) by setting the threshold value to 0.4 to improve the prediction accuracy. Simultaneously, the virus was selected as the target organism. The allergenicity analysis of the epitope candidates was conducted with the online tool AllerTOP v2.0, which has a predictive accuracy of 88.7%.[4] Prediction of the selected candidate epitope's transmembrane topology was performed with TMHMM v.2.0 (http://www.cbs.dtu.dk/services/TMHMM/). The output of this tool was in the form of a probability graph of the possible epitope presentation locations and a statement indicating one of three location choices: inside, outside, or transmembrane.[10]

Reverse translation of selected epitope sequences and codon optimization

Selected epitope sequences were back-translated to obtain nucleotide sequences in EMBOSS BackTranseq (https://www.ebi.ac.uk/Tools/st/embossbacktranseq/). The results of this back translation were then optimized with the NovoPro Labs online tool to obtain a nucleotide sequence with a codon adaptation index >0.8.

Construction of recombinant plasmid vectors

The plasmid selected as a vector was pcDNA3.1(+) N-GST (thrombin). The plasmid insert was designed in Snap-Gene software. The design result sequence was inserted via the Eco321 restriction site of the plasmid.

Results

Identification of SARS-CoV-2 sequence samples and mapping of spike protein-encoding genes

The list of accession codes of the complete genome sequences of Indonesian SARS-CoV-2 used in this study can be found in Table 1. The complete genome sequences were mapped with multiple sequence alignment methods to determine the gene sequences encoding the spike protein. These sequences were used as epitope candidates and were selected in this study on the basis of the Wuhan wild-type spike protein-coding reference sequences nucleotides (reference No. NC_045512.2).

Table 1: List of Indonesia SARS-CoV-2 complete genome sequence access codes downloaded from GISAID.

Number	Province	Accession number
1	East Java/Sidoarjo	EPI_ISL_956315
2	Banten/Tangerang	EPI_ISL_947327
3	Jakarta	EPI_ISL_953427
4	West Java	EPI_ISL_747241
5	Central Java	EPI_ISL_791988
6	Special Region of Yogyakarta	EPI_ISL_911709
7	Aceh	EPI_ISL_791981
8	Bangka Belitung Islands	EPI_ISL_747237
9	North Sumatra	EPI_ISL_756401
10	Lampung	EPI_ISL_791978
11	Riau Islands	EPI_ISL_791985
12	South Sumatra/Palembang	EPI_ISL_833039
13	West Sumatra	EPI_ISL_910014
14	Bengkulu	EPI_ISL_791979
15	Bali	EPI_ISL_775596
16	East Nusa Tenggara/Kupang	EPI_ISL_766048
17	West Nusa Tenggara	EPI_ISL_775598
18	South Kalimantan	EPI_ISL_753699
19	Central Kalimantan	EPI_ISL_538502
20	East Kalimantan	EPI_ISL_791983
21	North Kalimantan	EPI_ISL_803876
22	West Kalimantan	EPI_ISL_911750
23	North Sulawesi/Manado	EPI_ISL_574623
24	South Sulawesi/Makassar	EPI_ISL_833502
25	North Maluku	EPI_ISL_791986
26	West Papua	EPI_ISL_775597
27	Papua/Timika	EPI_ISL_574603
28	Jakarta	EPI_ISL_1118931
29	Jakarta	EPI_ISL_1118933

Translation of the nucleotide sequences encoding Indonesian SARS-CoV-2 spike protein, analysis of sequence continuity, and construction of the amino acid consensus sequence of spike protein

Nucleotide sequences identified as spike protein-encoding genes from the results of alignment to the reference sequences were translated for each Indonesian SARS-CoV-2 sample to obtain amino acid sequences of spike proteins, on the basis of the reference reading frame of the spike protein-coding sequences from the annotated complete genome of Wuhan SARS-CoV-2 wild type. Sequence sustainability analysis was conducted by using the alignment results of the amino acid spike sequences from the entire Indonesian SARS-CoV-2 sample.

The results were based on the analysis of the sustainability of known sequence regions containing amino acid residues with conserved and non-conserved regions (Figure 1). Sustainable areas of spike protein sequences are indicated in red on the chart, whereas unsustainable areas are in white. Amino acid residues in unsustainable regions were located at positions 74, 149, 249, 398, 513, 583, 51, 700–725, 775–795, 813, 838, 924, 1126, and 1298.

