PATHOGENS AND GLOBAL HEALTH
2020, VOL. 114, NO. 8, 463–470
https://doi.org/10.1080/20477724.2020.1838190

Taylor & Francis
Taylor & Francis Group

Immunodominant regions prediction of nucleocapsid protein for SARS-CoV-2 early diagnosis: a bioinformatics and immunoinformatics study

Yufeng Dai[a*], Hongzhi Chen[b*], Siqi Zhuang[a], Xiaojing Feng[a], Yiyuan Fang[a], Haoneng Tang[a], Ruchun Dai[c], Lingli Tang[a], Jun Liu[d], Tianmin Ma[e] and Guangming Zhong[f]

[a]Department of Laboratory Medicine, the Second Xiangya Hospital, Central South University, Changsha, Hunan, China; [b]National Clinical Research Center for Metabolic Diseases, Key Laboratory of Diabetes Immunology, Ministry of Education, Metabolic Syndrome Research Center, and Department of Metabolism and Endocrinology, the Second Xiangya Hospital, Central South University, Changsha, Hunan, China; [c]National Clinical Research Center for Metabolic Diseases, Hunan Provincial Key Laboratory for Metabolic Bone Diseases, Department of Metabolism and Endocrinology, the Second Xiangya Hospital, Central South University, Changsha, Hunan, China; [d]Department of Radiology, The Second Xiangya Hospital, Central South University, Changsha, Hunan, 410011, China; [e]Asian International Collaboration, Waitemata District Health Board, New Zealand, Level 1, Auckland, 15 Shea Terrace, 0622, New Zealand; [f]Department of Microbiology and Immunology, University of Texas Health Science Center at San Antonio, San Antonio,TX, 7703 Floyd Curl Drive, 78229, USA

ABSTRACT

COVID-19 caused by SARS-CoV-2 is sweeping the world and posing serious health problems. Rapid and accurate detection along with timely isolation is the key to control the epidemic. Nucleic acid test and antibody-detection have been applied in the diagnosis of COVID-19, while both have their limitations. Comparatively, direct detection of viral antigens in clinical specimens is highly valuable for the early diagnosis of SARS-CoV-2. The nucleocapsid (N) protein is one of the predominantly expressed proteins with high immunogenicity during the early stages of infection. Here, we applied multiple bioinformatics servers to forecast the potential immunodominant regions derived from the N protein of SARS-CoV-2. Since the high homology of N protein between SARS-CoV-2 and SARS-CoV, we attempted to leverage existing SARS-CoV immunological studies to develop SARS-CoV-2 diagnostic antibodies. Finally, $N_{229\text{-}269}$, $N_{349\text{-}399}$, and $N_{405\text{-}419}$ were predicted to be the potential immunodominant regions, which contain both predicted linear B-cell epitopes and murine MHC class II binding epitopes. These three regions exhibited good surface accessibility and hydrophilicity. All were forecasted to be non-allergen and non-toxic. The final construct was built based on the bioinformatics analysis, which could help to develop an antigen-capture system for the early diagnosis of SARS-CoV-2.

KEYWORDS
SARS-CoV-2; COVID-19; linear B-cell epitopes; nucleocapsid protein; antigen-capture diagnostics

1. Introduction

Since the outbreak of coronavirus disease 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), 33,249,563 cases of infection and 1,000,040 COVID-19 related deaths have been reported as of 29 September 2020 (https://www.who.int/). A large portion of infected people are asymptomatic, or with only mild symptoms, which pose a major challenge in controlling the COVID-19 pandemic since some asymptomatic patients are still contagious [1]. Thus, early diagnosis and screening of COVID-19 in the large population is urgently needed to contain the pandemic.

RT-PCR (short for real-time reverse transcriptase-polymerase chain reaction) is the primary method for the detection of SARS-CoV-2 as well as other respiratory viruses [2,3]. However, RT-PCR requires time-consuming and labor-intensive RNA preparation and professional operation, which increases the difficulty of on-site detection. Antibody testing for SARS-CoV-2 is another option to screen infected patients in the high prevalence areas. As it takes time for hosts to generate antibodies against viruses, antibody detection is suitable for population immunity investigation, but not for early diagnosis [4]. Hence, developing an appropriate antibody for viral antigen detection is highly valuable for the pandemic containment.

Structural proteins have been regarded as important targets for antigen detection, such as nucleoprotein of influenza virus, p24 antigen of human immunodeficiency virus (HIV), VP6 of Rotavirus (RV), etc. [5–7]. The coronavirus nucleocapsid (N) protein is a structural protein that plays a critical role in viral RNA replication [8]. According to the studies of severe acute respiratory syndrome coronavirus (SARS-CoV), high levels of circulating N protein in the serum of SARS-CoV patients could be caught as early as clinical symptoms appeared [9]. A comparison of detecting SARS-CoV RNA, specific IgG, and N protein during the early

CONTACT Lingli Tang ✉ linglitang@csu.edu.cn Department of Laboratory Medicine, the Second Xiangya Hospital, Central South University, Changsha, Hunan 410011, China

*Yufeng Dai and Hongzhi Chen contributed equally to this study.

