MOL2NET, International Conference Series on Multidisciplinary Sciences, 6th Edition
USINEWS-04: US-IN-EU Worldwide Science Workshop Series, UMN, Duluth, USA, 2020

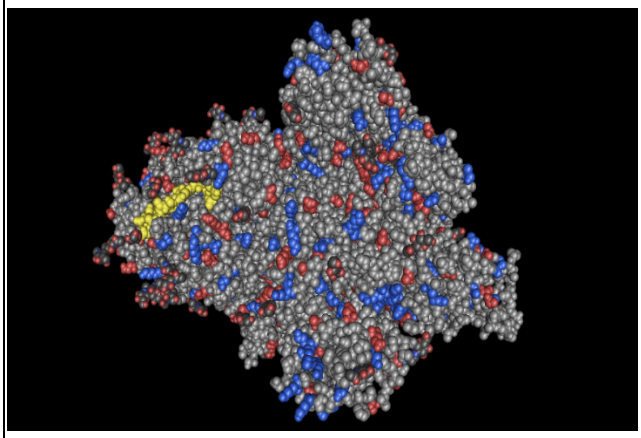
In Silico Approach for Peptide Vaccine Design for CoVID 19

Subhamoy Biswas [a], Shreyans Chatterjee [b], Tathagata Dey [c], Sumanta Dey [d], Smarajit Manna [d,e], Ashesh Nandy [d], Subhash C Basak [f]

[a]*Department of Electrical Engineering, Jadavpur University, Kolkata, West Bengal, India*
[b]*Department of Microbiology, St. Xavier's College, Kolkata, West Bengal, India*
[c]*Department of Computer Science Engineering, Government College of Engineering and Leather Technology, Serampore, West Bengal, India*
[d]*Centre for Interdisciplinary Research and Education, Kolkata, West Bengal, India*
[e]*Jagadis Bose National Science Talent Search, Kolkata, West Bengal, India*
[f]*Department of Chemistry and Biochemistry, University of Minnesota, Duluth, Minnesota, USA*

Email ID of corresponding author: *anandy43@yahoo.com*

Graphical Abstract



Abstract

The currently surging SARS-COV-2 (or CoVID-19) is challenging the public health authorities worldwide. As of now there is no approved vaccine or drug available for the control of the viral disease. Therefore, non-pharmaceutical interventions (NPIs) are being used around the world to manage the spread of CoVID-19. In this article we used a computer-assisted vaccine design (CAVD) approach to develop a set of most probable peptide vaccine candidates which can be tested for their efficacy by wet lab experiments.

Introduction

Coronaviruses consist of a group of enveloped, positive single-stranded RNA viruses that infect humans, but also a wide range of animals. Coronaviruses were first described in 1966 by Tyrell and Bynoe [1], who cultivated the viruses from patients with common colds. Based on the morphology of

the spherical virions with a core shell and surface projections resembling a solar corona, the authors names them coronaviruses.

SARS (Severe Acute Respiratory Syndrome), Middle East Respiratory Syndrome (MERS), and the currently raging SARS-COV-2 (or CoVID-19) are three coronaviruses which have emerged as important global pathogens in the past two decades.

The rapid spread of the recently emergent novel coronavirus, SARS-COV-2 and absence of any therapeutic remedies have caused global concern with the World Health Organization (WHO) declaring it as a as a pandemic on March 11, 2020 [2]. The virus is suspected to have made the jump from a zoonotic virus to attack human hosts from a wild animal market in Wuhan, Hebei province, in China in December 2019 and led to the CoVID-19 disease. (As per current practice, the terms SARS-COV-2 and CoVID-19 are used interchangeably).

Analogous to the SARS-CoV, clinical symptoms of COVID-19 infection include fever, myalgia, fatigue, and cough. More than 50% of the patients develop dyspnea. Some patients had radiographic ground-glass lung alterations and lower than average circulating lymphocyte and platelet numbers. Such severe symptoms sometimes lead to death.