The epitope sequences in unsustainable areas were eliminated according to determination of B cell and T cell epitopes. The consensus sequence from Indonesian SARS-CoV-2 spike protein amino acid sequences was used in the epitope determination stage. The results of the consensus sequence of spike protein amino acids are shown in Table 2.
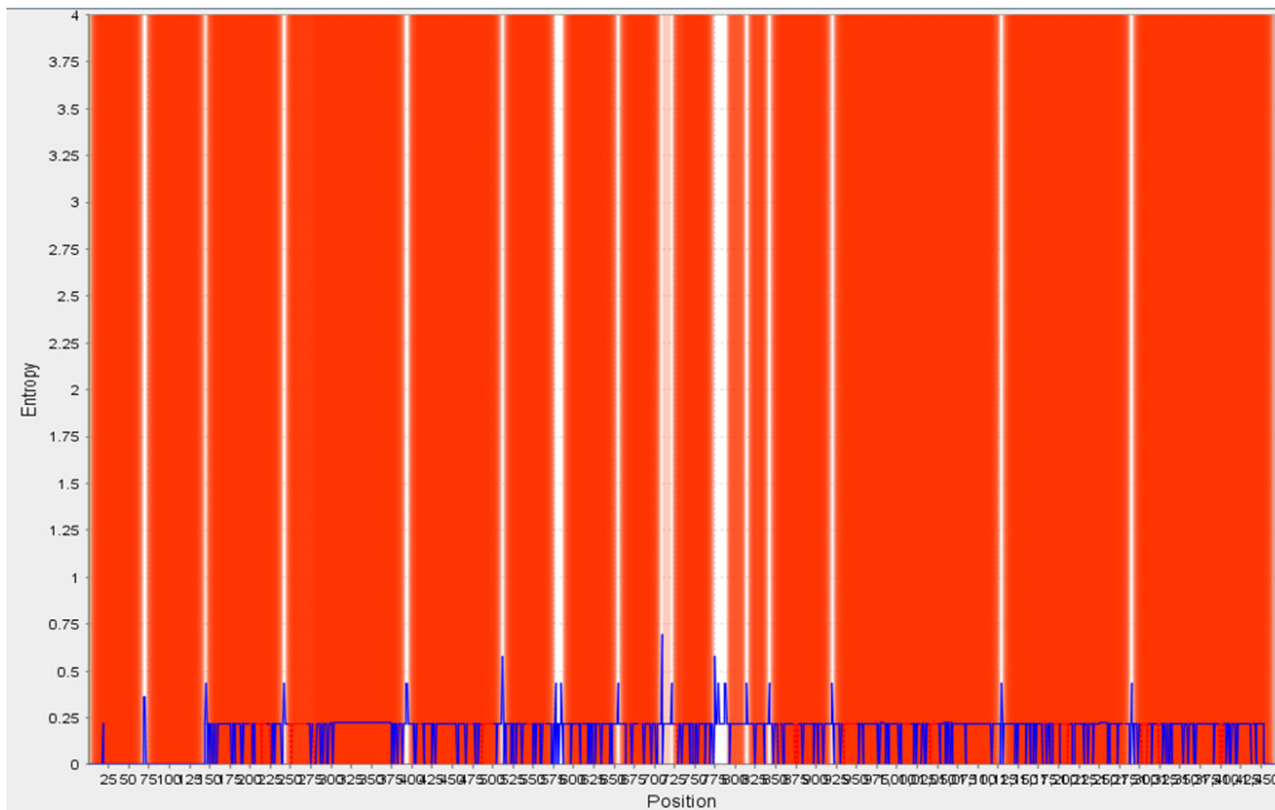

Figure 1: Determination of conserved regions of the spike protein sequence of Indonesian SARS-CoV-2. Conserved regions of spike protein sequences are marked in red, while unconserved regions are in white. Amino acid residues belonging to unconserved regions are at positions 74, 149, 249, 398, 513, 583, 51, 700-725, 775-795, 813, 838, 924, 1126, and 1298.

Table 2: Indonesia SARS-CoV-2 spike protein consensus sequence.

>Consensus
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWF
HAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDP
FLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPIN
LVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTIT
DAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRIS
NCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGC
VIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVG
YQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAV
RDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAG
CLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTT
EILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFG
GFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYT
SALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
LQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASA
NLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPRE
GVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDV
DLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCC
SCLKxGCCSCGSCCKFDEDDSEPVLKGVKLHYT

Table 3: Prediction of B cell epitope and analysis of its sustainability.

No.	Start	End	Peptide	Length	Conservancy analysis (based on AVANA)
1	13	37	SQCVNLTTRTQLPPAYTNSFTRGVY	25	Conserved region
2	59	81	FSNVTWFHAIHVSGTNGTKRFDN	23	Non-conserved region
3	97	98	KS	2	Conserved region
4	138	154	DPFLGVYYHKNNKSWME	17	Non-conserved region
5	177	189	MDLEGKQGNFKNL	13	Conserved region
6	206	221	KHTPINLVRDLPQGFS	16	Conserved region
7	250	260	TPGDSSSGWTA	11	Conserved region
8	293	296	LDPL	4	Conserved region
9	304	322	KSFTVEKGIYQTSNFRVQP	19	Conserved region
10	329	363	FPNITNLCPFGEVFNATRFASVYAWNRKRISNCVA	35	Conserved region
11	369	393	YNSASFSTFKCYGVSPTKLNDLCFT	25	Conserved region
12	404	426	GDEVRQIAPGQTGKIADYNYKLP	23	Conserved region
13	440	501	NLDSKVGGNYNYLYRLFRKSNLKP FERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTN	62	Conserved region
14	516	536	ELLHAPATVCGPKKSTNLVKN	21	Conserved region
15	555	562	SNKKFLPF	8	Conserved region
16	580	583	QTLE	4	Non-conserved region
17	602	606	TNTSN	5	Conserved region
18	617	632	CTEVPVAIHADQLTPT	16	Conserved region
19	635	643	VYSTGSNVF	9	Conserved region
20	656	666	VNNSYECDIPI	11	Conserved region
21	672	690	ASYQTQTNSPRRARSVASQ	19	Conserved region
22	695	710	YTMSLGAENSVAYSNN	16	Non-conserved region
23	748	748	E	1	Conserved region
24	773	779	EQDKNTQ	7	Non-conserved region
25	786	800	KQIYKTPPIKDFGGF	15	Non-conserved region
26	807	814	PDPSKPSK	8	Non-conserved region
27	828	842	LADAGFIKQYGDCLG	15	Non-conserved region
28	988	992	EAEVQ	5	Conserved region
29	1035	1043	GQSKRVDFC	9	Conserved region
30	1107	1118	RNFYEPQIITTD	12	Conserved region
31	1133	1172	VNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGI	40	Conserved region
32	1203	1206	LGKY	4	Conserved region
33	1252	1267	SCCKFDEDDSEPVLKG	16	Conserved region
34	1269	1269	K	1	Conserved region