Supplemental data for this article can be accessed here.

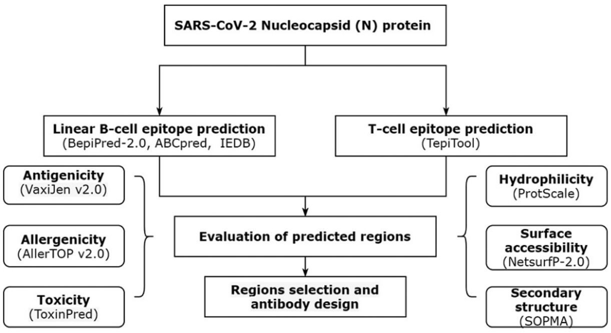
SARS-CoV-2 Nucleocapsid (N) protein
Linear B-cell epitope prediction (BepiPred-2.0, ABCpred, IEDB)
T-cell epitope prediction (TepiTool)
Antigenicity (VaxiJen v2.0)
Allergenicity (AllerTOP v2.0)
Toxicity (ToxinPred)
Evaluation of predicted regions
Regions selection and antibody design
Hydrophilicity (ProtScale)
Surface accessibility (NetsurfP-2.0)
Secondary structure (SOPMA)

Figure 1. Study workflow. First, three linear B-cell epitope prediction tools and one T-cell epitope prediction method were selected to forecast the potential epitopes of SARS-CoV -2 nucleocapsid (N) protein. Second, six bioinformatics tools were applied to evaluate the important characteristics of the predicted regions. Third, potential immunodominant regions were selected for antibody design.

period of illness showed that the detection capability of N protein was notably higher than the other two indicators [10]. These evidences suggested that the N protein could be an appropriate candidate for the early diagnostic testing and screening of SARS-CoV-2.

Bioinformatics is a scientific field combining biology, computing, and information technology. It organizes plentiful biological information to systematically and accurately interpret the information from genome transcriptome and proteome. It is extensively applied in immunodiagnostics, immunotherapeutics, and vaccine design [11–13]. Moreover, bioinformatics was used to identify the epitopes of SARS-CoV for raising neutralizing antibodies and diagnostic antibodies in previous studies [14,15]. Given the scarcity of biological data on the antigenic epitopes of SARS-CoV-2, bioinformatics is crucial in the early stages of exploring epitope information.

Therefore, we utilized a bioinformatics and immunoinformatics approach to comprehensively deduce the potential immunodominant regions on the N protein of SARS-CoV-2. The complete study workflow is presented in Figure 1. Our study could provide an important complementary strategy in the development of early diagnostic systems to combat the current pandemics.

2. Methods

2.1. Data retrieval and sequence alignment

N protein sequences of HCoV-NL63, HCoV-229E, HCoV-HKU1, HCoV-OC43, MERS-CoV, SARS-CoV, and SARS-CoV-2 were downloaded in FASTA format from NCBI database. Sequence alignment of N protein in these seven coronaviruses was performed on EMBL-EBI server Clustal Omega. Clustal Omega exploits seeded guide trees along with HMM profile-profile techniques to produce alignments between multiple sequences [16]. The phylogenetic analysis was also executed to find evolutionary ties among those seven coronaviruses, and the branch length represented the evolutionary distances between two nodes [17].

2.2. Linear B-cell epitope prediction

ABCpred, BepiPred-2.0, and Antibody Epitope Prediction online server in the Immune Epitope Database (IEDB) were adopted for linear B-cell epitope prediction. ABCpred server predicts the peptides according to the scores that acquired by the trained recurrent neural network, the higher the peptide score, the higher the prediction accuracy [18]. In our study, the cutoff of ≥0.80 (corresponding to 95.50% specificity) and the length of amino acids of 16 (default window length) of ABCpred server was employed [18]. BepiPred-2.0 server originates a random forest algorithm, which is derived from peptides annotated by antibody-antigen constructions, the residues with scores higher than the threshold are forecasted to be the segment of an epitope. We used a threshold value of ≥0.55 to achieve a specificity of 81.66% for epitope prediction [19]. IEDB is a depository of information associated with epitopes, which provides bioinformatics implements combined with algorithms [20]. The Antibody Epitope Prediction servers were accessible on the B-cell prediction tool page in IEDB, and the threshold was set at 0.35 (default threshold) [20].

2.3. Murine T-cell epitope prediction

To forecast murine T-cell epitopes, we utilized the TepiTool resource in IEDB, which employs SMM, ANN, and combinatorial library methods [21]. Here, we set the method as 'IEDB recommended'. For MHC (Major Histocompatibility Complex) class I binding prediction, the selected model exploits the consensus method comprising of CombLib, ANN, and SMM [22]. In this study, we set predicted consensus percentile rank ≤1 and the length of amino acids to 9. For MHC class II binding prediction, the selected model exploits the consensus method embracing of Sturniolo/ Combinatorial Library, NN_align, and SMM_align [23]. The predicted consensus percentile rank ≤10 and 15 residues in length were set.