The virus spreads by human-to-human contact through droplets from coughing or sneezing or contacts, which allows the virus to spread very fast. As of the writing of this report, 28thMarch 2020, barely three months from first identification, there have been reports of a total of 571,678 infections and 26494 deaths worldwide [3], where Italy has the highest fatalities and the United States is turning out to be the new epicenter of the infection. Public health administrations worldwide are planning various methods of containment, which China had taken to an extreme by locking down millions of people in their own homes and cities, showing some benefits in the absence of any other remedies.

The ongoing development of the CoVID-19 epidemic closely parallels the SARS epidemic of 2002-03, though with a fatality rate at about 2% against the SARS rate of around 10%. As of now there is no remedy for SARS, and CoVID-19 is totally new. However, attempts are being made to repurpose therapeutics existing or developed for other closely related diseases or develop new drugs and vaccines to combat the current epidemic [4]. Such emergency measures still will require several years for results. Other methods are required.

We have approached this issue using the recent concept of peptide vaccines. While traditional vaccines require years of trials and cost hundreds of millions of dollars, peptide vaccines hold the promise of quick development at significantly lower costs. These new vaccines also have the potential to scale up to meet demand requirements and possible modifications to the structure to better conform to local community requirements [5]. We have explored *in silico* the potential of peptide vaccines to combat influenza [6,7], rotavirus [8], human papillomavirus [9], Zika virus infections [10], and recently Ebola virus as well [11], demonstrating that individual sets of peptide vaccines can be

considered for wet lab experimentation. That peptide vaccines have shown great promise in cancer tumour treatment points to a possible role likewise for viral infections as well [12,13].

Materials and Methods

We downloaded the spike glycoprotein nucleotide and protein sequences from the NCBI GenBank database for the coronaviruses of the 7 different coronavirus types (Table 1). We are interested in the CoVID-19 virus complete sequences only and partial, duplicate and unannotated entries were left out. For purposes of this exercise, only 72 out of 106 full length spike glycoproteins of the SARS-CoV-2 (CoVID-19) coronaviruses available at this time in the database were taken into account.
Table 1: Examples of the 7 different human coronaviruses

Type	**Accession ID**	**Date of Collection**	**Host**
229E	AF304460.1	11.07.2001	*Homo sapiens*
NL63	AY567487.2	22.06.2004	*Homo sapiens*
HKU1	AY597011.2	27.01.2006	*Homo sapiens*
OC43	NC_006213.1	21.02.2019	*Homo sapiens*
SARS	AY274119.3	19.12.2017	*Homo sapiens*
MERS	JX869059.2	04.12.2012	*Homo sapiens*
CoVID 19	NC_045512.2	13.03.2020	*Homo sapiens*

Our method to design peptide vaccines follows a protocol we have developed over several applications to determine possible peptide vaccines targets. Briefly, we first choose the protein we want to design vaccines for, and then scan our database of virus protein sequences by a sliding window method to determine their sequence segment descriptors, determine the average solvent accessibility (ASA) and compare the two sets of results to determine where the proteins have the least variability and also have high ASA.

In our alignment-free method to compare sequences we adopt the graphical representation and numerical characterization methods that have been expounded in several of our papers [8, 14]. Briefly, we consider a 20-dimensional space where each amino acid is assigned to a particular axis. For the protein sequence we start at the origin and move one step along the axis represented by the first amino acid, then move another step in the direction designated by the second amino acid, and so on. Doing this for the entire sequence plots out a curve in the 20D abstract space where we can define weighted average for each axis (μ_{xi}) and a consolidated graph radius as follows:

$$\mu_{x1} = \Sigma_{i=1,\,N} x_{1i}\,;\ \mu_{x2} = \Sigma_{i=1,\,N} x_{2i}\,;\ \ldots;\ \mu_{x20} = \Sigma_{i=1,\,N} x_{20i}$$

$$p_R = \sqrt{(\mu_{x1}^2 + \mu_{x2}^2 + \ldots + \mu_{x20}^2)}$$

where N is the number of amino acids representing the length of the whole sequence or part thereof, and p_R is the graph radius. To ensure that no degeneracy takes place one can devise various scenarios;