Table 4: List of Class I and Class II HLA (human leukocyte antigen) are used[a].

56 HLA class I alleles possessed by the Indonesian population
HLA-A01:01,HLA-A02:01,HLA-A02:03,HLA-A02:06,HLA-A02:11,HLAA03:01,HLAA11:01, HLA-A11:04,HLA-A24:02,HLA-A24:07,HLA-A24:10,HLA-A26:01,HLAA29:01, HLA-A30:01, HLA-A32:01,HLA-A33:03,HLA-A34:01,HLAA74:01,HLA-B07:02,HLA-B07:05,HLA-B08:01, HLA-B13:01,HLA-B13:02,HLA-B15:01,HLA-B15:02,HLAB15:10,HLAB15:12, HLA-B15:13, HLA-B15:17,HLA-B15:21,HLA-B15:25,HLAB15:32,HLA-B18:01,HLA-B18:02,HLA-B27:06, HLA-B35:01,HLA-B35:02,HLA-B35:03,HLA-B35:05,HLA-B35:30,HLA-B37:01,HLA-B38:02, HLA-B39:15,HLAB40:01,HLAB40:02,HLA-B40:06,HLA-B41:01,HLA-B44:03,HLA-B48:01, HLA-B51:01,HLAB51:02,HLA-B52:01,HLA-B56:01, HLA-B56:02, HLA-B56:07, HLAB57:01, HLA-B58:01

22 HLA class II alleles possessed by the Indonesian population
DRB1_0101, DRB1_0301, DRB1_0401, DRB1_0402, DRB1_0403, DRB1_0405, DRB1_0406, DRB1_0701, DRB1_0802, DRB1_0901, DRB1_1001, DRB1_1101, DRB1_1201, DRB1_1301, DRB1_1302, DRB1_1401, DRB1_1404, DRB1_1405, DRB1_1407, DRB1_1454, DRB1_1501, DRB1_1602

[a] Reference:[8].

Determination of the epitopes of B cells, CD4 T cells, and CD8 T cells

The full-length spike protein consensus sequence based on 29 Indonesian SARS-CoV-2 sequences was mapped against B cell epitopes to determine the locations of linear epitopes recognized by B cells, on the basis of the annotated epitope database and Bepipred 2.0 parameters in the IEDB web-server. In this stage, a total of 34 peptide sequences recognized by B cells were obtained, with lengths varying from 1-mer to 62-mer (Table 3). After selection according to the sustainability analysis results, eight peptide sequences containing amino acid residues in non-conserved regions were obtained. In addition, a peptide sequence with at least 9-mer amino acids was required to proceed to the mapping stage against HLA class I and class II. Only 18 peptide sequences of B-cell epitopes with lengths varying from 9-mer to 62-mer, located in a conserved region, and mapping against the HLA I and HLA II alleles were found in the Indonesian population. Mapping of B cell epitope sequences against HLA class I and class II was performed to determine the epitopes recognized by CD4 T cells and CD8 T cells. The list of HLA class I and class II used is provided in Table 4.

On the basis of the mapping of class I HLA alleles in the Indonesian population, 63 epitope sequences had the

Table 5: Similarities to selfpeptide, hydrophobicity characteristics, and novelty of candidate epitopes.