2.4. Profiling and evaluation of predicted epitopes

The selected epitopes were submitted to the VaxiJen v2.0 with the given threshold of 0.40 (corresponding to 70% accuracy) for assessing the antigenic propensity [24]. VaxiJen is an alignment-free method for antigen prediction, it depends on auto cross-covariance (ACC) transformation of protein sequences into uniform vectors of principal amino acid properties, the prediction accuracy is between 70% and 89%. The higher the score, the higher the likelihood to induce immune response [24]. The hydrophilicity was evaluated with ProtScale. We chose the Kyte & Doolittle model as the amino acid scale. This program uses the approach of moving-segment to continuously determine the average hydrophilicity within a segment of a predetermined length in the process of sequence advance [25]. Surface accessibility of predicted peptides was evaluated with the NetsurfP-2.0, which uses an architecture comprised of convolutional along with long short-term memory neural networks trained on solved protein structures to forecast the relative exposure of amino acids [26]. The threshold was set to 25% exposure, but we filtered the regions where RSA (Relative Surface Accessibility) ≥50%. The secondary structure was analyzed by SOPMA. SOPMA has a success rate of 73.2% for a three-state (a-helix, β-sheet, and aperiodic states) description of secondary structure [27]. Toxicity was appraised via the ToxinPred server, which is established on the machine learning technique and quantitative matrix using numerous characters of fragments for forecasting the toxicity. The precision of the dipeptide-based model is 94.50% [28]. Allergenicity was assessed via the AllerTOP v2.0 server, it employs amino acid E-descriptors, auto- and cross-covariance transformation, and some machine learning methods for division [29]. A Protein BLAST search was carried out to determine the possibility of cross-reactivity among the final construct with other proteins. It can yield functional and evolutionary clues about amino acid sequences.

3. Results

3.1. Sequence analysis of N protein in 7 human-related coronaviruses

Human coronavirus (HCoV) includes α-coronaviruses and β-coronaviruses. HCoV-229E and HCoV-NL63 belong to the former, HCoV-HKU1, HCoV-OC43, the Middle East respiratory syndrome-related coronavirus (MERS-CoV), SARS-CoV, and the SARS-CoV-2 belong to the latter [30]. N protein is a relatively conservative protein in coronaviruses and has been successfully used as a diagnostic antigen [8,31]. Amino acid sequences of N proteins from these HCoV were obtained from the NCBI database, of which accession IDs were presented in Figure S1A. To better understand the divergence of N protein sequences between SARS-CoV-2 and other HCoVs, Clustal Omega was utilized to compare the full-length N protein sequences of the seven coronaviruses mentioned above. The result revealed that the N protein sequences between SARS-CoV-2 and SARS-CoV are highly similar and the evolutionary relationship of these species based on their N protein sequence information was presented in the phylogenetic tree (Figure S1A & S1B; File S1). As the close homology of N protein between SARS-CoV-2 and SARS-CoV, we analyzed the immunodominant antigenic regions of the SARS-CoV-2 N protein and compared them with the existing immunological studies of SARS-CoV. We abandoned the identical epitopes shared by both to enhance the specificity.

3.2. Linear B-cell epitope prediction of SARS-CoV-2 N protein

The full-length sequence of SARS-CoV-2 N protein was evaluated through ABCpred, BepiPred-2.0, and IEDB. The antigenicity was calculated via VaxiJen v2.0 with the given cutoff of ≥0.40. Using the ABCpred algorithm with the threshold value of ≥0.80, we identified 24 peptides (Table S1). Fourteen peptides were obtained via BepiPred-2.0 with a cutoff of ≥0.55 (Table S2). In parallel, a total of 16 peptides were identified by using IEDB (Table S3).

The short peptides (less than six amino acids) were discarded. To ensure the specificity, the peptides consistent with the SARS-CoV N protein sequence were excluded. Finally, 11, 8, and 7 potential linear B-cell epitopes forecasted by ABCpred, BepiPred-2.0, and IEDB, respectively, were obtained after stringent filtering (Table 1). After mapping the positions of peptides identified by those three servers, eight regions containing differently predicted epitopes were obtained (Figure 2).

3.3. Profiling and evaluation of hydrophilicity, surface accessibility and secondary structure of selected sequences

To further evaluate the potentiality of these eight antigenic regions as targets for antibody binding, their hydrophilicity, surface accessibility, and secondary structure were analyzed. Eight sequences were forecasted to be hydrophilic via ProtScale (Figure S2A); and when RSA ≥50%, four main regions with good surface accessibility were predicted in NetsurfP-2.0 (Figure S2B). After comparing the predicted peptides with the hydrophilic regions and the surface accessible regions, $N_{58\text{-}106}$, $N_{275\text{-}290}$, and $N_{327\text{-}349}$ were eliminated due to their surface inaccessibility, which might sterically hinder the approachability of antibody (Figure 2). At this point, we obtained five predicted regions ($N_{1\text{-}51}$, $N_{164\text{-}221}$, $N_{229\text{-}269}$, $N_{361\text{-}399}$, $N_{408\text{-}416}$).