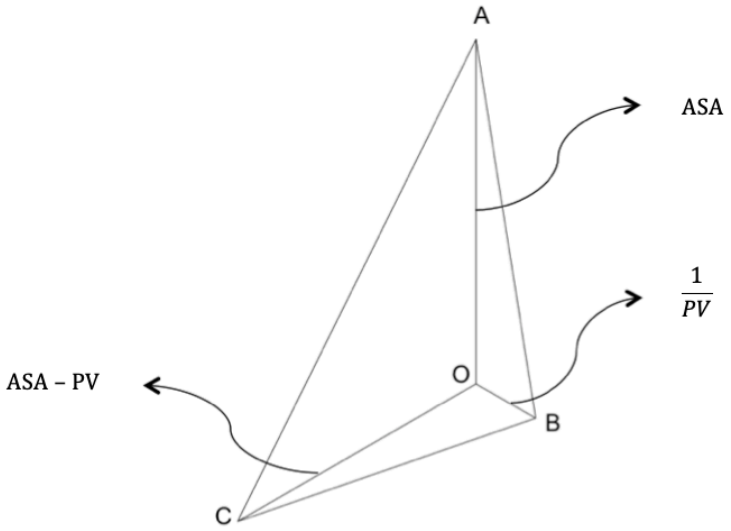
the ones we use are outlined in detail in our paper [8]. Apart from assigning a numerical value for any sequence by the p_R, which we may call the sequence descriptor, it is found that two identical sequences will have the same p_R; this property may be used to mark out duplicate sequences in any protein database. In the current exercise we take a small window of 12 amino acids and determine the p_R values of each of the sequences at the same aa number, then move the window by one amino acid and take the p_R values again and so on, from the beginning to the end of the sequence. Lining up the p_R values for each window over all the sequences, we can determine how many different p_R values, say PV, are there; the lower the PV, the more conserved this peptide segment will be in the protein sequence.

The use of conserved peptides serves the purpose of configuring a vaccine that works on those segments of the viral protein that will not mutate too fast. We also need to determine which of these conserved peptides are surface exposed, i.e., have high average solvent accessibility (ASA), using web based tools such as the SABLE server and then take a moving average over the 12 aa window size preferred for the PV so that they are compatible. Now we need to determine those segments where the PV is low and ASA is high to ensure as surface exposed conserved peptide domains as possible.

In our earlier endeavours of such determinations, we had used graphical methods and eye estimation. Here we have used a more rigorous mathematical method for a more robust result, the details of which are being communicated elsewhere. In this method we construct, method we construct, for each peptide stretch of length 12, with its ASA and PV values, a triangle with a vertical measure related to ASA value, a measure at 120 degrees for 1/PV value and another measure at 120 degree of ASA-PV value (see Fig.1). The area of this triangle is

$$a(ASA,PV) = \Delta ABC = \frac{\sqrt{3}}{4} * (ASA^2 + 2 * \frac{ASA}{PV} - ASA * PV - 1)$$

where $a(ASA,PV)_n > a(ASA,PV)_m$ when the distance between the ASA and PV values in the case of the



n-th amino acid of the sequence is greater than the ASA and PV values of the m-th amino acid. Doing this over the entire sequence we can get a list of the areas from the highest values to the lowest and take only the first 10 to 20 peptides that fit this criterion for the next step of the analysis. In this brief report we state only the results and discussions coming from application of this method without getting into further details.

Results and Discussion

As mentioned above, we used our 20D protein sequence model with window size 12 aa to compute the descriptors for all the 72 protein sequences. Window by window comparison allowed us to compute the number of variations in each window for all the sequences. Thus, we have protein variability (PV) record of each window demonstrating by way of low count which segments have low volatility, i.e., high conservativeness.

To get the segments with high access to the solvent we use the SABLE server (15) to determine the number over 12 counts. Comparison with the PV moving average in the same way will give us the best surface exposure with highest conservativeness. Applying now the 2D polygon representation technique mentioned above we get a set of peptide regions which satisfy this criterion. The first few results are shown in Table 2.

The recent availability of the crystal structure of the CoVID-19 spike glycoprotein, 6VYB, in the PDB allowed us to check that the above peptides were indeed surface exposed. Fig.2 shows one example where the segment marked in yellow, 696-711, is one result from our analyses. Unfortunately, due to experimental constraints not all residues of the glycoprotein could be displayed in the crystal structure.