Epitope candidates	Analysis of similarity to self-peptide	Hydrophobicity test		Epitope recency test	Selected epitopes (go to the next step)
		GRAVY score	Meaning		
VNLTTRTQL	No similarities in 7/9 or more were found without gaps	−0.200	Hydrophilic	New epitope (database no reference)	VNLTTRTQL
LTTRTQLPP	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
TRTQLPPAY	No similarities in 7/9 or more were found without gaps	−0.922	Hydrophilic	New epitope (2021 reference database-specifically SARS-CoV-2)	TRTQLPPAY
LVRDLPQGF	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
FTVEKGIYQ	No similarities in 7/9 or more were found without gaps	−0.200	Hydrophilic	New epitope (database no reference)	FTVEKGIYQ
GIYQTSNFR	No similarities in 7/9 or more were found without gaps	−0,822	Hydrophilic	New epitope (database no reference)	GIYQTSNFR
FASVYAWNR	No similarities in 7/9 or more were found without gaps	−0.044	Hydrophilic	New epitope (2020-specific SARS-CoV-2 reference database)	FASVYAWNR
FNATRFASV	No similarities in 7/9 or more were found without gaps	0.433	Hydrophobic	Not tested because it is hydrophobic	Eliminated
VFNATRFAS	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
TRFASVYAW	No similarities in 7/9 or more were found without gaps	0.267	Hydrophobic	Not tested because it is hydrophobic	Eliminated
YAWNRKRIS	No similarities in 7/9 or more were found without gaps	−1456	Hydrophilic	New epitope (database no reference)	YAWNRKRIS

Table 5 (*continued*)

Epitope candidates	Analysis of similarity to self-peptide	Hydrophobicity test		Epitope recency test	Selected epitopes (go to the next step)
		GRAVY score	Meaning		
VYAWNRKRI	No similarities in 7/9 or more were found without gaps	−0,900	Hydrophilic	New epitope (database no reference)	VYAWNRKRI
ATRFASVYA	No similarities in 7/9 or more were found without gaps	0.567	Hydrophobic	Not tested because it is hydrophobic	Eliminated
EVFNATRFA	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
FKCYGVSPT	No similarities in 7/9 or more were found without gaps	0.089	Hydrophobic	Not tested because it is hydrophobic	Eliminated
CYGVSPTKL	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
ASFSTFKCY	No similarities in 7/9 or more were found without gaps	0.267	Hydrophobic	Not tested because it is hydrophobic	Eliminated
STFKCYGVS	No similarities in 7/9 or more were found without gaps	0.178	Hydrophobic	Not tested because it is hydrophobic	Eliminated
FERDISTEI	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
LYRLFRKSN	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
IYQAGSTPC	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
CYFPLQSYG	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
YFPLQSYGF	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
YRLFRKSNL	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
FRKSNLKPF	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
YNYLYRLFR	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
YQAGSTPCN	No similarities in 7/9 or more were found without gaps	−0,833	Hydrophilic	New epitope (database no reference)	YQAGSTPCN
FNCYFPLQS	No similarities in 7/9 or more were found without gaps	0.133	Hydrophobic	Not tested because it is hydrophobic	Eliminated
DISTEIYQA	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
EIYQAGSTP	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
LLHAPATVC	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
TQTNSPRRA	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
QTNSPRRAR	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated
FKNHTSPDV	No similarities in 7/9 or more were found without gaps	−1133	Hydrophilic	New epitope (database no reference)	FKNHTSPDV
ELDSFKEEL	Highest similarity: 7/9 without gap	The sequences were eliminated because of their high similarity to self-peptide			Eliminated

Table 6: The antigenicity, allergenicity, and topology of the epitope candidate membrane.

No.	Epitope candidates	Antigenicity test		Allergenicity test	Topology	Potential epitope candidates
		Score	Meaning			
1	TRTQLPPAY	1.2923	Antigenic	Non-allergen	Inside	TRTQLPPAY
2	VNLTTRTQL	1.3468	Antigenic	Allergens	Inside	Eliminated
3	GIYQTSNFR	0.5380	Antigenic	Allergens	Inside	Eliminated
4	FTVEKGIYQ	−0.1987	Non-antigenic	Allergens	Inside	Eliminated
5	FASVYAWNR	0.0713	Non-antigenic	Allergens	Inside	Eliminated
6	YAWNRKRIS	0.8209	Antigenic	Non-allergen	Inside	YAWNRKRIS
7	VYAWNRKRI	0.5003	Antigenic	Allergens	Inside	Eliminated
8	FKNHTSPDV	0.4846	Antigenic	Non-allergen	Outside	FKNHTSPDV
9	YQAGSTPCN	0.4992	Antigenic	Allergens	Inside	Eliminated

Table 7: Characteristics of selected epitope candidates.

Characteristics	Information	Meaning
Similarity to self-peptide	No similarities in 7/9 or more were found without gaps	Not similar to self-peptide
Hydrophobicity	GRAVY Score: −1.133	Hydrophilic (dissolved)
Molecular mass	1044.13	
tpI	6.74	
Stability	Stable protein	
Recency (based on IEDB database)	The research has never been done and was not found in the database	New (novel)
Antigenicity	0.4846	Antigenic
Allergenicity	Non-allergen	
Topology	Outside	

Table 8: Selected epitope coding nucleotide sequences and codon optimization results.

Selected epitope (amino acid sequence)	Reverse translated sequence (nucleotide sequences)	Optimized sequences	Characteristics	
			CAI	GC content (%)
FKNHTSPDV	TTCAAGAACCACACCA GCCCCGACGTG	TTCAAAAACC ACACTTCTCC GGACGTA	0.87	44.44

potential to bind class I HLA and could potentially be CD8 T cell epitopes. The CD8 T cell epitope sequences were then mapped against HLA class II to determine the potential for introducing the same epitope sequence into CD4 T cells. According to the mapping results against HLA class II, the core peptides from epitope sequences predicted to be recognized by CD8 T cells but containing non-specific amino acid profiles (marked with the symbol X) were eliminated as vaccine candidates. After selection according to HLA class II mapping results, 35 CD4 T cell epitope sequences were obtained. The determination of CD4 T cell epitopes and CD8 T cells on the basis of the recognized list of HLA alleles can be found in Supplementary 1.