Considering that structures such as beta-turn and random coil are more conducive to bind with the specific B cell receptor (BCR), we adopted SOPMA online software to predict the secondary structure of N protein (Figure S2C). To ensure the structural integrity of the predicted epitopes, we adjusted the terminus of the selected regions by appropriately adding several amino acid residues at both ends. By synthesizing the results of hydrophilicity, surface accessibility, and secondary structure analysis, we altered $N_{164\text{-}221}$ to

Table 1. Summary of the linear B-cell epitopes of SARS-CoV-2 nucleocapsid (N) protein predicted via ABCpred, BepiPred-2.0, and IEDB.

Tools	Seq#	Position	Epitope sequence	Length	Score	Antigenicity
ABCpred	1	91–106	TRRIRGGDGKMKDLSP	16	0.94	1.1467
	2	24–39	TGSNQNGERSGARSKQ	16	0.91	0.6333
	3	327–342	SGTWLTYTGAIKLDDK	16	0.88	0.9425
	4	59–74	HGKEDLKFPRGQGVPI	16	0.87	0.6163
	5	182–197	ASSRSSSRSRNSSRNS	16	0.87	0.9557
	6	376–391	ADETQALPQRQKKQQT	16	0.86	0.6949
	7	77–92	NSSPDDQIGYYRRATR	16	0.85	0.4843
	8	362–377	TFPPTEPKKDKKKKAD	16	0.85	0.5442
	9	206–221	SPARMAGNGGDAALAL	16	0.83	0.4517
	10	334–349	TGAIKLDDKDPNFKDQ	16	0.82	1.8438
	11	275–290	GRRGPEQTQGNFGDQE	16	0.80	1.2359
Bepipred-2.0	1	5–13	GPQNQRNAP	9		1.2141
	2	61–67	KEDLKFP	7		0.9682
	3	71–81	GVPINTNSSPD	11		0.5067
	4	89–104	RATRRIRGGDGKMKDL	16		0.8755
	5	171–213	FYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGN	43		0.6201
	6	229–265	QLESKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTAT	37		0.6878
	7	361–399	KTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAAD	39		0.5321
	8	408–416	QQSMSSADS	9		0.7484
IEDB	1	1–51	MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTAS	51		0.4006
	2	58–85	QHGKEDLKFPRGQGVPINTNSSPDDQIG	28		0.5570
	3	93–104	RIRGGDGKMKDL	12		0.8771
	4	164–216	GTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD	53		0.5455
	5	232–269	SKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYN	38		0.5302
	6	338–347	KLDDKDPNFK	10		2.1298
	7	361–390	KTFPPTEPKKDKKKKADETQALPQRQKKQQ	30		0.5605

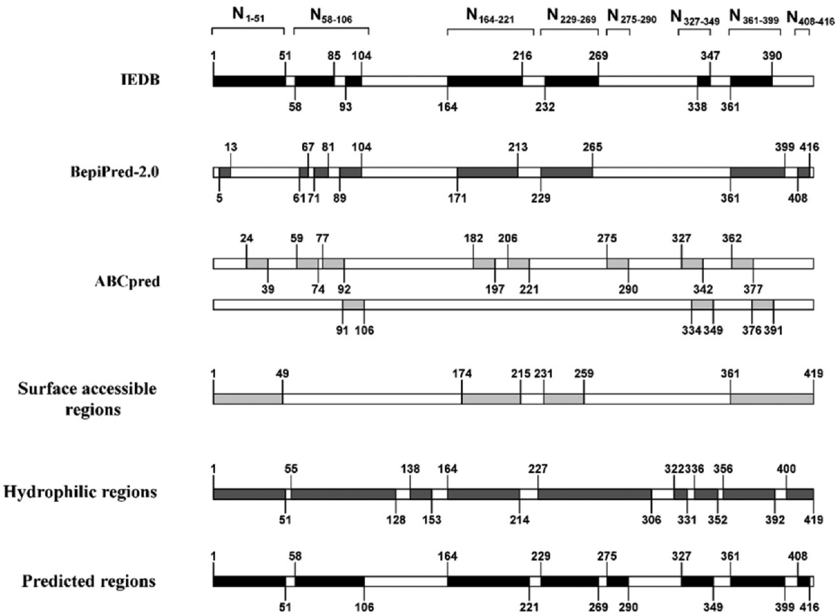

Figure 2. Distribution of predicted B-cell epitope regions, hydrophilic areas and surface accessible regions.

$N_{161\text{-}221}$; $N_{361\text{-}399}$ to $N_{354\text{-}399}$. Hereinabove, $N_{1\text{-}51}$, $N_{161\text{-}221}$, $N_{229\text{-}269}$, $N_{354\text{-}399}$, and $N_{408\text{-}416}$ were selected, which contained several potential B-cell epitopes with a high antigenicity score.