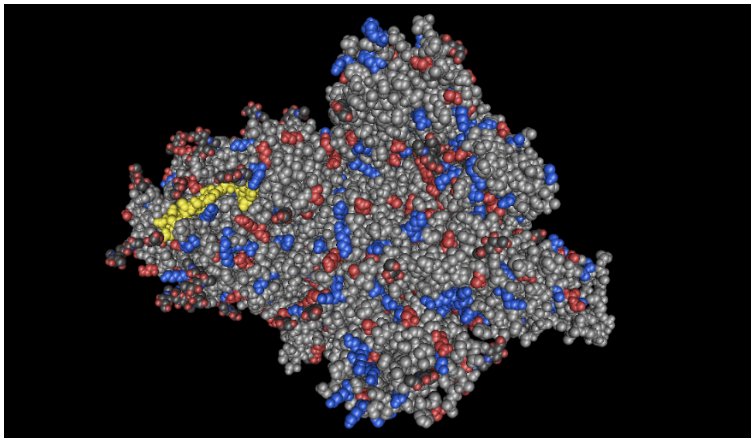


Fig.2. The crystal structure of the spike glycoprotein 6VYB showing the 696-711 segment.

Next, we analyze whether all the segments we have determined to be surface situated and well conserved have good epitope potential. For this we use Ellipro analysis tools from IEDB and another web based predictor, Bepipred, and a selection of the best results from Bepipred is shown in Table 3.

Table 2: List of peptides that gives the best surface exposure with highest conservativeness:

Rank	Starting Position	Score (Area of the Polygon)	Protein Variability	Peptide (Length=12)
1	16	586.9246612	1	VNLTTRTQLPPA
2	18	573.7087526	1	LTTRTQLPPAYT

3	1251	552.8760304	1	GSCCKFDEDDSE
4	1249	547.7279905	1	SCGSCCKFDEDD
5	1157	542.6040069	1	KNHTSPDVDLGD
6	1255	529.8992939	1	KFDEDDSEPVLK
7	1256	529.8992939	1	FDEDDSEPVLKG
8	15	519.8437768	1	CVNLTTRTQLPP
9	1252	519.8437768	1	SCCKFDEDDSEP
10	1250	514.8521025	1	CGSCCKFDEDDS

Table 3. List of candidates having appreciable epitope probability (≥0.45)

Sl No	Start	End	Peptide	Epitope Probability
1	13	29	SQCVNLTTRTQLPPAYT	0.475
2	1248	1270	CSCGSCCKFDEDDSEPVLKGVKL	0.471
3	1143	1173	PELDSFKEELDKYFKNHTSPDVDLGDISGIN	0.5103
4	802	818	FSQILPDPSKPSKRSFI	0.509
5	696	711	TMSLGAENSVAYSNNS	0.447
6	785	799	VKQIYKTPPIKDFGG	0.46
7	178	191	DLEGKQGNFKNLRE	0.454
8	1134	1151	NNTVYDPLQPELDSFKEE	0.466
9	437	450	NSNNLDSKVGGNYN	0.471

As a further check, we did more Ellipro analyses and determined that the above regions cover linear epitopes and form part of conformational epitopes also. A BLAST analysis of these possible epitope regions with respect to human proteins to eliminate autoimmune threats resulted finally in a short list (Table 4) of four peptides that may be tried out in a wet lab environment for trials for peptide vaccines.

Table 4: Final Shortlisted candidates for peptide vaccine design against CoVID 19

Sl No	Position		Protein Variability	Peptide	Epitope Probability
	Start	End			
1	1143	1173	1	PELDSFKEELDKYFKNHTSPDVDLGDISGIN	0.5103
2	802	818	1	FSQILPDPSKPSKRSFI	0.509
3	696	711	1	TMSLGAENSVAYSNNS	0.447
4	437	450	1	NSNNLDSKVGGNYN	0.471

Conclusion

Our alignment-free analyses of the SARS-CoV-2, more widely known by its disease name, CoVID-19, spike glycoprotein protein sequence analyses have yielded four regions that appear to be surface situated and well-conserved among the 72 sequences in our database. Developing peptide vaccines on these lines should be biochemically relatively easy task leaving ample ground and opportunity for testing among animal samples in the various phase trials. Many different experiments have shown that peptide vaccines are quite effective, especially in cancerous tumour cases, and although the first peptide vaccine was designed for canines, no peptide vaccines have been granted a license yet for use in human cases. Given the good experimental results, it is believed that eventually such a license will come by, and then peptide vaccines will enable combat emergent viral epidemics and pandemics in an effective way. Our results here are small endeavours in that direction.