Analysis of similarity to the human peptide, hydrophobicity characteristics, and novelty of the candidate vaccine epitope

The results of the epitope tests based on the three analysis parameters are presented in Table 5. According to the results of 35 B-cell and T-cell epitope sequences, 19 homologous identity sequences have no any gaps or mismatches with human peptides or self-peptides. At this stage, 16 epitope candidates were selected, which were considered to have low similarity to human peptides. Determination of the solubility or hydrophobicity properties of the epitope sequence was based on the scores resulting from the calculation of the average GRAVY score in the ExPASy tool. A result above or below zero indicates that a protein is hydrophobic or hydrophilic, respectively.[10]

On the basis of the results of the hydrophobicity analysis, nine epitope candidates were found to be hydrophilic. Epitope candidates that were hydrophobic were eliminated, because they are not considered ideal vaccine candidates. On the basis of epitope recency analysis performed on March 6, 2021, nine epitope candidates passing the hydrophobicity analysis were classified as new epitopes (not found in the database).

Analysis of antigenicity and allergenicity, and determination of membrane topology for the vaccine epitope candidate

On the basis of the analysis of the antigenicity of the nine candidates that passed the epitope vaccine parameters, seven epitopes had antigenicity scores above the threshold value set
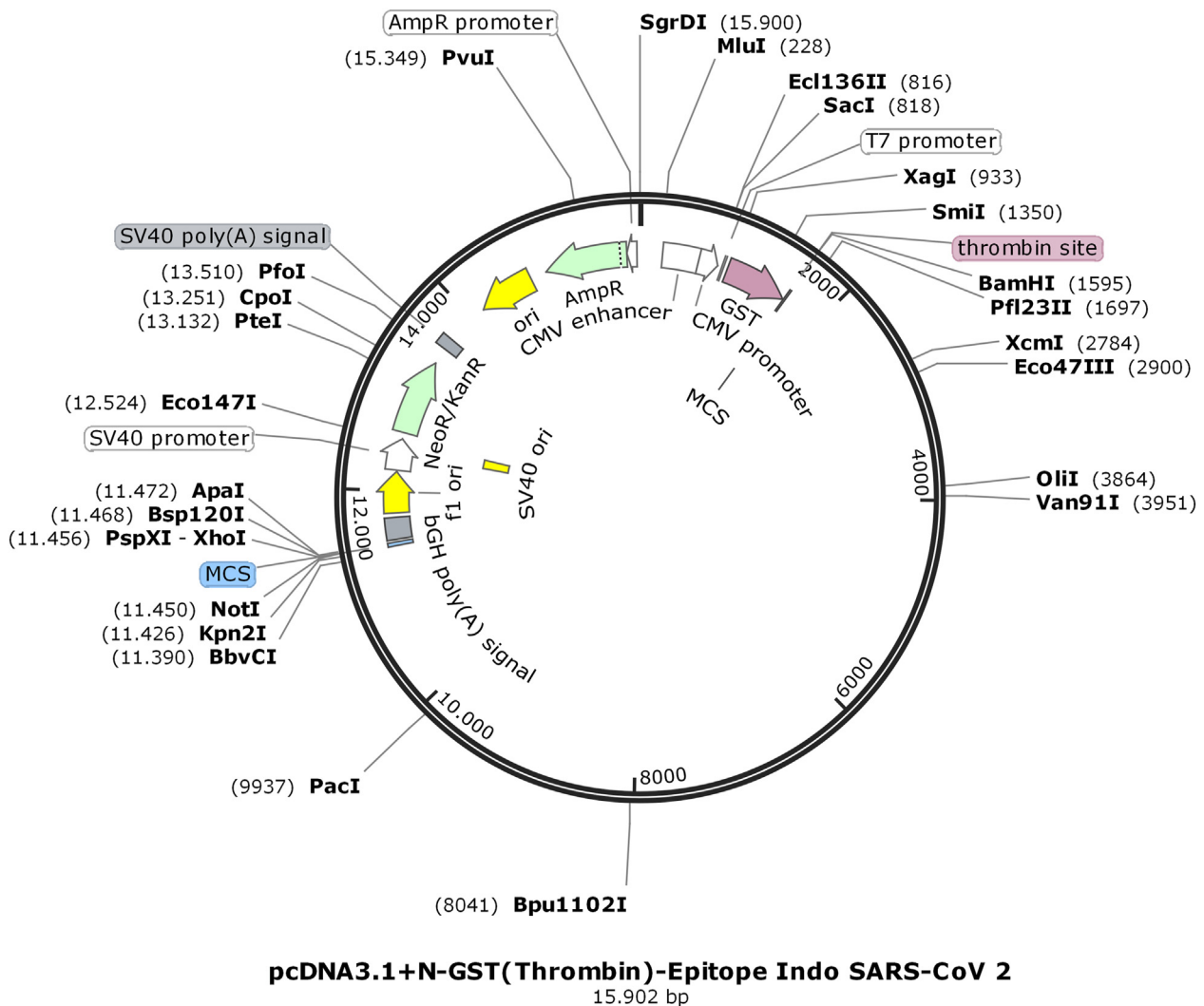
pcDNA3.1+N-GST(Thrombin)-Epitope Indo SARS-CoV 2
15.902 bp