3.4. Prediction of the murine T-cell epitopes in SARS-CoV-2 N protein

For potential murine T-cell epitopes prediction, the TepiTool resource incorporated in IEDB was utilized. Thirteen peptides of MHC class I binding epitopes along with 16 peptides of MHC class II binding epitopes were identified (Table S4). All predicted T-cell epitopes that overlapped with the selected B-cell epitope regions were displayed in Table S5. Theoretically, MHC-I molecules promote the activation of cytotoxic T lymphocyte (CTL) which kills virus-infected cells [32]. Helper T cells (Th) activated by MHC-II presenting epitopes could provide essential signals to B cells for antibodies production [15]. Predicted murine MHC class II binding epitope $N_{244\text{-}258}$ was included in our predicted B-cell epitope region $N_{229\text{-}269}$; $N_{357\text{-}371}$ and $N_{384\text{-}398}$ were contained in predicted region $N_{354\text{-}399}$; predicted B-cell epitope $N_{408\text{-}416}$ was contained in predicted murine

T-cell epitope $N_{405\text{-}419}$. Therefore, $N_{229\text{-}269}$, $N_{354\text{-}399}$, and $N_{405\text{-}419}$ were finally chosen because they contained potential murine MHC class II binding epitopes without murine MHC class I binding epitopes, which makes them more conducive to the production of antibodies.

Immunome Browser 3.0 in IEDB comprises the records of existing reference sequences. It can form a response frequency (RF) score to indicate the frequency of the residues in the positive epitopes together with the independent experimental records [33,34]. After scanned in Immunome Browser 3.0, none of the predicted murine MHC class II binding epitopes of SARS-CoV-2 N protein had been confirmed yet. Concurrently, we retrieved the murine MHC class II binding epitopes of SARS-CoV with Immunome Browser 3.0, we found that $N_{351\text{-}365}$ (Epitope ID: 69,035) and $N_{353\text{-}365}$ (Epitope ID: 985,589) in SARS-CoV N protein had been verified as murine MHC class II binding epitopes by experiments [15,35], they corresponded to the identical sequences $N_{350\text{-}364}$ and $N_{352\text{-}364}$ in SARS-CoV-2, respectively. Moreover, we noticed that Val^{350} and Leu^{352} were located in an extended strand structure. Thus, to keep the secondary structure integrity, we expanded the predicted region $N_{354\text{-}399}$ to $N_{349\text{-}399}$.

Allergenicity of the selected regions was assessed via AllerTOP v2.0, the results demonstrated that all identified regions were predicted to be non-allergen; The toxicity of the three predicted regions was examined by ToxinPred, all of them were forecasted to be non-toxin (Table 2).

3.5. Immunodominant regions selected for recombinant antigen generation

Three predicted regions ($N_{229\text{-}269}$, $N_{349\text{-}399}$, and $N_{405\text{-}419}$) were connected using flexible linkers $(GGGGS)_2$. Flexible linker $(GGGGS)_2$ is excellent in segmenting protein fragments, maintaining biological activity, and promoting protein expression [36]. It had been proved that the Pan DR epitope PADRE (AKFVAAWTLKAAA) functions as a universal T helper epitope, which can induce specific high titer antibodies and lasting antibody responses [37,38]. Hence, we added it to the N-terminal of our construct to boost the humoral immune response (Figure 3). The result of Protein BLAST showed little similarity between the final construct with any known encoded protein, it prompted that the antibody derived from the immune fragment designed in our study is not likely to cross-react with other peptides beyond SARS-CoV-2, which will be confirmed by experiments in the follow-up study.

Table 2. Distribution of the potential epitopes and physicochemical properties of the predicted immunodominant regions.

$N_{229\text{-}269}$	
BepiPred-2.0	**QLESKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTAT**KAYN
IEDB	QLE**SKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYN**
TepiTool	QLESKMSGKGQQQQG**QTVTKKSAAEASKKP**RQKRTATKAYN
Hydrophilicity	*QLESKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYN*
Surface accessibility	QLESKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYN
Secondary structure	hhhhhhcccccccttceeehhhhhhhtccccccccccchee
Antigenicity	0.5212
Allergenicity	Non-allergen
Toxicity	Non-toxin
$N_{349\text{-}399}$	
BepiPred-2.0	QVILLNKHIDAY**KTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAAD**
IEDB	QVILLNKHIDAY**KTFPPTEPKKDKKKKADETQALPQRQKKQQ**TVTLLPAAD
ABCpred	QVILLNKHIDAYK**TFPPTEPKKDKKKKAD**ETQALPQRQKKQQTVTLLPAAD QVILLNKHIDAYKTFPPTEPKKDKKKK**ADETQALPQRQKKQQT**VTLLPAAD
TepiTool	QVILLNKH**IDAYKTFPPTEPKKD**KKKKADETQALPQRQKKQQTVTLLPAAD QVILLNKHIDAYKTFPPTEPKKDKKKKADETQALP**QRQKKQQTVTLLPAA**D
Verified Mouse-MHC-II	QVI**LLNKHIDAYKTFP**PTEPKKDKKKKADETQALPQRQKKQQTVTLLPAAD Q**VILLNKHIDAYKTFP**PTEPKKDKKKKADETQALPQRQKKQQTVTLLPAAD
Hydrophilicity	*QVIL*LNK*HIDAYKTFPPTEPKKDKKKKADETQALPQRQKKQQTV*TLLPAAD
Surface accessibility	QVILLNKHIDAYKTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAAD
Secondary structure	eeeeehhhhhhtccccccccccccccchhcccccccccccceeeeccccc
Antigenicity	0.4168
Allergenicity	Non-allergen
Toxicity	Non-toxin
$N_{405\text{-}419}$	
BepiPred-2.0	KQL**QQSMSSADS**TQA
TepiTool	**KQLQQSMSSADSTQA**
Hydrophilicity	*KQLQQSMSSADSTQA*
Surface accessibility	KQLQQSMSSADSTQA
Secondary structure	hhhhhhhhhhccccc
Antigenicity	0.4771
Allergenicity	Non-allergen
Toxicity	Non-Toxin

h, Alpha helix; e, Extended strand; t, Beta turn; c, Random coil.