References

1. Tyrrell DA, Bynoe ML. Cultivation of viruses from a high proportion of patients with colds. Lancet 1966: 1: 76–77.

2. WHO declares CoVID-19 a pandemic.https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

3. WHO Coronavirus Disease Situation Report – 68. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200328-sitrep-68-covid-19.pdf?sfvrsn=384bc74c_2

4. Galvez J, Zanni R, Galvez-Llompart M. Drugs Repurposing for Coronavirus Treatment: Computational Study Based On Molecular Topology, https://revistas.ucv.es/index.php/Nereis/article/view/592

5. A Nandy, S Dey, P Roy, SC Basak. (2018) Epidemics and Peptide Vaccine Response - A Brief Review. Curr Top Med Chem. 18, 2202-2208. doi: 10.2174/1568026618666181112144745.

6. A Ghosh, A Nandy and P Nandy, Computational analysis and determination of a highly conserved surface exposed segment in H5N1 avian flu and H1N1 swine flu neuraminidase *BMC Structural Biology* 2010, **10**:6 doi:10.1186/1472-6807-10-6.

7. T Sarkar, S Das, A De, P Nandy, S Chattopadhyay, M Chawla-Sarkar, A Nandy. H7N9 influenza outbreak in China 2013: In silico analyses of conserved segments of the hemagglutinin as a basis for the selection of peptide vaccine targets. *Comput. Biol. Chem.***59** (**2015**) 8–15.

8. Ghosh A, Chattopadhyay S, Chawla-Sarkar M, Nandy P, Nandy A (**2012**) In Silico Study of Rotavirus VP7 Surface Accessible Conserved Regions for Antiviral Drug/Vaccine Design. *PLoS ONE***7(7)**: e40749. doi:10.1371/journal.pone.0040749).

9. *Dey S, De A, Nandy A. Rational Design of Peptide Vaccines Against Multiple Types of Human Papillomavirus. Cancer Informatics**2016**, 15(S1), 1-16.* doi: 10.4137/CIN.S39071.

10. *Dey S, Nandy A,* Basak SC, Nandy P, *Das S. (**2017**)* A Bioinformatics approach to designing a Zika virus vaccine. *ComputBiol Chem.* 68, 143-152.

11. Biswas S, Dey T, Chatterjee S, Manna S, Nandy A, Das S, Nandy P, Basak SC. A novel approach to Peptide Vaccine Design for Ebola virus. Published: 24 November 2019 by MDPI AG in MOL2NET 2019, International Conference on Multidisciplinary Sciences, 5th edition session USINEWS-03: US-IN-EU Worldwide Science Workshop Series, UMN, Duluth, USA, 2019. DOI: 10.3390/mol2net-05-06712. https://sciforum.net/paper/view/conference/6712.

12. Nandy A, Basak SC. A Brief Review of Computer-Assisted Approaches to Rational Design of Peptide Vaccines. *Int. J. Mol. Sci.***2016**,*17*,666; doi:10.3390/ijms17050666.

13. Basak SC, Majumdar S, Nandy A, Roy P, Dutta T, Vracko M, Bhattacharjee AK. Computer-Assisted and Data Driven Approaches for Surveillance, Drug Discovery, and Vaccine Design for the Zika Virus. *Pharmaceuticals* 2019, 12, 157; doi:10.3390/ph12040157.

14. A Nandy, A Ghosh and P Nandy, Numerical Characterization of Protein Sequences and Application to Voltage-Gated Sodium Channel Alpha Subunit Phylogeny, *In Silico Biology***9**, 77-87, 2009.

15. SABLE server: https://sable.cchmc.org/

16. BepiPred server: http://www.cbs.dtu.dk/services/BepiPred/