Figure 2: Construction results for Plasmid pcDNA3.1(+) N-GST (Thrombin)-Epitope of Indonesian SARS-CoV-2. The red part of the plasmid construct is the part designed in this study. The inserted gene was added with ATG as a start codon that encoded the amino acid methionine.

(>0.4). This result indicated that these epitopes are antigenic (Table 6), whereas the remaining epitopes are non-antigenic because they had scores below the threshold. The results of allergenicity analysis indicated that six epitopes were predicted to trigger allergic reactions (i.e., to be allergens), whereas the remaining epitopes were non-allergenic. According to the results of these two analyses, three potential epitope candidates were antigenic and non-allergenic. In addition, the prediction of the epitope candidate transmembrane topology results reported an inside location, meaning that the epitope tends to be present on the inner surface of the membrane.[11]

Reverse translation of selected epitope sequences and codon optimization

The best epitope was selected on the basis of the parameters of the previous analysis. The selected epitope had an amino acid sequence of FKNHTSPDV, which had relatively high solubility according to the GRAVY value. Thus, the

FKNHTSPDV sequence was chosen as the best epitope to potentially serve as a vaccine candidate in this study. The characteristics of the selected epitope candidates from the various stages of analysis performed are summarized in Table 7. The selected epitope sequences were then back translated and optimized to obtain a nucleotide sequence encoding the epitope with codons that could be optimally expressed in the human body. The results of nucleotide sequence optimization are shown in Table 8.

Construction of recombinant pcDNA3.1(+) N-GST (thrombin) plasmid

The results of the plasmid construction are shown in Figure 2. The red part of the plasmid construct is the part designed in this study. The inserted gene was added with ATG as a start codon, encoding the amino acid methionine. The addition of features to vaccine constructs, such as the gene encoding the enzyme glutathione S transferase (GST), which can increase the expression and

solubility of antigens,[12] and genes encoding proteins NSP 1–4, which are essential in replication,[13,14] conferred added value on the recombinant protein.

Discussion

The development of the SARS-CoV-2 vaccine in this study was based on a recombinant protein. Recombinant protein vaccines stimulate the production of antibodies that interact with antigen proteins or viral particles.[5] The process of identifying and initially testing vaccine candidates can start from in silico studies using bioinformatics tools. The use of in silico studies beforehand can decrease the likelihood of failure and the losses incurred.[15]

This research began with downloading of the complete genome sequences of SARS-CoV-2 in the GISAID database. A total of 29 sequences from provinces in Indonesia were used, which were considered to generally represent a complete genome sample of Indonesian SARS-CoV-2, because of their high similarity to one another. The proofreading activity, which is regulated by NSP14 3′-to-5′ exoribonuclease (NSP14-ExoN) in the SARS-CoV-2 genome, is a key determinant of both coronavirus replication and recombination. Thus, the genome sequence of viruses belonging to this group are conserved regions.[16] Sustainability analysis has indicated a high similarity among SARS-CoV-2 sequences reaching >99%.[17]

The target protein selected in this study was the spike protein or surface glycoprotein. On the basis of previous research, the spike protein shows a relatively high antigenicity value and has the potential to be a candidate for vaccine epitopes that can induce an excellent immune response.[18] The SARS-CoV-2 spike protein is also conserved, as compared with the spike protein in another human coronavirus (hCoV).[19] This conserved area has been predicted not to undergo mutation, and to be responsible for a particular function or provide a necessary structural characteristic.[20] Determination of conserved regions aims to ensure that the epitopes used as candidate vaccines show promise for efficacy and coverage of groups or clusters.[21]

Determination of a SARS-CoV-2 Indonesian spike protein peptide sequence able to bind B cell receptors or immunoglobulin was performed to obtain a specific B cell epitope to induce the activation and function of B cells. If this epitope is recognized, it should cause B cells to differentiate and proliferate into plasma B cells, which produce antibodies that directly attack the virus, and can become memory B cells, which support a faster and more effective immune response to infection.[22]

On the basis of the results of sequence conservation analysis, we eliminated the predicted peptide sequences that could be recognized by B cells (Table 3), contained non-conserved amino acid residues (non-conserved), and were less than 9-mers in length. To determine that the T cell epitopes were in a conserved region, we ensured that the peptide sequences mapped against HLA class I and class II. MHC class I is an APC ligand recognized by CD8 T cells. The ligand binds antigen, then presents it to be recognized by CD8 T cells as an antigenic determinant (epitope) with activity toward infected cells.[23] A series of analyses of epitopes that are potentially recognized by B cells, CD4 T cells, and CD8 T cells were performed to select the best epitope as a vaccine candidate before moving to the next stage.