The predicted epitopes, the hydrophilic regions, and the surface accessible areas were marked in Bold, Italic, and Underlined, respectively.

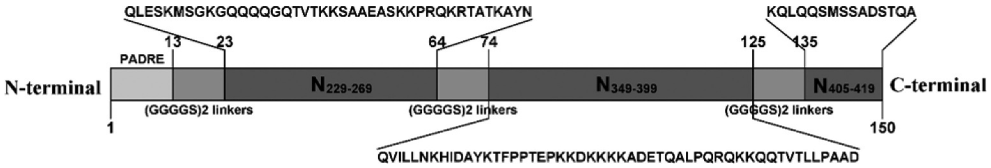

Figure 3. Schematic diagram of the selected immunodominant regions and the final construct for designing diagnostic antibodies.

4. Discussion

In this study, we utilized multiple bioinformatics and immunoinformatics approaches to forecast potential immunodominant regions of SARS-CoV-2 N protein. Though several bioinformatic predictions of potential epitopes for SARS-CoV-2 have been reported, these studies mainly focused on vaccine development [33,-39–41]. Compared to these studies, we employed distinctive strategies. As we aimed to develop diagnostic antibodies against SARS-CoV-2, mouse was selected as the host species for MHC class II binding epitopes prediction. Additionally, peptides that shared identical sequences with SARS-CoV were excluded to enhance the specificity. Nevertheless, the results of this study were derived from computational algorithms. Whether the antibody can effectively bind to SARS-CoV-2 N protein in clinical samples and the performance of the assay reach the national standard, that is, cross-reactivity, sensitivity, and specificity, remain to be verified by experiments in vitro and in vivo.

N protein was reported as a good diagnostic antigen because of its high immunogenicity and affluence during coronavirus infection [8,31]. N protein of the influenza virus is also the main target in antigen-detection tests [42,43]. The influenza antigen detection was useful especially for patients tested within the first 48 hours of illness, when the influenza viral load in the upper respiratory tract was high [44]. Accordingly, we attempted to develop an efficient and accurate N protein-detection assay for SARS-CoV-2. In the follow-up study, newly developed materials such as fluorescent dyes and nanoparticles could be adopted to improve the sensitivity of detection [5].

Influenza virus infection causes respiratory symptoms similar to COVID-19, which makes it difficult to distinguish the diseases by symptoms [45]. Hence, we compared the N protein sequences between the influenza virus and SARS-CoV-2 (File S2). The results showed that the sequence similarity is low, which suggested that direct detection of N protein could distinguish COVID-19 from influenza virus infection. Besides, we noticed that the full-length N protein of SARS-CoV-2 may cross-react with the serum of patients infected with SARS-CoV [4], while truncated protein was proved to reduce the cross-reactivity without reducing sensitivity [46,47]. Hence, we chose to use truncated recombinant protein rather than the full-length N protein for developing diagnostic antibodies. Recently, cladistic studies based on N protein sequence of SARS-CoV-2 have been reported [48–51]. None of the reported mutations are located in the selected immunodominant regions of the current study (data not shown). Nevertheless, attention should be paid to the diagnostic efficiency of mAbs derived from the fragments of SARS-CoV-2 N protein, which need to be evaluated by experiments.

SARS-CoV-2 N protein displayed almost the same distribution of hydrophilic regions as that of SARS-CoV N protein, which is easily understood owing to their sequence homology. Yu and colleagues demonstrated that the $N_{122-422}$ incorporated the main immunogenic sites of the SARS-CoV N protein and could be used for efficient diagnosis [52]. Interestingly, the selected fragments in our study were all located within the corresponding region of SARS-CoV-2 N protein, suggesting an advantage of these fragments in generating optimal diagnostic antibodies.

In conclusion, three potential immunodominant regions of SARS-CoV-2 N protein: $N_{229-269}$, $N_{349-399}$, and $N_{405-419}$ that contain both linear B-cell epitopes and murine MHC class II binding epitopes were identified. A construct with 150 amino acids was built (Figure 3). The final construct consists of seven B-cell epitopes, six murine MHC class II binding epitopes, and a PADRE sequence (Table 2). After cloning, expression, and purification of the recombinant protein derived from this study, we will immunize Balb/c mice to generate mAbs with the hybridoma technique. The cross-reactivity, reactivity, specificity, and titer of mAbs will be further evaluated. After confirming the biological functions, these mAbs would be utilized to develop an antigen-capture-based assay system for early diagnosis of SARS-CoV-2.

Acknowledgments

We appreciate the help of Dr. Hezhi Fang and Dr. Xiang Wu for the proofreading of the manuscript. Special thanks to Xiaoke Huang for her revision to the language of the manuscript.