Analysis of the similarity of epitopes to human peptides (self-peptides) was conducted to eliminate epitopes with the potential to induce autoimmune reactions. This analysis was fundamental to ensuring that the selected epitope did not have high similarity to human peptides, which could elicit antibodies or other immune responses against the host peptide itself. Hence, peptides with high similarity were eliminated.[9] The novelty aspect was assessed by cross-checking the epitope candidates against the IEDB database. Epitopes that had never been reported or studied before were selected as vaccine candidates in this study. Epitope solubility analysis based on the hydrophobicity value (GRAVY score) was performed to determine the epitope with the best solubility (hydrophilicity), to enable its interaction with the immune system in aqueous media.[24] The results of the analyses of similarity, recency, and hydrophobicity are presented in Table 1.

Antigenicity analysis is used to determine the ability of epitopes to be recognized as antigens by adaptive immune responses, particularly to stimulate the responses of B cells and T cells. Epitopes with good antigenicity can induce an adaptive immune response resulting in the production of memory cells, which "remember" viral antigens. Memory cells result in a faster and more effective immune response when an infection occurs.[21] Seven epitopes were found to have antigenicity scores exceeding the threshold value set (>0.4). Allergenicity analysis was used to eliminate epitopes that are potential allergens and might induce allergic reactions in the body. Epitopes used as vaccine candidates should not cause allergic reactions that are harmful to the body.[17] Thus, only epitope candidates with high antigenicity and were not allergenic were selected as vaccine epitope candidates. After the chosen vaccine epitope candidate was translated into the nucleotide sequence encoding the epitope, codon optimization was performed to ensure and increase the likelihood of the protein being expressed in humans.[25]

The selected epitope had an amino acid sequence of FKNHTSPDV, which has relatively high solubility, on the basis of the GRAVY value. Thus, this sequence was chosen as the best epitope that could potentially serve as a vaccine candidate in this study. This epitope has also been identified in a previous study,[26] which used a different sequence.

The plasmid pcDNA3.1(+) N-GST (thrombin) was used as a vector in this study, and the target gene was inserted via the Eco321 restriction site. Escherichia coli was the primary choice for producing recombinant protein in this study, because of its low cost and ease of culture, and the availability of accessible related technology.[2,27,28] Various recombinant proteins from bacteria, archaebacteria, and eukaryotes can be produced efficiently in E. coli.[29]

E. coli do not have disulfide isomerase protein; thus, the recombinant protein expressed in E. coli bacteria cannot fold completely, thus resulting in low solubility and activity of the protein produced. A strategy widely used to overcome this problem involves fusion of the protein with GST.[11] GST fusion also simplifies the protein purification process. Moreover, the addition of the genes encoding NSP 1–4, which are essential proteins in replication,[11,12] added value to the recombinant protein produced.

Conclusions

On the basis of the search results for an Indonesian SARS-CoV-2 spike protein epitope that can be recognized by selected B cells and T cells, we identified an epitope with the amino acid sequence FKNHTSPDV, which is hydrophilic, does not have the potential to induce autoimmune and allergic reactions, is antigenic, is classified as a stable protein, and is predicted to be present outside the cell membrane. The selected epitope sequence was inserted into the plasmid vector pcDNA3.1(+) N-GST (thrombin) with the addition of a GST sequence to increase the solubility and activity of the protein produced and the genes encoding NSP 1−4, which are essential in replication.

Source of funding

Conflict of interest

The authors have no conflict of interest to declare.

Ethical approval

This article does not contain any studies involving animals or human participants performed by any of the authors.

Recommendation

SARS-CoV-2 mutations can increase the number of COVID-19 cases, and some mutations can decrease vaccine effectiveness. Recent advances in the field of immunological bioinformatics have provided a viable vaccine development tool to significantly decrease the time, costs, and risk of trial error.

Authors contributions

RA and EN conceived and designed the study. RA conducted research, provided research materials, collected and organized data, and analyzed and interpreted data. EN supervised the research. RA and EN wrote the initial and final drafts of the article. All authors have critically reviewed and approved the final draft and are responsible for the content and similarity index of the manuscript.

References

1. Hubber JE, Ahlfeld J, Scheck MK, Zaucha M, Witter K, Lehmann L, et al. Dynamic changes in circulating T follicular helper cell composition predict neutralising antibody responses after yellow fever vaccination. **Clin Transl Immunol 2020**: 112−119.
2. Zheng Y, Renfeng L, Shunai L. Immunoregulation with mTOR inhibitors to prevent COVID-19 severity: a novel intervention strategy beyond vaccines and specific antiviral medicines. **J Med Virol 2020**: 1495−1500.
3. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. **Nature 2020**: 581.
4. Safavi A, Kefayat A, Mahdevar E, Abiri A, Ghahremani F. Exploring the out of sight antigens of SARS-CoV-2 to design A candidate multi-epitope vaccine by utilizing immunoinformatics approaches. **Vaccine 2020**; 38: 7612−7628.
5. Taboga O, Tami C, Carrillo E, Nunez JI, Rodriguez A, Saiz JC, et al. A large-scale evaluation of peptide vaccines against foot-and-mouth disease: lack of solid protection in cattle and isolation of escape mutants. **J Virol 1997**; 71: 2606−2614.
6. Pollet J, Chen W, Strych U. Recombinant protein vaccines, A proven approach against coronavirus pandemics. **Adv Drug Deliv Rev 2021**; 170: 71−82.
7. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. **Nucleic Acids Res 2017**: 1−6.
8. Gustiananda M, Sulistyo BP, Agustriawan D, Andarini S. Immunoinformatics analysis of SARS-CoV-2 ORF1ab polyproteins to identify promiscuous and highly conserved T-cell epitopes to formulate vaccine for Indonesia and the world population. **Vaccines 2021**; 9: 1459.
9. Lon JR, Bai Y, Zhong B, Chai F, Du H. Prediction and evolution of B cell epitopes of surface protein in SARS-CoV-2. **Virol J 2020**; 17(165): 1−9.
10. Huang H, Chen W, Wu J. Total protein extraction for metaproteomics analysis of methane producing biofilm: the effects of detergents. **Int J Mol Sci 2014**; 15: 10169−10184.
11. Möller S, Croning MD, Apweiler R. Evaluation of methods for the prediction of membrane-spanning regions. **Bioinformatics 2001**; 17: 646−653.
12. Koschorreck M, Fischer M, Barth S, Pleisse J. How to find soluble proteins: a comprehensive analysis of alfa/beta hydrolases for recombinant expression in *E. coli*. **BMC Genom 2005**; 6: 1−10.
13. Kowalski PS, Rudra A, Miao L, Anderson DG. Delivering the messenger: advances in technologies for therapeutic mRNA delivery. **Mol Ther 2019**; 27(4): 710−728.
14. Saxena SK, Kumar S, Baxi P, Srivastava N, Puri B, Ratho RK. Chasing COVID-19 through SARS-CoV-2 spike glycoprotein. **Virus Dis 2020**; 31(4): 399−407.
15. Ullah A, Sarkar B, Islam SS. Exploiting the reverse vaccinology approach to design novel subunit vaccine against ebola virus. **Immunobiology 2020**: 1519−1549.
16. Gribble J, Stevens LJ, Agostini ML, Anderson-Daniels J, Chappell JD, Lu X, et al. The coronavirus proofreading exoribonuclease mediates extensive viral recombination. **Plos Pathog 2021**; 17(1): 1−28.
17. Ceraolo C, Giorgi FM. Genomic variance of the 2019-nCoV coronavirus. **J Med Virol 2020**; 92(5): 522−528.
18. Sarkar B, Ullah MA, Johora FT, Taniya MA, Araf Y. Immunoinformatics guided designing of epitope-based subunit vaccine against the SARS coronavirus-2 (SARS-CoV-2). **Immunobiology 2020**: 1519−1555.
19. Huang Y, Yang C, Xu X, Xu W, Liu S. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. **Acta Pharmacol Sin 2020**: 1−9.
20. Nagar A, Hahsler M. Fast discovery and visualization of conserved regions in DNA sequences using quasi-alignment. **BMC Bioinf 2013**; 14(11): 1−12.
21. Bauer A, Podola L, Mann P, Missanga M, Haule A, Sudi L, et al. Preferential targeting of conserved gag regions after vaccination with A heterologous DNA prime-modified vaccinia virus ankara boost HIV-1 vaccine regimen. **J Virol 2017**; 91(18).

22. Murphy K, Weaver C. *Janeway's immunobiology*. Garland Science; 2016.
23. McDowall J, editor. *T cell receptors*; 2017 [monograph on the internet].[cited 2021 May 30]. Available at: https://www.ebi.ac.uk/interpro/potm/2005_3/Page1.htm.
24. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein identification and analysis tools on the ExPASy server. In: *The proteomics protocols handbook*. Totowa (NJ): Humana Press; 2005. pp. 571–607.
25. Mauro VP, Chappell SA. A critical analysis of codon optimization in human therapeutics. **Trends Mol Med 2014**; 20(11): 604–613.
26. Oyarzun P, Kashyap M, Fica V, Salas-Burgos A, Gonzales-Galarza FF, McCabe A, et al. A proteome-wide immunoinformatics tool to accelerate T-cell epitope discovery and vaccine design in the context of emerging infectious diseases: an ethnicity-oriented approach. **Front Immunol 2021**; 12: 598778.
27. Hu WG, Wei J, Yang X, Xia H, Li Z, Chang Z. Expression of overlapping pre-S1 fragment recombinant proteins for the determination of immunogenic domains in HBsAg PreS1 region. **Acta Biochim Biophys Sin 2004**; 36(6): 397–404.
28. Kristensen J, Petersen HUS, Mortensen KK, Sorensen HP. Generation of monoclonal antibodies for the assessment of protein purification by recombinant ribosomal coupling. **Int J Biol Macromol 2005**; 37: 212–217.
29. Lombardi A, Sperandei M, Cantale C, Giacomini P, Galeffi P. Functional expression of A single-chain antibody specific or the HER2 human oncogene in A bacterial reducing environment. **Protein Expr Purif 2005**; 44: 10–15.