Disclosure statement

The authors declare no conflict of interest.

Funding

This work was supported by [Changsha Science and Technology Bureau Special Project for the Emergency response of pneumonia infected by COVID-19] under Grant [NO. kq2001038], [Hunan Provincial Natural Sciences Foundation] under Grant [No. 2018JJ3740], and [the National Science Foundation of China] under Grant [No. 81970746].

Contributions

Yufeng Dai: Conceptualization; Methodology; Software; Validation; Formal Analysis; Resources; Data Curation; Writing-Original Draft Preparation; Writing-Review and Editing; Visualization. Hongzhi Chen: Conceptualization; Methodology; Software; Validation; Formal Analysis; Resources; Data Curation; Writing-Review and Editing; Visualization; Supervision; Project administration. Siqi Zhuang: Methodology; Software; Resources; Writing-Review and Editing; Visualization. Xiaojing Feng: Methodology; Software; Resources. Yiyuan Fang: Methodology; Software; Resources. Haoneng Tang: Methodology; Software; Resources. Ruchun Dai: Methodology; Software; Resources; Visualization. Lingli Tang: Conceptualization; Methodology; Software; Validation; Formal Analysis; Resources; Data Curation; Writing-Review and Editing; Visualization; Supervision; Project administration; Funding. All authors have read and agreed to the published version of the manuscript.

References

[1] Rothe C, Schunk M, Sothmann P, et al. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. N Engl J Med. 2020;382:970–971.

[2] Beck ET, Henrickson KJ. Molecular diagnosis of respiratory viruses. Future Microbiol. 2010;5:901–916.

[3] Chu DKW, Pan Y, Cheng SMS, et al. Molecular diagnosis of a novel Coronavirus (2019-nCoV) causing an outbreak of Pneumonia. Clin Chem. 2020;66:549–555.

[4] Okba NMA, Muller MA, Li W, et al. Severe acute respiratory syndrome Coronavirus 2-specific antibody responses in Coronavirus disease 2019 patients. Emerg Infect Dis. 2020;26:1478–88.

[5] Vemula SV, Zhao J, Liu J, et al. Current approaches for diagnosis of influenza virus infections in humans. Viruses. 2016;8:96.

[6] James VL, Lambden PR, Caul EO, et al. Enzyme-linked immunosorbent assay based on recombinant human group C rotavirus inner capsid protein (VP6) To detect human group C rotaviruses in fecal samples. J Clin Microbiol. 1998;36:3178–3181.

[7] Fitzgerald N, Cross M, O'Shea S, et al. Diagnosing acute HIV infection at point of care: a retrospective analysis of the sensitivity and specificity of a fourth-generation point-of-care test for detection of HIV core protein p24. Sex Transm Infect. 2017;93:100–101.

[8] McBride R, van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional protein. Viruses. 2014;6:2991–3018.

[9] Che XY, Hao W, Wang Y, et al. Nucleocapsid protein as early diagnostic marker for SARS. Emerg Infect Dis. 2004;10:1947–1949.

[10] Li YH, Li J, Liu XE, et al. Detection of the nucleocapsid protein of severe acute respiratory syndrome coronavirus in serum: comparison with results of other viral markers. J Virol Methods. 2005;130:45–50.

[11] Soria-Guerra RE, Nieto-Gomez R, Govea-Alonso DO, et al. An overview of bioinformatics tools for epitope prediction: implications on vaccine development. J Biomed Inform. 2015;53:405–414.

[12] Berglund L, Bjorling E, Jonasson K, et al. A whole-genome bioinformatics approach to selection of antigens for systematic antibody generation. Proteomics. 2008;8:2832–2839.

[13] Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics. 2016;32:511–517.

[14] Lin Y, Shen X, Yang RF, et al. Identification of an epitope of SARS-coronavirus nucleocapsid protein. Cell Res. 2003;13:141–145.

[15] Zhao J, Huang Q, Wang W, et al. Identification and characterization of dominant helper T-cell epitopes in the nucleocapsid protein of severe acute respiratory syndrome coronavirus. J Virol. 2007;81:6079–6088.

[16] Madeira F, Park YM, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47:W636–W41.

[17] Whelan S, Lio P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. 2001;17:262–272.

[18] Saha S, Raghava GP. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins. 2006;65:40–48.

[19] Jespersen MC, Peters B, Nielsen M, et al. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. 2017;45:W24–W9.

[20] Dhanda SK, Mahajan S, Paul S, et al. IEDB-AR: immune epitope database-analysis resource in 2019. Nucleic Acids Res. 2019;47:W502–W6.

[21] Paul S, Sidney J, Sette A, et al. TepiTool: a pipeline for computational prediction of T cell epitope candidates. Curr Protoc Immunol. 2016;114:189 1–9 24.

[22] Wang P, Sidney J, Kim Y, et al. Peptide binding predictions for HLA DR, DP and DQ molecules. BMC Bioinformatics. 2010;11:568.

[23] Wang P, Sidney J, Dow C, et al. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS Comput Biol. 2008;4:e1000048.

[24] Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinformatics. 2007;8:4.

[25] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157:105–132.

[26] Klausen MS, Jespersen MC, Nielsen H, et al. NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. Proteins. 2019;87:520–527.

[27] Geourjon C, Deleage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci. 1995;11:681–684.

[28] Gupta S, Kapoor P, Chaudhary K, et al.; Open Source Drug Discovery C. In silico approach for predicting toxicity of peptides and proteins. PLoS One. 2013;8:e73957.

[29] Dimitrov I, Bangov I, Flower DR, et al. AllerTOP v.2–a server for in silico prediction of allergens. J Mol Model. 2014;20:2278.

[30] Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019;17:181–192.

[31] Seo SH, Wang L, Smith R, et al. The carboxyl-terminal 120-residue polypeptide of infectious bronchitis virus nucleocapsid induces cytotoxic T lymphocytes and protects chickens from acute infection. J Virol. 1997;71:7889–7894.

[32] Janice Oh HL, Ken-En Gan S, Bertoletti A, et al. Understanding the T cell immune response in SARS coronavirus infection. Emerg Microbes Infect. 2012;1:e23.

[33] Grifoni A, Sidney J, Zhang Y, et al. A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. Cell Host Microbe. 2020;27:671–80 e2.

[34] Dhanda SK, Vita R, Ha B, et al. ImmunomeBrowser: a tool to aggregate and visualize complex and heterogeneous epitopes in reference proteins. Bioinformatics. 2018;34:3931–3933.

[35] Zhao J, Zhao J, Mangalam AK, et al. Airway memory CD4(+) T cells mediate protective immunity against emerging respiratory coronaviruses. Immunity. 2016;44:1379–1391.

[36] Chen X, Zaro JL, Shen WC. Fusion protein linkers: property, design and functionality. Adv Drug Deliv Rev. 2013;65:1357–1369.

[37] Alexander J, Del Guercio MF, Maewal A, et al. Linear PADRE T helper epitope and carbohydrate B cell epitope conjugates induce specific high titer IgG antibody responses. J Immunol. 2000;164:1625–1633.

[38] Del Guercio MF, Alexander J, Kubo RT, et al. Potent immunogenic short linear peptide constructs composed of B cell epitopes and Pan DR T helper epitopes (PADRE) for antibody responses in vivo. Vaccine. 1997;15:441–448.

[39] Ahmed SF, Quadeer AA, McKay MR. Preliminary identification of potential vaccine targets for the COVID-19 Coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. Viruses. 2020;12:254.

[40] Lee CH, Koohy H. In silico identification of vaccine targets for 2019-nCoV. F1000Res. 2020;9:145.

[41] Chen HZ, Tang LL, Yu XL, et al. Bioinformatics analysis of epitope-based vaccine design against the novel SARS-CoV-2. Infect Dis Poverty. 2020;9:88.

[42] Dziabowska K, Czaczyk E, Nidzworski D. Detection methods of human and animal influenza virus-current trends. Biosensors (Basel). 2018;8:94.

[43] Storch GA. Rapid diagnostic tests for influenza. Curr Opin Pediatr. 2003;15:77–84.

[44] Seki Y, Oda Y, Sugaya N. Very high sensitivity of a rapid influenza diagnostic test in adults and elderly individuals within 48 hours of the onset of illness. PLoS One. 2020;15:e0231217.

[45] Larsen JR, Martin MR, Martin JD, et al. Modeling the onset of symptoms of COVID-19. Front Public Health. 2020;8:473.

[46] Lee HK, Lee BH, Seok SH, et al. Production of specific antibodies against SARS-coronavirus nucleocapsid protein without cross reactivity with human coronaviruses 229E and OC43. J Vet Sci. 2010;11:165–167.

[47] Surjit M, Lal SK. The SARS-CoV nucleocapsid protein: a protein with multifarious activities. Infect Genet Evol. 2008;8:397–405.

[48] Zuckerman NS, Pando R, Bucris E, et al. Comprehensive analyses of SARS-CoV-2 transmission in a public health virology laboratory. Viruses. 2020;12:854.

[49] Banu S, Jolly B, Mukherjee P, et al. A distinct phylogenetic cluster of Indian SARS-CoV-2 isolates. Open Forum Infect Dis. 2020;ofaa434.

[50] Raghav S, Ghosh A, Turuk J, et al. SARS-CoV2 genome analysis of Indian isolates and molecular modelling of D614G mutated spike protein with TMPRSS2 depicted its enhanced interaction and virus infectivity. 2020. 2020.07.23.217430. Advance online publication. https://doi.org/10.1101/2020.07.23.217430

[51] Joshi M, Puvar A, Kumar D, et al. Genomic variations in SARS-CoV-2 genomes from Gujarat: underlying role of variants in disease epidemiology. 2020. 2020.07.10.197095. Advance online publication. https://doi.org/10.1101/2020.07.10.197095

[52] Yu F, Le MQ, Inoue S, et al. Recombinant truncated nucleocapsid protein as antigen in a novel immunoglobulin M capture enzyme-linked immunosorbent assay for diagnosis of severe acute respiratory syndrome coronavirus infection. Clin Vaccine Immunol. 2007;14:146–149.